Quality and Agreement in Multilabel Emotion Annotation: A Case Study and Evaluation Framework

Emily Ohman & Anna Koufakou
Waseda University, Florida Gulf Coast University
Tokyo, Japan; Fort Myers FL, USA
ohman@waseda.jp, akoufakou@fgcu.edu

Abstract

Emotion annotation is inherently subjective, yet most NLP pipelines still assume "gold" labels, typically produced by majority voting, and treat annotator variation as noise. In this paper, we present a multilabel emotion annotation case study and use it to examine how annotator behavior and aggregation choices affect both agreement estimates and downstream emotion classifiers. Rather than collapsing disagreement into a single label, we represent targets as soft vote-share labels (including an intensity-weighted variant) and evaluate models using both thresholded metrics (macro-/micro-F1) and probabilistic alignment (Bernoulli cross-entropy SoftBCE), alongside data-derived disagreement diagnostics. Across annotation regimes, we show that disagreement is structured and leaves measurable traces in model behavior: hard labels may maximize F1 metrics, while soft supervision yields predictions that better reflect empirical annotator variance and uncertainty. Our results provide practical guidance for designing, aggregating, and evaluating multilabel emotion datasets when multiple interpretations are plausible.

Keywords: emotion annotations, annotator disagreements, perspectives, evaluation metrics

1. Introduction

The difficulty of annotating for emotions is widely acknowledged in the field of emotion detection and sentiment analysis (see e.g. Andreevskaia and Bergler, 2007; Bermingham and Smeaton, 2009; Mohammad, 2016; Munezero et al., 2014; Strapparava and Mihalcea, 2010; Wiebe and Riloff, 2005; Öhman, 2021). Recent reviews also summarize the research landscape, available resources, challenges and gaps in this area (Plaza-del Arco et al., 2024; Koufakou and Nieves, 2025). In the following, we focus on specific aspects and challenges related to our work.

Overall, NLP research typically uses basic emotions (see e.g. Ekman, 1971; Plutchik, 1984) or simplifies fine-grained frameworks (see e.g. Cowen and Keltner, 2017; Öhman, 2020), relying on classification models with a small set of largely disjoint labels. This facilitates model training and improves performance: for example, the GoEmotions (Demszky et al., 2020) experiments showed higher performance when the authors merged and remapped fine-grained categories to basic emotions (Ekman). These techniques, however, fail to capture the full, fine-grained, and often overlapping nature of human emotional experience.In addition to discrete-category theories, affect is also commonly modeled in dimensional terms such as valence–arousal–dominance (VAD) and related circumplex accounts (Russell and Mehrabian, 1977).

Equally notable is that much of the related research in NLP relies on the notion of "gold" labels, typically derived through majority voting among annotators. Due to the highly subjective nature of interpreting emotions, let alone emotions in a medium like text, humans usually do not agree on emotion labels. When annotators disagree, the corresponding records are often discarded. As Plank (2022) notes, "human variation in labeling is often considered noise." Yet this raises an important question: *is such variation truly noise*? Aside from cases of genuinely careless or inconsistent annotators, whose contributions may indeed warrant removal, disagreement among humans often indicates that an item is open to interpretation. In such cases, rather than enforcing a single "correct" label, we might instead expect NLP models to reflect this nuance in their predictions.

Agreement and alignment among annotators is measured by Inter-Annotator Agreement (IAA). There are several metrics used to assess IAA; most common are Cohen's κ, Fleiss' κ, and Krippendorff's α. The choice of IAA metric is influenced by various factors, for example the complexity of the annotation scheme (e.g. binary vs multi-label), the number of annotators, and how the work is assigned to the annotators. Regardless of the IAA metric, highly subjective tasks such as perceived emotions might lead to moderate or even low IAA. On the other hand, high agreement does not always mean high-quality annotations, for example, annotators could consistently agree on the wrong label. Besides the agreement between annotators, there might also be questions as to removing or filtering certain labels based on annotator effort or quality, which are not as straightforward to answer as we show in our case study (see Section 2).

Instead of eliminating disagreement, embracing

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 1–15
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

it can reveal the range of emotional interpretations a text may evoke. *Perspectivism* in NLP (Basile et al., 2021; Cabitza et al., 2023; Frenda et al., 2025) challenges the assumption of a single, objective ground truth and recognizes that multiple valid interpretations arise from annotators' diverse backgrounds, experiences, and perspectives. By preserving disagreement, *perspectivism* enables more inclusive and representative emotion recognition systems and reframes annotation not merely as labeling but as a means of exploring the full spectrum of emotions relevant to a task. In the past, only a few datasets released full individual annotator labels, e.g. (Demszky et al., 2020), but this is starting to change[1](Barz et al., 2025; Muhammad et al., 2025). Additionally, there have been efforts to gain insights into the annotator disagreements, for example in environmental communications (Barz et al., 2025), while Weber-Genzel et al. (2024) showed the challenges of distinguishing between human label variation that may be important to preserve and annotation errors.

To preserve human disagreement, rather than focusing on traditional evaluation metrics defined over gold or *hard* labels, prior work has proposed the use of *soft* labels, for example, see earlier SemEval Tasks *Learning with Disagreement* (Leonardelli et al., 2023, 2025) based mostly on data with binary or Likert scale labels. To the best of our knowledge, soft-label approaches have not yet been explored specifically for emotion annotation.

In this paper, we present a multilabel emotion annotation case study and examine how soft labels can be used to preserve annotator disagreement, inform annotation quality assessment, and support appropriate NLP model evaluation. The case-study anchors the following contributions we make:

- A critical review of annotator agreement metrics and common practice of assigning hard labels for NLP
- A multilabel emotion annotation case study illustrating real disagreement patterns and questions of annotator quality
- A soft-label framework that preserves annotator disagreement, together with an evaluation framework and metrics specifically aligned with soft labels in a multilabel setting, covering both emotion categories and intensity annotations.

Section 2 introduces the case study, followed by our methodology for aggregating multi-annotator labels in Section 3.1, and then the evaluation framework for agreement analysis and downstream NLP experiments in Section 4. Finally, we provide our concluding remarks in Section 5.

[1]We release all processing code and derived annotation tables keyed by sentence index (without redistributing copyrighted text). The experiments can be reproduced by applying the scripts to a locally obtained copy of the novel. The code is available at https://github.com/esohman/SoftBCE_Emotions.

2. Case Study

We present a case study of three annotators annotating Hemingway's *The Old Man and the Sea* (in English) using an expanded version of Plutchik's wheel of emotions.

Field	Value
Resource name	The Old Man and the Sea emotion annotations (TOMATS)
Resource type	Sentence-level multilabel emotion annotations with intensity
Language	English
Unit of annotation	Roughly sentence-level text segments
Size	1,677 data points (sentences)
Annotators	3 expert annotators
Label set (core)	{anger, anticipation, disgust, fear, joy, sadness, surprise, trust}
Additional labels	Optional free-text "other" emotions/notes (not modeled)
Intensity	Pre-selected-emotion rating on 0–10; rescaled to [0,1] for modeling
Annotation format	Spreadsheet (one row per item; up to 3 emotions + intensities + other)

Table 1: Resource card for the TOMATS emotion annotation case study.

2.1. Dataset and Annotation Protocol

All three annotators were research assistants of the first author and native speakers of English, majoring in NLP or adjacent fields. Annotators 2 and 3 received most of their education in the US, while Annotator 1 attended international schools with English as the primary language of instruction. Although demographically similar, women in their mid-20s, they differ in cultural upbringing and educational exposure, which we expected to influence annotation style.

Ideally, a broader demographic profile (for example, including older and male annotators) would have improved diversity, but the present study uses the available annotators to highlight how even superficially "similar" annotators can diverge substantially in emotion annotation.

The annotators were asked to first read the text in full before considering the emotions and then start from the beginning to label each sentence[2] in context. They could assign multiple emotions or select

[2]roughly sentence level with minor splitting/merging where needed for coherence

"no emotion." For the core annotation, annotators selected from the eight Plutchik emotions: *anger, anticipation, disgust, fear, joy, sadness, surprise, trust*. Annotators could additionally enter free-text "other" emotions/notes when none of the core labels captured their interpretation. These "other" emotions are listed in Appendix B.

We also introduced Plutchik's dyad system to encourage reflection on secondary and tertiary emotions and allowed annotators to add additional categories or notes when they felt the taxonomy was insufficient. Annotators also rated the intensity of each selected emotion on a 0–10 scale. We used a 0–10 scale to provide annotators a familiar, high-resolution ordinal range that supports within-item comparisons (e.g., weak vs. strong evidence for a selected emotion). For modeling, we linearly rescaled intensities to [0,1] so they can be used directly as weights in vote-share aggregation while keeping all soft targets within the Bernoulli objective's natural range.

2.2. Annotator Behavior and Disagreement

Sum	primary	secondary	tertiary
A1	1593	1258	443
A2	184	13	3
A3	1369	1200	566

Table 2: Annotation counts for each annotator

Table 2 shows the number of data points annotated by each annotator. We can see that Annotator 1 and Annotator 3 annotated emotions at similar rates across the 1677 sentences and assigned comparable numbers of secondary and tertiary labels. Their primary-emotion agreement is 42%, but agreement rises to 62% when considering any overlapping emotion label.

Annotator 2 presents a different case. They only considered roughly 10% of the data to have any emotion-association and almost no secondary or tertiary emotion associations (13 and 3 respectively). A closer look at their annotations further reveal that not only did they not annotate many data points at all, the data points they did annotate were not, contrary to expectations, data points with clear high-intensity emotion-linked content. Taken together with similarly sparse completion patterns observed in other research tasks, this points to limited protocol engagement rather than an alternative emotional reading.

Preliminary agreement scores calculated using Krippendorff's α in Table 3 indicate low agreement. We include these metrics to highlight how unsuitable they are for multilabel emotions despite being commonly requested by reviewers.

Emotion	Krippendorff's α
Anger	0.1674
Anticipation	0.2160
Disgust	0.1632
Fear	0.1599
Joy	0.1984
Sadness	0.2793
Surprise	0.2075
Trust	0.0920

Table 3: Krippendorff's α for binary presence/absence of each Plutchik emotion.

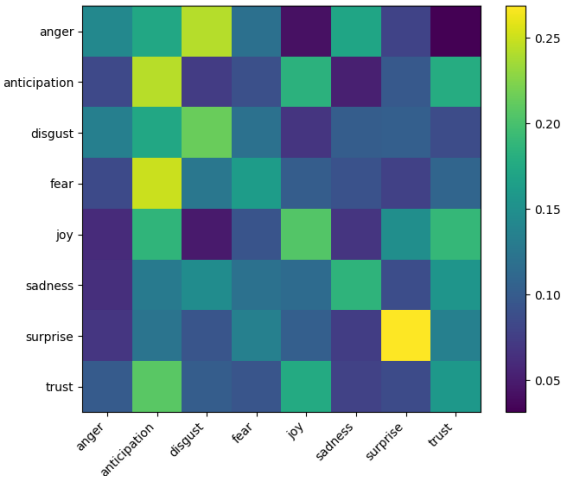

Figure 1: Annotation overlap (A1 vs A3)

To contextualize model confusion patterns, we additionally compute an annotation overlap matrix that captures systematic disagreement between annotators. We construct an emotion–emotion co-occurrence matrix (Fig. 1) by counting how often an emotion selected by one annotator overlaps with an emotion selected by another annotator for the same item, allowing up to three labels per annotator. The resulting matrix is row-normalized and visualized as a heatmap. We can see that *fear* is often confused with *anticipation*, and *anger* with *disgust* something that was also noted by e.g. Öhman et al. (2024).This representation does not encode correctness, but rather reflects structured ambiguity in the annotation process itself. Appendix A shows intra-annotator emotion overlap matrices.

2.3. Agreement Analysis and Implications

To assess annotation reliability, we compute pairwise agreement scores between the three annotators across all eight Plutchik emotions. Because the task is multilabel, each instance may exhibit zero or more emotions; classical inter-annotator agreement coefficients (e.g., Cohen's κ or Fleiss' κ) are therefore not directly applicable, as they assume

mutually exclusive categorical judgments. Instead, we adopt a per-emotion binary evaluation framework commonly used in multilabel affect annotation, where agreement is computed independently for each emotion (e.g., Mohammad and Turney, 2013; Mohammad and Bravo-Marquez, 2017) (see also Artstein and Poesio, 2008 for a general discussion of agreement in non-exclusive annotation tasks).

For each annotator pair (a, b), for emotion e, and item i, we consider binary labels

$$y_{i,e}^{(a)},\ y_{i,e}^{(b)} \in \{0, 1\}, \tag{1}$$

indicating the presence or absence of emotion e. We compute the per-emotion F1 score as this metric directly captures how consistently annotators agree on the presence of an emotion and is less sensitive to extreme class imbalance than accuracy. From the set of eight per-emotion F1 scores, we report:

- **macro-F1**: the unweighted mean of the per-emotion F1 scores;
- **micro-F1**: the F1 score computed after flattening all $8 \times N$ binary decisions into a single contingency table;
- **Per-emotion F1**: the individual agreement scores for each Plutchik emotion.

Annotator 2 left the majority of items unannotated ("none"). Here, "none"/blank entries denote no provided label (nonresponse/placeholder) rather than an explicit 'no emotion' judgment, which was available as a selectable option. Treating such blanks as negative judgments would artificially deflate agreement. To distinguish disagreement from missing information, we therefore measure agreement under two conditions: (i) on the full dataset, where blank entries count as negatives, and (ii) on a filtered subset consisting only of items where Annotator 2 produced at least one genuine emotion label. In the latter setting, we treat all other entries for Annotator 2 as missing.

In Table 4, we observe substantial differences in pairwise annotator agreement depending on whether the analysis is performed on the full dataset or restricted to the subset of items that Annotator 2 actually annotated with a genuine emotion label. On the full dataset, agreement between Annotator 2 and the other annotators is extremely low (macro-F1 $\approx$ 0.08), reflecting the fact that Annotator 2 left most items unannotated and thus systematically disagreed with the others. However, when we restrict the evaluation to the 184 items that Annotator 2 labelled, agreement rises sharply (macro-F1 $\approx$ 0.38-0.40), approaching the level of agreement observed between the two reliable annotators (Annotator 1 and Annotator 3: macro-F1 $\approx$ 0.49). This indicates that Annotator 2 is not inherently inconsistent; rather, the majority of their missing labels should be treated as missing data rather than as true negatives. Consequently, for all downstream modeling we treat Annotator 2's missing annotations as missing information rather than evidence of emotion absence, and only incorporate their annotations on items they explicitly rated.

As a sidenote, it should not be ignored that the low agreement scores are in part because the annotators, Annotator 3 in particular, utilized the "other emotion" column. For example, in one case Annotator 1 annotated the primary emotion as *anticipation* and *fear* as the secondary emotion. For that same data point Annotator 3 had left the primary slot empty and added *pessimism* as the other emotion. It could be argued that *pessimism* is indeed an overlap of *anticipation* and *fear*.

In another example, the sentence *"Chew it well, he thought, and get all the juices."* occurs during Santiago's (the protagonist) prolonged struggle at sea, he eats a small fish raw to maintain his strength. The internal monologue reflects pragmatic self-discipline and physical endurance rather than pleasure, emphasizing survival under extreme deprivation. Annotator 1 had selected *disgust, anger, anticipation* (in that order) and Annotator 3 selected *anger, anticipation*, and added *determination*. Again, the annotations seem aligned with a human interpretation, but harder to accurately show quantitatively.

3. Aggregating multi-annotator labels

Here, we describe our methodology for constructing hard and soft targets from human multilabel annotations.

Each instance is annotated by up to three annotators, who may select multiple emotions and optionally provide intensity ratings. Because Annotator 2 labels a smaller subset of instances, we treat missing annotations as missing rather than as explicit negative votes. For each instance i and emotion k, we compute a soft target as the empirical vote share among contributing annotators:

$$\tilde{y}_{i,k} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \mathbf{1}[k \in S_{a,i}] \tag{2}$$

where $S_{a,i}$ is annotator a's selected label set and $\mathcal{A}_i$ denotes annotators who provided at least one label for instance i. This produces multilabel targets $\tilde{y}_{i,k} \in [0, 1]$ that can be interpreted as per-label Bernoulli probabilities.

To incorporate intensity information, we additionally define an intensity-weighted soft target. For each annotator, selected emotions are weighted by their reported intensity (scaled to $[0, 1]$), and aggregated across contributing annotators. This preserves graded signal about strength of affect

Pair	Setting	macro-F1	micro-F1	Avg. Per-emotion F1	Min/Max Per-Emotion F1
A1–A2	Full	0.0788	0.0882	0.0795	0.0093 / 0.1760
A1–A2	Filtered	0.3949	0.4857	0.3932	0.1053 / 0.7015
A1–A3	Full	0.3794	0.3952	0.3804	0.2857 / 0.4980
A1–A3	Filtered	0.4899	0.5187	0.4870	0.3051 / 0.6410
A2–A3	Full	0.0829	0.0899	0.0820	0.0058 / 0.1469
A2–A3	Filtered	0.3797	0.4664	0.3830	0.0455 / 0.7706

Table 4: Summary comparison of pairwise annotator agreement in the full vs. filtered subsets.

while still producing multilabel targets in $[0, 1]$ suitable for Bernoulli objectives.

3.1. Targets and disagreement diagnostics

Let $y_{i,k}^{(a)} \in \{0, 1\}$ denote whether annotator a assigned emotion k to instance i, and let $\mathcal{A}_i$ denote the set of annotators who provided at least one label for instance i (i.e., missing annotations are treated as missing). We define the per-label vote share

$$p_{i,k} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} y_{i,k}^{(a)} \tag{3}$$

which yields multilabel soft targets $p_{i,k} \in [0, 1]$ that can be interpreted as Bernoulli probabilities. Hard targets are obtained by union-of-labels over contributing annotators:

$$y_{i,k}^{\text{hard}} = \mathbf{1}[p_{i,k} > 0] \tag{4}$$

To characterize ambiguity in the annotations independently of any model, we compute two data-derived disagreement measures from annotator label sets. First, we use mean per-label Bernoulli variance,

$$D_{\text{var}}(i) = \frac{1}{K} \sum_{k=1}^{K} p_{i,k}\big(1 - p_{i,k}\big) \tag{5}$$

which is maximized when annotators split evenly on a label and minimized when they agree ($p_{i,k} \in \{0, 1\}$). Second, we compute mean pairwise Jaccard disagreement between annotators' label sets,

$$D_{\text{Jac}}(i) = 1 - \frac{1}{\binom{|\mathcal{A}_i|}{2}} \sum_{a<b} \frac{|S_{a,i} \cap S_{b,i}|}{|S_{a,i} \cup S_{b,i}|} \tag{6}$$

where the sum is taken over all annotator pairs (a, b) contributing to instance i. These measures are computed from annotations alone and are therefore constant across model variants trained under the same annotator-inclusion regime.

Because our downstream models operate over the 8 Plutchik emotions, we filter free-text "other" labels and compute soft-label distributions over this fixed set. This effectively reallocates probability mass from custom labels to the remaining categories via renormalization, which can reduce apparent disagreement (lower Dvar) in cases where annotators expressed qualitatively different states outside the taxonomy. Future work could mitigate this by introducing an explicit "other" class, mapping free-text labels into a hierarchical taxonomy, or analyzing Dvar both with and without the "Other" bucket.

3.2. Bernoulli cross-entropy for multilabel soft targets

Because our task is multilabel, we model each emotion as an independent Bernoulli variable. Let $\hat{p}_{i,k} = \sigma(z_{i,k})$ denote the model's predicted probability for label k (with logit $z_{i,k}$). We quantify probabilistic alignment between model predictions and the aggregated soft targets $p_{i,k}$ using Bernoulli cross-entropy (binary cross-entropy; BCE):

$$\text{SoftBCE}(i, k) = -\Big(p_{i,k} \log \hat{p}_{i,k} + \big(1 - p_{i,k}\big) \log\big(1 - \hat{p}_{i,k}\big)\Big) \tag{7}$$

We report SoftBCE averaged across labels and instances as an auxiliary model–annotation alignment metric; lower values indicate closer alignment. In addition to SoftBCE, downstream models are evaluated using macro-/micro-F1 with decision thresholds tuned on a validation split. Similar cross-entropy-style diagnostics have been used in prior disagreement-oriented work (e.g., Leonardelli et al., 2023, 2025), but our setting differs in that targets are multilabel Bernoulli vote shares (with missingness treated explicitly), rather than single-label distributions.

Concrete stratified examples of the resulting hard and soft targets (low/medium/high disagreement) are provided in Table 5.

4. Evaluation Pipeline

Our evaluation pipeline is designed to examine how annotation behavior and disagreement structure affect both agreement estimates and downstream emotion classification. Rather than optimizing predictive performance, the goal is to assess how different assumptions about annotation completeness and disagreement propagate through the modeling process.

The pipeline consists of four stages.

(1) Label extraction and harmonization. Annotations are provided in separate spreadsheet sheets, one per annotator, with up to three Plutchik emotion labels per item and optional free-text notes. For each annotator a and item i, we extract the set of explicitly assigned Plutchik emotions, discarding "none", blanks, and non-Plutchik entries. This yields a mapping $(i) \rightarrow \{e_1, e_2, \ldots\}$, that reflects only positive emotion assignments.

(2) Multilabel representation. We convert the extracted labels into a binary multilabel tensor

$$Y \in \{0,1\}^{A \times N \times K}$$

where $Y_{a,i,k} = 1$ indicates that annotator a assigned emotion k to item i. This representation treats each annotator as an independent multilabel classifier and supports both pairwise agreement analysis and model training. Missing annotations (i.e., annotator–item pairs where no labels were provided) are encoded as empty label sets and tracked via a contribution mask, which is used in the next stage when aggregating vote shares over the set of contributing annotators $\mathcal{A}_i$.

(3) Agreement analysis with missing-label filtering. Because one annotator exhibits systematic sparsity, interpreting blank entries as negative judgments would conflate missingness with disagreement. We therefore compute agreement under two conditions:

1. **Full dataset:** all missing labels are treated as "no emotion".
2. **Filtered dataset:** items for which Annotator 2 assigned at least one genuine Plutchik emotion are retained; all other items are removed for all annotators.

This comparison allows us to disentangle structural disagreement caused by missing labels from genuine divergence in emotional interpretation.

(4) Downstream modeling and evaluation. To assess whether differences in annotation behavior propagate into downstream models, we evaluate a small set of representative classifiers under different annotation regimes. The goal is not to optimize predictive performance, but to use modeling behavior as a diagnostic lens on annotation consistency. We further introduce an evaluation framework for soft labels in multilabel emotion annotation and demonstrate its practical application.

We consider four classifier families:

(a) a transformer-based multilabel classifier trained with Bernoulli cross-entropy (binary cross-entropy) on aggregated hard targets;

(b) the same architecture trained with Bernoulli cross-entropy using vote-share soft targets $\tilde{y}_{i,k} \in [0,1]$, thereby preserving annotator disagreement in the supervision signal;

(c) the same architecture trained with intensity-weighted soft targets, where vote shares are modulated by annotator-provided intensity scores (scaled to $[0,1]$);

(d) a one-vs-rest SVM baseline using frozen transformer embeddings.

Each model is trained under two annotation conditions: using all available annotators, and using only the two annotators identified as reliable by the filtered agreement analysis. We report both thresholded classification metrics (macro-/micro-F1) and probabilistic alignment metrics. Data-derived disagreement measures (e.g., D_{var} and D_{Jac}) are computed from annotations only, while SoftBCE denotes the Bernoulli cross-entropy between model probabilities and the soft targets and is used as an auxiliary model–annotation alignment metric (see subsections 3.1 and 3.2 respectively).

4.1. Training objectives

We train all transformer models with a multilabel Bernoulli objective (`BCEWithLogitsLoss`). In the hard-label condition, targets are binary vectors aggregated from annotator label sets. In the soft-label conditions, we replace binary targets with either (i) vote-share soft targets $\tilde{y}$ or (ii) intensity-weighted soft targets, yielding a soft-target variant of BCE that exposes annotator uncertainty and graded affect strength during training. Examples are stratified by D_{Jac} tertiles (computed from annotator label sets).

We report both thresholded classification metrics (macro-/micro-F1) and probabilistic alignment metrics. While F1 evaluates discrete decisions after thresholding, SoftBCE evaluates the full predicted probability vector against the aggregated soft targets; improvements in SoftBCE therefore indicate better-calibrated probabilities that more faithfully reflect inter-annotator variation.

Table 6 lists implementation details for reproducibility.

All results are from a single run with fixed seed; we did not perform multi-seed averaging.

4.2. Evaluation

Table 7 compares BERT models trained with hard targets versus two soft-target variants, under settings that either include or exclude Annotator 2. Hard-label training yields the highest F1 overall when Annotator 2 is included (macro-F1 = 0.616; micro-F1 = 0.625), but the intensity-weighted

Example text	Annotations	Hard labels	Soft labels
Low disagreement			
"That's very kind of you," the old man said.	A1: Sa, Su, Tr A2: Jo, Tr A3: Jo, Tr	$[0,0,0,0,1,1,1,1]$	$[0,0,0,0,\frac{2}{3},\frac{1}{3},\frac{1}{3},1]$
"The bird is a great help," the old man said.	A1: Jo, Tr A2: Jo A3: Jo, Su	$[0,0,0,0,1,0,1,1]$	$[0,0,0,0,1,0,\frac{1}{3},\frac{1}{3}]$
Medium disagreement			
He was gone and the old man felt nothing.	A1: Sa A2: Sa A3: Su	$[0,0,0,0,0,1,1,0]$	$[0,0,0,0,0,\frac{2}{3},\frac{1}{3},0]$
Come up easy and let me put the harpoon into you.	A1: Ant, Jo, Tr A2: Ant A3: Ant, Di, Jo	$[0,1,1,0,1,0,0,1]$	$[0,1,\frac{1}{3},0,\frac{2}{3},0,0,\frac{1}{3}]$
High disagreement			
"I think perhaps I can too. But I try not to borrow. First you borrow. Then you beg."	A1: Fe, Sa, Tr A2: Di A3: Ant, Jo, Tr	$[0,1,1,1,1,1,0,1]$	$[0,\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3},0,\frac{2}{3}]$
"They say his father was a fisherman. Maybe he was as poor as we are and would understand."	A1: Ant, Jo, Sa A2: Tr A3: Su, Tr	$[0,1,0,0,1,1,1,1]$	$[0,\frac{1}{3},0,0,\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{2}{3}]$

Table 5: Stratified examples illustrating how hard and soft targets are derived from human annotations. Hard labels are the union-of-labels over annotators (1 if any annotator selected the emotion). Soft labels are vote-share targets (fraction of contributing annotators selecting each emotion). Vectors follow the fixed emotion order: [Ang:anger, Ant:anticipation, Di:disgust, Fe:fear, Jo:joy, Sa:sadness, Su:surprise, Tr:trust].

Component	**Setting**
Backbone	`bert-base-uncased`
Tokenizer	`AutoTokenizer` (uncased WordPiece)
Input length	Max length 256; truncation; padding to `max_length`
Optimizer	AdamW, learning rate 2×10^{-5}
Batch size	16
Epochs	Up to 10 with early stopping
Early stopping	Patience 2 on validation macro-F1; keep best checkpoint (min $\Delta = 10^{-4}$)
Threshold selection	Tune global threshold t on validation macro-F1; grid $t \in \{0.10, 0.15, \ldots, 0.90\}$
Data split	Random 80/20 train/validation split; shuffle=True
Random seed	42 (`numpy` + `torch`)
SVM baseline	One-vs-rest LinearSVC on frozen BERT [CLS] embeddings (embedding batch size 32; other params default)

Table 6: Model and training configuration used in all transformer experiments.

soft-label model remains competitive (macro-F1 = 0.607; micro-F1 = 0.617) while achieving substantially lower SoftBCE (0.548 vs. 0.681), indicating markedly better alignment between predicted probabilities and the aggregated soft targets. Uniformly weighted soft labels reduce F1 more noticeably, though they still improve SoftBCE relative to hard training. Excluding Annotator 2 slightly decreases F1 across objectives and also increases SoftBCE for the hard-label model, suggesting that Annotator 2 contributes useful signal despite partial coverage. Finally, the disagreement measures D_{var} and D_{Jac} are stable within each annotation regime, as expected because they are computed from the annotations themselves rather than model outputs; they confirm that the underlying level of annotator divergence is comparable across the compared training objectives.

Because macro-/micro-F1 are computed after binarizing predictions, they depend on a tuned decision threshold and primarily reflect ranking performance around that cutoff. SoftBCE, in contrast, evaluates the full vector of predicted probabilities against the aggregated soft targets without thresholding. The two metrics therefore capture complementary aspects of model quality: thresholded classification performance versus probabilistic alignment with annotator uncertainty.

Incorporating intensity information improves the soft-label objective. Intensity-weighted soft labels yield higher macro- and micro-F1 than uniformly weighted soft labels in both the *with A2* and *without A2* settings, and they also produce markedly lower SoftBCE. This suggests that intensity ratings provide a useful inductive bias that sharpens the

Setting	Model	Thr.	macro-F1	micro-F1	SoftBCE	D_{var}	D_{Jac}	$\overline{A}$
With A2	Hard labels	0.35	0.616	0.625	0.681	0.066	0.708	1.98
With A2	Soft labels (uniform)	0.10	0.589	0.599	0.577	0.066	0.708	1.98
With A2	Soft labels (intensity)	0.10	0.607	0.617	0.548	0.066	0.708	1.98
With A2	SVM baseline	-	0.503	0.519	-	-	-	-
Without A2	Hard labels	0.15	0.606	0.616	0.717	0.065	0.707	1.85
Without A2	Soft labels (uniform)	0.10	0.588	0.597	0.564	0.065	0.707	1.85
Without A2	Soft labels (intensity)	0.10	0.591	0.600	0.578	0.065	0.707	1.85
Without A2	SVM baseline	-	0.501	0.517	-	-	-	-

Table 7: Downstream results for BERT under different supervision targets and annotator-inclusion regimes. Thr. is the tuned decision threshold (validation set). SoftBCE is the Bernoulli cross-entropy between predicted probabilities and aggregated soft targets (lower is better). D_{var} denotes mean per-label Bernoulli variance $\frac{1}{K}\sum_k \tilde{y}_k(1-\tilde{y}_k)$, D_{Jac} denotes mean pairwise Jaccard disagreement (both data-derived; higher means more disagreement), and $\overline{A}$ is the average number of contributing annotators per item.

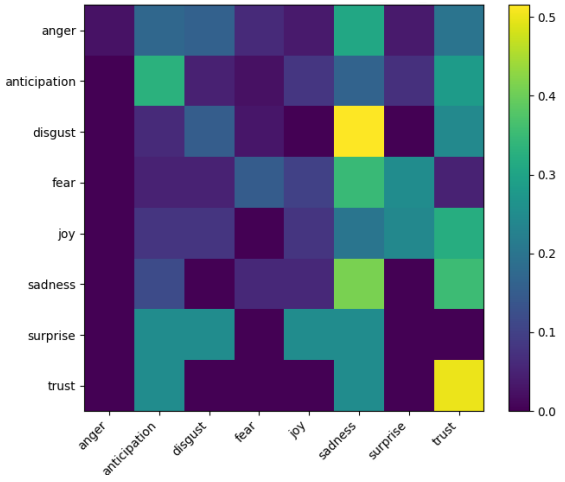

Figure 2: Confusion matrix for BERT with hard labels including A2 labels

soft targets by distinguishing weak affective cues from strong ones, improving both downstream decisions and probabilistic alignment with annotator uncertainty. Overall, these results support the use of soft labels as a diagnostic tool for understanding how annotation regimes propagate into model behavior, while highlighting that additional annotation detail does not automatically translate into improved downstream performance.

4.3. Confusion

Figure 2 presents a row-normalized single-label confusion matrix for the results from training BERT on hard labels with all annotators. Several emotions exhibit strong diagonal dominance, most notably *sadness* and *trust*, indicating that when these emotions are the primary label, the model predicts them consistently. The most prominent off-diagonal confusion occurs between *disgust* and *sadness*, with a substantial proportion of *disgust*-labeled instances predicted as *sadness*, suggesting overlap in negative-valence affective cues. This has also been shown in previous work (Rossi and Öhman, 2025). Other emotions, such as *anger* and *anticipation*, display more diffuse prediction patterns without a single dominant confusion partner. Overall, the matrix reveals clusters of related emotions rather than random error, supporting its use as a diagnostic tool for identifying structured affective overlap rather than discrete misclassification.

4.4. Error Analysis

To complement aggregate F1 scores, we inspect structured error patterns using a row-normalized single-label confusion matrix, obtained by projecting multilabel outputs to a dominant label per instance. Confusions are not random: the largest off-diagonal mass occurs between semantically and valence-adjacent emotions, including *sadness* vs. *fear*, *trust* vs. *anticipation*, and *anger* vs. *disgust*. These patterns align with the dataset's disagreement diagnostics (D_{var}, D_{Jac}), suggesting that model errors concentrate in regions of the label space that are also ambiguous for annotators. Predictive performance is weakest for rare emotions (e.g., *surprise*), consistent with challenges in multilabel learning under extreme imbalance.

5. Conclusions

This paper examined how annotation design choices and annotator reliability affect downstream modeling for multilabel emotion classification. Rather than treating annotation as a static preprocessing step, we used it as an object of analysis and asked how different aggregation strategies and training objectives interact with empirically observed disagreement. Across experiments, annotation decisions produced consistent, measurable differences in both thresholded performance (macro-/micro-F1) and probabilistic model-target

alignment (Bernoulli cross-entropy or SoftBCE).

The downstream results suggest that the benefits of soft supervision depend on what is being optimized. With all annotators included, hard-label training achieved the highest macro- and micro-F1 in our setup. Soft-label objectives, however, produced markedly lower SoftBCE while remaining broadly competitive in F1. In particular, intensity-weighted soft targets reduced SoftBCE the most under the full-annotator regime and narrowed the F1 gap to hard labels, indicating closer alignment with the uncertainty encoded in the aggregated targets. When we excluded Annotator 2, both hard and soft variants showed slightly lower F1, while the disagreement diagnostics (D_{var}, D_{Jac}) remained essentially unchanged, as expected for measures computed from the annotations alone.

These findings clarify a common claim in disagreement-aware learning. Preserving disagreement does not necessarily translate into higher F1, but it can yield probability estimates that better reflect the empirical label uncertainty. For affective tasks, where multiple interpretations can be simultaneously plausible and downstream applications often consume probabilities rather than hard decisions, this distinction is practically important. In this sense, SoftBCE provides a useful complementary lens to standard F1 when evaluating uncertainty-aware training signals.

At the same time, our results highlight limits to how additional annotation detail translates into downstream gains. Intensity ratings are not automatically helpful: their impact depends on how consistently annotators use the scale and on how intensities are normalized and aggregated. In this case study, incorporating intensity improved model-target alignment substantially but did not consistently dominate thresholded metrics, suggesting that intensity can encode graded uncertainty even when it does not directly improve binary decisions. This points to straightforward extensions, including annotator-specific calibration, learned normalization, and objectives that explicitly model individual differences in scale use.

The main contribution of this paper is methodological. Using downstream models as a diagnostic tool, together with disagreement measures computed directly from the labels, provides a practical way to surface design trade-offs in emotion annotation and to distinguish ambiguity in the data from annotator-specific behavior. Applying the same pipeline to additional texts and domains would help establish when these patterns generalize and when they are corpus- or task-specific.

We anticipate that applying this pipeline to short-form or informal text (e.g., posts or comments) will increase both overlap among emotions and annotator disagreement due to reduced context and higher ambiguity. Our approach is designed for precisely these conditions: missingness-aware aggregation separates nonresponse from negative judgments, and SoftBCE measures distributional alignment when "gold" labels are intrinsically uncertain. Future work can further test robustness under larger, crowd-sourced annotator pools and examine how disagreement patterns vary with domain and annotator diversity.

In conclusion, the results support treating annotation and modeling as coupled choices rather than sequential steps. By making disagreement explicit in both target construction and evaluation, and by reporting both thresholded and probabilistic alignment metrics, we provide a transparent framework for analyzing and improving multilabel emotion annotation practices.

Limitations

This work is intentionally framed as a case study, and several limitations are worth noting. First, the annotation effort involves a small number of annotators and a single literary text. This setting enables detailed qualitative and quantitative analysis of annotator behavior, but it limits the generalizability of absolute agreement values and downstream performance. Our conclusions therefore emphasize *patterns* of annotation behavior and their modeling consequences rather than population-level estimates of annotator reliability.

Second, the annotators were not drawn from a fully representative population. Although all were native speakers of English, they were relatively homogeneous and affiliated with the same research group. This reduces demographic variance, but also makes the observed divergence in annotation density and engagement noteworthy: substantial differences in labeling behavior emerged even within a narrow pool. Future work should test whether similar patterns hold with larger and more heterogeneous annotator groups.

Third, one annotator exhibited systematically sparse labeling behavior, which complicates both agreement measurement and target construction. While we address this through missingness-aware aggregation (treating non-annotations as missing rather than negative votes) and filtered analyses, the degree of missingness underscores the need for clearer guidelines and more explicit handling of abstentions in emotion annotation workflows. In particular, low agreement scores can reflect missing labels rather than conceptual disagreement, but distinguishing these cases is not always straightforward in practice.

Fourth, while we use Bernoulli cross-entropy (SoftBCE) as a model-target alignment metric for multilabel soft targets, there is no single universally

accepted evaluation measure for soft labels especially for multilabel settings, and different metrics emphasize different aspects of disagreement. Prior work has pointed out limitations of cross-entropy-style metrics in some soft-label settings (Rizzi et al., 2024), and recent shared-task work has also explored alternatives such as Manhattan and Wasserstein distances (Leonardelli et al., 2025). We therefore treat SoftBCE as an auxiliary diagnostic and report complementary data-derived disagreement measures; evaluating additional soft-label metrics for multilabel emotion annotation remains an important direction for future work.

Fifth, although the annotation scheme includes additional (rare) emotion categories beyond the main set analyzed here, we did not model these extra labels. Incorporating very low-frequency categories without destabilizing training or evaluation is non-trivial, and developing principled strategies for handling rare emotions (e.g., hierarchical labels, grouping, or targeted re-annotation) is a clear next step. Similarly, the intensity score reliability and coherence could be improved by adapting a best-worst-scaling approach (as demonstrated by Kiritchenko and Mohammad, 2017).

Finally, the classification models evaluated in this paper are deliberately simple and are used primarily as analytical tools rather than optimized systems. The goal is not to establish state-of-the-art performance, but to examine how different annotation regimes affect downstream modeling and evaluation. More sophisticated architectures or training strategies may yield higher performance, but they would not change the core methodological point: annotation behavior and aggregation choices can measurably shape both model outputs and how those outputs should be evaluated.

Ethical Considerations

This paper is primarily methodological: we analyze a small multilabel emotion annotation case study to understand how aggregation and disagreement-aware modeling choices affect evaluation and interpretation. The work does not involve user-generated content, sensitive personal data, or human subjects beyond a limited internal annotation exercise. Nonetheless, automated emotion recognition has broader societal implications that warrant explicit discussion.

Emotions are not directly observable from text, and the same utterance can support multiple plausible readings depending on context, culture, and reader perspective. Text-only emotion models are therefore best understood as predicting *annotated interpretations* rather than latent "true" emotions. This is especially important in settings where other signals (e.g., conversational context, prosody, facial expression, or situational metadata) are unavailable. Our results reinforce this point: disagreement is structured and often reflects genuine ambiguity rather than annotator error. Treating majority-voted labels as "gold" can mask uncertainty and overstate model reliability.

Emotion recognition models can be misapplied to surveillance, profiling, workplace monitoring, or high-stakes decision-making. These risks increase when outputs are presented as objective measures of an individual's internal state. We caution against such uses and recommend that, when emotion predictions are used at all, systems should expose uncertainty (e.g., calibrated probabilities) and be accompanied by clear documentation of what the labels represent, how they were collected, and where disagreement is expected (e.g., Mohammad, 2022).

Emotion categories, their lexical realizations, and their social meanings vary across languages and cultures. Models trained and evaluated on narrowly sampled annotator populations or single-language datasets may not transfer reliably to other contexts and may reproduce cultural biases. A practical implication of our work is that dataset documentation should report annotator composition, missingness patterns, and disagreement diagnostics, and that future annotation efforts should prioritize more diverse annotator pools and multilingual settings.

Because emotion annotation and interpretation are inherently tied to theories of affect, narrative, and communication, we encourage involving researchers beyond NLP (e.g., psychology, linguistics, HCI, digital humanities) in annotation design, guideline development, and evaluation choices. Disagreement-aware approaches should be informed by domain knowledge about when multiple interpretations are expected and meaningful, rather than treated as purely statistical noise.

Finally, any future release of materials derived from copyrighted literary sources will comply with licensing constraints. Where full text cannot be redistributed, we will document what can be shared (e.g., derived labels, code, and evaluation scripts) to support reproducibility while respecting rights holders.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K21058.

6. Bibliographical References

Alina Andreevskaia and Sabine Bergler. 2007. CLaC and CLaC-NB: Knowledge-based and

corpus-based approaches to sentiment tagging. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 117–120. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Christina Barz, Melanie Siegel, Daniel Hanss, and Michael Wiegand. 2025. Understanding disagreement: An annotation study of sentiment and emotional language in environmental communication. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 1–20.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Adam Bermingham and Alan F Smeaton. 2009. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 784–785.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

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

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Ernest Hemingway. 1995. The old man and the sea. 1952. *New York: Scribner*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Anna Koufakou and Elijah Nieves. 2025. Review of recent emotion-annotated text corpora and resources. *Language Resources and Evaluation*, pages 1–35.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *WASSA@ NAACL-HLT*, pages 174–179.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*, pages 65–77.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A.

Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

M. Munezero, C. Montero, E. Sutinen, and J. Pajunen. 2014. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing*, 5(02):101–111.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.

Emily Öhman. 2021. *The language of emotions: Building and applying computational methods for emotion detection for English and beyond*. Helsingin yliopisto.

Emily Öhman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. Emotionarcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 51–66.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives 2024*, pages 84–94. European Language Resources Association (ELRA).

Riikka Rossi and Emily Öhman. 2025. Combining qualitative and computational approaches for literary analysis of finnish novels. *Scandinavian Studies*, 97(3):27–51.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Carlo Strapparava and Rada Mihalcea. 2010. Annotating and identifying emotions in text. In *Intelligent information access*, pages 21–38. Springer.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.

A. Inter-annotator emotion co-occurence

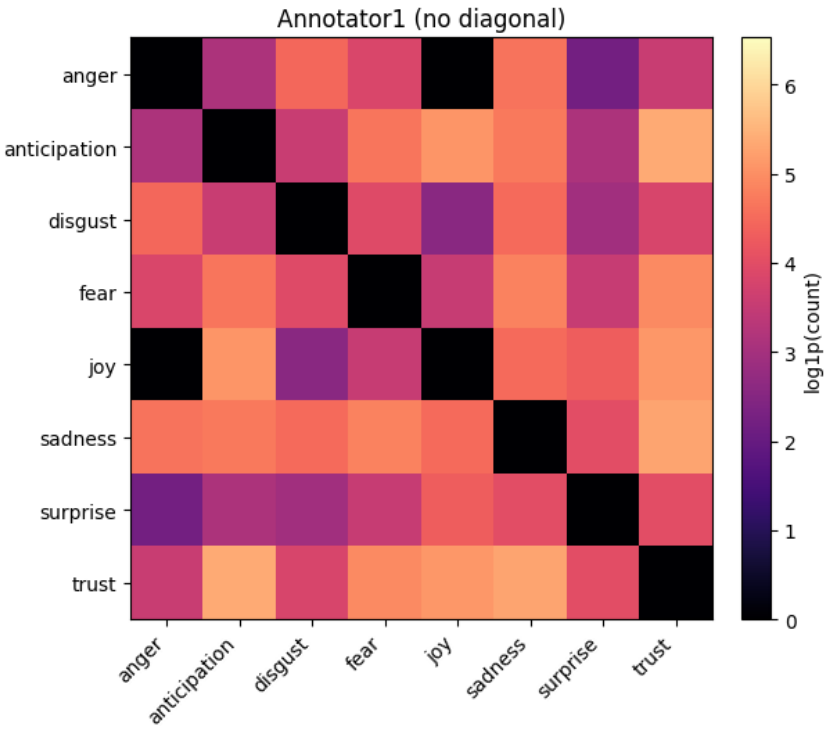

Figure 3: Annotator 1 emotion co-occurence

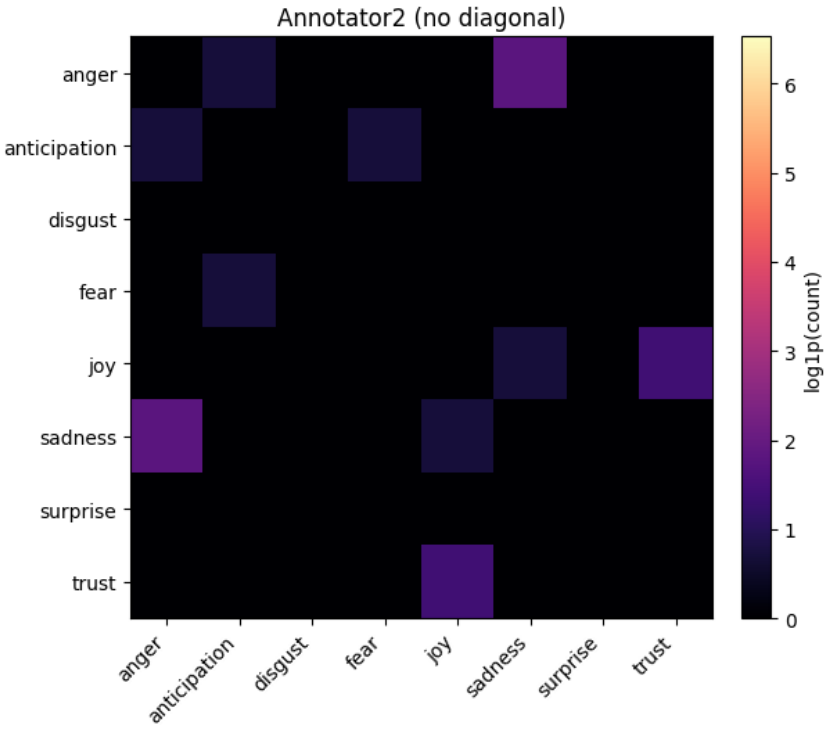

Figure 4: Annotator 2 emotion co-occurence

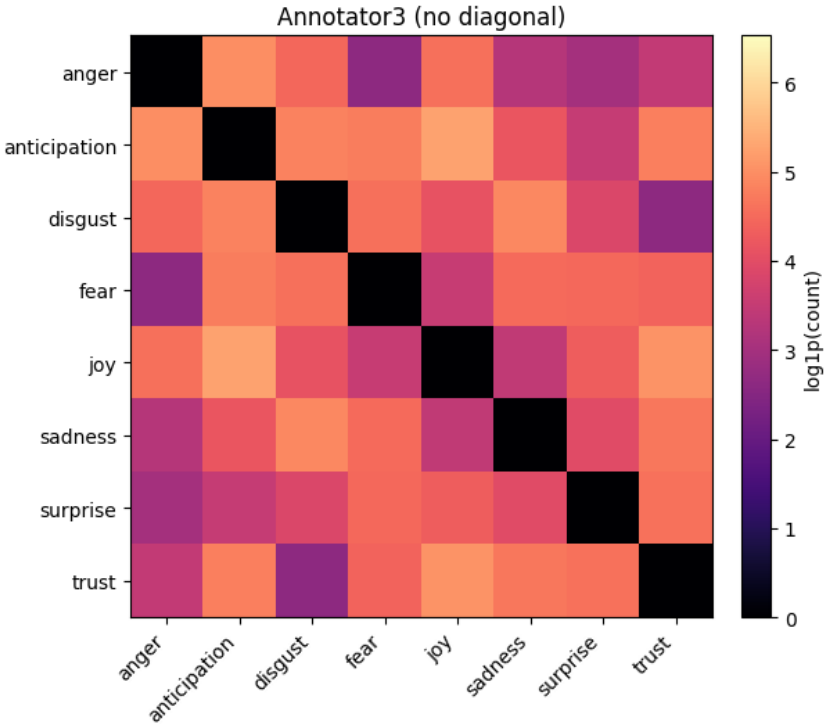

Figure 5: Annotator 3 emotion co-occurence

B. Top other emotions

Label	Count
sentimentality	6
appreciation	3
unbelief	3
confusion	3
pessimism	2
disapproval	2
contempt	2
mild delight	2
hope	2
mild despair	2
mild contempt	2
shame	2
mild submission	2
determination	2
love	1
optimism	1
caring	1
protectiveness	1
childish exasperation or mild annoyance	1
remorse	1
acceptance	1
submission	1
mild disapproval	1
mild submission and love	1
mild remorse	1
loving pride	1
guilt	1
mild aggression plus optimism	1
suspicion	1
doubtfulness	1

Table 8: Out-of-taxonomy emotion labels provided in the free-text *Other* field (frequency).

Table 8 shows the top 30 most commonly occurring other labels outside of Plutchik's categories.

Emotion pairs
disapproval; contempt
mild delight; hope
sentimentality; pessimism
love; optimism
caring; protectiveness
childish exasperation or mild annoyance
unbelief; confusion
shame, guilt
sentimentality; mild submission
mild aggression plus optimism

Table 9: Top emotion co-occurrences for *[Annotator X]* (off-diagonal counts).

Table 9 shows the top co-occurring other category items