Unimodalities Count as Perspectives in Multimodal Emotion Annotation

Quanqi Du[1,*], Sofie Labat[1], Thomas Demeester[2] and Veronique Hoste[1]

[1] LT3, Language and Translation Technology Team, Ghent University, Groot-Brittanniëlaan 45, 9000 Gent, Belgium

[2] T2K, Text-to-Knowledge Research Group, IDLab, Ghent University - imec, Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium

Abstract

Most datasets for multimodal emotion recognition only have one emotion annotation for all the modalities combined, which serves as a gold standard for single modalities. This procedure ignores, however, the fact that each modality constitutes a unique perspective that contains its own clues. Moreover, as in unimodal emotion analysis, the perspectives of annotators can also diverge in a multimodal setup. In this paper, we therefore propose to annotate each modality independently and to more closely investigate how perspectives between modalities and annotators diverge. Moreover, we also explore the role of annotator training on perspectivism. We find that for the different unimodal levels, the annotations made on text resemble most closely those of the multimodal setup. Furthermore, we see that annotator training has a positive influence on the annotator agreement in modalities with lower agreement scores, but it also reduces the variety of perspectives. We therefore suggest that a moderate training which still values the individual perspectives of annotators might be beneficial before starting annotations. Finally, we observe that negative sentiment and emotions tend to be annotated more inconsistently across the different modality setups.

Keywords

Multimodal versus unimodal emotion annotation, Annotator agreement, Emotion analysis, Perspectivism in NLP

1. Introduction

The study of emotion has expanded from philosophy, psychology to other research fields such as sociology, anthropology, neuroscience, and computer science. With the aim of achieving the so-called "emotional intelligence" [1], machines are expected to not only understand human language but also human's affect and emotions. In this context, the computational modeling of emotions has been studied both at the level of single modalities and at the multimodal level. While the former focuses on one single modality such as text [2], speech [3] and video [4], the latter considers the combination of two or more single modalities. Considering the fact that the essence of communication has always been multimodal [5], and given the rich characteristics and complex distribution of human emotions in different modalities, multimodal data has the potential to reflect emotional changes from multiple perspectives. As a result, multimodal models for the automatic detection of affect have also been shown to outperform unimodal models by aggregating complementary information across modalities [6, 7].

Currently, the success of multimodal emotion recognition (MER) is partly due to recent advancements in neural network architectures, which require large amounts of training data to learn useful representations. However, annotating multimodal data for emotions is not an easy task, as emotion is a subjective concept which depends on one's world knowledge, cultural or personal experiences. Annotators may thus attach different emotion annotations to the same expressions, due to different perspectives.

Just as emotions can be seen from different perspectives of the annotators, each single modality may also convey different polarities [8] and emotions than the multimodal setup. In the field of multimodal studies [9], it is assumed that the often narrow emphasis on one unimodality (e.g., text modality) in analyses is inadequate for fully comprehending meaning, since different modes of communication work in harmony, each serving a specialized function in the process of meaning-making. Therefore, the key to understanding every communication instance lies in understanding the relations between modes of communication [10]. Transferring these insights to the task of multimodal emotion analysis, we investigate the information each single modality contributes.

In this paper, we investigate emotion annotation at the multimodal level and at the level of single modalities, taking each modality as a unique perspective on emotion. We focus on the following research questions:

1. Within modalities: How often do unimodal and multimodal emotion annotations of the same video snippet share the same emotion states? Can

2nd Workshop on Perspectivist Approaches to NLP

[*] Corresponding author.

quanqi.du@uegnt.be (Q. Du); sofie.labat@ugent.be (S. Labat); thomas.demeester@ugent.be (T. Demeester); veronique.hoste@ugent.be (V. Hoste)

0000-0002-1698-1566 (Q. Du); 0000-0003-1675-8927 (S. Labat); 0000-0002-9901-5768 (T. Demeester); 0000-0002-9901-5768 (V. Hoste)

CEUR Workshop Proceedings (CEUR-WS.org)

we discern any unimodalities that dominate others and lead to the multimodal emotion state? (RQ1)

2. Beyond modalities: Perspectivists advocate that gold standards do not reflect the multiple perspectives through which annotations are collected. What is the effect of annotator training on the subsequent annotation behaviour for both unimodal and multimodal annotation? (RQ2)
3. Features of inconsistency: Inconsistency in emotion annotations across modalities is expected. Which tendencies can we discern in this inconsistency? (RQ3)

2. Related Research

2.1. Multimodal Emotion Annotation

Emotion annotation is not trivial. Emotions are too complex to have a universally accepted standard taxonomy or annotation scheme. Normally, emotions are either annotated along categorical or dimensional frameworks. In the categorical emotion description, *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* are usually seen as the six most basic universal emotions [11]. This model has been extended by Plutchik [12] to cover two more emotions (*anticipation* and *trust*). Dimensional models, on the other hand, project emotions in a multidimensional space with usually three axes, namely *valence*, *arousal* and *dominance* [13]. Sometimes less (e.g., *valence* and *arousal* only [14]) or more dimensions (e.g., *unpredictability* [15] or appraisal dimensions [16]) can also be used.

Most multimodal emotion corpora are annotated with either categorical or dimensional labels, or both. IEMOCAP [17] is one popular dataset with 10,039 turns of acted conversations. The recordings were manually segmented into conversation turns and then annotated with both discrete categorical emotion labels (i.e., Ekman's six basic emotions [11] complemented with *frustration*, *excited*, *neutral*), and continuous dimensional labels (i.e., *valence*, *arousal* and *dominance* on a five-point Likert-like scale). It was argued that the combination of both emotion frameworks could provide complementary information on how emotion was displayed in real life, since the categorical level could not give insights in the intensity level of emotions [17]. In MSP-IMPROV [18], the same two approaches were adopted, but a different set of discrete labels was used (i.e., *happy*, *angry*, *sad*, *neutral* and *other*) to collect emotion annotations on 8,438 sentences through crowdsourcing and majority voting.

Although annotating emotions along the different dimensions could relieve the burden of choosing the appropriate categorical framework, numerous studies [19, 20] have shown that the dimensions *arousal* and *dominance* are difficult to annotate due to their subjectivity, leading to low inter-annotator agreement (IAA). On the other hand, *valence* is less subjective and in some sense, it shares the same connotation with sentiment, where both of them are scaled into positive, negative, neutral and sometimes some intermediate values, e.g., lightly positive or very negative. Therefore, in some datasets, only sentiment was annotated. CMU-MOSEI [21] is such a dataset annotated by three crowdsourced judges with Ekman's emotions [11] on a [0,3] Likert-scale and sentiment on a [-3,3] Likert-scale. The 3,228 videos in CMU-MOSEI make it one of the largest datasets for sentiment analysis and emotion recognition. Taking the same annotation strategy as CMU-MOSEI, the multimodal emotion dataset MELD [22] evolved from the textual emotional dataset EmotionLines [23] and achieved higher agreement, suggesting that the additional modalities were instrumental for the annotation improvement.

While the previous datasets were annotated with emotion labels at the multimodal level, the Chinese CH-SIMS [8] dataset was annotated both at unimodal and multimodal level, albeit with less fine-grained annotations than the previously mentioned corpora since only sentiment was annotated. The dataset contains 2,281 video segments annotated by five independent students with integers from [-1,1] for negative, neutral and positive sentiment. For the purpose of regression and multiclass classification tasks, the annotations were then averaged and divided into five clusters as *negative*, *weakly negative*, *neutral*, *weakly positive* and *positive*. During a closer examination of the annotation results, it was found that the sentiment difference between the modalities was not distributed evenly, with audio and the multimodal setup showing a minimal difference while video and text exhibited maximal difference [8]. As a pioneering study in unimodal and multimodal sentiment annotation, CH-SIMS offers inspiring findings on quantified sentiment relationships among different modalities. Compared with sentiment, emotion annotations would give more fine-grained insights in the variety of emotions expressed in different modalities. In our study, we aim to tackle this challenge by annotating fine-grained emotions both at the unimodal and multimodal level.

2.2. Perspectivism in Emotion Analysis

Data perspectivism, as the name suggests, is a recently popular paradigm for data annotation, which advocates integrating the diversity of human subjects' opinions in annotations and in the knowledge representations machine learning models [24]. Traditionally, it is quite often the case that annotators have different opinions, but this disagreement is usually resolved through some aggregation methods, such as majority voting in MELD [22] or averaging in CH-SIMS [8]. The aggregation process is named ground truthing, where a ground truth or gold standard is constructed. However, in subjective tasks such as sentiment analysis and emotion recognition, there are often cases where there is no ground truth but just different perspectives, and the creation of a ground truth results in a loss of subtle, but valuable nuances in annotations [25, 26]. Also, annotation aggregation may unfairly cause an under-representation of certain annotators' perspectives [27].

To make full use of the different annotations and capture the contextual nuances, some machine learning researchers proposed to use different annotations as soft labels [28], instead of using the ground truth as hard labels. In this way, improvements of accuracy on speech emotion detection were reported by using soft labels that incorporated knowledge collected from all annotators [28]. While yielding performance improvement, it also helped to solve the paucity of training data by utilizing ambiguous emotional utterances without dominant targets [29].

Instead of considering multi-annotator modeling as a multi-label problem where each annotator's labels are seen as a perspective on the same task, the multi-task approach attempts to learn multiple perspectives as separate classification tasks. In computer science, multi-task learning aims to leverage useful information contained in multiple related tasks to help improve the general performance of all tasks [30]. In our case, different annotators' labels can be considered as input of different tasks. When there are enough annotations contributed by each annotator, the multi-task model shows significantly better performance and higher robustness with lower standard deviation, but without a significant drop in efficiency although the annotations as input are multiplied [31].

In the field of multimodal emotion recognition, when drawing an analogy between different annotators' perspectives and independent modality annotations, it is reasonable to consider multimodal emotion recognition as a multi-task learning problem containing as subtasks the detection of emotions in different modality combinations. In the experiments of Yu et al. [8] for sentiment analysis, it was found that multi-task models outperformed single-task models for most of the evaluation metrics.

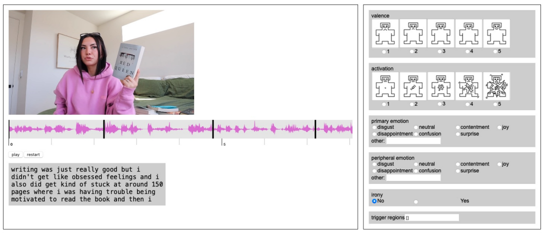
In what follows, we investigate how perspectives can be leveraged for emotion analysis. More precisely, we look at perspectives between (i) different modality levels and (ii) different annotators.

3. Method

To obtain high-quality data, we carefully designed a pilot study in which we monitored the annotation process. In this section, we motivate our selection of multimodal data, discuss the fine-grained annotation framework and outline the design of three annotation rounds.

3.1. Data Collection and Annotators

For our pilot study, we collected emotion-rich videos from YouTube. YouTube videos are more easily available than dramas, soap operas and movies. Furthermore, rather than acted emotions, these videos contain natural expressions of emotion. The collected dataset consists of 94 video clips of reviews, each of which last for about 10 seconds, which is longer than the average length in popular datasets, e.g., CH-SIMS (3.67 seconds) [32], CMU-MOSEI (7.28 seconds) [21], M^{3}ED (7.39 seconds)[33], MELD (about 8 seconds) [22]. We believe that 10 seconds is a sufficient time length to allow annotators to detect emotion states in each of the independent unimodalities. This pilot corpus covers 14.75 minutes in total. The clips were assigned to three annotators who each annotated the clips separately, eventually leading to three sets of annotations for the full corpus. These three annotators are students from Ghent University who are proficient in English. Before annotation started, the annotators received instructions on the chosen emotion framework and the custom-designed annotation interface, as shown in Figure 1.[1]



```
writing was just really good but i
didn't get like obsessed feelings and i
also did get kind of stuck at around 150
pages where i was having trouble being
motivated to read the book and then i
```

primary emotion
○ disgust ○ neutral ○ contentment ○ joy
○ disappointment ○ confusion ○ surprise
other:

peripheral emotion
○ disgust ○ neutral ○ contentment ○ joy
○ disappointment ○ confusion ○ surprise
other:

irony
● No ○ Yes

trigger regions []

Figure 1: The custom-designed annotation interface.

3.2. Annotation Method

For most multi-modal emotion datasets, there is only one unified emotion label for each video clip. Inspired by findings from [8] regarding the difference in sentiment annotations between different modalities, we decided to annotate each video clip on four levels, namely *text*, *audio*, *video* (without audio) and *all* (all three modalities combined). To ensure other modalities did not interfere in their judgements, annotators received the four setups of raw materials in a shuffled order, meaning that there were time gaps between an annotator seeing different setups of the same video clip. It should be noted that for the audio modality, annotators were instructed to

[1] The frame in the figure is taken from the following video: https://www.youtube.com/watch?v=kbn7gOXedSo.

mark the emotion by focusing on the clues in the audio only, and by ignoring the words in the speech as much as possible.

Two emotion frameworks were adopted in the annotation, namely a categorical and a dimensional framework. In order to cover a wide diversity of emotions, we opted for the 25 fine-grained emotion taxonomy proposed by Shaver et al. [34], including *anger*, *contentment*, *disappointment*, *disgust*, *enthrallment*, *enthusiasm*, *envy*, *fear*, *frustration*, *irritation*, *joy*, *longing*, *love*, *lust*, *nervousness*, *optimism*, *pity*, *pride*, *rejection*, *relief*, *remorse*, *sadness*, *suffering*, *surprise*, and *torment*. A *neutral* label was also added in case there is no emotion present. In case there was an emotion that did not match any of the provided labels, the annotators were allowed to customize an emotion label in their own words. For the dimensional framework, *valence* and *arousal* were annotated on an analogue-visual five-point Self-Assessment Manikin (SAM) scale [35]. *Dominance* was not included, as multiple studies on emotion annotation [19, 20] had shown that annotator agreement on dominance was too low to be useful.

3.3. Multiple Rounds of Annotation

Annotations were obtained in three sessions divided by the training of the annotators, which is the process of gold standard production by taking into account multiple perspectives. Before training, three annotators were given a minimum set of initial guidelines, and were then allowed to do the annotations freely on the 94 video clips, without interference from each other. During the training session, the annotators were gathered to jointly annotate a subset of the 94 video clips, with discussion and negotiation. This subset for training consisted of 11 video clips for which the annotators had to annotate the four modality levels, and 53 clips for which they only had to annotate the multimodal setup. While the former annotation setup focused on agreement between different modality levels, the latter targeted agreement over the video clips as a whole. It took four hours for the three annotators to go through this training session. During the first hour of the training session, the first author of this study sat with the annotators to guide the discussion. To maximally reduce interference from external factors (e.g., the author), for the next three hours, each annotator in turn took on the role of discussion leader. During the training session, they were supposed to learn more about each other's definition of emotions and evaluation of the expressed emotions. Since all annotators were involved and explained their views in the discussion, we might consider this process as a kind of weak perspectivism since all perspectives would be summarized into one single position or gold standard [24]. After training, the three annotators separately annotated another subset of the 94 video clips for the four modality levels. The annotation results before annotator training were grouped into three categories based on their valence score agreement, namely *full agreement* (3 annotators agreeing), *little agreement* (2 annotators agreeing) and *disagreement* (no one agrees). Ten video clips in each category were randomly selected as the test set for the after-training session, leading to a subset of 30 video clips to be annotated. These three sessions (i.e., before annotator training, joint gold standard annotation and after annotator training) serve unique functions. The annotator training session resulted in a set of fully agreed annotations and more insights into the way how the other annotators perceive emotions; the sessions before and after the training could be seen as annotation processes without and with knowledge of an adjudicated gold standard.

4. Perspectivism analysis in multimodal annotation

Although we obtained rich annotations for valence, arousal and categorical emotion labels, in the following part we first focus on valence analysis, which is less subjective than arousal and easier to quantify than categorical emotion labels. While the analysis on annotator level before and after training aims to explain the relationships among modality levels (RQ 1), the investigation on inter-annotator level tries to probe into the changes of annotators' perspectives after the training session (RQ 2). In the last part of this section, we focus on the inconsistency in emotion annotation for the different modality levels (RQ3).

4.1. Polarity annotation analysis before training

After the first annotation round, four subsets of annotations (i.e., one for each modality setup) were obtained from the three annotators separately. The three annotators had no communication with each other before and during the annotation. The independently obtained annotations were intuitively made by each annotator, and could serve as a proxy of their personal emotion model.

Since the valence annotations are 5-scale scores ranging from *positive*, *weakly positive*, *neutral*, *weakly negative*, to *negative*, we took into account these scores to account for the fact that, for example, the difference between *positive* and *neutral* should be greater than the difference between *weakly positive* and *neutral*. Inspired by Yu et al. [8], we calculated the difference in valence scores between the four modality levels, which is formulated as:

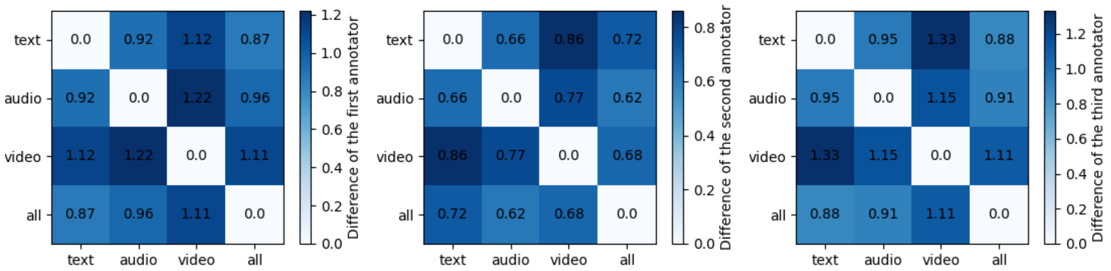


Figure 2: Valence difference score of the 3 annotators on the 94 video clips. The higher the score, the higher the difference in valence across modalities.

$$D_{xy} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2} \tag{1}$$

whereby $x, y \in \mathcal{M}$ are the considered modalities, $\mathcal{M} = \{text, audio, video, all\}$, N is the number of video clips, and x_n and y_n represent the assigned valence for clip n in modality x and y, respectively.

With this formula, we obtain the confusion matrices in Figure 2. We can see that for the second and third annotator, the maximal difference can be observed between the text and video modality. For the first annotator, the difference between the audio and video modality annotations is maximal. Averaging the difference scores across the three annotators leads to a 1.10 difference between text and video, a little higher than the difference score of audio-video (1.05). When we consider the minimal differences between the four modality levels, we can observe that the text (2 annotators) and audio (1 annotator) annotations are most in line with the "all" multimodal annotations.

To further investigate differences in valence annotations between the four modality levels, we averaged the valence score of each modality setup for each of the three annotators, as shown in Table 1. A first observation which can be made is that the annotations of annotator 2 show a lower variance compared to those of the other two annotators. The reason is that annotator 2 only once used a more extreme sentiment (viz., score 1), while the other two annotators used the full scale of valence scores. If we rank the four modalities by average valence scores, we can discern that the average valence score for the audio modality is consistently higher for all three annotators. As an important medium for emotional communication, audio can arouse the audience's emotional resonance through acoustic features, such as pitch or tone, which might explain the overall higher average valence scores. In comparison to audio, text as a form of literal expression may not be direct and authentic enough, resulting in lower average valence scores. Furthermore, when taking the average of the score of annotator 1 and annotator 3 on the video modality, we find that video has the lowest average valence scores. These results, however, are not corroborated by annotator 2.

Table 1
Average valence scores on 94 annotations before training from three annotators. μ represents the average scores for each modality level, while σ^2 is the variance. The higher the valence scores, the more positive the modalities. The higher the variance, the less similar the annotations are in valence.

	text	audio	video	all
	μ/σ^2	μ/σ^2	μ/σ^2	μ/σ^2
*Anno*1	3.13/0.97	**3.40**/0.86	3.07/1.08	3.23/0.97
*Anno*2	3.13/0.82	**3.37**/0.72	3.30/0.60	3.16/0.70
*Anno*3	3.06/1.26	**3.10**/1.09	2.93/1.17	2.97/1.10

In order to gain more insights into the valence annotations across the four modality setups, we more closely investigated the counts of negative (i.e., 1 and 2), neutral (i.e., 3) and positive (i.e., 4 and 5) annotations for the three annotators, as shown in Table 2. A first interesting observation to be made is that positive valence is dominantly annotated across all modalities. While annotator 1 and annotator 2 by a large margin more often recognize positive sentiment in the audio setup, this result is not supported by the annotations of annotator 3. In general, we found that the annotators detected more positive sentiment in audio than in other setups. At the same time, we could observe that the lowest values of positiveness are associated with the video modality. This indicates that annotators detect less often positive emotion states in the silent video than in other setups. As for negative emotion states, the lowest values lie in audio and the highest values lie in video (except annotator 2, who always has the lowest values for negativeness among the three annotators), suggesting that annotators tend to detect less negativeness in the modality of audio than others, and more negativeness in the modalities of silent video. When putting the annotations all together, it is found the all modality setup does not hold the maximum

Table 2
Relative sentiment counts across three annotators. neg, neu, and pos mean negative, neutral, and positive, respectively. $Anno1 - 3$ means annotator 1 to annotator 3, and Avg means the average relative count across the three annotators.

	text	audio	video	all
	neg / neu / pos	neg / neu / pos	neg / neu / pos	neg / neu / pos
$Anno1$	0.287 / 0.202 / 0.511	0.223 / 0.138 / 0.638	0.340 / 0.149 / 0.511	0.330 / 0.096 / 0.574
$Anno2$	0.191 / 0.362 / 0.447	0.106 / 0.362 / 0.532	0.106 / 0.543 / 0.351	0.117 / 0.457 / 0.426
$Anno3$	0.298 / 0.234 / 0.468	0.298 / 0.277 / 0.426	0.340 / 0.266 / 0.394	0.298 / 0.223 / 0.479
Avg	0.259 / 0.266 / 0.475	0.209 / 0.259 / 0.532	0.262 / 0.319 / 0.418	0.248 / 0.259 / 0.493

or minimum values in any type of sentiment, suggesting the existence of a subtle trade-off among the emotional contribution from each unimodality.

In the next part, we focus on the relationship between the different modality levels in a more quantified way. To this end, we conducted multiple regression analyses with the OLS (Ordinary Least Squares) model [36] with annotations combined from three annotators.

The regression results in Table 3 indicate that the variables text, audio, and silent video have a significant impact on the dependent variable, i.e., the multimodal level. With positive coefficients, the increase of valence scores in each unimodality is associated with an increase in the multimodality level. The t-value and the p-value (<0.05) indicate that these coefficient estimates are statistically significant and unlikely to be due to chance. The coefficient of text is bigger than that of audio, and also two times bigger than that of video, meaning each one-unit valence increase in the text modality is associated with a higher average increase of valence at the multimodal level than the other two unimodalities. Our results thus seem to confirm insights from previous studies [37] which reported a significant drop (30%) in binary accuracy when removing the text modality in multimodal sentiment analysis, a phenomenon which is termed "text predominance" [32].

Table 3
OLS regression results for annotations from three annotators. $coef$ means the estimated coefficients or slops, std_err means the standard error of the coefficient estimates, t means the t-statistics for each coefficient, $P > |t|$ means the p-values associated with the t-statistics, $[0.025\ 0.975]$ means the lower and upper bounds of the confidence interval for each coefficient. The higher the $coef$ score, the more impact the modality has on the multimodal label.

	coef	std_err	t	P>|t|	[0.025	0.975]
$const$	0.395	0.170	2.237	0.021	-0.061	0.728
$text$	0.376	0.051	7.406	0.000	0.276	0.475
$audio$	0.284	0.054	5.202	0.000	0.176	0.391
$video$	0.230	0.045	5.096	0.000	0.141	0.319

Inter-Annotator Agreement Since each annotator can be counted as having a unique perspective, it would be interesting to check the difference between their perspectives, which is the opposite of their agreement. In this study, we take the agreement of the annotations as an indication of perspective agreement. Higher agreement means less diversity in perspectives.

Table 4
Agreement among the three annotators in valence scores for different modalities. The higher the agreement score, the more similar perspectives the annotators have.

modality	text	audio	video	all
κ	**0.37292**	0.28132	0.19163	0.25808
$\alpha(nominal)$	**0.37516**	0.28387	0.19450	0.26071
$\alpha(interval)$	**0.74900**	0.55135	0.50379	0.52536

Table 4 shows that without exception, the three annotators reach the highest agreement on the text modality, no matter whether these scores are calculated with Fleiss' kappa [38] or Krippendorff's alpha [39]. It is as expected that annotators have the least agreement over silent videos, since emotion recognition on non-linguistic elements is the most difficult task compared with other modalities [40]. Furthermore, while we could argue that each single modality has its own unique emotion clues and that all this information together offers the annotators a rich multifaceted view on the valence expressed in a given video, we do not see this reflected in the agreement scores for the multimodal level: the agreement on the multimodal level is lower than the agreement on text and audio.

The results also indicate that the differences between annotators' perspectives vary among different modalities, and compared with the single text modality, the addition of the other two unimodalities makes the perspectives more varied.

4.2. Polarity annotation analysis after training

After the first annotation round, the annotators were invited for a training session in which they jointly discussed and annotated part of the data, with the goal of reaching an adjudicated gold standard annotation. As we hypothesized this would have an effect on the annotators' perspectives and subsequent annotations, we randomly picked thirty videos for an after-training session from the subsets of the fully annotated corpus with different agreement degrees (see Section 3.3).

Table 5
Valence difference scores calculated with Equation (1) on 30 annotations before and after training. t, a, v, A stand for $text$, $audio$, $video$ and all modality setup, respectively. $Anno1-3$ means annotator 1 to annotator 3, μ means the average of the scores, σ^2 means the variance of the scores. In each column the two numbers refer to the results **before/after** the training session. The higher the difference score, the bigger the sentiment gap between the modality setups.

	Anno1	Anno2	Anno3	μ	σ^2
$t-a$	0.93/1.21	0.71/0.73	1.02/0.91	0.89/0.95	0.02/0.04↑
$t-v$	**1.29/1.58**	**0.98/1.29**	**1.41/1.47**	**1.23/1.45**	0.03/0.01↓
$t-A$	0.88/1.20	0.75/0.95	1.20/0.86	0.94/1.00	0.04/0.02↓
$a-v$	1.29/1.20	0.73/1.13	0.98/1.00	1.00/1.00	0.05/0.01↓
$a-A$	0.88/0.95	0.58/0.61	0.86/0.66	0.77/0.74	0.02/0.02
$v-A$	1.20/1.10	0.82/1.05	1.05/1.22	1.02/1.12	0.02/0.01↓

As shown in Table 5, it can be observed that both before and after training, the text and video modalities have the highest difference in valence labels for all three annotators, while the audio and "all" multimodal level exhibit generally the highest consistency. In terms of the standard deviation, there is a general decrease across most modality pairs, indicating that the training session (or the gold standard) leads to a decrease in valence diversity among annotators.

Inter-Annotator Agreement In Table 6, it is clearly shown that there are increases in agreement over the four modality levels, especially in the video modality. Given the improvement on the video modality, the difference in agreement scores among modalities becomes smaller, and a more equal agreement across modalities is achieved. As for the interval Krippendorff's alpha [39] scores, increases are observed over the modality of video and the multimodal setup after training, while the agreement scores become slightly worse in the modalities of text and audio. We hypothesize that this is a result of annotator 2 choosing more extreme valence scores after training.

Table 6
Inter-annotator agreement on valence for the different modalities before (bf) and after (af) annotator training. α_c means categorical α, and α_i means interval α. The higher the agreement score, the more similar perspectives the annotators have.

	text		audio		video		all	
	bf	af	bf	af	bf	af	bf	af
κ	0.325	**0.437**	0.300	**0.337**	0.105	**0.342**	0.307	**0.315**
α_c	0.332	**0.443**	0.308	**0.344**	0.115	**0.349**	0.315	**0.323**
α_i	**0.702**	0.698	**0.700**	0.567	0.414	**0.528**	0.542	**0.705**

Although an increase in agreement implies a decrease in the variety of perspectives, we have to try to strike a subtle balance between the two. In our experiments, the kappa and alpha scores are modest before annotator training, and they remain modest even taking into account the small increases in agreement after the training, which means the perspectives are still represented in the annotations. Furthermore, we believe the training session could help to boost annotation agreement for the highly inconsistently annotated modalities, as shown especially in the modality of video, where the kappa was as low as 0.105 before training, and increased to a modest 0.342 after training. We could thus conclude that in order to keep enough diversity in annotation perspectives, while at the same time keeping the annotation quality high enough, it is necessary and helpful to keep a moderate training session before starting annotations.

4.3. Inconsistency Analysis

In addition to investigating the valence annotations of the different annotators for the four modality levels, we also performed an inconsistency analysis at the video clip level. The annotation inconsistency among modalities is considered to contain useful information as it might give more insights on what makes some video segments more difficult to label than others. Furthermore, related work [8] also suggests that the stronger the inconsistency, the better the complementarity of intermodal fusion. In this section, we briefly discuss the inconsistency distribution in our corpus. In doing so, we not only focus on valence, but also on more fine-grained emotions.

4.3.1. Inconsistency Distribution of Sentiment

For the inconsistency analysis, we had a closer look at the inconsistent annotations across modalities in the full corpus of 94 video clips; the full corpus was chosen for this analysis in order to have a sufficient amount of annotations for the analysis. We calculated the inconsistency score in a similar way to the difference score, formulated as:

$$I^n = \sqrt{\frac{1}{\binom{4}{2}} \sum_{(x,y)\in\mathcal{M}} (x_n - y_n)^2}, \quad n \in \mathcal{N} \tag{2}$$

whereby $\binom{4}{2}$ is the number of different ways to select two modalities from a set of four, $x, y \in \mathcal{M}$ are the considered modalities, $\mathcal{M} = \{text, audio, video, all\}$, N is the number of video clips, and x_n and y_n represent the assigned valence for clip n in modality x and y, respectively. The sentiment polarities are the labels assigned to the multimodal setup.

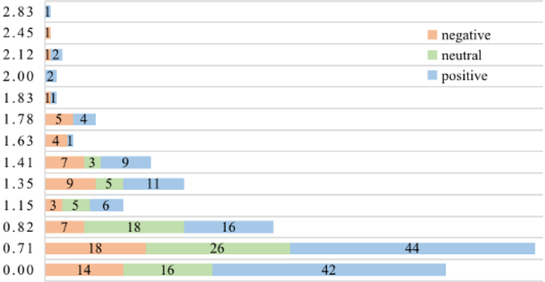


Figure 3: Distribution of inconsistency in 282 annotations from 3 annotators on 94 video clips. The vertical axis represents the inconsistency score, and the horizontal axis shows the number of annotations. The higher the inconsistency score, the more inconsistent the annotations.

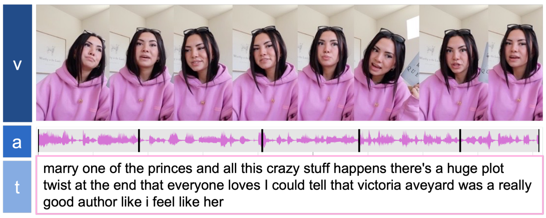
As shown in Figure 3, we can see there are some annotations with an inconsistency score of zero, which means they have the same valence score for the four modality levels. These consistent annotations account for 25.5% of the total number of 282 annotations across the three annotators, and among these annotations, the positive sentiment accounts for about 60%, while the remainder of the annotations is fairly equally spread over the other two sentiment states. The rest of the annotations were divided in two groups with different degrees of inconsistency: those with the top 25% (specifically 23.8%, as this is the nearest group of consistency scores) greatest inconsistency scores were considered as strongly inconsistent annotations, and the remainder of the annotations (50.7%) were considered as weakly inconsistent. The cutoff point for this division between the two groups was an inconsistency score of 1.35.

Table 7
Sentiment distribution in annotations in terms of consistency.

	negative	neutral	positive	total
total	70 / 24.8%	73 / 25.9%	139 / 49.3%	282
consistency	14 / 19.4%	16 / 22.2%	42 / 58.3%	72
inconsistency	56 / 26.7%	57 / 27.1%	97 / 46.2%	210
weak	28 / 19.6%	49 / 34.3%	66 / 46.1%	143
strong	28 / 41.8%	8 / 11.9%	31 / 46.3%	67

As shown in Table 7, when comparing the total annotations and the consistent annotations, it is found that the percentage of negative sentiment drops down from 24.8% in total annotations to 19.4% in consistent annotations, and the percentage of positive sentiment increases from 49.3% in total annotations to 58.3% in consistent annotations, while the percentage of neutral sentiment experiences nearly no change. Therefore, it is indicated that positive sentiment is associated with more consistent annotations across modalities. On the other hand, although there are no significant changes in sentiment distribution from total annotations to inconsistent annotations, we recognize changes in the sentiment distribution when following the above concept of strong inconsistency. It is noticed that the positive sentiment is distributed nearly equally in weakly and strongly inconsistent annotations, while negative sentiment is more located in strong inconsistency and neutral sentiment is more located in weak inconsistency. We hypothesize that people tend to fully show their positive sentiment across modalities, while they are less encouraged to fully show the negative sentiment, resulting in strong inconsistency across modalities. At the same time, we cannot rule out other possibilities due to the small size of our dataset.

4.3.2. Inconsistency in Emotions



marry one of the princes and all this crazy stuff happens there's a huge plot twist at the end that everyone loves I could tell that victoria aveyard was a really good author like i feel like her

Figure 4: Example of an inconsistent annotation. Annotator 1 labeled the text with *enthrallment & valence 4*, audio with *contentment & valence 4*, video with *disappointment & valence 1*, and multimodal setup with *irritation & valence 2*.

The strong inconsistency in valence scores (as illustrated in Figure 4) also gives insights into emotion inconsistency, since varied polarities are linked to varied emotions. The strongly inconsistent instances (as introduced in Section 4.3.1) were taken as a starting point for the investigation into the inconsistency in the emotion labels. To this end, we took the multimodal setup as standard and counted the number of times each unimodal modality (text, audio, and video) received the same categorical emotion annotation as the multimodal setup. The

Table 8
The number of cases in which emotion annotations at the unimodal level correspond to annotations at the multimodal level.

emotion	multimodal	text	audio	video
contentment	3	2	1	1
disappoint	2	1	1	1
disgust	2	2	1	1
embarrassment	1	0	0	1
enthusiasm	4	3	3	0
joy	**7**	**1**	**2**	**5**
longing	1	1	0	0
love	1	0	0	1
neureal	4	3	3	1
relief	1	0	1	0
sadness	3	3	0	0
suprise	5	2	3	1
total	**34**	**18**	**15**	**12**

results of this procedure are shown in Table 8. We found that in the 67 instances that were classified as strongly inconsistent, only 34 instances have at least one match in emotion label with the unimodal setups. In other words, in nearly half of the cases, each unimodal setup has a different emotion label than the multimodal setup. Upon further investigation, we found that there were 18 cases where the annotations of the text modality were consistent with the multimodal setup. The corresponding numbers for audio and video were 15 and 12, respectively.

Again, the text modality seems to share the highest consistency with the all modality setup in terms of emotion labels, which is in line with the results we obtained for the polarity annotations. Furthermore, some emotions seem to be more associated with one specific modality. For example, the emotion *joy* has a stronger association with the video modality, as shown in Table 8: the video modality has 5 consistent cases, while the other two modalities have 1 and 2 instances, respectively. This kind of association within inconsistent cases, though weak, gives a clue on how to explain the emotion labels of the multimodal setup.

5. Discussion

Each of the modalities in multimodal communication offers a unique perspective on the communication. Imagine that someone has a hearing impairment and suffers from prosopagnosia, meaning that this person can read but not hear nor recognize facial expressions, one blind who cannot see but hear, and one deaf-mute who cannot lipread. When standing in their shoes, we can experience the perspectives represented by the modalities of text, audio, and silent video in isolation.

One predominant unimodality? More than fifty years ago, Albert Mehrabian [41] presented an equation of feeling as *Total feeling = 7% verbal feeling + 38% vocal feeling + 55% facial feeling*, according to which facial expressions contribute the most to multimodal emotion. Recently, Liu et al. [42] verified that facial expressions indeed play a more dominant role than emotive markers of text in emotion perception. However, our results show that the modality of text has more effect on multimodal emotion recognition than the modality of silent video. This difference in results might be attributed to the data used for the experiments. First, the images of speakers with facial expressions in Liu's experiments [42] were reproduced from the Amsterdam Dynamic Facial Expression Set [43], in which the facial expression of a particular emotion was intentionally portrayed by actors. We can expect these acted emotions to be much more outspoken than emotional expressions in genuine interaction in real-life scenarios [44]. The facial expressions in our data come from non-acted recordings and we can assume that the genuine emotions conveyed in our data are far more complex and subtle than acted emotions. Furthermore, in our pilot study, we used dynamic displays (silent videos) of emotions instead of static pictures of facial expressions. It has been observed that the identification accuracy of facial expressions reaches near perfection when using images portraying fully developed and intense emotional states [45]. Finally, we also observed that, although the annotators were instructed to score the audio sentiment only with the acoustic clues rather than the language in spoken form, it was inevitable that the spoken language was automatically transcribed in their mind and exerted influence on the annotation. The annotation of audio is thus not totally independent from the modality of text and one of the possible reasons why the results of text and audio are intertwined, as shown in Figure 2.

To train or not to train? The training session in our annotation experiments could be seen as the construction of an adjudicated gold standard, the use of which is believed to have some negative effects as it ignores the diversity of opinions [24] due to the reduction of perspectives. To quantitatively measure this change of annotators' perspectives, we measured inter-annotator agreement before and after training. Our results show that after the training session, inter-annotator agreement indeed generally increased and was especially beneficial for the annotation of silent video. Overall, however, IAA rates remained modest for all unimodalities, safeguarding sufficient diversity in annotation perspectives.

Any benefits to inconsistency? Emotion expression varies in different modalities, and speech and facial expressions are under the control of different muscles. Facial expressions involve the movement of facial muscles, while speech involves the movement of vocal cords and muscles in the mouth. These different muscle combinations and physiological processes can lead to varying rates of emotional changes across different modalities. Therefore, by observing the rapid changes in emotional expression in one modality, it may be inferred that there will be corresponding changes in the emotional expression of another modality, even though these changes might not be as apparent or may occur at a slower pace. As is shown in Figure 5, sometimes the sentiment scores of audio and video are at different changing rates. The ones in blue circles show the different changing rates in positive sentiment, while the ones in red circles represent the difference in negative sentiment.

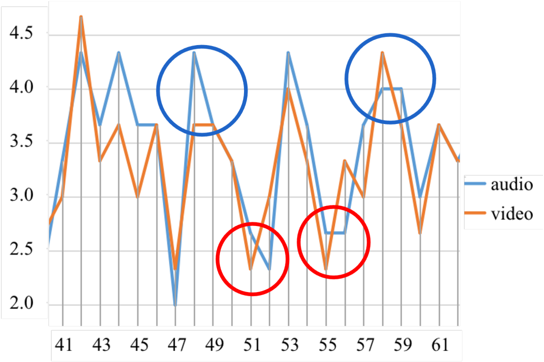


Figure 5: Valence scores of audio and video averaged from three annotators. X-axis is the video clip series and the Y-axis is the valence score. The earlier the valence scores change, the earlier the emotional state changes are detected.

Inconsistency might also give more insight into complex emotions. For example, text-only annotations do not allow to detect all types of irony, but an inconsistent audio or facial expression could account for the irony effect. Of course, basic emotions [11] can also be complicated when emotions in different modalities contradict each other, as shown in Figure 4. However, from a computational perspective, this inconsistency might help models to learn more differentiated information and improve the complementarity between modalities [8].

6. Conclusion

In this paper, we presented a pilot study on the basis of a newly collected video corpus which we annotated both at the level of the single modalities (text, speech, video) and the multimodal level. Both dimensional and categorical emotion annotation were provided for these four annotation setups, based on the assumption that unimodalities can also serve as unique perspectives in multimodal emotion recognition.

After a first annotation round, we organised a training session to build an adjudicated gold standard and to make the three annotators more acquainted with the other annotators' perspectives. From the valence annotations before training, we concluded that the text modality seems to dominate the other modalities and resembles most closely the multimodal annotations. After training, perspective diversity was reduced on the annotator level, as evidenced by the general decrease of standard deviation, and on the inter-annotator level, but the fairly modest IAA scores and more stability in IAA across the different modalities might advocate for a training session.

An analysis of the emotional annotation inconsistency among modalities showed that the inconsistency has almost equal distribution in positive and negative sentiments, and that inconsistency is most often located in the modality of silent video. We believe that this inconsistency is of special interest for future studies in more fine-grained emotion modeling. In future studies, apart from scaling up the dataset, we will also investigate in more depth the other annotation layers, such as the multilabel emotion annotations, the trigger of the emotions, and the annotation time.

Acknowledgments

This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174Z05623) and from the Research Foundation Flanders (FWO-Vlaanderen) with grant number 1S96322N. We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

References

[1] R. W. Picard, Affective Computing, MIT Press, 1997. doi:10.7551/mitpress/1140.001.0001.

[2] S. Labat, H. Amir, T. Demeester, V. Hoste, An emotional journey: Detecting emotion trajectories in Dutch customer service dialogues, in: Proceedings of the Eighth WNUT, ACL, 2022, pp. 106–112. URL: https://aclanthology.org/2022.wnut-1.12.pdf.

[3] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76. doi:10.1016/j.specom.2019.12.001.

[4] W. Mellouk, W. Handouzi, Facial emotion recognition using deep learning: Review and insights, Procedia Computer Science 175 (2020) 689–694. doi:10.1016/j.procs.2020.07.101.

[5] G. R. Kress, T. Van Leeuwen, Reading images: The

grammar of visual design, Psychology Press, 1996. doi:10.5860/choice.34-1950.

[6] S. K. D'Mello, J. K. Westlund, A review and meta-analysis of multimodal affect detection systems, ACM Computing Surveys 47 (2015) 1–36. URL: https://doi.org/10.1145/2682899. doi:10.1145/2682899.

[7] J. Chen, C. Sun, S. Zhang, J. Zeng, Cross-modal dynamic sentiment annotation for speech sentiment analysis, Computers and Electrical Engineering 106 (2023) 108598. doi:10.1016/j.compeleceng.2023.108598.

[8] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of ACL, ACL, Online, 2020, pp. 3718–3727. URL: https://aclanthology.org/2020.acl-main.343. doi:10.18653/v1/2020.acl-main.343.

[9] K. L. O'Halloran, Multimodal discourse analysis, Companion to Discourse. London and New York: Continuum (2011).

[10] C. E. Jewitt, The Routledge handbook of multimodal analysis, Routledge/Taylor & Francis Group, 2011.

[11] P. Ekman, An argument for basic emotions, Cognition & Emotion 6 (1992) 169–200. doi:10.1080/02699939208411068.

[12] R. Plutchik, A general psychoevolutionary theory of emotion, in: Theories of emotion, Elsevier, 1980, pp. 3–33. doi:10.1016/B978-0-12-558701-3.50007-7.

[13] A. Mehrabian, J. A. Russell, An approach to environmental psychology, MIT Press, 1974.

[14] J. A. Russell, A circumplex model of affect, Journal of personality and social psychology 39 (1980) 1161. doi:10.1037/H0077714.

[15] J. R. Fontaine, K. R. Scherer, E. B. Roesch, P. C. Ellsworth, The world of emotions is not two-dimensional, Psychological science 18 (2007) 1050–1057. doi:10.1111/j.1467-9280.2007.02024.x.

[16] E. Troiano, L. Oberländer, R. Klinger, Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction, Computational Linguistics 49 (2023) 1–72. URL: https://doi.org/10.1162/coli_a_00461. doi:10.1162/coli_a_00461.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (2008) 335–359. doi:10.1007/s10579-008-9076-6.

[18] C. Busso, S. Parthasarathy, A. Burmania, M. Abdel-Wahab, N. Sadoughi, E. M. Provost, MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception, IEEE Transactions on Affective Computing 8 (2016) 67–80. doi:10.1109/TAFFC.2016.2515617.

[19] L. De Bruyne, O. De Clercq, V. Hoste, Annotating affective dimensions in user-generated content: Comparing the reliability of best–worst scaling, pairwise comparison and rating scales for annotating valence, arousal and dominance, Language Resources and Evaluation (2021) 1–29. doi:10.1007/s10579-020-09524-2.

[20] S. Labat, T. Demeester, V. Hoste, Emotwics: A corpus for modelling emotion trajectories in dutch customer service dialogues on twitter, Language Resources and Evaluation. Accepted (2022). URL: http://hdl.handle.net/1854/LU-8769949.

[21] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of ACL, ACL, Melbourne, Australia, 2018, pp. 2236–2246. URL: https://aclanthology.org/P18-1208. doi:10.18653/v1/P18-1208.

[22] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of ACL, ACL, Florence, Italy, 2019, pp. 527–536. URL: https://aclanthology.org/P19-1050. doi:10.18653/v1/P19-1050.

[23] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, T.-H. Huang, L.-W. Ku, EmotionLines: An emotion corpus of multi-party conversations, in: Proceedings of the Eleventh ICLRE, ELRA, Miyazaki, Japan, 2018, pp. 1597–1601. URL: https://aclanthology.org/L18-1252.

[24] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, arXiv preprint arXiv:2109.04270 (2021). doi:10.48550/arXiv.2109.04270.

[25] L. Aroyo, C. Welty, Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard, in: ACM Web Science 2013, 2015. doi:10.6084/M9.FIGSHARE.679997.V1.

[26] H. M. Fayek, M. Lech, L. Cavedon, Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels, in: 2016 IJCNN, 2016, pp. 566–570. doi:10.1109/IJCNN.2016.7727250.

[27] V. Prabhakaran, A. Mostafazadeh Davani, M. Diaz, On releasing annotator-level labels and information in datasets, in: Proceedings of the Joint 15th LAW and 3rd DMRW, ACL, Punta Cana, Dominican Republic, 2021, pp. 133–138. URL: https://aclanthology.org/2021.law-1.14. doi:10.18653/

v1/2021.law-1.14.

[28] H. M. Fayek, M. Lech, L. Cavedon, Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels, in: 2016 IJCNN, 2016, pp. 566–570. doi:10.1109/IJCNN.2016.7727250.

[29] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, Y. Aono, Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification, in: 2018 IEEE ICASSP, 2018, pp. 4964–4968. doi:10.1109/ICASSP.2018.8461299.

[30] Y. Zhang, Q. Yang, A survey on multi-task learning, IEEE Transactions on Knowledge and Data Engineering 34 (2021) 5586–5609. doi:10.1109/TKDE.2021.3070203.

[31] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of ACL 10 (2022) 92–110. URL: https://aclanthology.org/2022.tacl-1.6. doi:10.1162/tacl_a_00449.

[32] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, K. Gao, Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module, in: Proceedings of the 2022 ICMI, ACM, New York, NY, USA, 2022, p. 247–258. URL: https://doi.org/10.1145/3536221.3556630. doi:10.1145/3536221.3556630.

[33] J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, H. Li, M3ED: Multi-modal multi-scene multi-label emotional dialogue database, in: Proceedings of the 60th Annual Meeting of ACL (Volume 1: Long Papers), ACL, Dublin, Ireland, 2022, pp. 5699–5710. URL: https://aclanthology.org/2022.acl-long.391. doi:10.18653/v1/2022.acl-long.391.

[34] P. Shaver, J. Schwartz, D. Kirson, C. O'connor, Emotion knowledge: Further exploration of a prototype approach, Journal of Personality and Social Psychology 52 (1987) 1061. doi:10.1037//0022-3514.52.6.1061.

[35] M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry 25 (1994) 49–59. doi:10.1016/0005-7916(94)90063-9.

[36] R. A. Fisher, Statistical methods for research workers, Springer, 1992.

[37] X. Li, M. Chen, Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language, in: Proceedings of the 19th CNCCL, CIPSC, Haikou, China, 2020, pp. 1089–1100. URL: https://aclanthology.org/2020.ccl-1.101. doi:10.1007/978-3-030-63031-7_26.

[38] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378. doi:10.1037/h0031619.

[39] K. Krippendorff, Content analysis: An introduction to its methodology, Sage Publications, 2018.

[40] E. G. Krumhuber, A. Kappas, A. S. Manstead, Effects of dynamic aspects of facial expressions: A review, Emotion Review 5 (2013) 41–46. doi:10.1177/1754073912451349.

[41] A. Mehrabian, Silent messages, Wadsworth Belmont, 1971.

[42] M. Liu, J. Schwab, U. Hess, Language and face in interactions: Emotion perception, social meanings, and communicative intentions, Frontiers in Psychology 14 (2023) 1146494. doi:10.3389/fpsyg.2023.1146494.

[43] J. Van Der Schalk, S. T. Hawk, A. H. Fischer, B. Doosje, Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set, Emotion 11 (2011) 907. doi:10.1037/a0023853.

[44] E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, C. Cox, Multimodal databases of everyday emotion: Facing up to complexity, in: Ninth ECSCT, 2005, pp. 813–816. doi:10.21437/Interspeech.2005-381.

[45] J. M. Carroll, J. A. Russell, Facial expressions in hollywood's protrayal of emotion, Journal of Personality and Social Psychology 72 (1997) 164. doi:10.1037/0022-3514.72.1.164.