Analyzing the Wild-West of Interrater Agreement in Affective Content Analysis on Text

A Systematic Literature Review

Macsim Violeta-Mara[1]
Supervisor: Bernd Dudzik[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Macsim Violeta-Mara
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Catharine Oertel

Abstract

Human-computer interaction has long been the focus of technological evolution; however, in order for this type of system to reach its peak potential, machines must recognize that humans are constantly influenced by emotions. Text affective content analysis models are one attempt to integrate human psychology into computers, trying to detect the emotion transmitted by written input. There are numerous approaches to implementing such systems, with supervised learning still popular. The challenge of creating textual affective datasets is not in the availability of records, as humanity has reached a peak in data production, especially text, but in ensuring the consistency of the annotations provided by humans when included in the process. This study conducts a systematic literature review focused on providing details of published corpora. The annotation process, as well as any trends in how it evolved, will be examined to obtain dataset particularities. The ultimate intent is to lay the groundwork for an ample study aimed at analyzing the relationship between interrater agreement levels and performance scores of models trained on these datasets. The relevant literature was extracted from 3 search engines: Scopus, IEEE Xplore, and Web of Science, with a focus on manually labeled written records that are not part of multimodal systems, resulting in an analysis of 41 datasets. According to the aggregations, when humans are recruited to perform this task, researchers are more likely to use multiple annotators and calculate the degree of agreement between them to ensure the data's reliability before using it. The conducting researchers are inclined to either train these people before the procedure or tailor the set of labels to potentially increase the uniformity of ratings. As a result, this paper highlights the variety of annotation process characteristics and points towards standardizing this task.

1 Introduction

With the advancement of technology providing means of easily sharing information, text has been a prominent method of voicing opinions and emotions. As more and more people join social media platforms and almost all industries are transitioning to digital environments, large quantities of data are being generated [1, 2] that can be easily manipulated for research purposes. This surge in development has not only increased data accessibility but also has also put human-computer interaction under the spotlight for a long time. While this is not a new topic, it only began to reach its full potential after engineers were able to incorporate human-specific characteristics, such as affections, into the system's rationality.

In psychology, affect refers to all unconsciously experienced emotions, such as feelings or sentiments [3]. Affective content analysis models assess the underlying emotions that various types of input attempt to express [4]. Text-based affect analysis systems extract emotions conveyed through written material. They are integrated into human-computer interaction systems, such as customer service chatbots [5, 6], to enhance the way people perceive communication with a non-human agent. Because text records were so simple to acquire and parse, this type of affect prediction technology was the first iteration of modern emotion prediction models, that now analyze more complex data such as audio or video.

The quality of data used to train or test models significantly impacts the performance of learning systems [7, 8, 9]. In a supervised setting, records must be encoded with the emotion they represent to serve as a ground truth example for the model. Human annotators are still being used as the main method for executing this process. This, however, faces multiple problems: emotions bring the issue of ambiguity [10], while including humans introduces subjectivity of interpretation [11]. These factors can distort dataset quality, potentially affecting model performance [12]. The term *interrater agreement* (IRA), also referred to as *interrater reliability* (IRR), was introduced in research to help reduce these discrepancies by quantifying how much the answers of multiple raters coincide [13].

There is still no standardized method for manually annotating records, but only proposals of good practices. Because there is no clear guide, the provided results must be weighted differently, which is why label uniformity is an important factor to consider when measuring. IRA can be calculated using a variety of methods, including Cohen's κ [14], Fleiss' κ [15], and ANOVA [16]. With so many options available, researchers have yet to determine the best formula for calculating label reliability, and there is no clear indication of when a specific coefficient should be used.

Studies have repeatedly highlighted the inconsistency of human coders as one of their drawbacks [13, 17]. However, this paper proposes to shift the focus on the effects of these inconsistencies when putting to use manual-labeled datasets, seeking to provide the foundation for answering the question "*How does interrater agreement influence the performance of text affect prediction models?*". Such research would imply **a two-step process**: first, identifying datasets from which particularities would be extracted, and then pinpointing models that use the corpus to form comparisons in performance scores and draw conclusions. This endeavor is divided into several sub-questions that focus on targeted feelings, various emotion representation schemes, details of the annotation process, and model performance when using specific datasets. The precise list of sub-questions is displayed in Table 1. As the work is constrained to only 9 weeks, this study will focus on the first 5, while preparing the scene for subsequent research to conclude the answer to the 6^{th}.

A systematic review will be conducted to assess the current state of the literature on the practical application of datasets for affect content analysis models. This allows researchers to obtain the most common characteristics of manually annotated

Table 1: List of sub-questions derived from the main research question and the reasoning behind them

	Sub-question	Motivation
SQ1	What types of affective states have been targeted by datasets?	Affect is an umbrella term for multiple elements which can be differently represented. An annotator's response will vary based on what they are requested to label.
SQ2	What different affect representation schemes have been used in these datasets and what is the motivation behind them?	Different representation schemes yield different sets of labels from which annotators can choose when preparing the training data.
SQ3	What are the settings of the annotation process (i.e. number of raters, interrater agreement, facilitation of agreement)?	Annotation specificities can have a significant impact on the outcome of models trained on these corpora, as it all comes down to how uniform the encoding is to the actual label.
SQ4	Is there a change in how datasets measure interrater agreement over time?	As there is no standardized procedure for annotation, a trend in specific settings chosen by researchers might become apparent.
SQ5	Is there a relationship between the affect representation scheme used by datasets and their interrater agreement?	Giving a larger set of vague notions as labels might make it harder for the annotators to decide on one, compared to when there is a limited number of emotions to choose from.
SQ6	Is there a relationship between the interrater agreement in datasets and the empirical performance of affect prediction systems using them for training and evaluation?	Models learn what they receive, which means that the uniformity of training data labels might affect how good the model is. However, this falls out of the scope of this research due to time limitations and will only be tackled through small observations derived from the obtained literature.

datasets, providing a solid foundation for analyzing the relationship between training data and model performance in the context of text affect prediction. The paper will begin in section 2 by providing a general overview of relevant information for this study and related work, followed by the methodology for conducting such a literature review in section 3, and then continue in section 4 by presenting the results of the findings in bibliographic databases while adhering to the methodology and answering the questions raised earlier. Section 5 will discuss the ethical implications of the research. Section 6 will provide a general interpretation of the results, followed by section 7, which will resume the research and propose future recommendations.

2 Background & Related Work

Annotating The usual approach for creating such datasets is through *expert annotation* [18, 19], where people with specialized knowledge perform the task. *Self-reporting* [20, 21] removes the interpretation step and asks individuals to assess their emotions in relation to the records. Technological advancements and the widespread use of social media created another method of self-reporting by introducing *distant supervision labeling* through hashtags. Posts are labeled by their creators when adding tags. *Crowdsourcing* uses remote platforms like Amazon Mechanical Turk[1] or Figure Eight[2] (formerly known as *CrowdFlower*) to gather manual encodings, but these tend to be less structured than those obtained in a professional setting.

Expert annotation is by far the most researched procedure, despite being more controlled given that it is guided by experts. M. Finlayson and T. Erjavec have proposed specific tools that can aid certain steps of this procedure and defined some actions that researchers can take to accomplish this more efficiently [22]. In 2021, M. Paquot and S. Gries published specialized instructions for linguistics corpus [23], discussing how to statistically handle such procedure, and Niladri Dash released a book [24], focusing on how to structure useful guidelines for annotators to follow when labeling.

Computing agreement Agreement is a metric used in different domains and settings for statistical analysis of categorical data. Most of the initial proposals were catering to medical research, but have been adopted in other fields, such as machine learning, later on. Simpler versions can only take into account the number of times the labels match, but chance-correcting ones are more common due to their accuracy. Typically, these metrics have values ranging from -1 (no agreement/reliability) to 1 (complete agreement/reliability). In 1955, William Scott proposed a formula for calculating IRA by counting the number of uniform labels and adjusting it with the probability of agreeing by chance [25], now known as *Scott's Pi*. In 1969, Ole Holsti simplified Scott's formula by only focusing on the number of times annotators agreed out of the total number of coding decisions [26]. Scott's Pi was further developed to also take into account the distribution of values among annotators by

[1] Amazon Mechanical Turk: https://www.mturk.com
[2] Figure Eight: https://www.figure-eight.com

Cohen's κ [14]. This became known in statistics as the "*golden kappa*" for computing any type of consistency or reliability in studies, with statistical studies spanning 40 years that suggest *Cohen's* κ as the measure of choice [27, 28]. In 1971, Joseph Fleiss generalized the classical coefficient through the *Fleiss'* κ [15], allowing to compute agreement between groups larger than two. A later proposal was *Krippendorff's* α [29], which considers multiple types of agreement (e.g. nominal, ordinal, ratio). Bhowmick et al. [30] proposed a new method for measuring this by fine-tuning the classical κ coefficient. Despite formula differences, Gisev et al. [31] found that different methods of calculating the IRR produce similar results.

Literature reviews Yadollahi A. et al. [32] conducted a survey on sentiment and affect content analysis datasets, creating a general picture of available resources and emphasizing the relationship between classification methods and corpus. Nandwani P. and Verma R. [33] evaluated unsupervised emotion classification systems and compared them to corpus-based systems. Wang Y. et al. [34] conducted a detailed study of affective computing publications, analyzing various datasets (visual, audio, text, physiological, and multimodal) and classification models. Oberländer L.A.M. and Klinger R. constructed a small standardized database containing widely recognized manually annotated text corpora and analyzed how well models perform when evaluating a different dataset from the test one [35]. While there are publications that examine the impact of dataset quality on model performance, there is a significant lack of research that includes annotation quality as a defining factor.

3 Methodology

The main method for answering the research questions has been determined to be a **systematic literature review**, which can help assess the current state of expert-developed corpora and analyze their characteristics. This type of procedure emphasizes systematic searching, as well as detailed decision documentation, in order to maintain transparency and reproducibility. *PRISMA* is the main framework for conducting and reporting such reviews, providing template documents that increase standardization. The newer 2020 version of PRISMA [36] was selected as it is more tailored for studies outside of the medical field compared to its predecessor. To obtain a comprehensive methodology, the official *PRISMA 2020 Checklist* [36, p. 4-5] will be followed. Section 3.1 will present the main exclusion and inclusion criteria that will determine whether or not a paper will be considered for the final result list, and section 3.2 will explain the search strategy. Section 3.3 will present the process of assessing inclusion, strengthened by additional constraints, presented in 3.4, to make the research feasible for only 9 weeks of work. Finally, section 3.5 will show the results after completing this process, and section 3.6 will list what information will be extracted from the included records.

3.1 Eligibility criteria

The first step is to define the set of useful literature for answering the research questions by establishing inclusion and exclusion criteria. To conduct a thorough analysis, the intersection of all elements in Table 2 should include only relevant documents in a sufficiently large number. Each criterion has a clear motivation behind the decision to use it as an eligibility basis for the choice of literature that will answer the research question.

3.2 Search strategy

The main literature search engines that will be used are Scopus[3], IEEE Xplore[4], ACM Digital Library[5], and Web of Science[6], as they are all popular large databases that either contain a significant quantity of technical literature or only focus on this domain of science. Their reliability is also endorsed by the Delft Technical University Databases recommendation of literature search engines for the field of Computer Science.

The next step is to determine exactly what needs to be searched. To accomplish this, the main research question is broken down into concepts that fully describe what is being studied. These terms ended up being *text*, *affect analysis*, *dataset* and *manual labeling*. However, the goal of this step is to collect as many sources as possible, which means that each term must be broken down into all possible occurrences within a piece of literature. For each term, all relevant concepts, synonyms, and word derivations that could represent the desired content will be identified. Table 3 shows the more specific search terms.

The final search query contains all of these terms. To ensure that all possible publications are included in this study, these will be used in queries across all metadata (i.e. search terms within the title, abstract, and full-text). Because *Web of Science*

[3]Scopus: https://www.scopus.com
[4]IEEE Xplore: https://ieeexplore.ieee.org/Xplore/home.jsp
[5]ACM Digital Library: https://dl.acm.org/
[6]Web of Science: https://webofscience.com

Table 2: Grounds of inclusion and exclusion of literature from the research

Inclusion Criteria	**Reasoning**
Literature discusses corpus designed for text affect prediction models	To analyze corpora, specifications for the label selection stage and annotation process must be available. A paper must either briefly mention an already-published and structured dataset or describe in detail their proprietary corpus.
Records are manually labeled by humans	To discuss interrater agreement, records need to be labeled by people.
Exclusion Criteria	**Reasoning**
Literature discusses datasets designed only for sentiment analysis	Sentiments are emotion polarities acquired through experience, but affect represents complex and precise emotions triggered unconsciously [3]. This leads to different sets of labels, which can influence the difficulty of the encoding task.
Literature discusses corpus designed for multimodal affect prediction	The research focuses only on datasets with text input. Furthermore, having multiple types of input sources, rather than just one, can affect performance by adding context to the situation, potentially leading to higher scores, and thus an unfair comparison.
Paper is not in English	Reproducibility is impeded by using papers that are not understandable by everybody.

Table 3: Concepts and synonyms related to each broad concept. The asterisk (*) shows that any valid word derived from the body before the symbol will be considered.

Concept	**Search terms**
TEXT	text*, textual content, post, social media, news, Twitter, Reddit, blog
AFFECT ANALYSIS	affect analysis, affect recogni*, affect predict*, affective computing, affect detect*, affective analysis, affective predict*, emotion analysis, emotion recogni*, emotion predict*, emotion detect*
DATASET	dataset, corpora, corpus, records
MANUAL LABELLING	raters, multiple annotators, multiple annotations, human annotators, human annotation, human-annotated, human-rated, human raters, interrat*, inter-rat*, multiple raters

uses a different interpretation order than others, the expression's structure will be designed to be universally applicable across all search engines. Other constraints are dictated by *IEEE Xplore*. First of all, it restricts the use of wildcards to 9, so some terms with a limited number of variations were explicitly listed rather than using a generic body form. Another feature to consider is the way the search engine parses the query and filters. Compared to the other three literature databases, *IEEE Xplore* filters sequentially until the first conjunction. This means that it must narrow down the search based on the size of the topic, beginning with the broader ones and progressing to the more niche ones, such as only talking about manually labeled datasets. The final search query will be as follows:

```
("affect analysis" OR "affect recogni*" OR "affect predict*" OR "affective computing"
    OR "affect detect*" OR "affective analysis" OR "affective predict*" OR "emotion analysis"
    OR "emotion recognise*" OR "emotion predict*" OR "emotion detect*")
AND (text* OR "textual content" OR post OR "social media" OR news OR Twitter OR Reddit OR blog)
AND (corpus OR dataset OR records OR corpora)
AND (raters OR "multiple annotators" OR "multiple annotations" OR "human annotators"
    OR "human annotation" OR "human-annotated" OR "human-rated" OR "human raters" OR interrat*
    OR inter-rat* OR "multiple raters")
```

3.3 Selection process

After retrieving the whole poll of relevant literature, there needs to be a more thorough selection strategy due to the large number of results. While *Web of Science* retrieved only 16 papers and *IEEE Xplore* 19, *Scopus* identified 870 results and *ACM Digital Library* 2293. The search needs to be followed by a screening process that will look at the inclusion and exclusion criteria. To accomplish this, the following will be executed:

1. **Run the base query:** Use the query presented in section 3.2 in each of the literature database engines to collect data
2. **Filter by title:** The title will, in most cases, explain the content and will indicate whether it fits any of the criteria from Table 2; when the content of the title is not too explicit, it will be transferred to the next filtering step

3. **Filter by abstract:** The abstract contains the main points of a paper, so reading it will help in removing the non-eligible ones; if the abstract is too ambiguous regarding content, the paper will be passed to the next phase
4. **Filter by full-text:** This process will be done during the synthesis step; while reading the complete text, some records can be excluded as they might discuss multimodal datasets or unsupervised-annotated datasets

3.4 Feasibility filtering

While the first two sets of results can be manually screened, the results provided by *Scopus* and *ACM Digital Library* require an additional layer of stream-lined filtering, given the short time to complete the research. These additional conditions will be used before the manual screening step on all search engines. They will still present an objective viewpoint because there is no cherry-picking involved.

1. **Filter by ease of scanning through:** *ACM Digital Library* does not provide automated filtering through proposed keywords, which would reduce the amount of literature to later manually parse. As potential trends will be analyzed, it is not possible to limit the publication date to a specific interval, and selecting a subset for each year can introduce subjectivity into the process. This search engine will be excluded from the screening process because it returned 2293 results, which could not be reduced to a manageable number in 9 weeks.
2. **Filter by keywords:** *Scopus* allows to exclude based on descriptive keywords. This search engine lists all possible keywords, along with the number of occurrences inside the whole list of results from the query. The list of selected keywords is presented in Appendix A.
3. **Filter by field of expertise:** While there is literature that discusses the pitfalls of manual annotation of emotion and affective responses in a variety of domains, due to the limited amount of time the search will be limited to only the field of Computer Science. Fortunately, *Scopus* and *IEEE Xplore* allow for such filtering.

3.5 Search results

In the initial phase of the literature identification, 3198 records were obtained by applying the search query on the 4th of May, as mentioned in section 3.3. However, 2293 records provided by *ACM Digital Library* had to be discarded, as previously mentioned in section 3.4. This resulted in only 905 results, with only 4 being in a language other than English. Only 30 duplicates were identified and 6 others were removed given that the literature search engines identified them as posters, notes, editorials, or short surveys rather than scientific publications. The feasibility steps followed, namely the keyword filtering on *Scopus* and by field of expertise, prior to the manual screening phase. Out of 866 results from *Scopus*, 219 were excluded because they were not related to Computer Science. This was followed by the removal of 378 documents by keyword filtering, reducing the amount of literature to be manually scanned to 268. After comparing the titles to the inclusion and exclusion criteria, 157 were removed, and the remainder were analyzed using the abstract. After reviewing the abstract, 53 records were excluded. The final stage of the screening was fingering by full-text reading, from which 17 others were found to not match the inclusion criteria as they discussed sentiment analysis or used automated-annotated records. Finally, the literature review includes **41 records**. However, some of the literature did not specify details of the datasets they were using, only stating the name. For this reason, **10 additional papers** were added which introduce the corpus creation process mentioned in the included studies. Figure 1 visually summarizes these steps.

3.6 Data Extraction

Data extraction was done during the full-text screening. References were stored into a management tool called *Mendeley*[7] and relevant data that aids in answering the research questions were extracted and kept in a spreadsheet in *Microsoft Excel*[8]. The information stored from each paper is as follows:

- **Dataset name:** the records the model uses, including the proposed novel ones
- **Year of publication:** the year when the dataset was officially published, which might differ from the publication year of the analyzed paper; this value helps identify any trend in the annotation process over the years
- **Targeted affective states:** what the emotion describes in the context of the input and whose emotional perspective is being recorded, as this will correlate with what the annotators will be requested to label

[7]Mendeley: https://www.mendeley.com/
[8]Microsoft Excel: https://www.microsoft.com/en-us/microsoft-365/excel

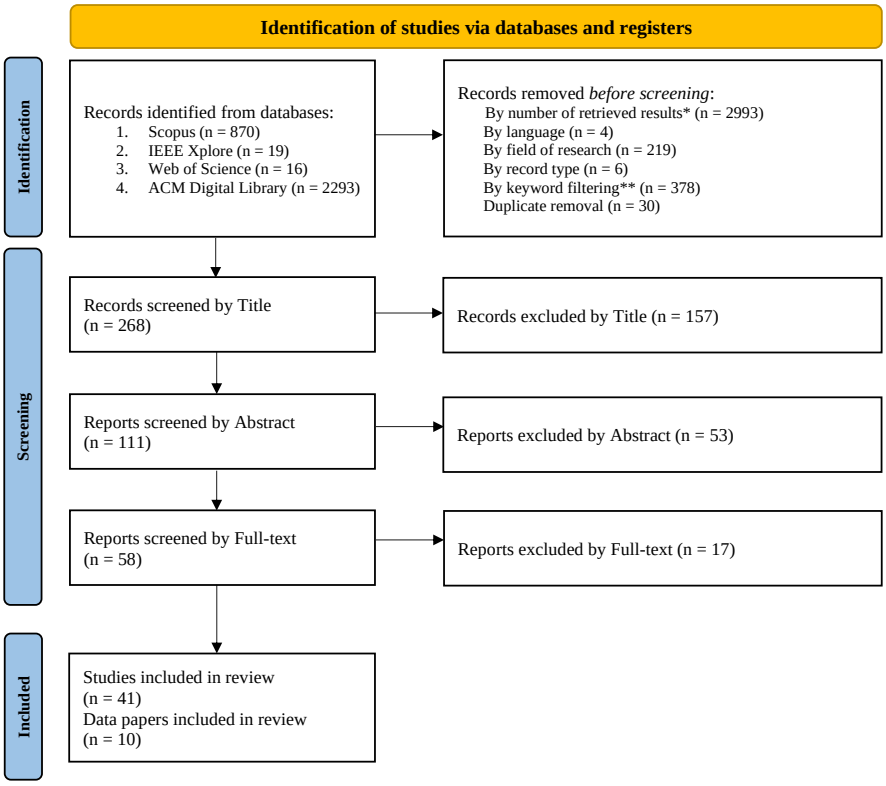


Figure 1: PRISMA Flow Diagram [36, p. 10] summarizing the filtering and screening of 3198 initial results, leading to 51 references included in the analysis

- **Affective representation scheme:** what type of emotion scheme the model uses; depending on how many emotions annotators have to choose, they might not find it so easy to choose from the poll of options
- **Number of annotators:** how many people labeled the records, which can determine whether IRA needs to be calculated or not
- **IRA:** the method of viewing consistency in labeling, in case there is more than one annotator, and the values of agreement
- **Agreement facilitation:** methods of increasing the agreement rate before or during the annotation process, if it is the case
- **Model performance evaluation:** accuracy scores of the model trained using the respective dataset or baselines presented by researchers to motivate the relevancy of their proposal

4 Results

A literature review is continued by the synthesis step, where the collected data is aggregated to answer the research questions. This process resulted in the identification of **41 datasets**, as a couple of corpora are mentioned more than once in different records. All their relevant information is schematically presented in Appendix B or in a digital version. What also needs to be taken into consideration is that datasets are being used by multiple pieces of literature. However, not all datasets are completely manually annotated; some are hybrid: a small subset of the total number of records is annotated manually, while the rest is streamlined and uses the subset as ground truth for learning. Therefore, only the portion of the corpus that is manually labeled by people will be taken into consideration during analysis. Furthermore, some datasets included not only

labels for affect prediction tasks but also actions such as content prediction or sentiment analysis. The study will only look at the annotation process used for affect prediction.

In this section, corpora spanning almost 30 years will be analyzed from multiple perspectives. Section 4.1 addresses **SQ1** by overviewing the affective states targeted by these datasets. Section 4.2 categorizes by the affect representation scheme (ARS), answering **SQ2**. Section 4.3 examines how human annotators are used, from the number of people asked to label one record, to how they measure agreement when engaging multiple people, and how they facilitate agreement between them, answering **SQ3**. Section 4.4 will discuss the trends of different IRA methods over time (corresponding to **SQ4**), section 4.5 will analyze a possible relationship between ARS and the degree of agreement (answering **SQ5**), and finally, section 4.6 will transition to **SQ6**, looking at the depths of how the identified literature can be a start for a potentially more ample study on the effects of the annotation process on a model's performance.

4.1 Targeted affective states

Annotators' jobs vary depending on the task they are asked to perform. The specific questions they are asked to answer influence what the labels mean. This section discusses what each dataset represents in terms of affective state, as well as whose perspective of emotion the labelers were asked to encode for the records, addressing **SQ1**.

Table 6 lists the datasets grouped by targeted affective state. The identified representations are *emotion*, *mood*, and *opinion*. At first glance, it shows that the majority of the literature focuses on representing emotions, with 92% of the corpus portraying emotions. Scherer K. [37] concluded that people misuse the term "emotion", which may explain this commonality. According to his definition, this type of affect must be associated with a trigger, an event that causes the organism to react in a specific way. The stimuli include both external factors, such as actions in the individual's surroundings, and internal ones, like memories. Moods are at the other end of the spectrum, defined as low-intensity emotional states that do not require a stimulus and can last longer than emotions. The analyzed datasets use only written records, making it difficult to identify lower-intensity emotions that may not be linked to events when the only context is words. This may explain why mood is conveyed only by **C1** and **T2**. In terms of opinions, it appears that there is no problem distinguishing the meaning behind the ambiguous word, as the definitions align to some extent across versions. Kim S. and Hovy E. [38] organized this construct into four components: topic, opinion holder, claim, and sentiment. **CLARIN-Emo** is the only dataset focusing on this construct, using more than just polarities as the sentiment component; each record is also labeled for affect.

Although there are minor differences in what the labels express, using a third party to encode a piece of text that is not their own raises a new question: whose emotions are these annotators labeling - their own or the general public's? This point of view is rarely discussed in literature. **SM2**, **FB-SEC-1** and **XED** appear to be the only datasets that specify what raters were tasked with doing. For the first corpus, annotators were asked to respond to the question "How might someone feel after seeing the following post?" [39], suggesting a more objective representation. A similar case is for **FB-SEC-1**, where annotators looked at "the point of view of a typical reader" [40]. The request for the latter was to annotate from the perspective of the person who was saying the movie line (i.e. what the speaker was supposed to feel) [41]. For **CLARIN-Emo**, one can speculate that the annotators were asked to label the records with the opinion of the text author to reduce the level of subjectivity, but the other option of encoding as them being the opinion holder is also possible.

4.2 Emotion representation schemes

The method by which emotions are expressed influences the labeling process. This section will provide an overview of the various affect representation schemes used, with a focus on **SQ2**. The table of Appendix B contains an aggregation by scheme, including the emotion label sets which are detailed in the legend. The main categories are *Categorical*, *Dimensional*, and *Hybrid*, with 90% fitting the first group.

A small fraction of the datasets don't use only categorical representation. **TWISCO** utilizes the three-dimensional way of representing affect, valence-arousal-dominance, alongside discrete labels. The authors' approach differs from usual practices in that they do not encode these dimensions using non-discrete values, but instead use predefined labels [42], such as *positive*, *neutral*, or *negative* for valence. The **Emotion-Stimulus** dataset distinguishes itself from other categorical datasets by identifying the source of emotion associated with a record, in addition to the actual affect. The authors of [43] intended to convey even more complex emotions, so they used the triggers and a more restricted set of labels, distinguishing between the reasons behind an emotion as they can result in varying intensities or durations. **SemEval-2007** and **Affect Database** augment the limited poll of basic emotions with numerical value for valence. **J1** appears to be the only dataset deriving the emotions based on continuous values of polarity and arousal. Its encoding scheme spans values between -5 and 5, with three thresholds for each dimension.

There are 29 different emotion label sets used; however, 8 of them are variations on Ekman's basic emotions and 2 of Plutchik's wheel of emotions. Some others share a lot of similarities in how they represent a full range of emotions, with minor differences in label selection. At first glance, most datasets appear to convey the same set of emotions, but in a synonymous format. However, due to the use of datasets for precise emotion classification in the study, the decision was made to retain the subtle differences in intensity or duration of the emotions, rather than merging representation schemes with many similarities when aggregating the data. For example, **7A** and **7S** have the same set of labels, with one exception: *Amusement* and *Entertained*. While both are a positive response to an activity, *Amusement* encompasses the latter, while entertainment does not always lead to feeling amused.

The most prevalent representations are *Ekman's 6 Basic Emotions* [44] and slight derivations of them, like additional labels for emotions (e.g.: *Shame* or *Contempt* or *Calm*), two emotions exchanged for others with bigger intensities (implemented by **T2**), namely *Rage* instead of *Anger* and *Joy* instead of *Happiness*, two labels for neutral emotions or for when the record doesn't fit with any of the proposed emotions (implemented by **SDTC**), an additional dimension to encompass an even larger variety of emotions without the need of too many labels (used by **SemEval-2007**), or even a version with only 5 emotions (used by **FB2**), where *Disgust* is removed. **TWISCO**, **U1**, **R1** and **Kannada-English** use the plain version. Other psychological theories are *Plutchik's Wheel of Emotions* [45], *Shaver's Taxonomy* [46], and *Izard's Theory of Emotions* [47], but they do not constitute a significant majority.

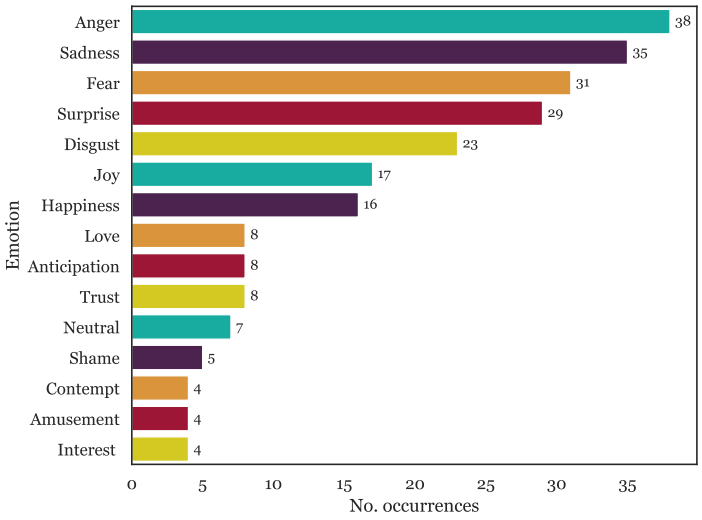


Figure 2: Diagram showing the frequency of the top 15 emotions that appear in the affect representation schemes of the identified datasets. The rest of the emotion labels that were too specific to a dataset (i.e. appeared too infrequently) can be overviewed in Table 6.

Most literature lacks actual justifications for how to represent the most important emotions for their specific case. As social media platforms already provide some kind of labels for emotions, **FB-SEC-1** uses the reactions each post can get on Facebook, with the exception of *Like*, which they deem as being used ambiguously. In this case, the label was translated into the actual emotion conveyed by the reactions. **CARE** seems to have solid justification as to the choice of labels, stating that the most used 4 emotions (i.e. *Excitement*, *Anger*, *Sadness* and *Scared*) are combined with some which are deemed as important for the context of emotions in social media.

Figure 2 illustrates the popularity of emotions as labels. *Sadness*, *Anger*, and *Fear* appear more than 30 times. The most frequently used positive emotions are *Surprise* and *Joy*, with 29 and 17 usages, respectively, a significant difference in appearances for the leading two. From the top 6 downwards, a wide range of positive emotions can be observed, which may lead to the belief that humans use a broader range of words to describe positive sentiments than negative ones, being limited to a few unpleasant ones that encompass a less nuanced feeling. Everyday vocabulary appears to employ more complex discourse to express happy feelings. A fourth of the datasets include the *Neutral* emotion tag, which can be used when none of the provided labels are appropriate for the context. For example, according to [48], if annotators are unable to select a predefined label, they can choose *Neutral*. However, **SDTC** contains both *No emotion* and *Not sure* in the ARS, but doesn't differentiate the situations when one would be better fitting than the other.

4.3 Annotation process

In this section, the annotation process for the 41 datasets will be analyzed, taking a look at how many raters are used, how interrater agreement is calculated, and whether the researchers take any steps to improve agreement. The information below will focus on answering **SQ3**.

Size of annotator groups

Interestingly, even if the search query took into consideration manually labeled datasets through terms like "*human annotation*" or "*human-rated*" that do not necessarily point towards more than one person employed, the majority of the identified literature uses multiple annotators to encode the records used in the corpus. **CrowdFlower** is an exception as it is a crowdsourcing effort to create a dataset. **TEC** also uses one annotator per record, as the authors extracted Twitter posts along with hashtags that are followed by emotions, and the author of the tweet is the rater. Therefore, when researchers introduce subjective human work, the majority of them seek multiple opinions rather than relying solely on one.

Table 4: Number of annotators used for designating label to one record

No. annotators	No. datasets	Datasets
1	3	CrowdFlower, TEC, ISEAR
2	3	GitterCom, FB1, Kannada-English
3	13	Thai Hate Speech, CM-MEC-21, CARE, XED, Affect Database, FB-SEC-1, Indonesian Amazon Reviews, R1, DD1, C1, J1, TWISCO, SDTC
4	3	Emotion-Stimulus, U2, SDTC
5	2	ArECTD, SM2
6	3	SemEval-2007, CLARIN-Emo, SM1
7	4	SemEval-2018, EmoContext, C2, FB2
10	1	PERC
11	1	T1
Varying	4	GoEmotions, SSEC, RomEmoLex, JIRA Database
Not specified	3	EmoInt-2017, U1, T2

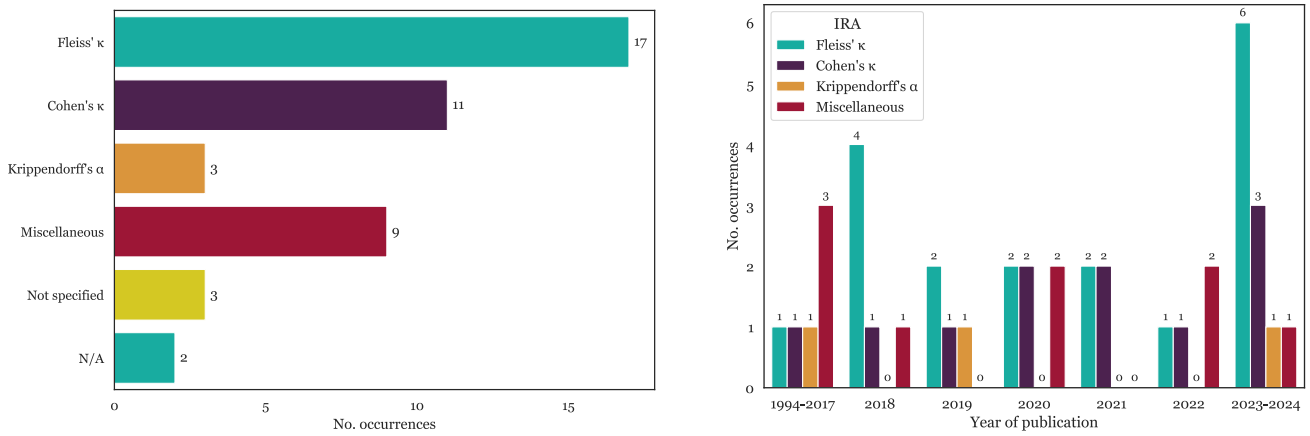
Table 4 categorizes datasets according to the number of annotators used. The *Varying* category includes corpora without a fixed number of annotators per record. **GoEmotions** employs 3-5 annotators, **SSEC** uses 3-6, and **RomEmoLex** utilizes 2-3. For those labeled as *Not specified*, the exact number of people employed is not specified, but it is known that the ground truth is a collaborative work of multiple individuals. **JIRA Database** takes a more complex approach to this than the other three in the same category, as this dataset is annotated in three phases with different affect representation schemes that have common emotion labels, where 2 use 3 annotators and one uses 16. Researchers prefer a small number of experts because it would be more difficult to organize the process with a larger group of people. It's common to seek labels from three people. **ISEAR** uses a complex approach to determine a golden truth for each record, using self-reports through surveys and one country acting as an annotator when computing the label. However, the number of individuals from each country varies. The study asked 2921 people to complete a questionnaire, and they came from 37 different countries, so the author considers 37 annotators per label. Unfortunately, none of the pieces of literature provide justification as to the number of raters used.

Interrater agreement

36 out of the 41 datasets mention computing the degree of agreement between the annotators. The **CrowdFlower** and **TEC** datasets don't require agreement computation because they use self-annotated records. **SM3**, **T2**, and **Emotion-Stimulus** don't specify whether the creators thought of seeing if the labels are consistent or not between the employed experts. Despite the interesting approach of **ISEAR**, uniformity between labels between countries is being calculated, but not on the individual level as it would be irrelevant.

Figure 3a depicts the popularity of methods for calculating agreement among annotators when assessing label reliability. The *Miscellaneous* category includes multiple approaches that were mentioned less frequently, namely the Central Limit Theorem, ANOVA, Matthew's Correlation, Spearman's Correlation, Pearson's Correlation, Siegel & Castelan's κ, and the number of full agreements. The *Not specified* label contains both the datasets that didn't specify whether they compute IRA (**SM3**) as it cannot be accurately said that they did not compute that, and the datasets that didn't explicitly state what formula or method they have used for computing agreement (**SM2** and **Emotion-Stimulus**). People clearly prefer computing using *Fleiss'* κ, with 17 out of 41 datasets using it. It is worth noting that some use multiple ways to calculate agreement as a more thorough verification method, rather than combining them. *Cohen's* κ seems to also be favored by researchers. An interesting

observation is that *Krippendorff's* α, which can be considered the most complete method of calculating agreement due to the incorporation of multiple ways of agreeing, is rarely used in the literature. However, this could be due to the proposal's recent date, which was only in 2013, leaving insufficient time for the academic community to become accustomed to it.



(a) (b)

Figure 3: Popularity of IRA calculation methods identified in the datasets (a) in general and (b) over time

12 datasets rate the level of agreement as high or very high, while 4 rate it as moderate-high. **TWISCO** and **FB-SEC-1** show extremely low agreement. However, these indicators can be highly subjective. For example, **T3** recorded a *Fleiss'* κ score of 0.26, but the authors describe the inter-annotator agreement as "quite satisfactory" [49] when this number would hardly qualify for moderate agreement. On the opposite spectrum, **U2** has a high agreement, stating a *Cohen's* κ score of 0.82, but partially diminishing the results by saying it is a "very decent rate"[50]. The perceptions of a fairly acceptable agreement level seem to differ a lot between the authors. It is also difficult to compare when the researchers do not provide a score, but only an indicative mark ranging from *very low* to *very high*. **SM2**, **Thai Hate Speech**, and **SSEC** belong to this category, where it cannot be said for sure if the writer isn't overstating or understating.

Facilitating agreement

While determining how uniform the record encoding is, simply calculating a value does not affect the work that the dataset will be put to. Even after the procedure, researchers who organize any data collection activity can help to facilitate agreement, increasing the likelihood of achieving a higher level of agreement. Looking at what people in this field have done, a positive outcome can be observed. Unfortunately, there is no way to compare how effective these actions were and how much value they added to the annotation process because there is no analysis of agreement if no bonus activities were carried out. In most cases, preventative measures are implemented to reduce the likelihood of annotators labeling randomly because they are unsure what to choose. A few datasets used multiple measures in the same stage of the process. However, **EmoInt-2017**, **RomEmoLex**, **Indonesian Amazon Reviews**, **SM1**, **SM3**, and **J1** don't specify any measures. This does not imply a lack of interest in achieving high label-uniformity. Figure 4 presents the different methods and their frequency. The *Miscellaneous* category includes multiple actions: creating the dataset in three phases, starting with a large number of emotions and narrowing it down after each step to the most agreed-upon ones, as done by **JIRA Database**, providing a question that the label should answer to with regards to the record and designing an iterative annotation process until achieving unanimous consensus, both implemented by **SM2**), and narrowing down to the most expressive and comprehensible labels for emotions (**Emotion-Stimulus**).

Finding a workaround for the labeling systems appears to be the most popular. In most cases, raters can select multiple emotions per record, but they are limited to a certain number or are asked for a precedence order. **GoEmotions** does not have a fixed number of labels per record, but the instructions stated to put only those that are relevant and stand out the most. To avoid forcing annotators to choose at random when they are unsure, they can choose the *Neutral* emotion. This decision can sometimes backfire, as in some cases (e.g. **Thai Hate Speech** or **U1**), this label is the most common. Another idea of enhancing the label set was used by **ArECTD**, where the researchers discussed with the group of raters about the selected labels that they will be using and ended up realizing that they weren't broad enough, so two more emotions were added.

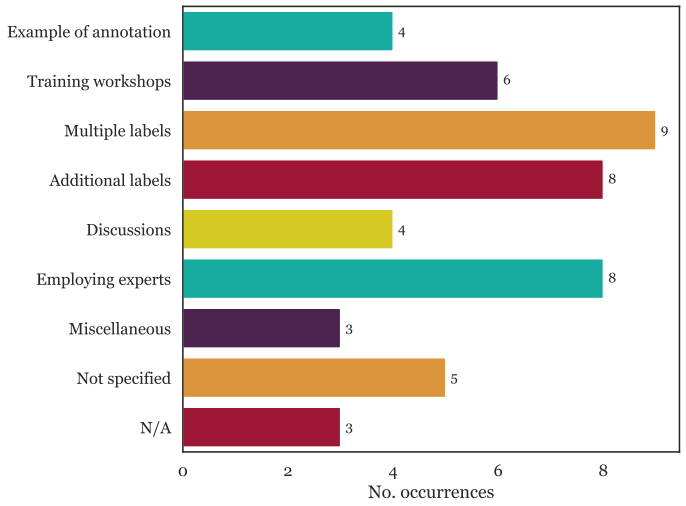


Figure 4: Categorizing of datasets by measures taken to increase the likelihood of agreement

Competence is also considered prior to the annotation process, either by hiring psychologists and linguists to label the records or by preparing encoders in proprietary workshops.

4.4 IRA calculation popularity

This section will discuss the frequency of the methods used to compute IRA identified in the 51 literature pieces, looking at the overall frequency of calculation methods as well as categorizing them by dataset release year. The identified literature dates from 1994 to 2024, spanning almost 30 full years of advancement in the field of supervised affect prediction.

Figure 3b shows the IRA calculation options grouped by the dataset's release year. Due to a lack of literature between 1994 and 2017, they were combined into a single period, with the same holding for 2024, which was merged with the year 2023. While there are a few corpora before the 2000s, the majority of them are after 2018, when all of the methods listed in the legend were published by researchers. Given that Krippendorff's α was first used in 2013, it is understandable that the identified literature does not specify its use until 2019. Fleiss' κ appears to have maintained its popularity over the years, with the exception of 2022 and between 1994 and 2017, when other methods appear to be more prevalent. The *Miscellaneous* category appears to be more prevalent in the years 1994-2017, even though the formulas that are considered "standard" practices to this date were already decades old at that time. Surprisingly, researchers do not appear to have a linearly increasing preference for a specific method, with dips in each category of years for every calculation method.

4.5 Relationship between agreement and ARS

After analyzing the various methods for computing the IRA, some patterns between multiple features of the dataset can be observed. This section will look at the possibility of a link between the affect representation scheme and the degree of agreement, to answer **SQ5**. To accomplish this, the recorded levels of agreement as stated by the authors will be compared, categorizing them by the type of labels used for representing emotions and the variety of emotions used to portray the most important feelings in the context of the dataset's records. As observed in section 4.2, there is an abundance of ways that were used to convey emotions from the encoding point of view, talking about roughly 30 representations, which means that correlation is difficult to back up just by a combination of emotions.

In datasets with a *very high* or *high* expressed label of agreement, the most prominent ARS is *Ekman's basic emotions*, specifically **BET-5**, **BET-6**, **BET-6S** and **BET-6NN**, which appears 5 times, 4 times with a high level, and once with a very high level of agreement. Another interesting observation is that only 4 out of the 12 corpora with a significant level contain a *Neutral* label, which would have helped in situations of uncertainty and removed the option of selecting randomly. This may imply that not providing an alternative option doesn't necessarily mean that the annotators will be unable to assess the records "correctly".

Only three datasets, **J1**, **Affect Database** and **TWISCO**, use non-categorical labeling systems, rendering it insufficient

to link a preference for one calculation method to the affect representation scheme used. The first two have a high degree of agreement (0.72 and 0.93, respectively), while **TWISCO** has values between 0.14 and 0.38. All three use Fleiss' κ, but there aren't enough data points to conclude that this method works better for a hybrid (categorical and continuous) or pure continuous representation.

4.6 Relationship between model performance and dataset specifications

The ultimate goal of this research is to investigate any links between the annotation process used in corpus creation and the performance of models trained on them. However, as previously stated, the limited time budget prevents the entire process from being completed in this study. Without a separate analysis of supervised models trained on datasets, it would be impossible to accurately compare performance scores because the examples are not sufficiently well represented - 20 datasets are proprietary, tailored for a specific use, and can provide different results if used by another learning architecture. Additionally, some data papers, such as [51] or [52], do not provide a "*Benchmark*" section where the most basic integration of the dataset is being put to the test with a standard language model.

Table 5: Accuracy computed as F-1 scores obtained for datasets included in at least 2 pieces of literature

Dataset	F-1 score	Related paper
GoEmotions	0.5	[48]
	≈0.647	[53]
	≈0.23	[54]
ISEAR	0.71	[55]
	<0.4	[48]
EmoInt-2017	0.86	[48]
	0.59	[53]
Emotion-Stimulus	0.9	[48]
	0.74	[53]
SemEval-2007	≈0.5	[53]
	0.5	[56]
Affect Database	0.72	[57]
	0.62	[58]

According to Table 6, affect prediction models use only 7 datasets multiple times in the identified literature. The datasets that will be analyzed are **GoEmotions**, **ISEAR**, **EmoInt-2017**, **Emotion-Stimulus**, **SemEval-2007**, and **Affect Database**. Unfortunately, not all papers include a "*System Evaluation*" section that assesses the model's performance, situation encountered for **PERC**. Reviewing the literature that describes how datasets are being used for training and testing, the common metric used for analyzing performance is the **F-1 score**. The obtained values displayed in Table 5 are not generally consistent and vary between occurrences. The smallest variance is 0.1 (**Affect Database**), while the most significant one is seen in **ISEAR**, where the model in [55] scores an F-1 of 0.71, but in [48] it records less than 0.4 for each emotion. However, this dataset has previously shown it has a novel approach to label-making, which may explain the drastic variation. **SemEval-2007** appears to be the only one consistent, with an F-1 score of 0.5 in both implementations.

Due to a poor representation of datasets in third-party scenarios and a lack of benchmarks to assess the quality of the proposed corpora, **SQ6** cannot be answered as conclusions may not hold in a more comprehensive setting with more samples of literature. This can be addressed in future studies that examine the body of literature that describes the model-creation and assessment process, aggregating data that would convey a potential correlation between a performance metric and a group of datasets, or even a set of corpora-specific characteristics.

5 Responsible Research

Publications are always used to influence study results. All parties involved in research must recognize the potential ethical risks and understand the moral responsibilities that such work entails. The current study must adhere to the universal set of principles of responsible research [59]. For these reasons, this section will review how the research methodology complies with best practices and the ethical implications of affective content analysis systems when applied in everyday activities.

5.1 Reflection upon the study

Transparency is a fundamental value that researchers must uphold. A systematic review improves this quality; however, the extent of reproducibility may vary slightly. The uncertain nature of the discrepancy stems from the possibility of errors propagating into the research or the dynamic state of the search engines. The number of results displayed in Figure 1 may not match the identification date, as literature can be added or retracted during the process. Bias caused by a single person conducting research can be problematic, especially in the context of a literature review. Every possible unscientifically justified decision has been provided, allowing for an assessment of decision subjectivity. However, the used framework, PRISMA, improves the process of reproducing experiments by requiring strong motivation for each decision as well as a detailed description of the steps taken to arrive at the end result. Additionally, providing comprehensive descriptions, such as the PRISMA Guidelines, aids in maintaining a consistent process.

Another form of uncertainty might come from the selected subset of the literature. Due to time constraints, it was decided to limit the number of search engines used and apply additional constraints to the list of results. The set of literature excludes *ACM Digital Library*, an engine that contains proceedings of important conferences in this domain, as it did not provide any form of automated filtering, making it infeasible to finish manual screening on all 2293 results. As a consequence, it may not be the most complete representation of how datasets have performed throughout the development of affective systems. For example, as illustrated in section 4.4, there is a significant gap in the representation of how datasets were constructed between 1994 and 2007, with only one example in particular years between 2008 and 2017. However, even on a smaller scale, this study provides a good indication of how supervised affect prediction system performance correlates with annotators' agreement.

5.2 Ethical viewpoint on affect prediction

While other forms of input can express information more directly, text is more easily interpretable and vague than videos or audio snippets. Affect prediction models rely heavily on labeled data. While annotators must select from a predefined set of affections, each sentiment is influenced by the individual's experience, upbringing, and opinions [60]. This issue is raised due to the contrasting polarity of emotions depending on an annotator's background [61]. For example, if a dataset that uses news reports includes the headline "Switzerland won the 2024 Eurovision Song Contest", fans of the Swiss representative would label this as *Happiness*, but others who do not watch the performance would say *Neutral*, or those who would have wanted someone else to win would label as *Disappointment* or *Anger*. This results in a large number of uncontrollable variables (e.g., ethnicity or gender) that can influence the integration of such datasets into practical uses, namely the development and advancement of affect prediction systems. In a homogenous setting, these technologies can provide erroneous results for categories of people other than the ones included during labeling, especially when working with culturally contextualized data.

6 Discussion

When creating a new corpus, it is common to use multiple people to annotate records because self-reporting is unreliable and can result in a noisy dataset, such as **CrowdFlower**. Manual annotation focuses on label uniformity, as evidenced by the fact that 97.5% of datasets with multiple labelers compute this metric. As shown in section 4.3, there are many ways to calculate interrater agreement, assuming it is tracked. More "unconventional" methods were used to compute affect prediction in the early stages. Despite literature demonstrating concrete favoritism for *Cohen's* κ [27, 28], it appears that *Fleiss'* κ started becoming an unofficial norm for assessing uniformity in labeling. One possible explanation is that larger studies may prefer *Fleiss'* κ for convenience in calculating as it is suitable for groups of any size. Due to a lack of diversity, only 2 out of 41 datasets included some form of continuous dimension and one derives emotions from valence and arousal, the method chosen to quantify IRA could not be linked to any affect representation scheme.

Interestingly, there are publications of corpora with low IRAs being pursued, as the authors believe they're still useful for affect prediction models. Although there are some outliers with κ scores below 0.25, the majority of datasets register values above 0.45. Even if low agreement was detected, there was no second iteration of annotation with improved factors to obtain a better IRA. Those that did not specify any additional measures to improve agreement are distributed across a wide range of confidence levels, without linking a low level of uniformity to a refusal to facilitate agreement. The datasets that did not specify any measures also do not talk about any IRA values, which may infer that this part of corpus creation was not taken into consideration at all. 14% have a low score of agreement, but also 14% have a high score, with the rest recording moderate to high-moderate agreement.

Indeed, 36 different affect representation schemes are used across the 41 datasets returned by the search, which are not very different in terms of the polarity they convey, but rather in the set of emotions they choose to use. Most of the sets do not rely on a studied and well-known psychological theory but are just randomly chosen by the researchers when composing the procedure guidelines. This yields an unreliable conclusion about the relationship between a type of ARS and the degree of agreement. In some cases, one ARS is insufficiently represented to associate it with a high/lower core for IRA. The dispersed nature of the datasets also contributes to representation issues. There are 20 proprietary datasets, making it difficult to determine whether the dataset performs well because of a specific combination with the model developed by the same author or because of the corpus's characteristics.

When compared to other forms of content such as audio or video, text is the least descriptive and provides the least amount of context. Reading requires some prior knowledge of the subjects to which a text refers, as well as some imagination. The datasets include various types of content, such as news headlines (**SemEval-2007**), Twitter posts (**TEC**, **ArECTD**), poetry (**PERC**), and ancient metaphors (**C1**). The ease with which the underlying emotions can be understood is determined by the directness of the text, which researchers do not consider and cannot quantify when comparing levels of agreement.

Another limitation is represented by the researcher's opinion on a label for what an IRA value means. Cohen and Fleiss have presented in their publications [14, 15] how the values of their metrics should be interpreted, but in some cases, a writer's way of judging the significance of the computations doesn't match with how it should be done. We've seen this situation with the **T3** which considers a κ score of 0.26 to be acceptable, while the interpretation table would judge it as barely "fair agreement".

7 Conclusions and Future Work

The purpose of this study was to investigate the effect of human annotation during the corpus design process on the performance of text affect content analysis models, as well as to see if assessing interrater agreement (IRA) improves the end product. The survey examined who the emotions represent, what affect representation schemes (ARS) the identified datasets use, how the annotation process was carried out, the popularity of computing methods for IRA, whether there is a correlation between the ARS and the level of agreement, and whether good model performance scores and high IRA values are interdependent.

By conducting a systematic literature review adhering to the methodology introduced by the PRISMA 2020 framework, 41 papers were included in the study and an additional 10 were added to obtain detailed specifications of some mentioned datasets. The search query used on four literature engines, namely *Scopus*, *IEEE Xplore*, *ACM Digital Library*, and *Web of Science*, contains the intersection of the main topic of review: affect prediction, text, datasets, and manual labeling. Non-English literature that was used for multimodal affect prediction or sentiment analysis was manually excluded. On top of these, a couple more constraints were included due to the limited amount of time, such as the entire exclusion of *ACM Digital Library* and the limitation to only publication in the field of Computer Science.

The study outlines a connection between the introduction of humans into annotation jobs and the tendency to employ multiple people. The interrater agreement is almost all the time computed in non-self-reporting settings, but there is no description of how the values influence the corpus creation process in the end 84% of the datasets also include preventative measures that increase the chances of obtaining uniformity in labeling. Overall, most identified corpora have a high level of agreement. Based on the limited data available, there appears to exist significant variance in F-1 scores measured by different models trained on the same dataset, with some minor exceptions. This conclusion cannot be considered consistent because the number of usages per dataset is scarce, providing insufficient justification to consider one dataset superior to another.

Results should be evaluated with the presence of limitations in mind, such as the nature of the input and the intricate ways it can convey emotions, opposing views on a "good" rate of agreement, and what is entailed by such a fast-paced systematic review conducted by only one person. Future studies can address these factors to provide a comprehensive picture of the state of such corpora. It should be remembered that this endeavor constitutes only the initial stage towards answering the question "*How does interrater agreement influence the performance of text affect prediction models?*" due to the tight allocation of time and reduced number of people involved. The current efforts can be used to conduct a complete larger-scale research that assesses the relationship between the process of manually annotated corpus and the performance scores of models trained on such datasets by going more in-depth on **SQ6**. Future research can also look into how manual annotation affects multimodal affect prediction, which includes text as input, and compare it to the impact of singular-type input. Finally, the concept of standardization for computing and reporting the IRA should be proposed to easily assess worthy datasets and potentially suggest additional actions to improve the procedure if low values are encountered.

Addressing gaps in the study of corpus creation can potentially increase the efficiency of affect prediction models, resulting in more robust and accurate systems. Despite tackling a simple task like labeling text records, the process of creating corpora

illustrates numerous inconsistencies. This literature review's contributions invite for discovering the unknown regarding the importance of standardizing a lesser-known aspect of artificial intelligence that has a significant impact on the outcomes of the systems to which humans are exposed.

Acknowledgments

This endeavor would not have been possible without the assistance of my supervisor, who served as both an academic guide and mentor throughout the project. Your excitement for the field of psychology in intelligent systems has been really beneficial in understanding the significance of the research and has motivated me to delve deeper into this topic, despite the fact that I had little prior knowledge of it. The kindness with which you interact has made me look forward to our meetings and appreciate working with you. Your valuable comments have helped me keep the right trajectory and contributed to a great improvement of this thesis.

References

[1] David Reinsel-John Gantz-John Rydning, John Reinsel, and John Gantz. The digitization of the world from edge to core. *Framingham: International Data Corporation*, 16:1–28, 2018.

[2] Shazia Sadiq and Marta Indulska. Open data: Quality over quantity. *International Journal of Information Management*, 37(3):150–154, 2017.

[3] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111, 2014.

[4] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, 2020.

[5] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual affect sensing for sociable and expressive online communication. In *Affective computing and intelligent interaction: second international conference, ACII 2007 lisbon, Portugal, September 12-14, 2007 proceedings 2*, pages 218–229. Springer, 2007.

[6] Michal Ptaszynski, Pawel Dybala, Shinsuke Higuhi, Wenhan Shi, Rafal Rzepka, and Kenji Araki. Towards socialized machines: Emotions and sense of humour in conversational agents. *Web intelligence and intelligent agents*, 173, 2010.

[7] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3561–3562, 2020.

[8] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 4040–4041, 2021.

[9] Hima Patel, Fuyuki Ishikawa, Laure Berti-Equille, Nitin Gupta, Sameep Mehta, Satoshi Masuda, Shashank Mujumdar, Shazia Afzal, Srikanta Bedathur, and Yasuharu Nishi. 2nd international workshop on data quality assessment for machine learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4147–4148, 2021.

[10] Pero Subasic and Alison Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy systems*, 9(4):483–496, 2001.

[11] Flor Miriam Plaza-del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. *arXiv preprint arXiv:2403.01222*, 2024.

[12] Atul Soni, Chirag Arora, Rajkumar Kaushik, and Vaishali Upadhyay. Evaluating the impact of data quality on machine learning model performance. *Journal of Nonlinear Analysis and Optimization*, 14(01), 2023.

[13] Nasif Imtiaz, Justin Middleton, Peter Girouard, and Emerson Murphy-Hill. Sentiment and politeness analysis tools on developer discussions are unreliable, but so are people. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*, pages 55–61, 2018.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[15] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[16] R.A. Fisher. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd, 1925.

[17] Emily Öhman. Emotion annotation: Rethinking emotion categorization. In *DHN post-proceedings*, pages 134–144, 2020.

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

[19] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74, 2007.

[20] Klaus R Scherer and Harald G Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310, 1994.

[21] Bennett Kleinberg, Isabelle Van Der Vegt, and Maximilian Mozes. Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*, 2020.

[22] Mark A Finlayson and Tomaž Erjavec. Overview of annotation creation: Processes and tools. *Handbook of Linguistic Annotation*, pages 167–191, 2017.

[23] Magali Paquot and Stefan Th Gries. *A practical handbook of corpus linguistics*. Springer Nature, 2021.

[24] Niladri Sekhar Dash. *Language corpora annotation and processing*. Springer, 2021.

[25] William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

[26] Ole R Holsti. Content analysis for the social sciences and humanities. *Reading. MA: Addison-Wesley (content analysis)*, 1969.

[27] Michael E Dewey. Coefficients of agreement. *The British Journal of Psychiatry*, 143(5):487–489, 1983.

[28] Matthijs J Warrens. A comparison of Cohen's kappa and agreement coefficients by Corrado Gini. *International Journal of Research and Reviews in Applied Sciences*, 16:345–351, 2013.

[29] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2013.

[30] Plaban Kumar Bhowmick, Anupam Basu, and Pabitra Mitra. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65, 2008.

[31] Natasa Gisev, J Simon Bell, and Timothy F Chen. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338, 2013.

[32] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[33] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81, 2021.

[34] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[35] Laura Ana Maria Oberländer and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th international conference on computational linguistics*, pages 2104–2119, 2018.

[36] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372, 2021.

[37] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[38] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, page 1367âes, USA, 2004. Association for Computational Linguistics.

[39] Jane Dwivedi-Yu, Yi Chia Wang, Lijing Qin, Cristian Canton-Ferrer, and Alon Y. Halevy. Affective signals in a social media recommender system. pages 2831–2841. Association for Computing Machinery, 8 2022.

[40] Abdullah Alsaedi, Stuart Thomason, Floriana Grasso, and Phillip Brooker. Fb-sec-1: A social emotion cause dataset. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2023.

[41] Kaisla Kajava and Jorg Tiedemann. Xed: A multilingual dataset for sentiment analysis and emotion detection, 2020.

[42] Annika M. Schoene, Lana Bojanic, Minh Quoc Nghiem, Isabelle M. Hunt, and Sophia Ananiadou. Classifying suicide-related content and emotions on twitter using graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 14:1791–1802, 7 2023.

[43] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16*, pages 152–165. Springer, 2015.

[44] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[45] Plutchik Robert. *EMOTION: A Psychoevolutionary Synthesis*. Harper Row Publishers, 1980.

[46] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[47] Carroll E Izard. *Human emotions*. Springer Science & Business Media, 1977.

[48] Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. Text-based fine-grained emotion prediction. *IEEE Transactions on Affective Computing*, 2023.

[49] Despoina Chatzakou, Athena Vakali, and Konstantinos Kafetsios. Detecting variation of emotions in online activities. *Expert Systems with Applications*, 89:318–332, 12 2017.

[50] Muhammad Umair Arshad, Muhammad Farrukh Bashir, Adil Majeed, Waseem Shahzad, and Mirza Omer Beg. Corpus for emotion detection on roman urdu. 2019.

[51] Chris Van Pelt and Alex Sorokin. Designing a scalable crowdsourcing platform. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 765–766, 2012.

[52] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, and B. Adams. The emotional side of software developers in jira. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 480–483, Los Alamitos, CA, USA, may 2016. IEEE Computer Society.

[53] Sanghyub John Lee, Jong Yoon Lim, Leo Paas, and Ho Seok Ahn. Transformer transfer learning emotion detection model: synchronizing socially agreed and self-reported emotions in big data. *Neural Computing and Applications*, 35:10945–10956, 5 2023.

[54] Madhav Manohar Suresh, Nitish Chooramun, and Mhd Saeed Sharif. Real-time customer emotion analysis in e-commerce based on social media data: Insights and opportunities, 2023.

[55] Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. Learning from multi-annotator data: A noise-aware classification framework. *ACM Transactions on Information Systems*, 37, 2 2019.

[56] Hassan Hayat, Carles Ventura, and Agata Lapedriza. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22, 7 2022.

[57] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual affect sensing for sociable and expressive online communication, 2007.

[58] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Affect analysis model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17:95–135, 1 2011.

[59] Netherlands code of conduct for research integrity, 2018.

[60] Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22, 2022.

[61] Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, 2020.

[62] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

[63] Apoorva Singh and Sriparna Saha. Graphic: A graph-based approach for identifying complaints from code-mixed product reviews. *Expert Systems with Applications*, 216, 4 2023.

[64] Saif Mohammad. # emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, 2012.

[65] Sanja Stajner. Exploring reliability of gold labels for emotion detection in twitter. pages 1350–1359. Incoma Ltd, 2021.

[66] Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 13–23, 2017.

[67] Saif Mohammad and Felipe Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September 2017.

[68] Md Rakibul Islam and Minhaz F. Zibran. Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, 145:125–146, 11 2018.

[69] Abdulmohsen Al-Thubaity, Saif Alqahtani, Mohammed Alharbi, and Abdulrahman Aljandal. A saudi dialect twitter corpus for sentiment and emotion analysis. 2018.

[70] Dongyu Zhang, Hongfei Lin, Liang Yang, Shaowu Zhang, and Bo Xu. Construction of a chinese corpus for the analysis of the emotionality of metaphorical expressions. pages 144–150, 2018.

[71] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[72] Ankush Chatterjee, Nath Narahari, Meghana Joshi, and Puneet Agrawal. Semeval-2019 task 3: Emocontext contextual emotion detection in text, 2019.

[73] Mihaiela Lupea and Anamaria Briciu. Studying emotions in romanian words using formal concept analysis. *Computer Speech Language*, 57:128–145, 9 2019.

[74] Vimala Balakrishnan, Vithyatheri Govindan, Noreen Izza Arshad, Liyana Shuib, and Ernest Cachia. Facebook user reactions and emotion: An analysis of their relationships among the online diabetes community. *Malaysian Journal of Computer Science*, 2019:87–97, 2019.

[75] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[76] Sreeja P.S. and G.S. Mahalakshmi. Perc-an emotion recognition corpus for cognitive poems. 2019.

[77] P. S. Sreeja and G. S. Mahalakshmi. Using weighted directed graphs for identification of flow of emotions in poems. *Journal of Intelligent and Fuzzy Systems*, 39:2213–2227, 2020.

[78] P. S. Sreeja and G. S. Mahalakshmi. Emotion recognition in poetry using ensemble of classifiers. volume 922 CCIS, pages 77–91. Springer, 2019.

[79] Appidi Abhinav Reddy, Krishna Srirangam, Suhas Darsi, and Manish Shrivastava. Creation of corpus and analysis in code-mixed kannada-english twitter data for emotion prediction, 2020.

[80] Al-Khafaji Ali, J Askar, Nilam Nur, and Amir Sjarif. Annotated corpus of mesopotamian-iraqi dialect for sentiment analysis in social media.

[81] Zishan Ahmad, Asheesh Kumar, Asif Ekbal, and Pushpak Bhattacharyya. Emotion driven crisis response: A benchmark setup for multi-lingual emotion analysis in disaster situations. volume 2021-July. Institute of Electrical and Electronics Engineers Inc., 7 2021.

[82] Ahmed El-Sayed, Shaimaa Lazem, and Mohamed Abougabal. An arabic egyptian dialect covid-19 twitter dataset (arectd). pages 179–182. Institute of Electrical and Electronics Engineers Inc., 2021.

[83] Rhio Sutoyo, Harco Leslie Hendric Spits Warnars, Sani Muhamad Isa, and Widodo Budiharto. Emotionally aware chatbot for responding to indonesian product reviews. *ICIC Express Letters*, 19:861–876, 6 2023.

[84] Iqra Ameer, Grigori Sidorov, Helena Gomez-Adorno, and Rao Muhammad Adeel Nawab. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789, 2022.

[85] Jane Dwivedi-Yu, Meta Ai, and Alon Y Halevy. "that's so cute!": The care dataset for affective response detection, 2022.

[86] Farhat Ullah, Xin Chen, Syed Bilal Hussain Shah, Saoucene Mahfoudh, Muhammad Abul Hassan, and Nagham Saeed. A novel approach for emotion detection and sentiment analysis for low resource urdu language based on cnn-lstm. *Electronics (Switzerland)*, 11, 12 2022.

[87] Krishanu Maity, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. Exthaihate: A generative multi-task framework for sentiment and emotion aware hate speech detection with explanation in thai. volume 14174 LNAI, pages 139–156. Springer Science and Business Media Deutschland GmbH, 2023.

[88] Amirali Sajadi, Kostadin Damevski, and Preetha Chatterjee. Towards understanding emotions in informal developer interactions: A gitter chat study. pages 2097–2101. Association for Computing Machinery, Inc, 11 2023.

[89] Daniyar Amangeldi, Aida Usmanova, and Pakizar Shamoi. Understanding environmental posts: Sentiment and emotion analysis of social media data. 11, 2023.

[90] Usman Malik, Simon Bernard, Alexandre Pauchet, Clement Chatelain, Romain Picot-Clemente, and Jérôme Cortinovis. Pseudo-labeling with large language models for multi-label emotion classification of french tweets. *IEEE Access*, 2024.

[91] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Attitude sensing in text based on a compositional linguistic approach. *Computational Intelligence*, 31(2):256–300, 2015.

[92] Md Rakibul Islam and Minhaz F. Zibran. Deva: Sensing emotions in the valence arousal space in software engineering text. pages 1536–1543. Association for Computing Machinery, 4 2018.

[93] Bartłomiej Koptyra, Anh Ngo, Łukasz Radliński, and Jan Kocoń. Clarin-emo: Training emotion recognition models using human annotation and chatgpt. In *International Conference on Computational Science*, pages 365–379. Springer, 2023.

[94] Chang Su, Junchao Li, Ying Peng, and Yijiang Chen. Implicit mood computing via lstm and semantic mapping. *Soft Computing*, 24:15795–15809, 10 2020.

[95] T. Dhiliphan Rajkumar, Palagiri Yallareddy, Ediga Yoganand, Damera Rajkumar, and Gundlapalli Likith. Emotion detection in online social network- a multilabel learning process. Institute of Electrical and Electronics Engineers Inc., 2023.

A Scopus Filter Keywords

- "Affect Recognition"
- "Affective Computing"
- "Affective State"
- "Annotated Datasets"
- "Annotation"
- "Data set"
- "Dataset"
- "Emotion"
- "Emotions"
- "Emotion Analysis"
- "Emotion Classification"
- "Emotion Detection"
- "Emotion Prediction"
- "Emotion Recognition"
- "Emotional State"
- "Human Annotations"
- "Large Dataset"
- "Text Classification"

B Summary of Results[9]

LEGEND FOR TABLE 6:
3: Joy, Love, Sadness
3N: Happiness, Sadness, Anger + Neutral
4F: Joy, Sadness, Fear, Anger
4L: Joy, Sadness, Anger, Love
4DN: Excitement, Stress, Depression, Relaxation + Neutral
5H[10]: Anger, Amusement, Love, Surprise, Sadness
5ON: Sadness, Anger, Fear, Surprise, Optimism + Neutral
5DO: Anger, Trust, Sadness, Anticipation, Disagreeable + Other
6: Joy, Love, Sadness, Anger, Fear, Surprise
7G: Anger, Disgust, Fear, Joy, Sadness, Shame, Guilt
7A: Adoration, Entertained, Excitement, Sadness, Fear, Anger, Approving
7S: Adoration, Amusement, Approving, Excitement, Anger, Sadness, Fear
12: Warmth, Chill, Peace, Remoteness, Horror, Enthusiasm, Melancholy, Nostalgia, Romance, Magnificence, Elegance, Vitality
9: Love, Sadness, Anger, Hate, Fear, Surprise, Courage, Joy, Peace
9SN: Sarcasm, Hope, Joy, Surprise, Sympathy, Love, Sadness, Anger, Fear + Neutral
12N: Sadness, Relief, Worry, Hate, Enthusiasm, Happiness, Love, Amusement, Surprise, Boredom, Anger, Empty + Neutral
BET-6 (Ekman's Basic Emotion Theory [44]): Anger, Disgust, Fear, Happiness, Sadness and Surprise
BET-6S: BET-6 + Shame
BET-6IN: BET-6 + Unexploitable + Neutral
BET-6RJ: BET-6, but Anger is Rage and Happiness is Joy
BET-6NN: BET-6 + No emotion + Not sure
BET-7 (Ekman's Initial Basic Emotion Theory): Fear, Anger, Joy, Sad, Contempt, Disgust, Surprise
BET-5(Ekman's Basic Emotions without *Disgust*): Joy, Sadness, Fear, Anger, Surprise
BET-12: BET-6 + Enthusiasm, Rejection, Shame, Anxiety, Calm, Interest + Neutral
ST-6 (Shaver's Taxonomy [46]): Anger, Fear, Sadness, Joy, Love, Surprise
27N: 27 basic emotions identified by researchers at University of California [62] + Neutral
WoF-8 (Plutchik Wheel of Emotions [45]): Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust
WoF-11N: WoF-8 + Love + Optimism + Pessimism + Neutral
IToE (Izard's Theory of Emotions): Fear, Anger, Shame, Contempt, Disgust, Guilt, Distress, Interest, Surprise, and Joy

[9] A digital version is available.
[10] The original labels were: Angry, Haha, Love, Wow, Sad

Table 6: Summary of collected data included in the report grouped by targeted affective state and ARS

Targeted affective state	ARS	Introductory data paper	Dataset	Year	Emotion labels	No. raters	IRA measurement	Related papers
Emotion	Categorical	[20]	ISEAR	1994	7G	1/37	ANOVA	[55], [63]
		[64]	TEC	2012	BET-6	1	N/A	[65]
		[51]	CrowdFlower	2012	12N	Crowdsourcing	N/A	[53]
		[43]	Emotion-Stimulus	2015	BET-6S	4	Not specified	[63], [53]
		[66]	SSEC	2017	WoF-8	3-6	Cohen's κ	[53]
		[67]	EmoInt-2017	2017	4F	Not specified	Spearman & Pearson Correlation	[63], [53]
		N/A	Proprietary Twitter (T3)	2017	BET-12	4	Fleiss' κ	[49]
		[52]	JIRA Database	2018	3/3N/6	16/3	Cohen's & Fleiss' κ	[68]
		N/A	SDTC	2018	BET-6NN	3	Siegel & Castellan κ	[69]
		N/A	Proprietary Chinese (C2)	2018	BET-7	7	Fleiss' κ	[70]
		[71]	SemEval-2018	2018	WoF-11N	7	Fleiss' κ	[53]
		N/A	EmoContext	2019	3N	7	Fleiss' κ	[72]
		N/A	RomEmoLex	2019	ST-6	2-3	Fleiss' κ	[73]
		N/A	Proprietary Urdu (U2)	2019	3N	4	Cohen's κ	[50]
		N/A	Proprietary Facebook (FB2)	2019	BET-5	7	Krippendorff α	[74]
		[75]	GoEmotions	2020	27N	3-5	Mathew's Correlation	[48], [53], [54]
		N/A	XED	2020	ST-6	3	Cohen's κ	[41]
		N/A	PERC	2020	9	10	Fleiss' κ & Central Limit Theorem	[76], [77], [78]
		N/A	Kannada-English	2020	BET-6	2	Cohen's κ	[79]
		N/A	Proprietary Facebook (FB1)	2021	BET-7	2	Cohen's κ	[80]
		N/A	Proprietary Disaster Data (DD1)	2021	5ON	3	Cohen's κ	[81]
		N/A	ArECTD	2021	9SN	5	Fleiss' κ	[82]
		N/A	Indonesian Amazon Reviews	2021	ST-6	3	Fleiss' κ	[83]
		N/A	CM-MEC-21	2022	WoF-11N	3	Cohen's κ & % pairwise agreement	[84]
		N/A	CARE	2022	7S	3	Fleiss' κ	[85]
		N/A	Proprietary Urdu (U1)	2022	BET-6	Not specified	Discard record if raters didn't agree	[86]
		N/A	Proprietary Social Media (SM2)	2022	7A	5	Not specified	[39]
		N/A	Ex-ThaiHate	2023	5DO	3	Fleiss' κ	[87]
		N/A	GitterCom	2023	ST-6	2	Cohen's κ & iterative label reevaluation	[88]
		N/A	FB-SEC-1	2023	5H	3	Cohen's & Fleiss' κ	[40]
		N/A	Proprietary Social Media (SM1)	2023	ST-6	6	Cohen's κ	[89]
		N/A	Proprietary Review (R1)	2023	BET-6	3	Fleiss' κ	[63]
		N/A	Proprietary Twitter (T1)	2024	BET-6IN	11	Krippendorff α, Fleiss' κ & heuristics	[90]
	Hybrid	[19]	SemEval-2007	2007	BET-6	6	Pearson Correlation	[53], [56]
		N/A	Affect Database	2007	IToE	3	Fleiss' κ	[57], [58], [91]
		N/A	TWISCO	2023	BET-6	3	Fleiss' κ	[42]
	Dimensional	N/A	Proprietary Jira (J1)	2018	4DN	3	Fleiss' κ	[92]
Opinion	Categorical	N/A	CLARIN-Emo	2023	WoF-8	6	Krippendorff's α	[93]
Mood	Categorical	N/A	Proprietary Chinese (C1)	2020	12	3	Fleiss' κ	[94]
		N/A	Proprietary Twitter (T2)	2023	BET-6RJ	Not specified	Not specified	[95]

C Use of Large Language Models

C.1 ChatGPT

The free version of ChatGPT[11] was used mostly for formatting purposes. As Latex is quite a powerful text editor with lots of libraries, the prompts it received were questions or tasks that would have required searching for specific combinations of commands that would not generate any type of related content. Other tasks are related to generating graphs or verifying data inside tables for duplicity.

EXAMPLE OF PROMPT GIVEN TO CHATGPT:

"Decrease the font size of a caption in Latex"

RESPONSE[12] GENERATED BY CHATGPT ON 17TH OF MAY 2024:

```
\usepackage{caption}

\begin{minipage}{0.5\textwidth}
    This is some text inside a minipage.
\end{minipage}%
\begin{minipage}{0.5\textwidth}
    \centering
    \includegraphics[width=\textwidth]{example-image-a}
    \captionsetup{font=footnotesize} % Set the font size of the caption
    \caption{A figure inside a minipage}
    \label{fig:minipage}
\end{minipage}
```

C.2 QuillBot

QuillBot[13] was used during the writing of this report for rephrasing or finding synonyms in order to avoid repetition. This LLM is only capable of preserving the same content it was given, without generating any new information. The settings used were *Frequency* with the minimum level of synonyms.

EXAMPLE OF PROMPT GIVEN TO QUILLBOT:

"The authors' approach is different from usual practices, as they do not encode these dimensions using non-discrete values, but instead continue to use predefined labels."

TEXT GENERATED BY QUILLBOT ON 7TH OF MAY 2024:

"The authors' approach differs from usual practices in that they do not encode these dimensions using non-discrete values, but instead use predefined labels."

[11] ChatGPT: https://chatgpt.com

[12] The response only contains the most important information, namely the content that was requested, as ChatGPT generates all the necessary lines for rendering a valid Latex template; however, as the paper is being formatted in Latex, if it were to add the full response, the compiler would think the example response is part of the template we are currently editing and crash.

[13] QuillBot: https://quillbot.com/paraphrasing-tool