Another Approach to Agreement Measurement and Prediction with Emotion Annotations

Quanqi Du, Véronique Hoste
LT3, Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

Emotion annotation, as an inherently subjective task, often suffers from significant inter-annotator disagreement when evaluated using traditional metrics like kappa or alpha. These metrics often fall short of capturing the nuanced nature of disagreement, especially in multimodal settings. This study introduces Absolute Annotation Difference (AAD), a novel metric offering a complementary perspective on inter- and intra-annotator agreement across different modalities. Our analysis reveals that AAD not only identifies overall agreement levels but also uncovers fine-grained disagreement patterns across modalities often overlooked by conventional metrics. Furthermore, we propose an AAD-based RMSE variant for predicting annotation disagreement. Through extensive experiments on the large-scale DynaSent corpus, we demonstrate that our approach significantly improves disagreement prediction accuracy, rising from 41.71% to 51.64% and outperforming existing methods. Cross-dataset prediction results suggest good generalization. These findings underscore AAD's potential to enhance annotation agreement analysis and provide deeper insights into subjective NLP tasks. Future work will investigate its applicability to broader emotion-related tasks and other subjective annotation scenarios.

1 Introduction

Despite the significant progress in multi-modal NLP (Garg et al., 2022), such as GPT-4o[1], accurately recognizing and interpreting human emotions across different modalities (Zhang et al., 2024) remains a substantial challenge. This difficulty primarily arises from the complexity and variability of emotional expressions (Lindquist and Barrett, 2008; Barrett, 2009), which often manifest themselves differently across modalities. Consequently, there is a growing demand for fine-grained and reliable datasets to support the training and evaluation of emotion recognition systems (Yang et al., 2023; Ridley et al., 2024).

As a common and popular practice, the use of evaluation metrics like the kappa/alpha family has almost become a standard step in dataset construction (Zhao et al., 2018). However, even with careful dataset design, many annotated (multimodal) emotion datasets exhibit low kappa/alpha scores (Busso et al., 2008, 2016; Zadeh et al., 2018; Zhao et al., 2022; Du et al., 2025), and few studies have explored the reason behind these low scores. Given that the interpretation of kappa/alpha values can be significantly influenced by factors such as the numbers of annotators and categories(Antoine et al., 2014), and considering the inherently subjective nature of emotion annotation (Chou et al., 2024; Plaza-del Arco et al., 2024; Maladry et al., 2024), we propose the complementary use of the Absolute Annotation Difference (AAD) as an intuitive metric to better measure and examine agreement and disagreement patterns, particularly in datasets with low kappa/alpha scores.

To validate this proposal, we conducted two experiments. The first is a pilot study on a small multimodal emotion dataset, where (dis)agreement was assessed using both kappa/alpha and AAD. The findings suggest that AAD provides a distinct perspective on (dis)agreement and effectively uncovers annotation patterns. Building on these insights, the second experiment applied AAD to (dis)agreement modelling and prediction, achieving an accuracy improvement of nearly 10%. Together, these experiments highlight the added value of AAD in enhancing the analysis and prediction of (dis)agreement in emotion annotation tasks.

By offering a complementary view to conventional metrics, our work contributes to a more nuanced understanding of annotation reliability. We hope this research can inspire further methodological innovation in dataset evaluation and design.

[1] https://openai.com/index/hello-gpt-4o/

2 Related Work

Many tasks in natural language processing and computer vision sometimes suffer from disagreement (Basile, 2020; Uma et al., 2021; Mostafazadeh Davani et al., 2022), as they involve tasks (e.g. emotion detection, hate speech detection) which are difficult to define and influenced by an annotator's cultural, social, ethnic, and other backgrounds. In addition, annotation differences might also just be caused by attention slips (Beigman Klebanov et al., 2008). In their survey paper, Uma et al. (2021) identified several sources of disagreement, including annotator errors, annotation schemes, ambiguity, subjectivity and item difficulty. Although disagreement is sometimes undesirable, there are also scholars embracing disagreement and proposing to preserve disagreement as different perspectives to the same stimuli (Akhtar et al., 2020; Plepi et al., 2022; Cabitza et al., 2023).

2.1 Disagreement Measurement

Irrespective of the provenance of this disagreement, annotation disagreement is usually measured with statistical approaches, such as Cohen's kappa (1960), Fleiss' kappa (1971) or Krippendorff's alpha (2007). According to Landis and Koch (1977), for categorical data, kappa values smaller than 0 are regarded as poor agreement, and these values can increase from slight (0.01 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60) and substantial agreement (0.61 to 0.80), up until 0.81 to 1.00 as almost perfect agreement. Kappa is usually used for categorical ratings, while Krippendorff's alpha is more adaptive with different levels of measurement (Stevens, 1946), able to measure agreement in nominal, ordinal, interval and ratio data (Krippendorff, 2011). As for Krippendorff's alpha, it is suggested to rely on data when the alpha is greater than 0.8, discard data when the alpha is smaller than 0.667, and only draw tentative conclusions when the alpha is in-between (Krippendorff, 2004).

Although the use of such metrics has become the de facto standard for agreement measurement – offering a single, comprehensive score to summarize overall agreement across a dataset – these metrics have notable shortcomings. For Kappa, the primary concerns are the prevalence problem and the bias problem (Di Eugenio and Glass, 2004), two major paradoxes that complicate its interpretation (Wang and Xia, 2019). Specifically, kappa values fluctuate significantly when category distributions are imbalanced or when annotators favour certain categories. Similarly, Krippendorff's alpha is not only affected by skewed category distributions but it is also highly sensitive to the choice of distance function and levels of measurement (Krippendorff, 2011).

In emotion annotation tasks, these limitations are even more pronounced. Emotion datasets often exhibit a natural skew toward more frequently used categories (Zadeh et al., 2018), and defining the appropriate levels of measurement for emotion annotations poses additional challenges. Emotions are commonly annotated using both categorical and dimensional labels (Busso et al., 2016; Labat et al., 2024), which can be interconverted under specific conditions(Park et al., 2021). While Antoine et al. (2014) advocate for the use of weighted Krippendorff's alpha as a more reliable metric for ordinal annotations, achieving the commonly accepted threshold of 0.667 (Landis and Koch, 1977) in emotion annotation remains elusive in empirical studies (Antoine et al., 2014; Wood et al., 2018). This difficulty has led to increased scrutiny of these metrics, particularly in subjective domains such as emotion annotation, where the interpretation of scores often comes into question(Wong et al., 2021).

To address these challenges, we propose the use of the intuitive Absolute Annotation Difference (AAD) method as a complementary approach to measure agreement and examine (dis)agreement patterns in emotion annotation tasks. As the name suggests, AAD refers to the absolute difference between two or more sets of annotations. For dimensional annotations, AAD can be straightforwardly calculated as the absolute difference between two annotations, which can be formulated as

$$D^i = |x_i - y_i|, \quad i \in \mathcal{M} \tag{1}$$

whereby x_i and y_i represent the assigned dimensional labels (i.e., valence values) respectively for the instance i in the dataset $\mathcal{M}$. For categorical annotations, we propose converting them into two- or multi-dimensional representations and computing Euclidean differences, as suggested by Antoine et al. (2014). For example, when categorical annotations are projected into the valence-arousal space, the absolute difference will be formulated as

$$D^i = \sqrt{(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2}, \quad i \in \mathcal{M} \tag{2}$$

whereby x_{i1} and x_{i2} correspond to the projected valence values and y_{i1} and y_{i2} denote the projected arousal values for the instance i in the dataset $\mathcal{M}$, respectively. This ADD approach offers another perspective on agreement and provides deeper insights into (dis)agreement patterns, particularly in datasets with low kappa or alpha scores.

2.2 Disagreement Prediction

In addition to measuring agreement after emotion annotation, an equally compelling question is whether, and to what extent, it is possible to predict disagreement before the annotation process. While previous studies have focused on predicting individual annotators' ratings or the label distributions within a group (Fleisig et al., 2023; Weerasooriya et al., 2023), these approaches address disagreement only indirectly. To the best of our knowledge, direct disagreement prediction has been explored in only one prior study, specifically on sentiment analysis, conducted by Wan et al. (2023).

In their work, Wan et al. (2023) fine-tuned a RoBERTa model (Liu et al., 2019) on the DynaSent dataset (Potts et al., 2021) to predict disagreement using both binary disagreement labels and continuous disagreement rates. Additionally, they incorporated demographic information, such as age, gender, and ethnicity, to enhance the model's predictive performance. However, the inclusion of demographic data raises significant concerns related to annotator privacy and the potential for misrepresentation or underrepresentation of diverse social values and opinions (Weerasooriya et al., 2023).

We propose an alternative approach that leverages AAD to quantify disagreement and predict annotator disagreement based solely on textual features within the task, without relying on additional demographic information. This approach ensures privacy preservation and avoids biases associated with demographic-based selection, while providing an effective framework for disagreement prediction.

3 Data

To thoroughly investigate annotator disagreement within and across modalities and identify factors that make certain data types (textual, audio, silent video, or multimodal) challenging to annotate, we designed a two-session annotation study.

In the first session, four annotators independently annotated a small dataset across four modality setups: text, audio, silent video, and multimodal, providing distinct sets of annotations for each modality to assess inter-annotator agreement.

In the second session, one annotator re-annotated the dataset twice – 114 and 290 days later. These additional annotations enabled intra-annotator agreement analysis by comparing the three sets over time. The annotator reported vaguely remembering the content of some instances but stated not to have a recollection of the previous annotations.

Data collection and annotators Following Du et al. (2025), we use a subset of their Unic dataset, consisting of 94 YouTube video clips featuring authentic emotional expressions, unlike the exaggerated portrayals common in movies or TV series. Each video clip spans about 10 seconds, which was deemed sufficient in preliminary tests for identifying emotional states across modalities (Du et al., 2025). Four annotators (two male, two female college students proficient in English) participated after training on the annotation method and tools, ensuring consistent and informed annotations.

Annotation method All 94 video clips were annotated across three separate modalities – text, audio, and silent video – and also received a holistic multimodal emotion annotation. To capture emotional states as comprehensively as possible, both categorical and dimensional approaches were employed. For the categorical framework, we adopted the same labels as Du et al. (2025): *disgust*, *disappointment*, *confusion*, *surprise*, *contentment*, *joy*, and *neutral*. These categories were curated by clustering a larger set of emotions to reduce potential noise. For example, *love* is grouped under *joy* due to its lower frequency and closely related meaning. In the dimensional framework, emotional states were rated based on *valence* and *arousal*, using a 5-point scale ranging from very negative or very calm (1) to very positive or very excited (5), respectively. The dataset is available upon request.

4 Annotation Difference Analysis

To evaluate the annotations across annotators and modalities, we performed significance tests using the four sets of annotations from the first annotation session. Chi-Square test results suggest that both the categorical and dimensional emotion annotations are significantly influenced by the modality ($p = 6.068e^{-6}$, $p = 0.002$), and the annotators ($p = 3.669e^{-25}$, $p = 2.660e^{-42}$).

		text	audio	video	all
	e_4	.32	.27	.19	.29
κ	v_4	.33	.23	.21	.27
	a_4	.04	.06	.11	.09
α	$v_4 - nominal/unweight$	.33	.23	.22	.27
	$v_4 - ordinal/weight$	.64	.48	.46	.52
	$v_4 - interval/weight$	.64	.48	.46	.52
	$v_4 - ratio/weight$	.59	.42	.38	.46
	$a_4 - nominal/unweight$	.05	. 07	.12	.09
	$a_4 - ordinal/weight$	.01	.21	.32	.23
	$a_4 - interval/weight$	.01	.17	.30	.21
	$a_4 - ratio/weight$	<.01	.08	.19	.12

Table 1: Agreement with Fleiss' kappa and Krippendorff's alpha for the 4 annotation setups and in which *all* refers to the multimodal setup. v_4, a_4,and e_4 refer to the agreement of valence, arousal and emotion across 4 annotators.

As a common practice in dataset construction, we calculated both Fleiss' kappa and (weighted) Krippendorff's alpha. For emotion and valence, the kappa results, ranging from 0.19 to 0.33, suggest low agreement in the annotations, and similarly, the Krippendorff's alpha results, ranging from 0.22 and 0.64, reflect the same conclusion. This holds true even when considering different levels of measurement (e.g., ordinal and interval, etc.) or using weighted versus unweighted approaches valence annotations. Note that in our experiments, valence is scaled as integers from 1 to 5, which can be interpreted as very negative, negative, neutral, positive and very positive, making it a hybrid of multiple data types (Stevens, 1946). Default weights were applied in the calculation across these data types. For arousal, the results indicate less agreement.

The results in Table 1, along with similarly low agreement scores from other datasets, such as $\kappa = 0.27$ in IEMOCAP (Busso et al., 2008) or $\alpha = 0.25$ in CMU-MOSEI (Zadeh et al., 2018), prompted us to further investigate emotion annotation differences in the following sections.

4.1 Inter-annotator agreement across modalities

In addition to the common agreement statistics used to evaluate inter-annotator agreement among the four annotators, we also calculated the absolute annotation difference (AAD) between each pair of annotators. This approach allowed us to gain deeper insights into the specific areas where annotators agreed or disagreed, and to investigate whether any systematicity could be identified in these disagreements.

We begin with the valence annotations. Recall that valence was annotated on a scale of 1 to 5, ranging from very negative, weakly negative, neutral, over weakly positive to very positive. A valence difference of 0 or 1 between a pair of annotators indicates that they share the same or a similar assessment of the valence of a given fragment. However, when the valence difference is 2 or greater, it suggests that annotators hold a significantly different interpretation of the polarity (i.e., weakly negative versus weakly positive, neutral versus positive) expressed in the fragment.

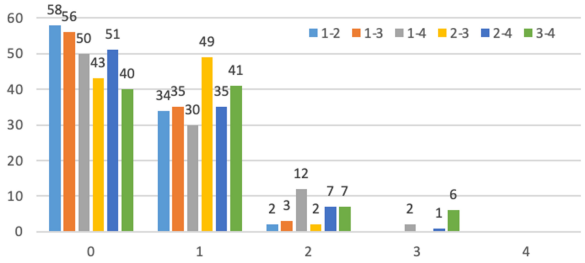

Figure 1: Absolute valence difference in texts between each pair of annotators (represented with different colours). The X-axis and Y-axis stand for valence difference and frequency respectively. Results for the other modalities are available in Figure 7 in Appendix B.

Diff	Text/%	Audio/%	Video/%	All/%
0	52.84	50.35	45.74	49.65
1	39.72	39.36	42.91	38.48
2	5.85	8.33	10.28	10.46
3	1.60	1.77	1.06	1.06
4	0.00	0.18	0.00	0.35

Table 2: Valence difference distribution in percentage across modalities, averaged from the six pairs of annotators.

As shown in Figure 1 and Table 2, the valence difference highlights (dis)agreement patterns among annotators. Figure 1 indicates that most of the valence differences between the six pairs of annotators are indeed limited to 0 or 1, with this tendency being consistent across the text, audio, video and multimodality setups. Table 2 confirms this, showing that in 52.84%, 50.35%, 45.74% and 49.65% of the text, audio, video and multimodality annotations, respectively, annotators selected the same valence score. Additionally, in around 40% of the cases, annotators chose a valence score in the nearest neighbouring category. This suggests that approximately 90% of the annotations show a strong agreement, with annotators consistently selecting the same or similar sentiment labels.

An interesting observation is that, according to

the kappa scores for valence, the agreement in the multimodal setup (0.52) is higher than in the audio setup (0.48). However, based on the results in Table 2, fewer annotators choose the same or similar labels in the multimodal setup (49.65% and 38.48%) compared to the audio setup (50.35% and 39.36%). One possible explanation is that the same or similar choices (diff = 0, 1) focus solely on agreement, whereas kappa combines both agreement and disagreement (diff > 1) into a single score. This suggests that while there is a greater degree of overall agreement in the multimodal setup, the higher kappa/alpha score may reflect less frequent or less severe disagreement compared to the audio setup.

Diff	Text/%	Audio/%	Video/%	All/%
0	33.51	37.41	41.67	35.28
1	38.65	45.39	44.50	43.44
2	22.34	15.07	13.12	18.26
3	4.96	2.13	0.71	2.84
4	0.53	0.00	0.00	0.18

Table 3: Arousal difference distribution in percentage acorss modalities, averaged from the six pairs of annotators.

Similarly, the absolute arousal differences, as presented in Table 3, suggest that annotators generally select the same or similar arousal labels with consistency. However, the frequency of identical choices is lower compared to valence.

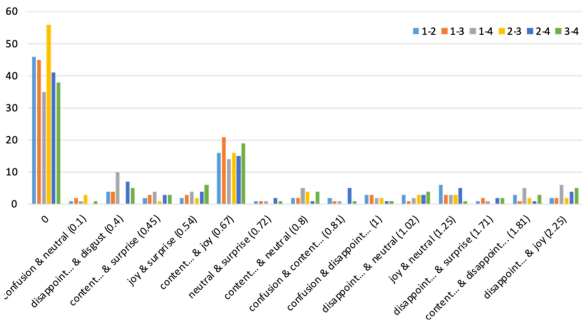

Figure 2: Emotion difference on the text modality between each pair of annotators. The X-axis and Y-axis stand for emotion (Euclidean) difference and frequency respectively. Results for the other modalities are available in Figure 8 in Appendix B.

As for the emotion annotations, we projected the different categorical emotion labels into a two-dimensional space as a vector, using their averaged valence and arousal scores (Table 8 in Appendix A). The Euclidean distance between the two vectors is the difference between two emotions. Then we plotted the distribution of emotion differences among the four annotators for the same instance.

Diff	Text/%	Audio/%	Video/%	All/%
0	46.28	44.68	35.46	43.62
0.1	1.42	0.35	1.95	2.13
0.4	5.32	1.06	2.3	5.67

Table 4: Distribution of top 3 minimum differences in percentage for different modalities, averaged from the six pairs of annotators. *Diff* stands for the absolute difference value in ascending order, ranging from 0 to 2.25.

As expected, the results in Figure 2 and Table 4 suggest a relatively high inter-annotator agreement. About 46.28%, 44.68%, 35.46% and 43.62% of the instances in the text, audio, video and multimodality setups, respectively, are annotated with identical emotions. Meanwhile, the most common confusing emotion pairs were $contentment$ and joy, accounting for more than 10% of the instances in all modality setups. This indicates that it is more challenging to differentiate emotions with similar valence values.

Based on the results of the valence, arousal and emotion analysis across modality, we can conclude that rather than relying solely on a single and comprehensive score provided by kappa/alpha, the absolute annotation difference (AAD) reveals valuable and insightful phenomena in emotion annotation. For instance, we found that most of the disagreement occurs between labels in the nearest neighbouring categories. Specifically, for valence, confusion frequently arose between labels with the same polarity but varying intensity. In the case of emotion annotations, disagreement often stemmed from emotions with similar valence but different arousal levels.

4.2 Intra-annotator agreement across modalities

Given the complexity of emotion annotation, we also calculated the absolute valence, arousal and emotion differences between three sets of annotations from the same annotator, who annotated the same dataset 114 days and 290 days after the initial annotation. The results as shown in Figure 3 confirm our earlier insights with respect to inter-annotator agreement. However, as expected, since inter-annotator differences in cultural and emotional background were minimized, the number of instances with identical annotation between the two annotation rounds was higher.

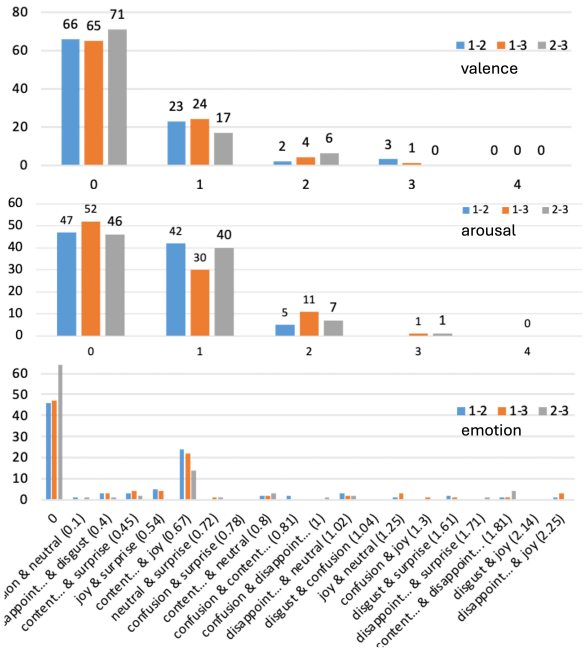

Figure 3: Absolute valence, arousal difference and emotion difference in text from three sets of annotations from the same annotator. Results of other modalities are available in Figure 9 and Figure 10 in Appendix B.

4.3 Qualitative analysis

The previous analysis focused on each modality setup individually, but it is also valuable to examine all setups together. Therefore, we further investigated the annotations with the most and least inter-annotator agreement on valence across all modality setups. This allowed us to gain a broader understanding of the patterns of (dis)agreement when considering all modalities simultaneously.

No.	text	audio	video	all
12	0,50	0,50	1,17	0,50
13	0,50	0,67	0,67	0,50
14	1,50	1,00	1,17	0,50
15	0,67	1,17	1,00	1,17
16	0,00	0,50	0,00	0,00
17	0,67	0,00	1,17	0,50
18	0,00	0,00	0,00	0,50

Figure 4: Part of the valence difference heatmap across modalities. Adequate agreement (≤ 0.5) is in blue while poor agreement (> 0.5) is in orange.

To identify the annotations with the most and least inter-annotator agreement, we first calculated the averaged valence difference score for each instance across all modality setups, ranging from 0 to 2.5, as shown in Figure 4. Since no instance has a full agreement (diff = 0) across all modality setups, we set a difference score of 0.5 (e.g. at most two annotator pairs showing a minimal annotation difference of 1) as the cut-off between adequate and poor agreement. As a result, weobserved that, 19% of the 94 instances exhibit adequate agreement across all four modality setups, 8.5% show poor agreement, while the large majority of the instances reside in between. Therefore, the top 19% (18 instances) and the bottom 8.5% (8 instances) were selected for further analysis as the high-agreement and high-disagreement annotations, respectively.

Although there is no actual gold standard annotation for the dataset, we assumed the emotion annotations obtained in the second annotation session (114 days after the first annotation) as silver standard to match the averaged valence difference score of each instance with a corresponding categorical emotion label.

With the emotion labels attached to the instances, it is found that for the 18 instances with adequate agreement in all four modality setups, only 2 negative emotion labels (two *disappointment*) appeared out of 72 labels, accounting for 2.8%. In contrast, for the 8 instances with poor agreement across all four modality setups, 12 negative emotion labels were recorded (11 *disappointment* and 1 *disgust*) out of 32 labels. This trend was also observed in the instances with adequate/poor agreement in three out of four modality setups (27 and 21 instances respectively), where the negative labels account for 22.2% and 40.5%, respectively.

This interesting finding suggests that, in our dataset, annotators tend to agree more on non-negative emotion states, but exhibit greater disagreement on negative emotions. One possible explanation for this phenomenon is that people tend to express positive emotions more openly, while they may feel less inclined to fully reveal negative emotions (Du et al., 2023).

5 Disagreement Prediction

Based on the insights from our agreement analysis, we also explored the potential of using AAD to model and predict disagreement, with the goal of identifying instances where annotators exhibit diverse interpretations, which can reveal valuable insights into the data. However, there are only a few studies on disagreement prediction, particularly concerning modalities such as audio or video. One recent research that caught our attention is the work of Wan et al. (2023) who performed dis-

agreement prediction on a dataset of over 100,000 textual instances (Potts et al., 2021). Given the constraints of data availability and computation cost, we conducted our initial investigation on texts, taking the research of Wan et al. (2023) as a starting point.

5.1 A novel rating strategy

We began by defining and scaling disagreement, as there are varying degrees of disagreement that we intend to investigate in greater detail. In the experiment of Wan et al. (2023), labels agreed by more than half of the annotators are considered the majority labels, while labels different from the majority are viewed as minority labels without looking at the nature of the underlying label. Since 5 annotators were involved in the annotation, Wan et al. (2023) calculated their disagreement rate as the number of minority labels divided by 3, where 3 is the borderline of minority labels in case of a majority, as formulated in the following:

$$D = \frac{\frac{n_{minority}}{N_{total}}}{\frac{3}{N_{total}}} = \frac{n_{minority}}{3} \tag{3}$$

Annotation distribution	Binary label	Wan's	Ours
🙂🙂🙂😐😐	disagree	0.67	0.77
🙂🙂🙂😐☹️	disagree	0.67	1.26
🙂🙂🙂😕☹️	disagree	0.67	N/A
☹️-negative 😐-neutral 🙂-positve 😕-mixed			

Figure 5: Comparison of two disagreement rating strategies on the same annotation distributions.

For example, as shown in Figure 5, there are three sets of annotations where the majority labels share the same sentiment *positive*, but the minority labels differ. The first minority labels are both *neutral*, and while the second are *neutral* and *negative*, both sets of annotations are assigned with a disagreement rate of 0.67. Considering the fact that the distance between *positive* and *negative* is much greater than that between *positive* and *neutral*, it is not appropriate to assign them the same level of disagreement.

As an alternative to the disagreement rating method of Wan et al. (2023), we propose to utilize the information from the absolute annotation difference (AAD) to evaluate the disagreement rate. Specifically, we take a variant of the root mean square error (RMSE) of the label distribution, which compares the differences between every two annotations (of an annotation set) that may vary. This approach is useful because, in practice, there are no "truth" annotations and aggregated annotations should not be considered as the "truth" (Cabitza et al., 2023). The variant is formulated as:

$$D^i = \sqrt{\frac{1}{\binom{n}{2}} \sum_{(x,y)\in\mathcal{N}} (x_i - y_i)^2}, \quad i \in \mathcal{M} \tag{4}$$

whereby n is the annotator number of the annotator set $\mathcal{N}$, $\binom{n}{2}$ is the number of different ways to select two annotators from the annotator set $\mathcal{N}$, $x, y \in \mathcal{N}$ are the considered annotators, and x_i and y_i represent the assigned sentiment labels respectively for the instance i in the dataset $\mathcal{M}$. Figure 5 provides further examples of the formula's application.

Our rating strategy considers sentiment annotation more like ordinal/interval variables rather than nominal ones. If we assign different sentiments with distinctive values, for example, {*negative* : -1, *neutral* : 0 and *positive* : 1}, we would derive more fine-grained disagreement rate scores, as shown in Figure 5, which effectively represent the sentiment distance among all the labels. Since it is difficult to assign a value to the *mixed* label and our evaluation dataset does not contain the *mixed* label, we excluded the instances with this label from the original DynaSent (Potts et al., 2021) dataset. The remaining instances, annotated with *negative*, *neutral* and *positive* labels, were mapped to -1, 0, and 1, respectively. The final reduced DynaSent dataset contained 75,127 instances, which was split into training, validation and test datasets with a ratio of about 6:2:2.

5.2 Experiment and results

Following the study of Wan et al. (2023), disagreement prediction was framed as both a binary classification task and a regression task, to represent different levels of disagreement. The experiments were conducted by fine-tuning a RoBERTa-base model (Liu et al., 2019) with a fixed learning rate 1e-5, and batch size 8 for 10 epochs, using NVIDIA Tesla V100-SXM2-16GB GPUs. Also, a DeBERTa-base (He et al., 2020) and DeBERTaV3-base (He et al., 2022) were investigated for the sake of comparison. Since Wan et al. (2023) used 4 scales for the regression task, we mapped the input RMSE scores into 4 scales. Additionally, to evaluate the accuracy and f1 score for the regression task,

we also mapped the regression output into 4 scales based on their absolute distance, leading to the disparity compared with the binary classification task as shown in Table 5 and Table 6.

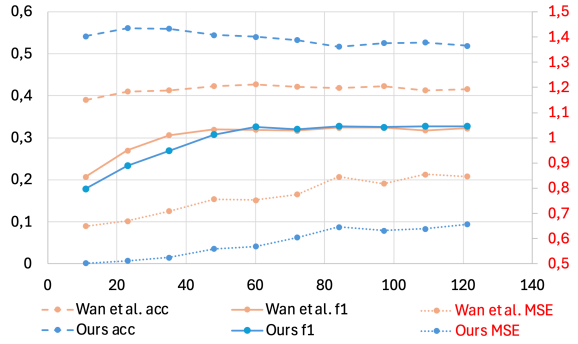

Figure 6: Comparison of two disagreement rating strategies during the training process of regression models with 4-scale outputs. Accuracy and f1 are plotted against the primary axis in black on the left, while MSE is plotted against the secondary axis in red on the right.

Task	Source	DynaSent	**acc** (↑)	**f1** (↑)	**MSE** (↓)
Bin.	Wan et al.	original	N/A	74.9	0.361
Reg.	Wan et al.	original	N/A	11.8	0.114
Bin.	Reproduced	original	73.89	57.7	0.261
Reg_4	Reproduced	original	37.46	31.4	0.111
Reg_4	Reproduced	reduced	**41.71**	32.1	**0.097**

Table 5: Results based on the rating strategy of Wan et al. reported in Wan et al. (2023) (upper) and reproduced by us (bottom) on the test dataset of the original and reduced DynaSent. Reg_4 refers to the regression output evaluated on a scale of 4.

Task	Model	Lr	**acc** (↑)	**f1** (↑)	**MSE** (↓)
Bin.	RoBERTa-base	1e-5	69.37	60.9	0.306
Reg_4	RoBERTa-base	1e-5	**51.64**	32.3	**0.072**
Reg_4	RoBERTa-base	5e-6	51.55	32.0	0.067
Reg_4	RoBERTa-base	1e-6	55.98	25.4	0.055
Reg_4	DeBERTa-base	1e-5	52.55	33.2	0.071
Reg_4	DeBERTaV3-base	1e-5	51.11	31.5	0.074

Table 6: Results based on the RMSE rating strategy with different models and learning rates on the test dataset of the reduced DynaSent. Reg_4 refers to the regression output evaluated on a scale of 4.

Figure 6 shows the RoBERTa-base model performance during the training process (10 epochs) on the validation dataset of the reduced DynaSent. During training, our disagreement rating strategy outperformed the other in terms of accuracy and MSE. For accuracy, higher values are better, while for MSE, lower values are preferred. Despite an overfitting warning during the 10 epochs training, it does not matter significantly when our main focus is the comparison of the two disagreement rating strategies.

The increase from 41.71% to 51.64% in accuracy and the drop from 0.097 to 0.072 in MSE in the final results on the test dataset, as shown in Table 5. and Table 6, reaffirms the better model performance based on our disagreement rating strategy. This suggests that using the AAD-based RMSE for rating disagreement yields improved performance in the task of sentiment annotation disagreement prediction. Additional experiments with other setups, as shown in Table 6, confirm these results.

5.3 Cross-dataset generalization

To test the model on our 94 instances of video subtitles, a fifth annotator was invited to independently annotate the subtitles, allowing for a similar experiment as in the previous section. We applied the AAD-based RMSE regression model, and the results are shown in Table 7.

	Instances	acc	f1	precision	recall
Reg_2	94	60.64	58.57	64.26	61.14
Reg_4	94	45.74	30.97	34.07	32.89
label-1	31	N/A	50.57	39.29	70.97
label-2	13	N/A	24.00	25.00	23.08
label-3	2	N/A	0	0	0

Table 7: Results of the regression task when the predictions are evaluated on a scale of 2 and 4, respectively, and the result breakdown, with label 1 to 3 for increasing disagreement.

In general, the results indicate the feasibility of predicting annotator (dis)agreement before annotation, even when the model was transferred to a new test dataset. Specifically, when evaluated with two polarities, i.e., agreement and disagreement, the models showed an accuracy of 60.64% and an f1 of 58.57%. When further breaking down the disagreement into three levels (label 1-3), unbalanced performance across levels of disagreement was observed, which might be caused by the imbalance of the label distribution in the training dataset with a ratio of 54:17:2.

6 Conclusion

While traditional IAA measures are favoured for providing a single comprehensive score that summarizes overall agreement across a dataset, they often complicate the interpretation of low scores and fail to capture finer (dis)agreement patterns. Prior research (e.g., Basile et al. (2021)) has highlighted these limitations, but effective solutions remain an

open area of research. Our study contributes a systematic exploration of AAD as a more interpretable measure of annotation variations, particularly in subjective tasks like emotion recognition. Rather than presenting AAD as a completely novel metric, we demonstrate its potential to complement existing agreement measures by providing richer insights into (dis)agreement.

We first applied AAD to analyze both inter- and intra-annotator (dis)agreement with a multimodal dataset, which enables us to observe how these (dis)agreements manifest differently depending on the input channel, proving a more comprehensive understanding of (dis)agreement across modalities. Furthermore, a nearly 10% increase in accuracy in the disagreement prediction task demonstrates the advantages of our AAD-based approach.

Due to the scarcity of available (multimodal) emotion datasets with sets of annotations for agreement study, we conducted our study on the most suitable dataset currently accessible. While a larger dataset could further validate our findings, our dataset is representative of real-world annotation challenges, and the observed improvements in disagreement prediction align with prior work. We would extend this research when new datasets become available, but the current results already demonstrate the effectiveness and potential impact of AAD.

7 Limitations

Although the database used in this study is relatively small, it provides valuable insights and lays a foundation for future research with larger datasets.

8 Acknowledgments

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (FAIR) (174K02325). We are grateful to Professor Thomas Demeester for his valuable guidance and support throughout the development of this work. We also extend our thanks the anonymous reviewers for their insightful and constructive feedback.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 151–154.

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559.

Lisa Feldman Barrett. 2009. Variety is the spice of life: A psychological construction approach to understanding variability in emotion. *Cognition and emotion*, 23(7):1284–1306.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8:67–80.

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.

Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N Salman, Chi-Chun Lee, and Carlos Busso. 2024. Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule. *IEEE Transactions on Affective Computing*.

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

Barbara Di Eugenio and Michael Glass. 2004. Squibs and discussions: The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2023. Unimodalities count as perspectives in multimodal emotion annotation. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025. Unic: a dataset for emotion analysis of videos with multimodal and unimodal labels. *Language resources and evaluation*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

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

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for NLP-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847, Marseille, France. European Language Resources Association.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage publications.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2024. EmoTwiCS: A corpus for modelling emotion trajectories in dutch customer service dialogues on twitter. *Language Resources and Evaluation*, 58(2):505–546.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Kristen A Lindquist and Lisa Feldman Barrett. 2008. Emotional complexity. *Handbook of emotions*, 4:513–530.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aaron Maladry, Alessandra Teresa Cignarella, Els Lefever, Cynthia van Hee, and Veronique Hoste. 2024. Human and system perspectives on the expression of irony: An analysis of likelihood labels and rationales. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8372–8382, Torino, Italia. ELRA and ICCL.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic. 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.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. Dynasent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404. Association for Computational Linguistics.

Harrison Ridley, Stuart Cunningham, John Darby, John Henry, and Richard Stocker. 2024. The affective audio dataset (aad) for non-musical, non-vocalized, audio emotion research. *IEEE Transactions on Affective Computing*.

Stanley Smith Stevens. 1946. On the theory of scales of measurement. *Science*, 103(2684):677–680.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Juan Wang and Bin Xia. 2019. Relationships of cohen's kappa, sensitivity, and specificity for unbiased annotations. In *Proceedings of the 4th International Conference on Biomedical Signal and Image Processing*, pages 98–101.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with disco. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695. Association for Computational Linguistics.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7053–7065, Online. Association for Computational Linguistics.

Ian Wood, John P. McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018. A comparison of emotion annotation schemes and a new annotated data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2024. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5699–5710.

Xinshu Zhao, Guangchao Charles Feng, Jun S Liu, and Ke Deng. 2018. We agreed to measure agreement–redefining reliability de-justifies krippendorff s alpha. *China Media Research*, 14(2):1–16.

A Categorical emotion labels and their averaged valence and arousal scores

Emotion	valence	arousal	vector
confusion	3.0	2.9	(3.0, 2.9)
contentment	3.8	3.0	(3.8, 3.0)
disappointment	2.0	2.8	(2.0, 2.8)
disgust	2.0	3.2	(2.0, 3.2)
joy	4.1	3.6	(4.1, 3.6)
neutral	3.0	3.0	(3.0, 3.0)
surprise	3.6	3.4	(3.6, 3.4)

Table 8: Categorical emotion labels and their averaged valence and arousal scores.

B Valence and Emotion Difference in Three Other Modality Setups

Figures 7 through 10 present the results of valence and emotion differences across audio, (silent) video and multimodal setups.

C Distribution of Disagreement

As shown in Figure 11, the distribution of disagreement rate changes with the rating strategies. One notable change is that more instances, regardless of sentiment polarity, are labelled as weak disagreement (0.33) instead of the stronger one (0.67). In both rating strategies, a larger proportion of negative instances receive strong disagreement (0.67) than neutral and positive ones, aligning with our findings in Section 4.3 that disagreement tends to happen more in negative instances.

D Discrepancy between Original and Reproduced Results

As shown in Table 5, there is quite some discrepancy between the F1 scores reported in Wan et al. (2023) and those of our reproduced experiments, while the MSE scores remain in the same range. For the sake of comparison, we believe that the results on the reduced dataset are better compared to our reproduced experiments following the same experimental set-up.

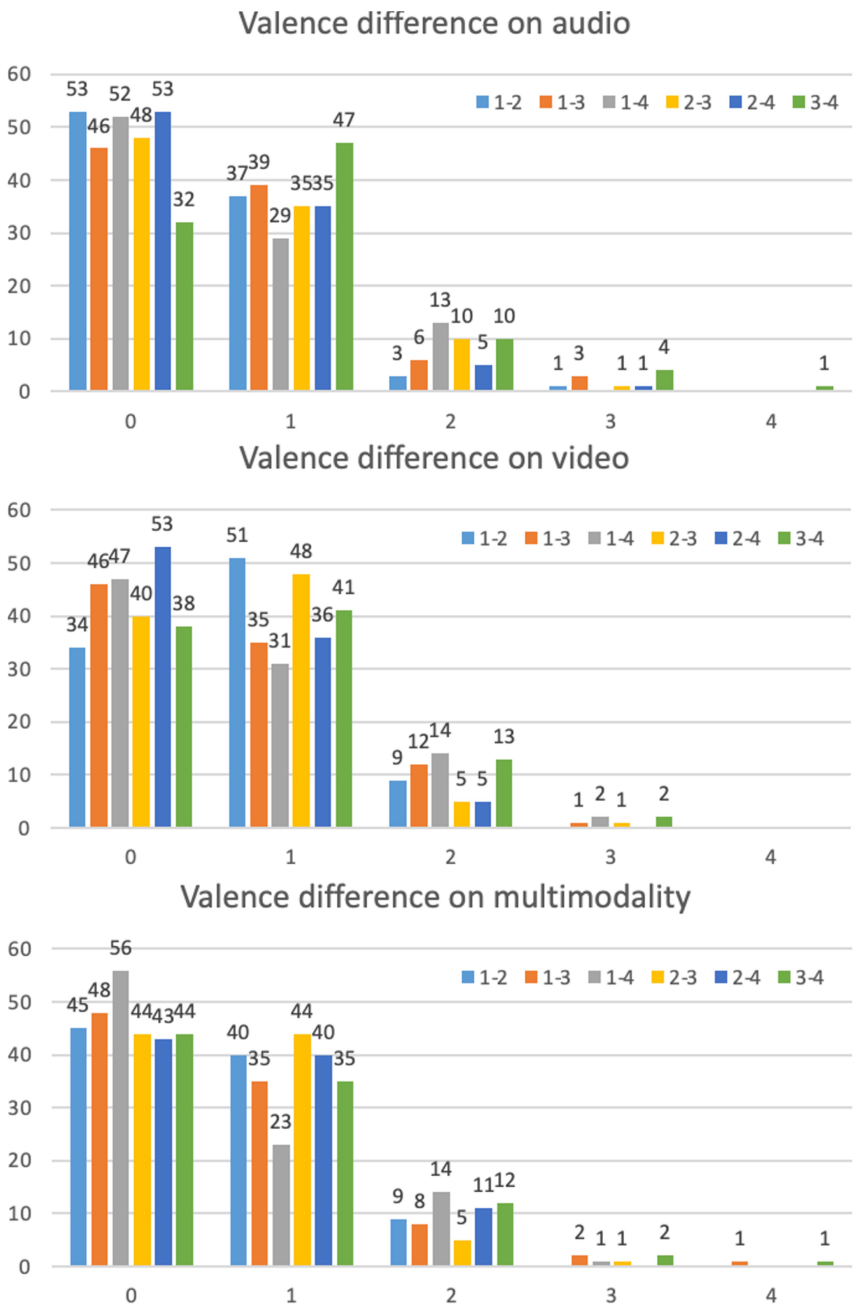

Figure 7: Absolute valence difference in audio, video and multimodal setups between each pair of annotators (represented with different colours). The X-axis is the absolute difference in valence; the Y-axis stands for the frequency of the difference values in the data.

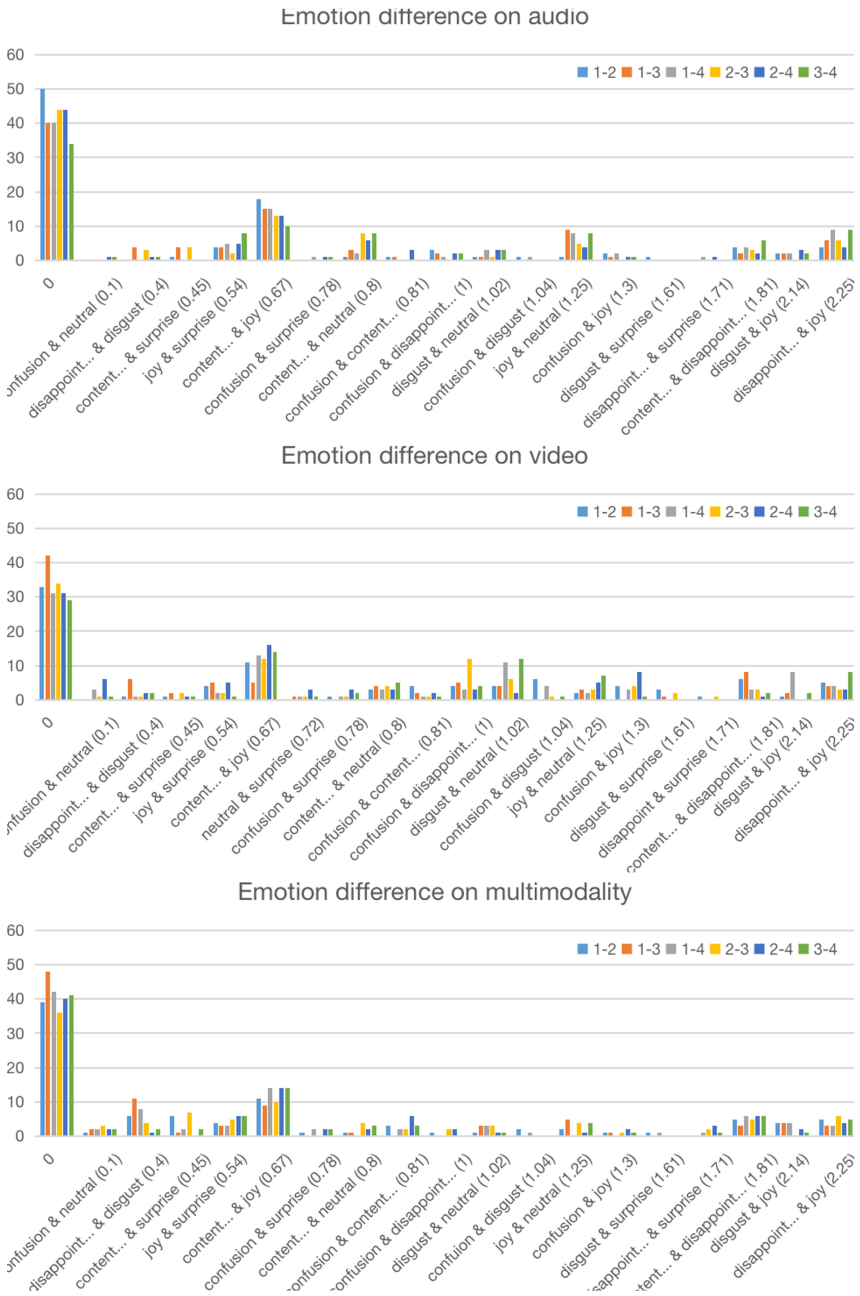

Figure 8: Emotion difference in audio, video and multimodal setups between each pair of annotators. The X-axis is the Euclidean distance between emotion vectors, while the Y-axis stands for the frequency of the difference values in the data.

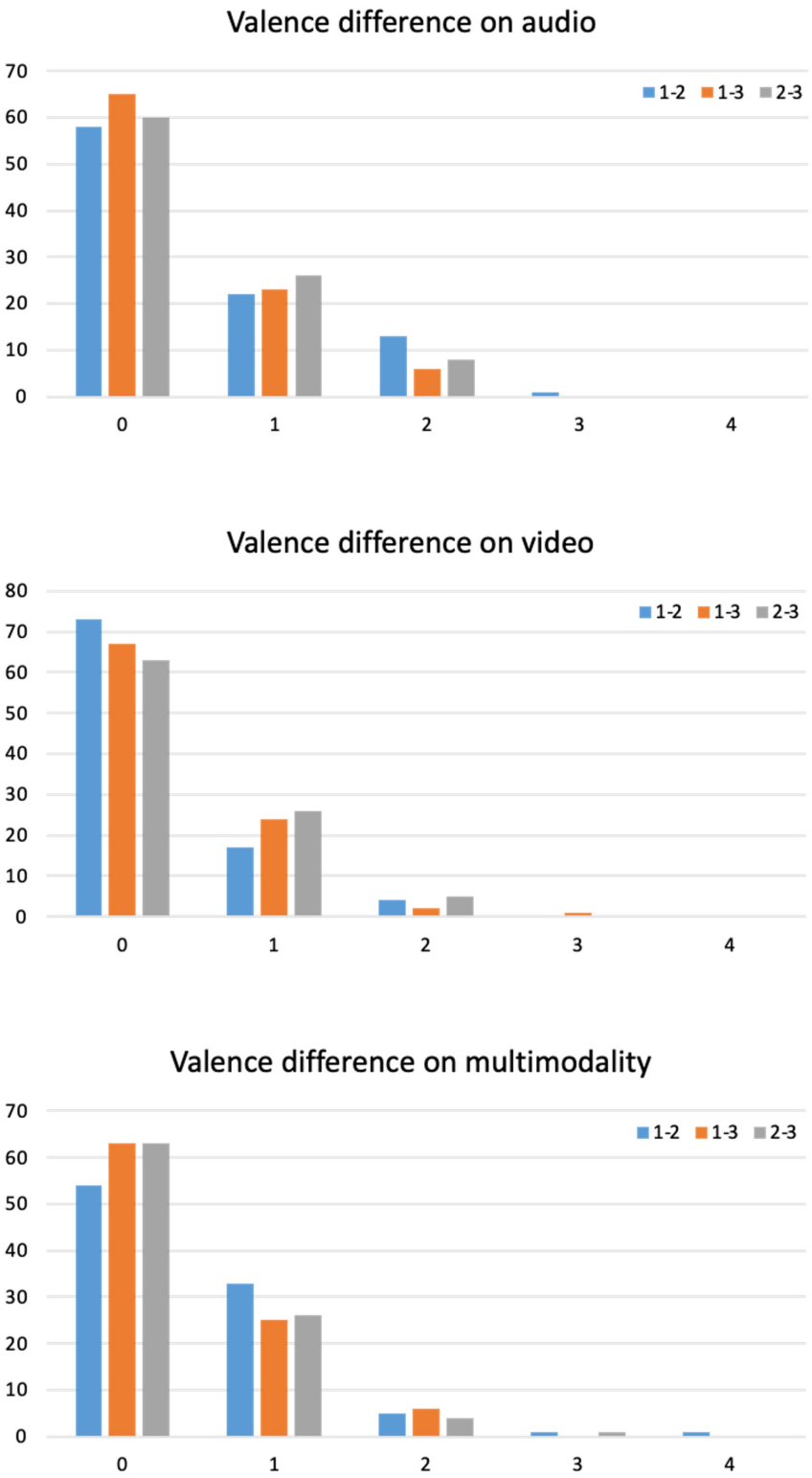

Figure 9: Absolute valence difference in audio, video and multimodal setups from three sets of annotations from the same annotator. The X-axis is the absolute difference in valence; the Y-axis stands for the frequency of the difference values in the data.

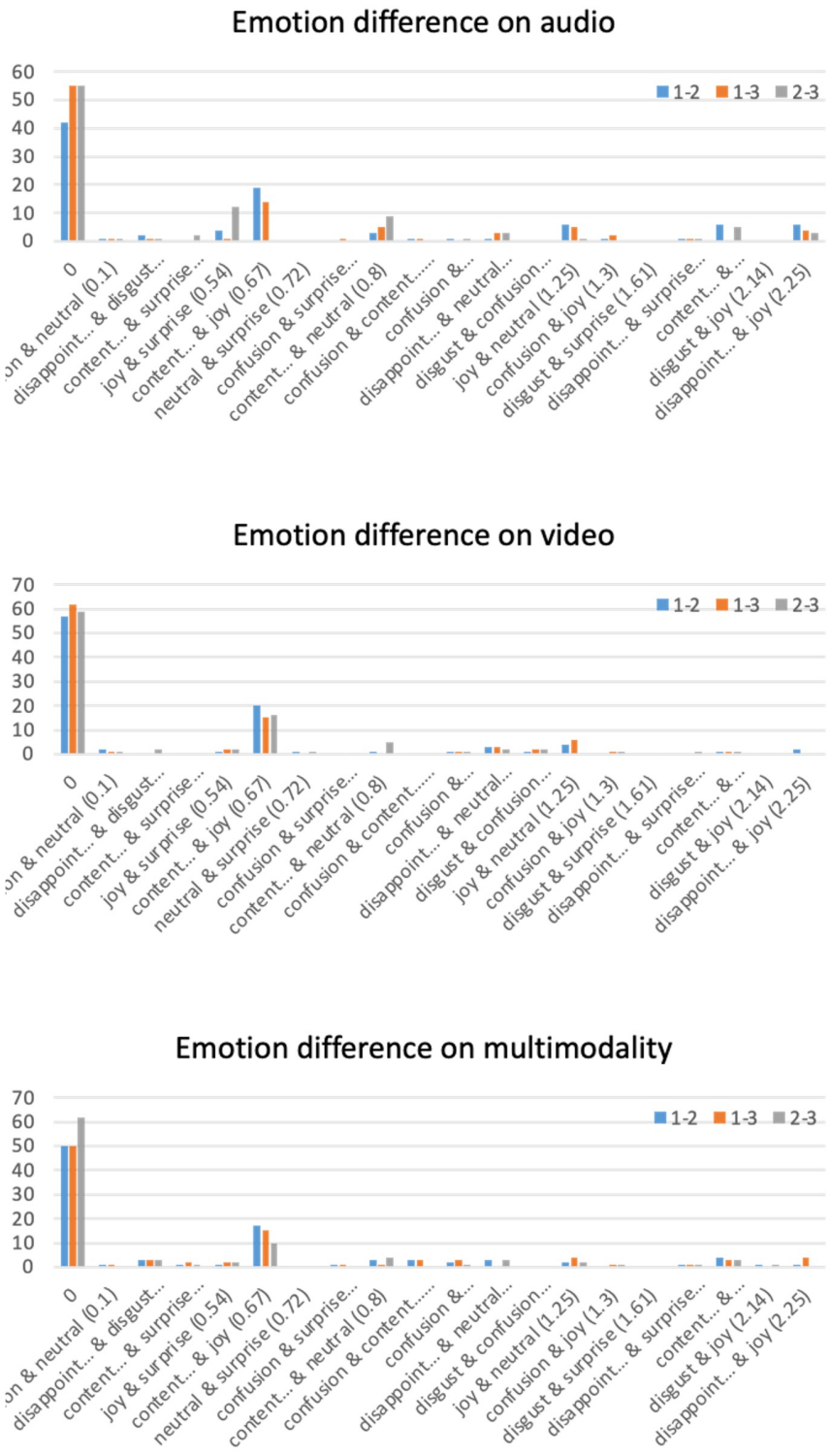

Figure 10: Absolute valence difference in audio, video and multimodal setups from three sets of annotations from the same annotator. The X-axis is the absolute difference in valence; the Y-axis stands for the frequency of the difference values in the data.

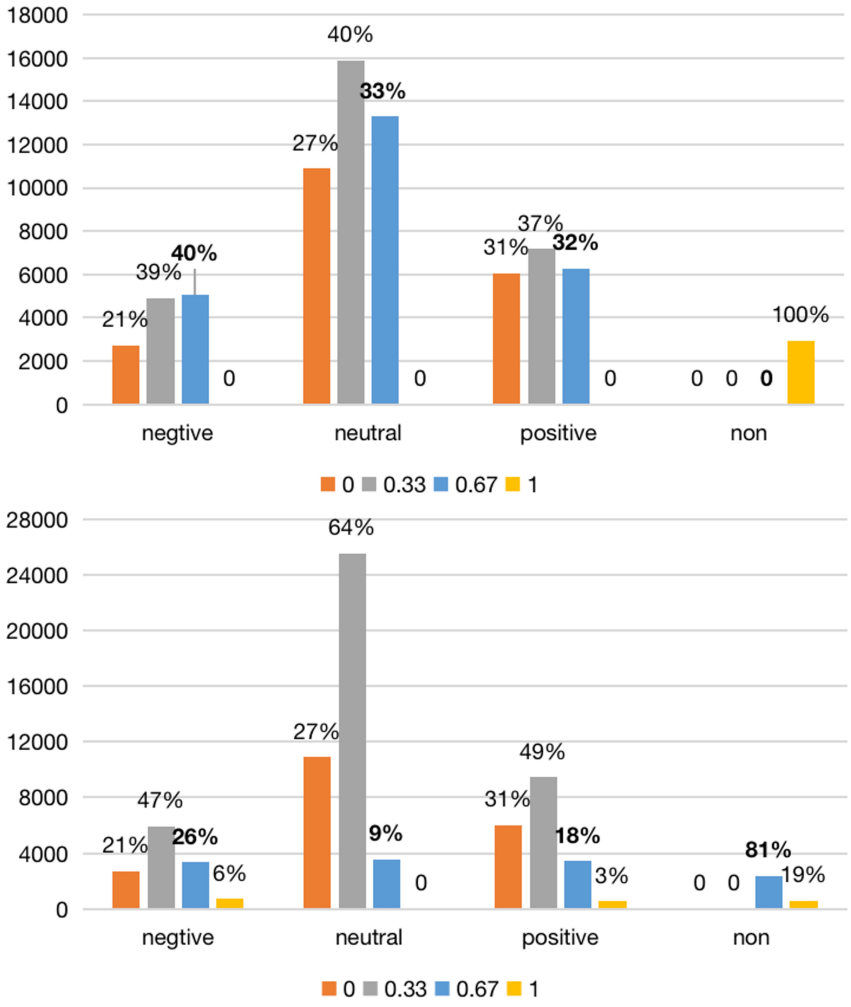

Figure 11: Distribution of disagreement rate across sentiment polarities in the reduced DynaSent dataset with different rating strategies. The first is based on the number of disagreement labels, while the second is mapped with RMSE scores. The X-axis represents the major sentiment polarities, with *non* referring to no majority.