UniC: a Dataset for Emotion Analysis of Videos with Multimodal and Unimodal Labels

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

[1*]LT3, Ghent University, Groot-Brittanniëlaan 45, Ghent, 9000, Flanders, Belgium.
[2]IDLab, Ghent University - imec, Technologiepark-Zwijnaarde 126, Ghent, 9052, Flanders, Belgium.

*Corresponding author(s). E-mail(s): quanqi.du@ugent.be;
Contributing authors: sofie.labat@ugent.be;
thomas.demeester@ugent.be; veronique.hoste@ugent.be;

Abstract

Emotion is a key characteristic that differentiates humans from machines. It is intricate, encompassing a wide variety of emotional states, and is expressed through both verbal and non-verbal communication channels. Different modalities contribute in unique ways to the integrated expression of emotion. However, in most of the existing multimodal datasets, there is only one unified emotion label for the various modalities, ignoring the heterogeneity and complementarity of the different modalities. To bridge this gap, we introduce UniC, a novel multimodal emotion dataset featuring both integrated multimodal labels and independent unimodal labels. UniC is comprised of 965 emotion-rich video clips selected from YouTube, annotated in text, audio, silent video, and multimodal setups with both categorical and dimensional labels. We outline the steps taken to construct the dataset and analyze different modality perspectives in UniC. Our findings indicate that while in most cases the modality of text shares more emotional resemblance with the multimodal setup, other modalities can exhibit different, sometimes even opposite emotions that might contribute more to the overall emotion state. This dataset offers a modality-specific perspective on multimodal emotion analysis and has the potential to provide valuable insights for further research in human emotion understanding.

Keywords: Unimodal and Multimodal Labels, Text, Speech, Video, Sentiment and Emotion Modelling

1 Introduction

We chat with big smiles, question with frowned eyebrows, and argue with glaring eyes. "People are their emotions" (Denzin, 1984). Thanks to verbal and non-verbal aids, we can express our emotions and understand those of others.

Since the computational study of emotion was explicitly suggested (Pfeifer, 1982), scholars have explored various methods for analyzing and modelling emotion to enable machines to better understand human affect and emotions, with the ultimate goal of "emotional intelligence" (Picard, 1997). The emergence of single-modal corpora has significantly advanced emotion recognition research, such as the SemEval-2018 Affect in Tweets Dataset (Mohammad, Bravo-Marquez, Salameh, & Kiritchenko, 2018), the EMO-DB (Burkhardt et al., 2005) and the FER2013 dataset (Goodfellow et al., 2013) for textual, audio and facial emotion recognition, respectively. While single-modal research on emotion recognition has yielded promising results, it focuses on only one aspect of emotion. In daily life, emotions often manifest as multimodal, dynamic patterns of behaviour (Keltner, Sauter, Tracy, & Cowen, 2019), involving vocalization, facial expression, and more.

By considering multiple modalities together, multimodal emotion modelling aims to jointly model these different expressions of emotions, ultimately striving for better performance in automatic emotion detection (D'Mello & Westlund, 2015). However, this task is still hindered by the challenge of fusing information from multiple modalities. Often, consistent emotion information across modalities can be represented by a unified single emotion label. Yet, there are many cases where the different unimodal emotion expressions complement each other, or where the emotion expressed in one modality is different from or even contradicts another. For example, imagine a person enthusiastically saying "This movie was awesome!" Both the text and audio signals clearly indicate a positive emotion, but if the same person also rolls their eyes upward while expressing their review, the visual modality reveals irony, indicating a negative emotion.

In most multimodal emotion datasets, only a single emotion label is annotated for all modalities, thus ignoring the independent emotions which may be expressed through each unimodality. To better understand the added value and interaction between each unimodality in multimodal emotion labelling, this paper introduces *UniC*, a multimodal multilabel emotion dataset with independent unimodal labels. The name stands for "Unimodality Counts", highlighting the unique contributions of different unimodalities to tasks such as emotion recognition and emotion modelling.

The remainder of this paper is structured as follows. In Section 2, we start with a brief introduction to NLP approaches for emotion detection from both unimodal and multimodal perspectives. We also discuss various prominent multimodal emotion datasets. Given the lack of datasets with natural emotion expressions and more fine-grained annotations, Section 3 introduces a novel approach for building a multimodal multilabel emotion dataset with independent unimodal labels based on YouTube video material. Section 4 presents the results of a detailed inter-annotator agreement study, including comparisons before and after annotator training, a methodology to enhance agreement through emotion label clustering, and a general agreement study on the entire dataset. Using the independent unimodality and multimodality annotations as

input, Section 5 investigates the emotion complexity in the data by exploring the relationship between emotion states, modalities, and annotators. Section 6 provides an example of the dataset to illustrate differences in emotion across modalities. Finally, we conclude the paper in Section 7.

2 Related Work

From a computational perspective, the study of emotion has gained more and more attention over the last decades. To enable machines to understand human affect and emotion, achieving so-called "emotional intelligence" (Picard, 1997), researchers have worked on different modalities, including text, audio, and facial expressions, to model human emotions. Scholars in natural language processing (NLP) specifically focus on the text modality. Their research ranges from investigating and modelling the overall sentiment or polarity in a given text (Nasukawa & Yi, 2003; L. Yang, Li, Wang, & Sherratt, 2020) to the fine-grained detection of specific polarities associated with different aspects of a product or event (Nazir, Rao, Wu, & Sun, 2020; Pontiki et al., 2016; Zhong et al., 2023), emotion detection in text from tweets or other sources (De Bruyne, De Clercq, & Hoste, 2021b; Ghafoor et al., 2023), and other emotion-related topics such as humour and irony (Barbieri & Saggion, 2014; Maladry, Lefever, Van Hee, & Hoste, 2022; Winters, Nys, & De Schreye, 2018).

Closely related to text, speech can be transcribed into text for emotion analysis, while paralinguistic information such as prosodic and spectral features also aids in emotion classification (Anagnostopoulos, Iliou, & Giannoukos, 2015; Lu, Cao, Zhang, Chiu, & Fan, 2020). Extracting this information from speech is challenging, as features like pitch and energy contours are easily affected by factors such as speakers, speaking styles, and speaking rates. Furthermore, spontaneous speech often features authentic emotion expressions that are more subtle and harder to distinguish than acted emotions (S. Zhang, Zhao, & Tian, 2019). Since speech is always intertwined with text, they typically appear together in emotion-related tasks (Hajek & Munk, 2023).

As for facial emotion recognition, the conventional approach is composed of three major steps: face detection, feature extraction (converting facial features into vectors), and expression classification. The combination of powerful spatial feature extraction of convolution neural networks (CNNs) and temporal feature detection with long short-term memories (LSTMs) (and their variants) has become state-of-the-art (Canal et al., 2022), achieving an accuracy of 78.2% on the FER2013 image dataset (Ming, Qian, Guangyuan, et al., 2022). Transfer learning has also been introduced in facial emotion recognition (Akhand, Roy, Siddique, Kamal, & Shimamura, 2021; Chowdary, Nguyen, & Hemanth, 2023) to speed up training and address the lack of big datasets. This shortage of datasets is especially problematic for the detection of spontaneous emotions in video data.

Due to the rich characteristics and complex distribution of human emotions, it is challenging to identify emotion expressions through unimodal information only. Therefore, multimodal emotion recognition (Gao, Li, Chen, & Zhang, 2020) was introduced to capture this rich emotion information by jointly modelling text, speech, and facial

expressions (Mittal, Bhattacharya, Chandra, Bera, & Manocha, 2020). Recent multimodal models, such as GPT-4V (vision) (Z. Yang et al., 2023) and MultiModal-GPT (Gong et al., 2023), have demonstrated the huge potential of combining text with other modalities, e.g., audio, images, and even videos. Some studies also take electroencephalogram (EEG) and electrocardiogram (ECG) signals as emotion indicators (Pan et al., 2023; H. Zhang, 2020).

Since all the mentioned approaches, both unimodal and multimodal, are data-driven methodologies requiring annotated data for training or fine-tuning, a number of emotion-annotated corpora have been created over the past decades. Given the scope of this paper, we will focus exclusively on multimodal emotion datasets, which can be categorized into acted emotion datasets and natural emotion datasets (see Table 1 for an overview). Acted emotion datasets often include soap operas and movies, which are performed by professional actors with scripted lines and plots, while natural emotion datasets usually consist of recordings of spontaneous speeches from non-professional actors.

IEMOCAP (Busso et al., 2008) is an acted dialogue dataset in English, featuring seven professional actors and three drama students, with five women and five men. The subjects used fixed scripts in the scripted sessions and improvised their own words in the spontaneous sessions. Two schemes were used for emotion annotation: a categorical scheme with labels like anger, sadness, happiness, frustration, and a neutral state, and a continuous scheme using valence, activation and dominance (VAD) attributes with the Self-Assessment Manikin (SAMs) (Fischer, Brauns, & Belschak, 2002). IEMOCAP contains 10,039 dialogue turns.

MSP-IMPROV (Busso et al., 2016) is another acted multimodal emotional database in English, featuring 12 theatre students. Unlike IEMOCAP, the actors were free to use their own language in the set scenarios, as long as they also included specific lexical content that was designed to express a target emotion. In this sense, it might be better to classify MSP-IMPROV as a half-acted dataset. The scenarios were based on 15 sentences believed to be generic enough to trigger target emotions (anger, happiness, sadness, and a neutral state). The corpus was manually segmented into dialogue turns and annotated using crowdsourcing. Emotion annotation used a multilabel strategy, including one primary emotional label and one or more secondary emotion labels, along with VAD attributes. The database consists of 8,438 utterances.

CHEAVD 2.0, also called MEC 2017 (Li et al., 2018), is a multimodal emotion database with 7,030 clips selected from Chinese movies, soap operas and TV shows. For emotion labelling, a discrete emotion annotation scheme was adopted with eight labels: happiness, sadness, worry, anger, anxiety, surprise, disgust, and neutral.

CMU-MOSEI (Zadeh, Liang, Poria, Cambria, & Morency, 2018) is a collection of 3,228 videos with 23,453 sentences from YouTube. To broaden the range of topics, 250 commonly used topics in online videos were selected as seeds and the number of videos from each channel was limited to 10. After several filtering rounds, including quality control and gender balancing, the final video pool contains 3,228 videos, with the most frequent topics being reviews (16.2%), debates (2.9%), and consulting (1.8%). The videos were annotated by three crowdsourced judges from Amazon Mechanical Turk with both sentiment and emotion labels.

Table 1 Summary of available datasets for multimodal sentiment/emotion recognition, featuring audio-visual data. For each dataset, the source (src), language, (lang), average segment length in second (avglen), agreement (agr) and used annotation scheme (categorical = cat; VAD; sentiment = sent) is mentioned.

Dataset	Src	Lang	Segments	Avglen	Agr	Annot
IEMOCAP (Busso et al., 2008)	lab	EN	10039	4.5s	$\kappa = 0.35$	cat + VAD
MSP-IMPROV (Busso et al., 2016)	lab	EN	8438	1.9-3s	$\kappa = 0.49$	cat + VAD
MEC 2017 (Li et al., 2018)	TV	CN	7030		$\kappa = 0.43$	cat
M^{3}ED (Zhao et al., 2022)	TV	CN	24449	7.39s	$\kappa = 0.59$	cat
CMU-MOSEI (Zadeh et al., 2018)	YT	EN	23453	7.28s	N/A	cat + sent
CH-SIMS (Yu et al., 2020)	TV	CN	2281	3.67s	N/A	sent
CH-SIMS v2.0 (Liu et al., 2022)	TV	CN	4402	3.63s	N/A	sent

M^{3}ED (Zhao et al., 2022) is a multimodal, multiscene, multilabel emotion-annotated dialogue dataset, including 990 emotional dialogues from Chinese TV series. Emotion labels cover the six basic emotions from Ekman (1992) and one additional label *neutral*. Each utterance could be annotated with multiple emotion labels, resulting in 11% of the utterances having multiple labels, where the emotion *anger* often co-occurs with other negative emotions, such as *disgust* and *sadness*.

Whereas the previously mentioned datasets have a unified emotion label for all modalities, the *UniC* multimodal multilabel emotion dataset with independent unimodal emotion labels draws inspiration from the CH-SIMS (Yu et al., 2020) dataset construction. CH-SIMS (Yu et al., 2020) employs an annotation strategy in which each modality receives an independent label, along with a label for the multimodality. The authors did not provide fine-grained emotion annotations, and the 2,281 video segments in CH-SIMS were annotated with *negative*, *neutral* and *positive* polarity information. The authors found that the multimodality layer did not always have the same sentiment status as one or more of the unimodality layers. They reported that 20% - 30% of unimodal annotations had inconsistent sentiment polarities compared to the multimodal ones. To enhance and extend the dataset, an additional 4,402 video segments were annotated and added to CH-SIMS, along with 10,161 unsupervised (not annotated) video segments, leading to the release of CH-SIMS v2.0 (Liu et al., 2022).

3 Dataset Collection and Annotation

In this section, we discuss our methodology for both data collection and annotation. Section 3.1 describes the collection and selection of videos using three filtering strategies. In Section 3.2, we detail our approach to labelling the videos, including a discussion of emotion frameworks, multiple rounds of multimodal annotation, and the redesigned emotion labels with cluster analysis.

3.1 Data Collection and Filtering

To obtain English multimodal data potentially rich in emotions, we collected videos from YouTube, one of the biggest online video-sharing platforms. Unlike traditional video sources (e.g., dramas, soap operas, and movies) which are largely composed of acted emotions, YouTube offers easier access to videos with natural and authentic emotion expressions. To collect suitable YouTube videos for the dataset, we defined the following guiding principles:

- Each frame should feature only one subject showing their face clearly in the video, allowing for the detection of non-verbal emotion expressions such as facial expressions and eye movements.
- The subject in the video should speak clearly so that their speech can be easily recognized by human ears.
- The subject should ideally express as much emotion as possible.

Since selecting a suitable dataset among a large amount of videos on YouTube is not trivial,[1] we set three filters to search more efficiently, namely a **keyword filter**, a **subtitle filter**, and a **manual filter**. The keyword filter narrows down the domains from which to sample data, targeting domains we hypothesised to contain emotion-rich content. The subtitle filter identifies emotion-rich samples through sentiment analysis of video transcripts, based on the hypothesis that the emotions expressed in text are accompanied by emotions in the other modalities. Finally, a manual filter ensures the quality and diversity of the resulting dataset.

Keyword Filter

Since we aimed to collect videos featuring clearly visible facial expressions, we started our search by feeding the YouTube search engine with specific topic-related keywords which we thought might involve more positive and negative sentiments. This led us to choose the following five keywords: *book review*, *movie review*, *review* (excluding book review and movie review), *interview*, and *psychological counseling*. The different review-related keywords were selected because reviews, such as those of hotels (Ray, Garain, & Sarkar, 2021), movies (Rehman, Malik, Raza, & Ali, 2019) and airlines (Kwon, Ban, Jun, & Kim, 2021), are popular text types for research in the field of customer sentiment analysis. Furthermore, the topic of *review* is frequent in another popular YouTube video dataset, CMU-MOSEI (Zadeh et al., 2018), accounting for 16.2% of over 3,000 videos. While review videos are mainly monologues, interview and counseling videos often feature two characters interacting, thus covering dialogue scenarios. We aimed to collect 100 videos for each keyword, but due to factors like advertisements, we ended up with a total of 93, 88, 94, 101, and 97 videos for the five keywords, respectively.

[1] As of 2024, YouTube has more than 2.70 billion monthly active users who upload 720,000 hours of video and consume 1 billion hours of video on a daily basis. See https://www.demandsage.com/youtube-stats (accessed March 18, 2024).

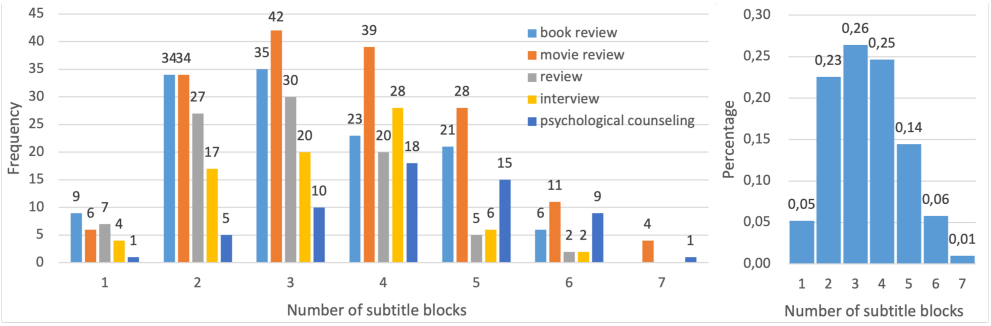

Fig. 1 The number of subtitle blocks containing manually annotated emotions, with one sample video for each keyword. The X-axis represents the number of subtitle blocks, and the Y-axis shows the absolute frequency/percentage. The left plot displays the distribution of 5 videos separately, while the right plot combines the data of these videos. The plots suggest that more than 90% of the emotional signal can be captured within 5 subtitle blocks.

Subtitle Filter

The dataset of 473 videos, resulting from the keyword filtering, was further refined to focus on those videos featuring expressions of emotions. To evaluate the emotion richness in these videos, we ranked them based on the sentiment exhibited in the subtitles, a modality more easily accessible than others (e.g., audio and facial expressions). Subtitles of all videos were downloaded from YouTube, with manually created subtitles preferred over automatically generated ones. These subtitles were subsequently processed using a RoBERTa-base model trained on approximately 58M tweets from the SemEval2017 dataset (Rosenthal, Farra, & Nakov, 2017) and fine-tuned for sentiment analysis (Barbieri, Camacho-Collados, Espinosa Anke, & Neves, 2020).

To evaluate the system on our data, we randomly selected a video (referred to as *Video-Zero* hereafter) and found that about 80% of the subtitles were classified as neutral. A manual check revealed that the subtitle blocks in *Video-Zero* had an average duration of about 1.89 seconds with on average 5.4 words, which might be too brief for accurate sentiment classification. To determine how many original subtitle blocks are needed to identify an emotion, we randomly selected one video for each of the other keywords and manually annotated them along with *Video-Zero*. The results, shown in Figure 1, suggest that about 93% of the emotions can be captured when considering video clips up to 5 subtitle blocks. With 5 subtitle blocks, the average video clip length is about 10 seconds, similar to the average video length in CMU-MOSEI (Zadeh et al., 2018). Therefore, we decided to use a combination of 5 subtitle blocks as input to the sentiment prediction model.

Table 2 shows two cases where sentiment predictions shift after merging additional subtitle blocks. In the first case, a new subtitle block with a positive sentiment is added to four neutral blocks and retains its positive sentiment after combination. In the second case, a variety of polarities at the subtitle block level results in a negative sentiment after merging. These sentiment changes may stem from context variations,

Table 2 Changes in sentiment prediction as the span extends from 1 to 5 subtitle blocks. P1, P5 and M1, M5 refer to sentiment predictions and manual annotations for 1 and 5 subtitle blocks, respectively.

Time	Subtitles	P1	M1	P5	M5
0:07:34	and we're given very clear very clear	neutral	positive	positive	positive
0:07:38	toolbox of strategies that	neutral	neutral		
0:07:41	have been highly researched and have	neutral	positive		
0:07:44	been proven as being	neutral	neutral		
0:07:45	incredibly effective to help people with	positive	positive		
0:09:02	problems so in terms of skills	negative	negative	negative	negative
0:09:04	a counselor is not going to really be	neutral	neutral		
0:09:07	that highly skilled	positive	positive		
0:09:09	in terms of offering really good quality	positive	positive		
0:09:11	solutions	neutral	neutral		

as the absence of context can lead to inaccurate predictions, whereas the proper application of context information can improve the accuracy of sentiment analysis (Kumar & Garg, 2020).

To evaluate the proposed subtitle filter more precisely, we selected five sample videos, each chosen based on one of the five different keywords, and manually annotated the joint subtitle blocks for sentiment. With accuracy scores ranging between 75% and 80%, we deemed these results sufficient to support the filtering process. Consequently, all videos were sorted in ascending order by their percentage of neutral classifications.

Manual Filter

The third and final filter consisted of a manual check, where one of the authors served as the human validator. This process included not only reviewing key elements of the character (e.g., face, voice) in the videos, but also ensuring character diversity to prevent any single subject from dominating the topic or dataset. In this process, we found that videos collected with the keywords *interview* and *psychological counseling* often violated our selection criteria due to the frequent use of "reaction shots" (or "cross-cutting"). These shots cut away from the speaker to show audience's facial expressions and body language (Brown, 2016). While reaction shots help video viewers better understand the emotional content of a scene (Brown, 2016), they also cause a mismatch between the speech of one character and the facial expressions (or body language) of another, making the data unsuitable for our dataset. Consequently, we decided to exclude videos under the keywords *interview* and *psychological counseling*, leaving the dataset with only monologues.

After applying the three filtering steps, we obtained a dataset of 965 video clips, each containing 5 subtitle blocks, for the annotation process. While we clearly described the filters used for selecting video material, we acknowledge that each selection or filtering step during corpus creation inherently introduces some kind of selection bias. For instance, by focusing on English video material across various domains, we inevitably

introduce some cultural and domain bias. The subtitle filtering may have led to some textual emotion bias, while the manual filter might have introduced bias based on the educational, social, or personal background of the human validator. Furthermore, by focusing on the collection and annotation of emotion-rich video material, the dataset naturally has a higher distribution of emotional content. However, the videos still contain non-emotional ("neutral") segments, which may mitigate this bias. This distribution should be taken into account when designing emotion detection systems intended for randomly sampled video data.

3.2 Data Annotation

The collection of manual annotations is the primary and most time-consuming aspect of dataset construction. In this section, we introduce the emotion taxonomy and the multiple modality setups to be annotated. Furthermore, we provide a brief overview of the customized annotation tool and the background of our annotators. We also explain our decision to conduct multiple rounds of annotations to balance diversity and agreement in the emotion annotation task. Finally, we discuss how the initial set of 26 emotion labels was reduced to 7 through a clustering analysis of the categorical emotion labels.

Emotions are typically annotated using categorical or dimensional frameworks. In the categorical framework, anger, disgust, fear, happiness, sadness, and surprise are considered the six basic universal emotions (Ekman, 1992). The dimensional model, on the other hand, projects emotions in a multidimensional space along the axes of valence, arousal, and dominance (Mehrabian & Russell, 1974). Both frameworks are applied in this work. For categorical emotion annotation, we decided not to start with the basic universal emotion framework. Instead, to capture a broad range of relevant emotions, we began with an elaborate set of 25 emotions (anger, contentment, disappointment, disgust, enthrallment, enthusiasm, envy, fear, frustration, irritation, joy, longing, love, lust, nervousness, optimism, pity, pride, rejection, relief, remorse, sadness, suffering, surprise, torment) from Shaver, Schwartz, Kirson, and O'connor (1987), plus a neutral label. Following a pilot study in which we annotated 94 video clips, this set of 26 emotion labels was reduced to a smaller emotion taxonomy through clustering analysis. We will discuss this clustering in Section 3.2.5. For the dimensional framework, we decided to annotate only valence and arousal, since previous studies showed that annotator agreement on dominance was too low for use in machine learning (De Bruyne, De Clercq, & Hoste, 2021a; Labat, Demeester, & Hoste, 2024).

3.2.1 Multimodal Annotation

When communicating feelings or attitudes, emotion information is conveyed not only through verbal communication, but also to a large extent via nonverbal modalities (Mehrabian, 1971). While these different modalities are considered when labelling data with a unified emotion label (as is the case for most multimodal emotion datasets discussed in Section 2), the interplay of these modalities remains unclear with unified labels. Since studies have shown that emotion states in unimodalities may differ from those in a multimodal setup (Du, Labat, Demeester, & Hoste, 2023; Liu et al., 2022),

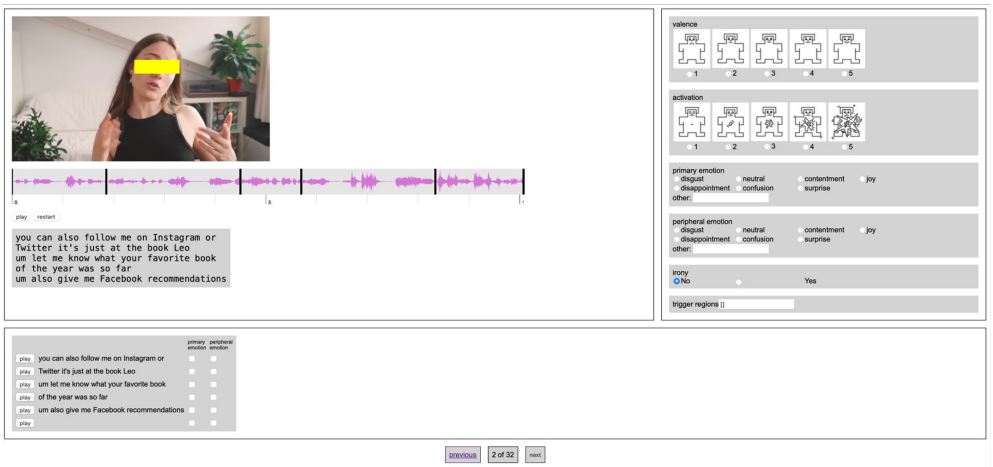

Fig. 2 Annotation interface for the all modality (multimodality) setup with a reduced set of categorical labels following the cluster study. For audio or silent video modalities, the lower part of the interface will not display text for the emotion trigger annotation.

we decided to annotate each video clip in four modalities, namely text, audio, silent video, and all (combining all three modalities). To minimize interference from other modalities, annotators received the four modalities in a shuffled order. For example, the text of clip A is followed by the audio of clip B, the silent video of clip C, and the all modality setup of clip D. We ensured that the four modality setups of one video clip were never presented in neighbouring slots. For the audio modality, annotators were instructed to focus on all audio clues while ignoring the content of the utterances as much as possible.

3.2.2 Annotation Tool

To carry out the annotation work, we built our own in-house annotation tool, as existing tools did not support the complexity of both unimodal and multimodal emotion annotation. While ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) is a widely used annotation tool for audio and video recordings that allows for unlimited textual annotations, it was not suitable for this project for several reasons. First, ELAN is designed for multimodal annotations as a whole and lacks the flexibility to consider each unimodality independently. Additionally, the text-based annotations in ELAN are more time-consuming to enter via the keyboard. To make the annotation process as efficient as possible, we tried to minimize the required human input.

Our annotation tool, shown in Figure 2, is an online tool that facilitates the distribution of assignments across different modalities to annotators and allows for monitoring the annotation process. In the upper left corner of the interface, the raw material (text, silent video, sound, or all modalities combined, as in Figure 2) is presented. In the upper right corner, valence and arousal (activation) are represented using two series of visuals with 5-scale Self-Assessment Manikin (SAM) scores (Bradley

& Lang, 1994). The scores are intended to depict the primary emotion, which is the emotion most clearly conveyed in the data. In addition to the primary emotion, annotators can also annotate peripheral emotions within the categorical framework. While there can be multiple peripheral emotions in the data, annotators must select the one that is most obvious. Each raw material is divided into five parts, as shown in the lower left section, to be clicked as the emotion trigger. The emotion trigger is defined as the word(s), phrase(s), audio, and video span(s) in the data that most strongly convey the emotion. Multiple triggers can be selected. There is also a button for irony, defaulted to "no", to mark the presence of irony for future research.

3.2.3 Annotators

To reduce the noise in the annotation process, we decided to work with expert annotators rather than crowdsourced workers. We hired three students from Ghent University, selected from 14 candidates based on their performance in an annotation test on a transcribed speech sample from a video. All were proficient in English and were paid based on their working time rather than the number of annotations produced. We imposed a time limit of 300 seconds per annotation to ensure that annotators had sufficient time without rushing through the task. Before the annotation task started, annotators received a set of initial guidelines, including dictionary definitions of the different emotions and instructions on using the annotation tool.

3.2.4 Multiple Rounds of Annotation

Emotion annotation is an inherently subjective task, influenced by an individual's background, personal experiences, and world knowledge. To minimize discrepancies among annotators and establish a standardized approach, while preserving the diversity and subjectiveness of emotion annotation, we structured the annotation process into three sessions:

- Session 1: Independent annotation without interference of other annotators. This session was carried out on a subset of the full dataset, referred to as *subset-A*, which includes 94 video clips. These clips were annotated in four modality setups, namely text, audio, silent video, and the multimodal setup combining all three, abbreviated as t, a, v and m, respectively.
- Session 2: A training session of approximately 4 hours, where the three annotators worked together on an annotation assignment. The first two hours were dedicated to jointly annotating the four modality setups for each video clip, which resulted in the annotation of 11 clips (named *subset-A1*). The remaining time was spent jointly annotating 53 clips in the multimodal setup (named *subset-A2*). The initial part aimed to consolidate the annotators' understanding of each modality, while the latter part focused on reaching a consensus on unified emotion evaluation criteria across modalities.
- Session 3: Return to independent annotation. This session involved annotating the remaining 30 clips of *subset-A*, referred to as *subset-A3*, and continued with all other video clips in the *UniC* dataset (named *subset-A'*). Both *Subset-A3* and *subset-A'* were annotated in the four modality setups.

Table 3 Annotation sessions. In total, there are 965 clips in *UniC*, composed of *subset-A*, *subset-A'* and *subset-A"*, while *subset-A* contains 94 clips, further composed of *subset-A1*, *subset-A2* and *subset-A3*. *Set(s) of annotation result* means the number of unique annotation results from individual annotators. *T*, *A*, *V*, *M* stand for the modality setups of text, audio, silent video and the multimodality setup, respectively.

	Session	Clips	Annotated jointly or independently	Set(s) of annotation result	Modalities annotated
subset-A	1	94	independently	3	4 (T, A, V, M)
subset-A1	2	11	jointly	1	4 (T, A, V, M)
subset-A2	2	53	jointly	1	1 (M)
subset-A3	3	30	independently	3	4 (T, A, V, M)
subset-A"	3	60	independently	3	4 (T, A, V, M)
subset-A'	3	811	independently	1	4 (T, A, V, M)

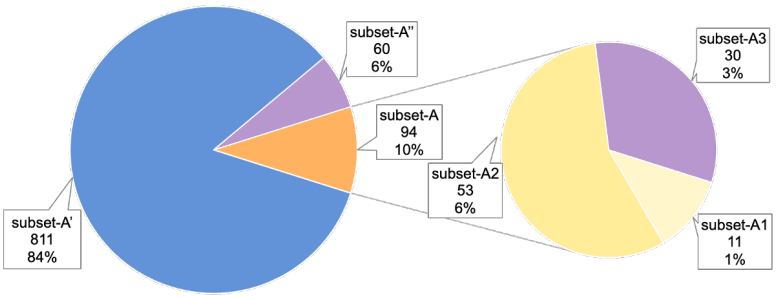

Fig. 3 Dataset *UniC* segmentation, with *UniC* containing 965 clips and composed of *subset-A*, *subset-A'* and *subset-A"*, while *subset-A* containing 94 clips and further composed of *subset-A1*, *subset-A2* and *subset-A3*.

During annotation sessions 1 and 3, annotators worked independently, labelling instances according to their own knowledge and cognitive competence, with or without the influence of the training session. The training session aimed to produce a set of annotations agreed upon by all annotators (gold standard) by considering and summarizing each annotator's perspectives into a single consensus. To ensure equal weight for each perspective (Cabitza, Campagner, & Basile, 2023), we controlled the discussion and negotiation process by appointing one annotator as the chair to lead the session, with the role of chair rotating among annotators.

Subset-A was used in all three sessions, making it possible to compare annotation differences before, during, and after training. In a pilot study on the annotations of *subset-A*, *subset-A1*, *subset-A2* and *subset-A3*, we found that the training in session 2 positively influenced annotator agreement while maintaining sufficient diversity in perspectives (Du et al., 2023). We will revisit this agreement study in Section 4.

3.2.5 Redesigned Categorical Emotion Labels

To investigate whether we could reduce the initial set of 26 emotion labels into a more manageable taxonomy, we performed a clustering analysis using the multimodal annotations of the 94 video clips (*subsets A1-3*), which were fully agreed upon by the

three annotators. By matching the emotion labels from the categorical framework with their average valence and arousal scores from the dimensional framework, we obtained a set of triads. Plotting these triads into a two-dimensional Valence–Arousal space results in a distribution plot of emotions. In this space, we calculate the Euclidean distance between two emotions.

The smaller the distance between two emotions, the more similarities they share. This distance is used as the input for K-means and hierarchical clustering, with the goal of reducing the large set of 26 fine-grained emotion labels into a smaller taxonomy.

The clustering analysis was conducted with the primary emotion labels from the multimodal annotations of 94 video clips obtained from the first and the second annotation sessions. Figure 4 shows the clustering space when partitioning these 94 triads of emotions into 3, 4, 5, 6, and 7 clusters, with each pair grouped to the cluster with the nearest mean (MacQueen, 1967). In these K-means clustering results, each dot represents an emotion label, with the size (diameter) of the dot indicating the frequency of the emotion labels in the 94 clips. Dots of the same colour are clustered in the same group. For example, when the cluster number is set to 4 (upper right part of Figure 4), the emotion *rejection* occurred 6 times and is clustered with *disappointment*, which occurred 12 times. When setting the number of clusters to 3, 4, 5 and 6, the results in Figure 4 reveal some clear clustering patterns, such as (1) *disgust* – *frustration*, (2) *rejection* – *irritation* – *disappointment*, (3) *neutral* – *confusion*, (4) *enthrallment* – *enthusiasm* – *joy* – *love*. However, the clustering of *surprise*, *optimism*, and *contentment* remains unclear. These three emotions are grouped in the same cluster when the cluster number is increased to 5, but *optimism* clusters with *surprise* while *contentment* is excluded when the cluster number is set to 6, and the reverse occurs when increasing the cluster number to 7.

To corroborate the insights from the K-means clustering analysis, we also performed hierarchical clustering. Unlike in K-means clustering, hierarchical clustering does not require pre-defining the number of clusters, and the hierarchical structure of the generated clusters shows the relationships between different data points in an interpretable way, especially when visualized with dendrograms. Figure 5 presents the hierarchical clustering results on the 94 video clips using average agglomerative clustering, where clusters are built bottom-up by treating each data point as its own cluster and iteratively combining the most similar pairs of clusters until all data points are included in a single cluster. We observe that the clustering of negative emotions *rejection* with *disappointment* and *irritation*, and *disgust* with *frustration* confirms our earlier findings from the K-means analysis. The same holds for positive emotions *enthrallment* – *enthusiasm* – *joy* – *love* and for the cluster *neutral* – *confusion*. For the emotions *optimism*, *contentment* and *surprise*, which had scattered results in the K-means clustering, we observe that *optimism* is grouped with *joy*, whereas *contentment* and *surprise* are combined into one cluster. To confirm this grouping, we ran further tests with other linking methods[2]; since six out of seven linking methods grouped *contentment* and *surprise*, we decided to separate *optimism* from that group and put it in the group with *joy*.

[2]The tested methods were average, centroid, complete, median, single, Ward, and weighted linking.

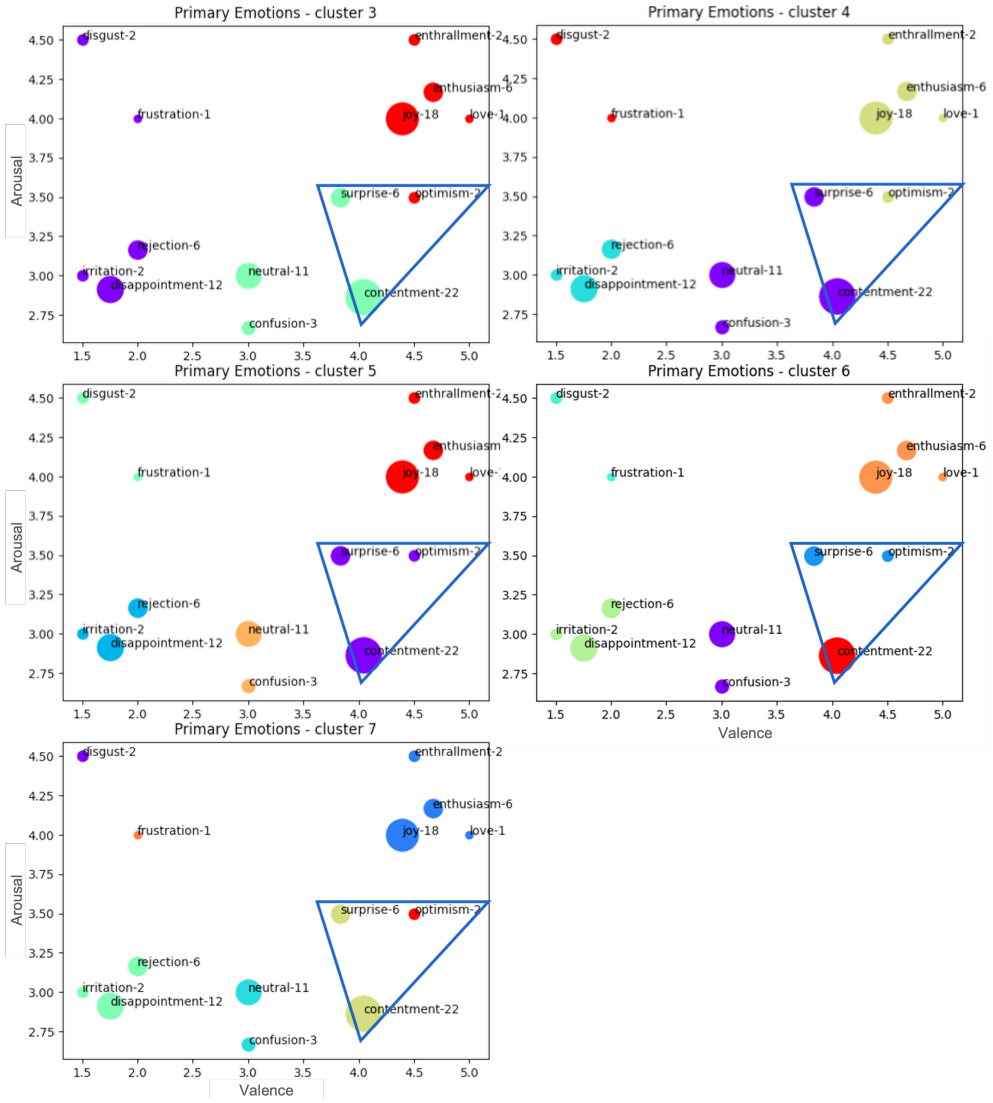

Fig. 4 K-means clustering of primary emotions with the number of clusters ranging from 3 to 7. Emotions are represented as dots, with the diameter indicating the frequency and the colour denoting the cluster. Note that emotions within the blue triangle change clusters more often than others.

With the complementary information from the two clustering analyses, we identified 5 clusters, as shown in the upper part of Figure 6. However, upon further review, we found some improper clusters. Firstly, although *neutral* and *confusion* were clustered together in both clustering methods, they are conceptually different. Despite their proximity in the Valence–Arousal space, the semantic difference between *confusion* and *neutral* is too large to cluster them together, so each was assigned to a separate

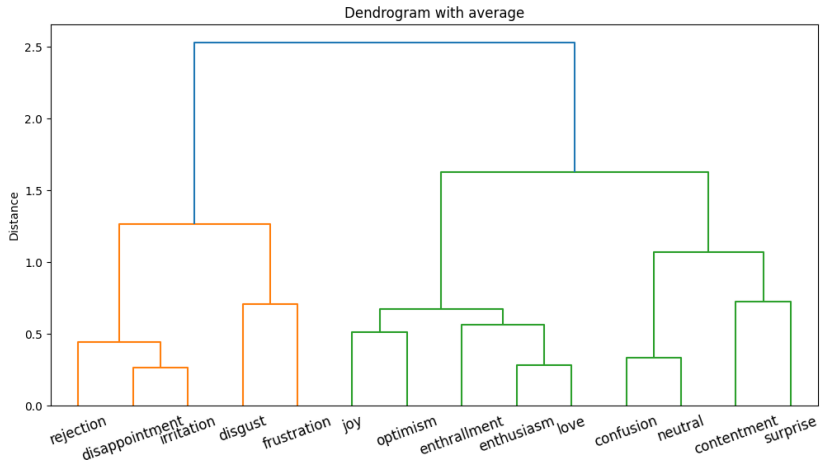

Fig. 5 Hierarchical clustering of primary emotions using average linking. Emotions linked to the same node are considered members of the same cluster.

Disgust Frustration	Disappointment Rejection Irritation	Neutral Confusion	Contentment Surprise	Joy Enthrallment Enthusiasm Love Optimism

Disgust Frustration	*Disappointment* Rejection Irritation	Neutral	Confusion	Surprise	Contentment	**Joy** Enthrallment Enthusiasm Love Optimism

Fig. 6 Initial (upper) and final (lower) sets of emotion clusters

cluster. Additionally, *surprise* and *contentment* were initially combined in one cluster. *Surprise*, one of the basic emotions in Ekman's framework (Ekman, 1973), is the feeling experienced when something unexpected happens, whether good or bad, and is linked to high arousal. In contrast, *contentment* is a feeling of quiet happiness and satisfaction, always has a positive polarity, and is characterised by low arousal. Given these differences, we decided to split the cluster. The final set of seven emotion clusters used for further annotation of the full dataset is shown in the lower part of Figure 6.

To select umbrella terms for the three clusters that contain multiple emotion labels, we first chose labels from Ekman's well-known basic emotion set (Ekman, 1992) (indicated in bold in Figure 6). For one cluster, however, none of the basic emotions could be chosen. In this case, we selected the emotion class with the highest frequency (De Bruyne, De Clercq, & Hoste, 2019) (indicated in italics in Figure 6). The final set of seven emotion clusters exhibits a balance in the polarity of the different emotion labels: *disgust* and *disappointment* share a negative polarity, while *contentment* and *joy* have a positive polarity. Both *confusion* and *surprise* can be linked to more than

one polarity. Notably, in the 94 annotations of the pilot study, all *confusion* annotations had a valence score of 3, while all *surprise* annotations had a valence score of no less than 3, indicating that *confusion* is more neutral and *surprise* is more positive in our dataset.

Emotion clustering from a large set of labels has been previously investigated by, e.g., De Bruyne et al. (2019) on a corpus of emotional tweets and Labat et al. (2024) on customer service dialogues from Twitter. After clustering, De Bruyne et al. (2019) identified a set of 5 emotions, while Labat et al. (2024) developed a taxonomy of 9 emotions that best represented the emotions expressed in their data. When we compare our emotion labels with those from previous studies, we find that the emotions *confusion* and *surprise* are unique to our set. The difference likely arises from the domain of our data, which focuses on reviews, whereas, for example, Labat et al. (2024) focused on customer service interactions. Another potential explanation is that emotions like *confusion* and *surprise* are less easily detected in textual data, and the complementary information from audio and video helps in identifying these emotions. This is further supported by the annotation results discussed in Section 5.1.

Since the seven clusters were obtained from the emotion labels assigned in the multimodal setup, we also mapped the labels of the unimodal setups into these clusters. This was necessary because, in the first annotation session before training, 75 out of 846 unimodal annotations were not covered by the clusters.[3] For instance, a label named *distraction* with a valence of 3 and an arousal of 3 assigned to an audio instance was classified into the cluster of *confusion*, and a label named *anger* with a valence of 2 and an arousal of 2 assigned to a silent video was grouped into *disappointment* instead of *disgust*.

By reducing the initial set of categorical emotion labels to seven clusters, the agreement value κ between the annotators (discussed further in Section 4) in the pilot study increased from 0.212 to 0.318. This improvement led us to use the reduced categorical emotion labels obtained from the cluster study for the annotation of the full dataset.

4 Agreement Study

It has become a common practice to measure inter-annotator agreement (IAA) during or after the annotation process. By comparing annotations on a given instance across different annotators, IAA gives an indication of the reliability of the annotation process, which is essential for proper annotations (Artstein, 2017). It also offers insights into the difficulty of the annotation work and the variety of annotator perspectives (Du et al., 2023). In this section, we investigate the inter-annotator agreement in different settings.

Previous research shows that valence tends to have higher agreement than arousal in the dimensional annotation framework (De Bruyne et al., 2021a; Labat et al., 2024). Therefore, we choose valence as a representative measure of agreement in dimensional annotations. In our experiment, valence annotations are aligned with sentiment

[3] 846 annotations by 3 annotators individually carried out on 3 separate modalities of the 94 clips. Among the 75 annotations not covered by the cluster set, there are 17 primary emotions, with the most frequent emotions being *sadness* (21), *longing* (10) and *nervousness* (9).

polarities, allowing us to evaluate the agreement in sentiment polarities with valence annotations. The following agreement study was conducted on valence and emotion annotations using Fleiss' kappa (κ) (Fleiss, 1971) and Krippendorff's alpha (α) (Krippendorff, 2018) as evaluation metrics. Since sentiment is represented on a 5-point scale and not as discrete values, Krippendorff's alpha with an interval distance function was also calculated for sentiment annotations, considering that the difference between 1 (negative) and 5 (positive) should be bigger than that of 1 (negative) and 3 (neutral).

4.1 Agreement Before and After Training

The three annotators individually annotated the 94 clips before the training session, and after the training, they re-annotated 30 of these clips. The goal of the training session was to establish a common ground for evaluating emotions among the annotators. As shown in Table 4, the overall agreement improved after the training session for both sentiment and emotion annotations. For the annotation of 26 emotion labels, the audio, silent video, and all modality setups clearly show an increase in agreement post-training, while the text modality experienced a small drop in IAA. Before training, the text modality already exhibited the highest agreement scores across different modalities. When we consider the agreement values for sentiment annotations, we observe a notable difference between interval α and categorical α; the higher interval α values suggest greater agreement between adjacent labels. Furthermore, the small difference between interval α for the 5-point scale and the 3-point scale (positive, neutral, negative) indicates that most disagreement occurs between different polarities rather than finer distinctions within the same polarity.

Table 4 Inter-annotator agreement across different modalities before (bf) and after (af) annotator training. α_c refers to Krippendorff's α with categorical distance, and α_i to Krippendorff's α with interval distance. The subscripts 26, 5 and 3 refer to the 26 categorical emotions, and sentiment labels on the 5- and 3-point scales, respectively.

Modality		Text		Audio		Video		Multimodality	
		bf	af	bf	af	bf	af	bf	af
Sentiment(5 labels)	κ_5	.325	**.437**	.300	**.337**	.105	**.342**	.307	**.315**
	α_{c5}	.332	**.443**	.308	**.344**	.115	**.349**	.315	**.323**
	α_{i5}	**.702**	.698	**.700**	.567	.414	**.528**	.542	**.705**
Sentiment(3 labels)	κ_3	.510	**.656**	**.442**	.413	.229	**.536**	.423	**.560**
	α_{c3}	.515	**.660**	**.448**	.419	.237	**.541**	.429	**.565**
	α_{i3}	**.701**	.648	**.705**	.502	.316	**.509**	.640	**.736**
Emotion (26 labels)	κ_{26}	**.296**	.279	.193	**.202**	.175	**.221**	.155	**.325**
	α_{c26}	**.301**	.287	.205	**.214**	.180	**.230**	.164	**.332**

4.2 Agreement Before and After Clustering

Through the process of clustering (Section 3.2.5), the number of categorical emotion labels was reduced from 26 to 7. As the set of categorical emotion labels becomes

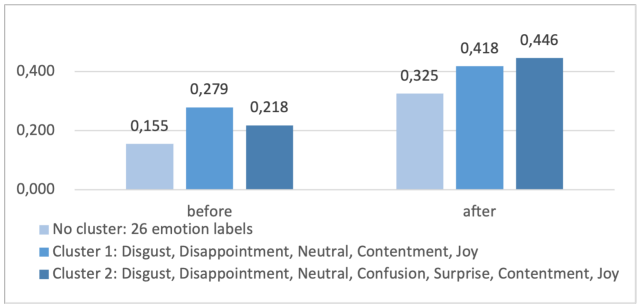

Fig. 7 Inter-annotator agreement with Fleiss' kappa on the multimodality setup for different cluster sizes on *subset-A3*, before and after training. The *no cluster* setup refers to the full set of 26 categorical emotion labels. For emotion label clustering 1, there are 5 clusters: *disgust*, *disappointment*, *neutral*, *contentment*, and *joy*. Compared to clustering 1, clustering 2 has seven clusters by separating *confusion* from the *neutral* cluster and *surprise* from *contentment*.

smaller, inter-annotator agreement should increase accordingly, since the disagreement caused by fine-grained emotion categories is minimized. As shown in Figure 7, the agreement scores of the initial set of 26 emotion labels (no clustering) are $\kappa = 0.155$ and $\kappa = 0.325$ on *subset-A3* before and after the annotator training, respectively. When grouping the 26 labels into 5 clusters (cluster type 1), the kappa scores increase to 0.279 and 0.418 on the two subsets, respectively. However, when separating *confusion* and *surprise* from the *neutral* and *contentment* clusters, which leads to 7 clusters (cluster type 2), the agreement score on the dataset before training drops from 0.279 to 0.218, while the agreement score on the dataset after training increases from 0.418 to 0.446. After training, annotators clearly have a better understanding of the distinctions between the different emotions, and isolating more ambiguous emotions as separate clusters even seems to enhance inter-annotator agreement.

4.3 Overall Agreement

Considering the entire dataset, each of the three annotators was responsible for annotating a portion of the 965 video clips. There is an overlap of 115 video clips, including the 94 clips from the pilot study (see Table 4 in Section 4.1) and an additional 61 clips annotated thereafter. The 61 video clips, annotated by all three annotators after the training session, were randomly sampled from the whole dataset. The sampling was done in two stages: 32 clips were taken in the first half of annotation session 3, and the remaining clips were taken in the second half of the same session. This approach allows us to calculate the agreement throughout the whole annotation process, as shown in Table 5.

It is observed that the modality of text has the highest agreement score for both sentiment and emotions, followed by the video and multimodality setups, while the modality of audio has the lowest agreement score. One possible reason for the higher agreement on the text modality might be that the segment length of the video fragments was based on textual transcriptions, making them better suited for textual emotion annotation. Additionally, this higher agreement score may be attributed to

the fact that sentiment and emotions are often conveyed through clear lexical information, such as positive ("awesome") or negative ("boring") words. People tend to attach the same emotion to a given word or phrase because we share a common understanding of word meanings, while this is less evident in the modalities of video and audio. The use of voice and facial expressions may be more subjective than text, partly due to the lack of clear definitions for their usage.

Table 5 Inter-annotator agreement across different modalities. α_c refers to Krippendorff's α with categorical distance, and α_i to Krippendorff's α with interval distance. The subscripts 7, 5 and 3 refer to the 7 categorical emotions, and sentiment annotations on 5- and 3-point scales, respectively. *T*, *A*, *V*, *M* stand for the modality setups of text, audio, silent video, and the multimodality setup, respectively.

Modality		T	A	V	M	T+A+V	T+A+V+M
Sentiment (5 labels)	κ_5	.404	.226	.304	.247	.320	.305
	α_{c5}	.407	.231	.308	.251	.321	.306
	α_{i5}	.588	.420	.578	.504	.532	.525
Sentiment (3 labels)	κ_3	.548	.315	.449	.395	.444	.436
	α_{c3}	.550	.319	.452	.398	.445	.437
	α_{i3}	.645	.397	.515	.507	.510	.511
Emotion (7 labels)	κ_7	.468	.297	.334	.311	.380	.365
	α_{c7}	.470	.301	.337	.315	.381	.366

When comparing the agreement on the 7 emotion labels with the agreement on sentiment using a 5-point scale (the actual scale used during the annotation process), we could hypothesize that agreement on emotion annotation would be lower since the emotion labels (7) outnumber the sentiment labels (5). However, as shown in Table 5, this does not seem to be the case. Regardless of whether κ or categorical α are used as evaluation metrics, agreement is consistently higher for emotion annotation than for sentiment annotation. However, when joining the *slightly negative* and *slightly positive* labels with *negative* and *positive*, respectively, the agreement on the 3-point sentiment scale annotation improves compared to the initial 5-point sentiment scale and even surpasses the agreement with 7 emotion labels. When we consider the interval α scores, they remain relatively stable across 5- and 3-point sentiment scales, indicating that the disagreement among annotators mostly stems from ambiguity between different polarities rather than ambiguity between different intensities of the same polarity. The results in Table 5 also show higher agreement for the combination of the three unimodal setups compared to the multimodal annotations. This suggests that focusing annotators on the emotional information contained in single modalities leads to a more accurate assessment of the emotional information in the video fragments.

We also compared the agreement scores on our dataset with those from other multimodal datasets, as shown in Table 6. To the best of our knowledge, our dataset is the only multimodal emotion dataset with independent labels for different unimodalities. Therefore, we only considered the agreement on the all modality setup for the comparison. At first sight, the overall agreement on our dataset ($\kappa = 0.31$) seems to be on

Table 6 Inter-annotator agreement on different multimodal (audio-visual) datasets evaluated with Fleiss' κ and/or Krippendorff's α, and the categorical emotion labels used. *Src* means the source of the dataset.

	Src	κ/α	Emotion labels
IEMOCAP	lab	.27/-	9: excited, happy, surprise, neutral, sad, frustration, anger, disgust, fear
IEMOCAP	lab	.35/-	6: happy, neutral, sad, frustration, anger, other
MSP-IMPROV	lab	.49/-	5: happy, neutral, sad, angry, other
MELD	TV	.43/-	7: joy, surprise, neutral, sad, disgust, anger, fear
M^3ED	TV	.59/-	7: happy, surprise, neutral, sad, disgust, anger, fear
CMU-MOSEI	YT	-/.25[1]	6: happy, surprise, sad, disgust, anger, fear
UniC	YT	.31/.32[2]	7: joy, contentment, surprise, confusion, neutral, disappointment, disgust
UniC		.37/.37[3]	

[1] The agreement of the CMU-MOSEI dataset is calculated on each emotion label with Krippendorff's α being 0.41 for *happy*, 0.09 for *surprise*, 0.12 for *sad*, 0.21 for *disgust*, 0.18 for *anger*, and 0.02 for *fear*, with a weighted mean being 0.25 in the multimodal setup (Liang, Salakhutdinov, & Morency, 2018). To the best of our knowledge, the originally reported agreement by the authors is no longer accessible on arXiv.

[2] For annotations of the multimodal setup in UniC, the values for Fleiss' κ and Krippendorff's α are 0.31 and 0.32, respectively (see also Table 5).

[3] When combining annotations of the four setups in UniC, the values for κ and Krippendorff's α are 0.37 and 0.37, respectively (see also Table 5).

the lower end of the spectrum. However, several factors contribute to this moderate agreement score.

First, we believe that authentic and natural emotional expressions are more challenging to annotate than acted emotions. Professional actors in TV series (e.g., MELD and M^3ED datasets) and drama students at university (e.g., IEMOCAP and MSP-IMPROV) are trained to express emotions in a clearer and more explicit way than in natural settings, such that the audience can understand them easily. In contrast, YouTube bloggers (e.g., CMU-MOSEI and UniC) are not professionally trained actors and tend to show their emotions more naturally. Moreover, the quantity and quality of the emotion label set influence the agreement. As shown in IEMOCAP (Busso et al., 2008), reducing the emotion label set from 9 labels to 6 labels leads to a noticeable increase in agreement, with the all-inclusive label *other* alleviating disagreement. Our clustering study in Section 4.2 shows that agreement increases as the number of emotion labels decreases. These findings indicate that a fine-grained emotion label set impacts inter-annotator agreement on emotion annotations.

Considering all these factors, we believe the agreement on our dataset is fairly good, especially given the genuine input data with natural emotion expressions. This will support the development of systems to recognize genuine and non-acted emotions.

5 Unimodal and Multimodal Perspective on Emotion Expression

Our dataset features both multimodal emotion annotations and independent annotations for each modality, providing a rich dataset to assess the complexity of emotions from both unimodal and multimodal perspectives. Figure 8 gives an overview of the annotated dataset for all 965 video clips.

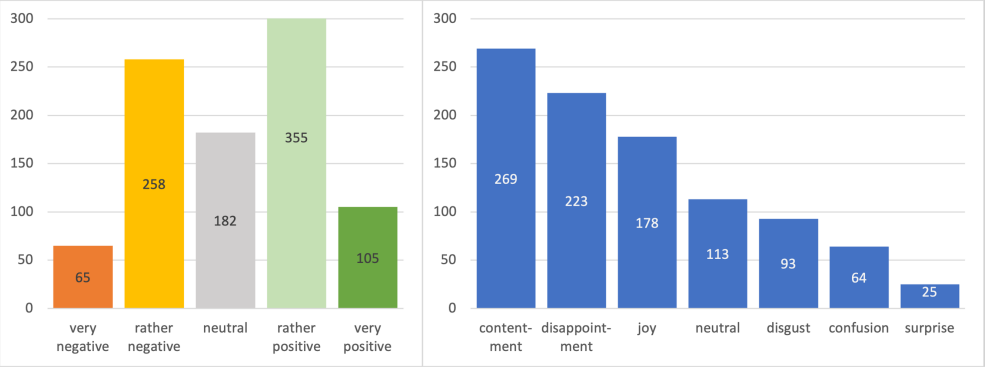

Fig. 8 Distribution of sentiment and emotions as primary emotional states in the *UniC* dataset.

In total, 323 instances are annotated with negative sentiment, while 460 instances are annotated with positive sentiment. As for emotions, *contentment* is the most frequent with 269 instances, followed by *disappointment* with 223 instances. Together with *joy*, these three emotions account for about 70% of the dataset. *Surprise* is the least frequent among the seven emotions, with only 25 instances. For peripheral emotion annotations, about 27% of the 965 video clips were annotated with a non-neutral peripheral emotion, making these annotations relatively sparse compared to the primary emotions.

5.1 Emotion Interaction Across Modalities

The independent emotion annotations on each unimodal layer allow us to investigate how emotions interact across modalities. As a first step in this analysis, we quantified the difference in sentiment annotations between any two modalities $m_1, m_2 \in \mathcal{M} = \{T, A, V, M\}$ (i.e., text, audio, video, and all modalities combined). This was done by calculating the root mean squared difference of valence scores for the considered modalities over each clip x in the dataset $\mathcal{D}$, similar to Du et al. (2023):

$$D_{m_1,m_2} = \sqrt{\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}(v_{m_1} - v_{m_2})^2} \tag{1}$$

As shown in Figure 9, the text modality has the lowest distance score when compared to the multimodal setup, whereas silent video has the highest. More specifically, the sentiment distance between video and text is almost 1.5 times greater than that between audio and text. Furthermore, the sentiment annotations of silent video exhibit similarly large distance scores with both text and audio modalities, which themselves have a moderate distance score of 1.33 when evaluated on a 5-point scale.

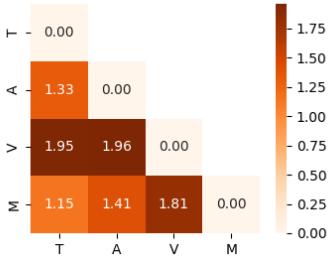

Fig. 9 The distance scores across modalities for sentiment, evaluated on a five-point scale. The patterns remain consistent, even when transforming the five-point scale to a three-point scale. *T*, *A*, *V*, and *M* refer to the modality of text, audio, silent video, and the multimodal setup.

Unlike sentiment annotations, where integer scores are on a 5-point scale, emotions are categorical and cannot be easily assigned numerical values. Instead of directly comparing emotion values, we compared the frequencies of emotion labels in one modality setup to their frequencies in the multimodal setup. We considered the frequency of two co-occurring emotions as a measure of their relative overlap. The results are shown in Figure 10. The X-axis stands for the emotion labels in the multimodal setup, while the Y-axis represents the emotion labels in each unimodal setup. The values in each cell indicate the relative frequencies of an emotion in a unimodal setup compared to the multimodal setup, averaged per column, with absolute frequencies in brackets.

Based on Figure 10, the largest overlap between emotion annotations is, unsurprisingly, found on the diagonal, with averaged co-occurrence frequencies reaching up to 0.73. However, this pattern is less clear for *disgust*, which is often confused with *disappointment* across all modalities. Considering that both *disgust* and *disappointment* share a negative valence score (as shown in Figure 4), one possible reason for the inconsistency of *disgust* across modalities could be its lower prominence when considering a single modality. This leads to the annotation of another less intense negative emotion, *disappointment*. By combining information from each modality, the more intense emotion *disgust* tends to emerge.

Another noteworthy finding in Figure 10 is that *surprise* proves a difficult emotion to annotate, although this should be interpreted with caution due to the limited number of instances labelled as *surprise*. Among the three unimodal setups, the highest frequency of co-occurring with itself is found in the audio modality, suggesting that *surprise* is more clearly expressed through speech. In contrast, it is often confused with *neutral* in text and with *confusion* in silent video. The video clip displayed in Figure 11 supports this finding: the plain text transcript gives little evidence of

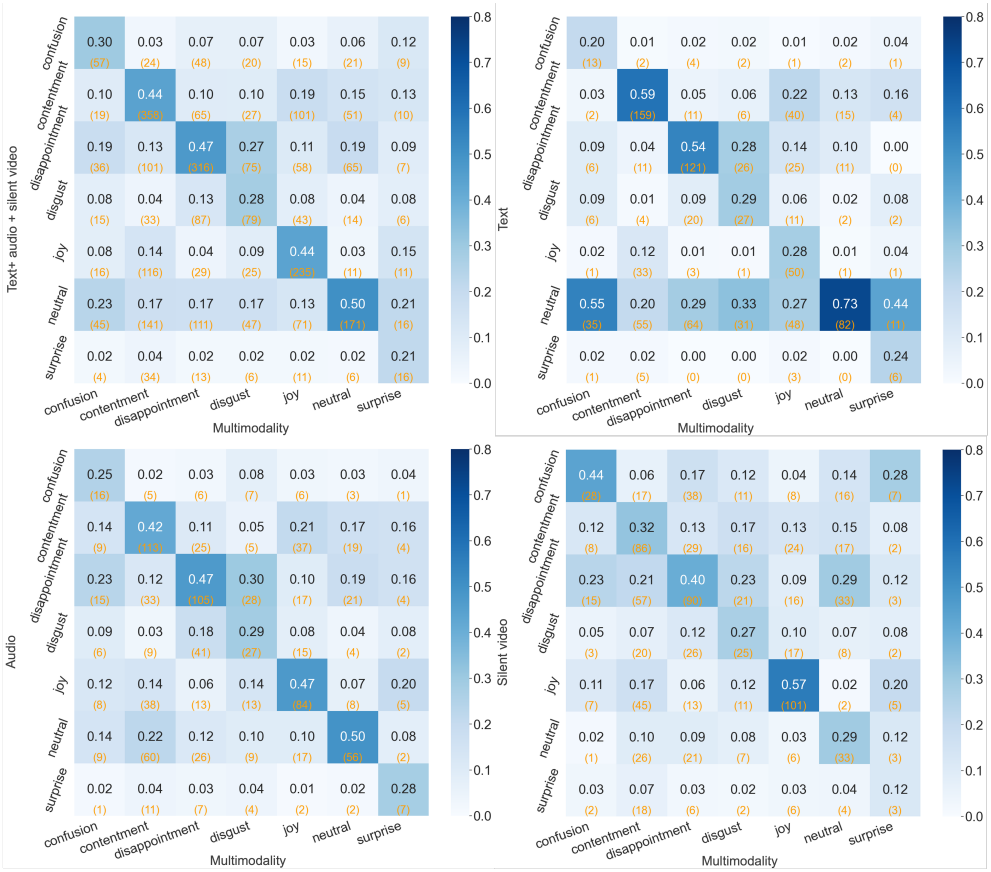

Fig. 10 Emotion overlaps across modalities. The upper left figure represents the combination of three modalities, namely text, audio, and silent video, the upper right represents the text modality, the lower left is audio, and the lower right is for silent video.

surprise, but with the addition of acoustic features (e.g., pauses, stress, tone and, intonation), the emotion *surprise* becomes more discernible. Regarding facial expressions, studies have found that people rarely exhibit a fully surprised facial expression (Reisenzein, Horstmann, & Schützwohl, 2019), so eyebrow-raising is often labelled as a sign of *confusion*. Consequently, a video labelled as *surprise* at the multimodal level may be labelled as *neutral*, *surprise*, and *confusion* in the text, audio, and silent video modalities, respectively.

When comparing emotion coherence across the three modality setups, we found that emotions in the multimodal setup more frequently co-occur with *neutral* in text than in audio and silent video. More specifically, the frequent co-occurrence of *neutral* with *confusion* or *surprise* suggests that these emotions are less efficiently expressed in the text modality. This highlights the importance of multimodal signals in accurately detecting emotions such as *confusion* and *surprise*.

Fig. 11 A video clip labelled as *surprise* at the multimodal level was annotated as *neutral* in text, *surprise* in audio, and *confusion* in silent video. Acoustic features (such as pauses and stress) were added to the transcription for this case study, but were not available during the annotation of the text modality.

5.2 Primary and Peripheral Emotion: The Mixture of Emotion Expression

During annotation, annotators were asked to label the most prominent emotion expressed in each clip as the primary emotion and to identify a peripheral (or secondary) emotion. As a result, about 27% of the video clips were labelled with non-neutral peripheral emotions. Table 7 gives an overview of the top five combinations of primary and peripheral emotions in each of the four modality setups. The co-occurrence of positive and negative emotions (e.g., contentment & disappointment, disappointment & joy) indicates that emotional mixtures are not limited to emotions with the same polarity and suggests a complexity beyond simple combinations of similar emotions.

Table 7 also shows differences in the distribution of blended emotions across modalities. The video modality has the highest number of emotion mixtures, with 287 cases (about 30% of the 965 instances), followed by the multimodal setup and the audio modality. The modality of text has the fewest co-occurrences of primary and peripheral emotions, with 125 cases (about 13%), less than half of those in the video modality. We believe that this unbalanced distribution is related to differences in emotion variety and ambiguity across modalities. In particular, the ambiguity in silent video and the difficulty of understanding individual variations in facial expressions can be high (see for example Ekman and Friesen (2003); Liliana, Basaruddin, Widyanto, and Oriza (2019)). Interestingly, when combining the three modalities in the multimodal setup, the number of blended emotions seems to decrease compared to the video modality alone. This could be a result of the complementary information provided by the text and audio modalities, which helps to disambiguate emotions in silent video.

Table 7 Top five frequent mixtures of emotions in the four different modality setups, excluding combinations with the label *neutral*. It should be noted that the primary emotion dominates the overall emotional state, so the mixture of *contentment-disappointment* refers to *contentment* as the predominant emotion and *disappointment* as the secondary emotion.

Modality	Top 5 primary	Top 5 peripheral	Frequency	Total	Percentage
text	contentment	disappointment	26	125	0.13
	disappointment	contentment	21		
	joy	contentment	10		
	contentment	surprise	9		
	joy	surprise	8		
audio	contentment	disappointment	26	172	0.18
	disappointment	joy	17		
	disappointment	contentment	15		
	disgust	joy	11		
	joy	disappointment	11		
video	disappointment	confusion	30	287	0.30
	joy	confusion	25		
	contentment	confusion	19		
	contentment	disappointment	19		
	disappointment	contentment	19		
all	contentment	disappointment	24	264	0.27
	joy	disappointment	23		
	disappointment	joy	23		
	contentment	surprise	22		
	disappointment	contentment	17		

To gain more insights into the relationship between emotions in unimodal setups and blended emotions in the multimodality setup, we selected the top five frequent mixtures of emotions in the multimodality setup: *contentment* & *disappointment*, *joy* & *disappointment*, *disappointment* & *joy*, *contentment* & *surprise*, and *disappointment* & *contentment*. An interesting finding is that the multimodal setup shares the same most frequent primary emotion with the unimodal setups. Furthermore, the peripheral emotion in the multimodal setup generally aligns with the second most frequent primary emotion in the unimodal setups. More details can be found in Figure A1 in Appendix A.

5.3 Response Times for Emotion Labelling

In addition to emotion annotations, our annotation tool also tracks the time taken to annotate each instance, allowing us to investigate the response times across emotions and annotators for each modality. Since the average length of the instances in our dataset is about 10 seconds, we focused on response times exceeding 10 seconds. We also set a ceiling of 300 seconds, which is considered sufficient to annotate a clip for emotions.

Figure 12 shows that annotators' response time varies across different modalities, which is expected since different organs perceive verbal and non-verbal emotions, and

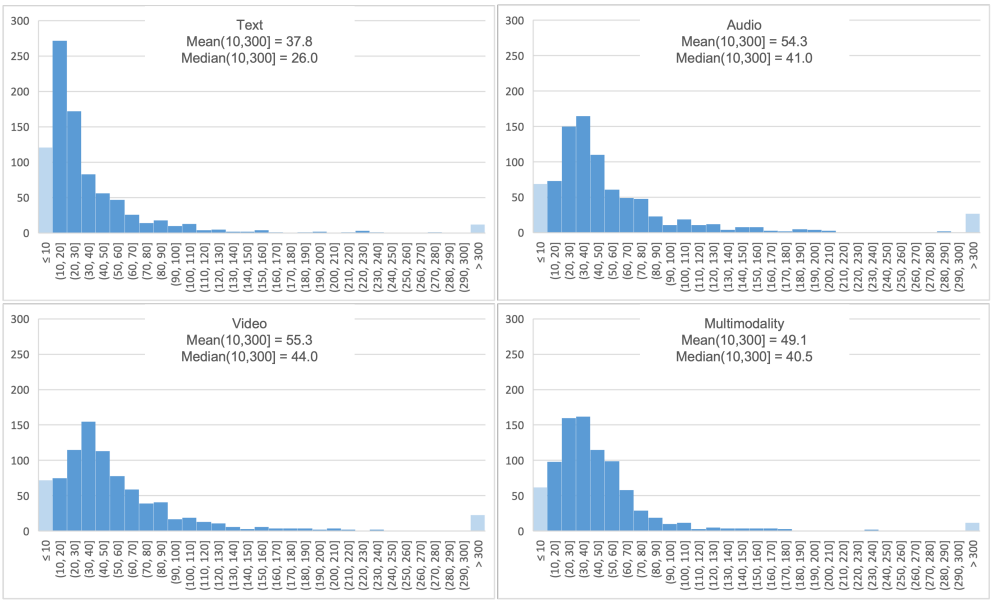

Fig. 12 Response times across modalities.

different brain regions are involved in non-verbal emotion processing (Schirmer & Adolphs, 2017). Comparing the response time for the four modality setups, emotions in the text modality are perceived more quickly, with an average response time of about 37.8 seconds. In contrast, emotions in the other three modality setups are recognized at a slower pace, with response times ranging from 49.1 to 54.3 seconds. Reading is generally faster than listening to or watching a video, implying that at least the average clip length (about 10 seconds) is required to process audio and video, while less time is needed for reading. The response time gap between the text modality and other setups is over 11 seconds, suggesting a faster response pace for text modality annotations. With an average response time of 55.3 seconds, the annotation of the silent video modality takes the longest, suggesting the difficulty of emotion recognition without the cues from text and audio.

Based on the general distribution of response times across modalities, we added the categorical emotion labels to our analysis to investigate how response times vary across emotions in different modalities. As shown in Figure 13, *surprise* takes the longest average time to be detected. Furthermore, the response time for *surprise* in the text modality is the highest across all emotions in each modality setup. One of the possible reasons is that the state of surprise requires a relatively long emotional preparation for a brief but startling response. In the audio modality, *joy* and *disgust* have shorter response times than other emotions. These two emotions have higher arousal scores, as shown in Figure 4, suggesting that the more "extreme" emotions are more easily perceived than the "non-extreme" emotions such as *contentment* and *disappointment*. For the silent video modality, the response time for all non-neutral

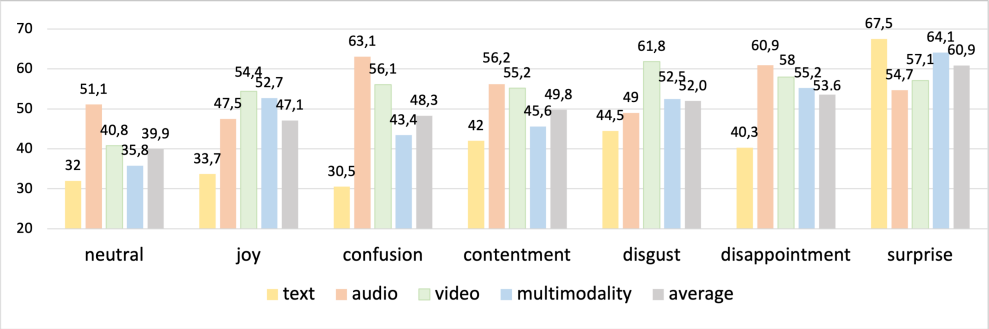

Fig. 13 Average response times for each emotion label in the four modality setups. Response times between more than 10s and less than 301s were considered, with response times outside this range considered as outliers.

emotions is no less than 54 seconds, while the response time for the neutral state is on average 41 seconds. This confirms the findings of Section 5.2, indicating that visual emotions are more ambiguous than emotions in other modalities.

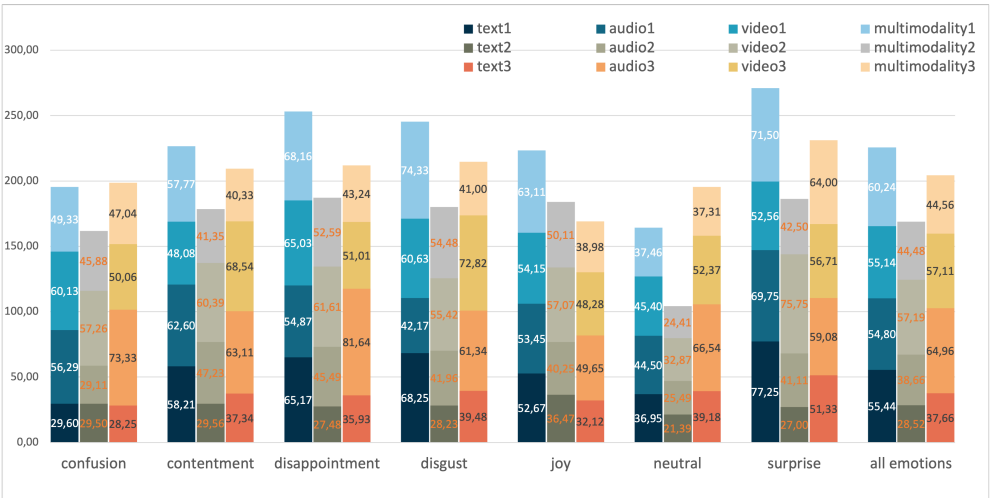

Fig. 14 Average response time for each emotion in four modality setups across three annotators, along with the overall average for the seven emotions. Only response times between 10 and 301 seconds were considered, with times outside this range treated as outliers. Each group of colours (i.e., blue, grey, and yellow) corresponds to a different annotator.

The previous analysis evaluated the emotional perception by averaging data from three annotators. To account for individual differences in response times, Figure 14 shows the response times of each annotator for the seven emotional states across the four modality setups. Generally, annotator 2 responds faster than the other two for most emotions, except for *joy*. In contrast, annotator 1 has the longest response time

for most emotions. All three annotators take the longest for *surprise*, while two have the shortest response time for *neutral* and one has the shortest response time for *joy*. This again suggests that *surprise* is the most challenging emotion for annotators to detect. Another interesting finding is that annotators seem to recognize emotions in specific modalities at different speeds. When we consider the text modality, annotator 1 spends more time (sometimes even double) than the other annotators. Annotator 2 labels emotions in audio faster than the others, while annotator 3 is the slowest in this modality.

In summary, the response time analysis supports the findings from our previous sections, highlighting that the text modality consistently stands out from the other modalities. The difficulty of detecting emotions varies, with *surprise* being more challenging to identify than other emotions.

6 Qualitative Analysis: Emotion Difference across Modalities

To gain more insights into differences in emotion across modalities, we selected one instance per primary emotion at the multimodal level from all video clips in *UniC*. We then exemplified how the emotion varied across modalities using one video clip. As shown in Figure 15, the text and audio for this video clip are both annotated as *joy*, partly because of the phrase *so much fun* and the influence of the intensifier *so much*, which gives the highest valence score of 5. However, as the person lowers their eyebrows and glowers in the silent video, the annotator marks the clip as disgust. When combining the three modalities, the lowered eyes and abrasive voice make the emotional signal come across as *disgust* rather than *joy* to the annotator.

Fig. 15 A *disgust* valence-2 video clip labelled *joy* valence-5, *joy* valence-5, and *disgust* valence-1 in the modality of text, audio, and silent video, respectively.

7 Conclusion

Most of the available multimodal emotion datasets only have a single unified emotion label for all modalities, ignoring the unique contributions of each modality. The multimodal multilabel emotion dataset, *UniC*, is the first dataset to include both independent unimodal labels and multimodal emotion labels.

UniC stands out in terms of dataset size, emotion authenticity, and inter-annotator reliability. While it is not the largest dataset in the field of multimodal emotion modelling, it contains 965 video clips from YouTube, with a total duration of about 160 minutes. After three rounds of filtering, we made sure to have a dataset which is rich in emotional content. The process of filtering undeniably leads to some selection bias, as is the case in any corpus selection process. However, we made sure to be very transparent on each of the filtering steps and the inherently created cultural, topical or emotional bias during corpus creation. We hope the corpus in the future will serve as a stepping-stone to add more languages, more topics and a more realistic distribution of emotional versus neutral content.

Since the individuals in these videos are not professional actors, the dataset captures natural and real-world emotional expressions, which can aid in the development of systems that recognize genuine, non-acted emotions. Our dataset covers a wide range of emotions with varied valence and intensity, and it demonstrates fairly good inter-annotator agreement. While this agreement is not as high as that of lab- or TV-based datasets featuring often exaggerated, acted emotions, it surpasses that of other YouTube-based datasets, such as the popular CMU-MOSEI dataset (Zadeh et al., 2018). Interestingly, we found higher agreement for the combination of the three unimodal setups than for the multimodal annotations. This suggests that focusing annotators on the emotion information in each single modality leads to a more accurate assessment of the emotional content in the video fragments.

We believe the *UniC* dataset has significant potential for computational emotion modelling. Previous studies have shown that using a set of unified emotion labels for both single modalities and multimodality can potentially mislead models into learning the intrinsic features of individual unimodal representations. By leveraging unimodal annotations, models can acquire more differentiated information across modalities and enhance the complementarity among them (Yu et al., 2020). The independent unimodal annotations in this dataset can help models to learn from a rich representation space.

To balance accuracy and efficiency, scholars have proposed a novel multimodal fusion approach that sets a gate to adaptively fuse inputs during inference (Xue & Marculescu, 2023). It saves computational costs for the "easy" inputs which can be correctly recognized using information from only a subset of modalities. Since the definition of "easy" inputs is not clear, the independent unimodal annotations in *UniC* could give insights into the emotional complementarity of different modalities. Similarly, understanding the emotional relationships between modalities can improve robustness in cases where information from one or two modalities is of poor quality or unavailable. Finally, the emotion trigger annotations in the dataset allow to track emotion shifts in videos, which might be useful for generating multimodal artificial affective speech, another hot topic in the field of human-machine interaction.

Acknowledgements. This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174K02325) 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.

Declarations

- Funding: This research received funding from the Flemish Government under the Research Program Artificial Intelligence (174K02325) and from the Research Foundation Flanders (FWO-Vlaanderen) with grant number 1S96322N.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): The authors have no conflicts of interest to declare that are relevant to the content of this article.
- Ethics approval and consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Data availability: The data is available on https://huggingface.co/LT3.
- Materials availability: The data will be made available upon request.
- Code availability: Not applicable.
- Author contribution: Q.D. developed the original research idea, designed and conducted experiments, and wrote and reviewed the manuscript. S.L. contributed to the development of the original research idea, the study's conception and design, and reviewed the manuscript. T.D. contributed to the study's conception and design, and reviewed the manuscript. V.H. contributed to the study's conception and design, reviewed the manuscript and managed funding.

Appendix A Primary and Peripheral Emotion Distribution Across Modalities

First, it is clear that the multimodal setup shares the same most frequent primary emotion with the unimodal setups, as shown in Figure A1.

Second, the peripheral emotion in the multimodal setup generally aligns with the second most frequent primary emotion in the unimodal setups. For instance, in the first, third and fifth case in Figure A1, the peripheral emotion in the multimodal setup is exactly the second most frequent primary emotion in the unimodal setups. However, the pattern is slightly different in the second and fourth cases. In the second case (*joy* & *disappointment*), the second most frequent primary emotion in the unimodal setups is *contentment*, followed by *disgust* and *disappointment*. When grouping *joy* and *contentment*, *disgust* and *disappointment* into two clusters, respectively, the same trend is found. Finally, in the fourth case (*contentment* & *surprise*), *surprise* ranks third in the unimodal setups but it is only outnumbered by the second emotion (*joy*) with a narrow margin.

Overall, the emotion annotations in unimodal setups have shown their potential to help understand and explain mixtures of emotion expression. More insights can be expected with larger datasets that include peripheral emotions.

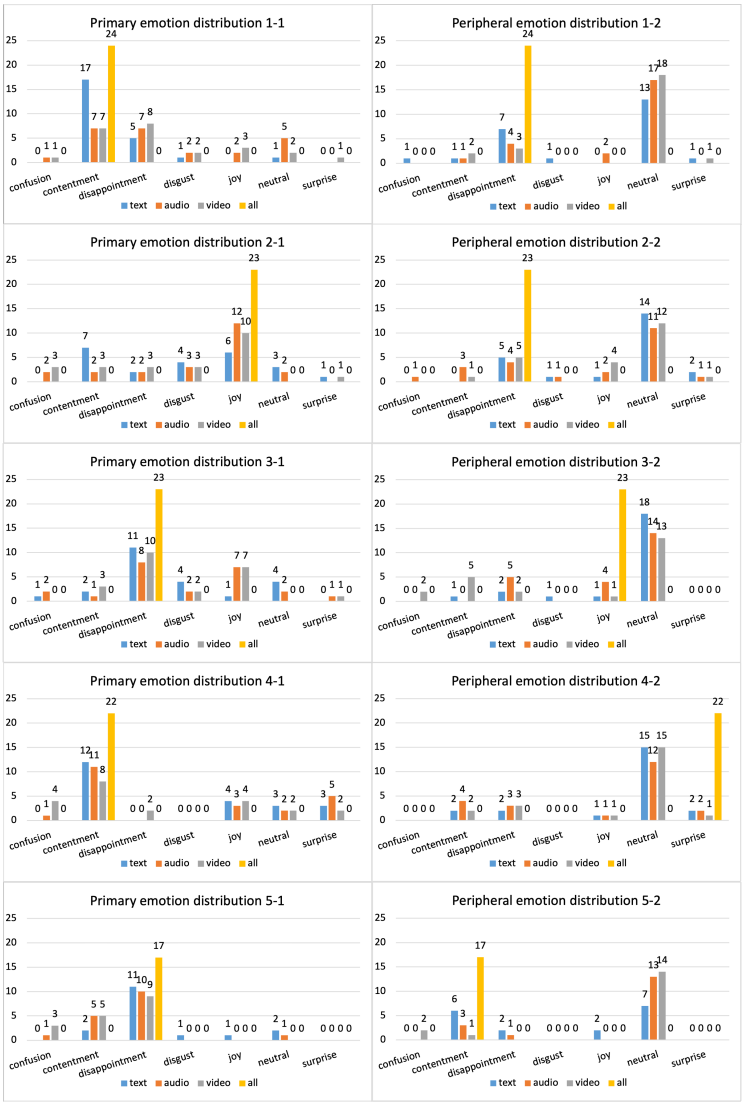

Fig. A1 Primary and peripheral emotion distribution across modalities when mixtures of emotions (primary emotion & peripheral emotion) in the multimodal setup are *contentment* & *disappointment* (1-1 & 1-2), *joy* & *disappointment*(2-1 & 2-2), *disappointment* & *joy* (3-1 & 3-2), *contentment* & *surprise* (4-1 & 4-2), *disappointment* & *contentment* (5-1 & 5-2).

References

Akhand, M., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics*, *10*(9), 1036, https://doi.org/10.3390/electronics10091036

Anagnostopoulos, C.-N., Iliou, T., Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, *43*, 155–177, https://doi.org/10.1007/s10462-012-9368-5

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297–313, https://doi.org/10.1007/978-94-024-0881-2_11

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L. (2020, November). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1644–1650). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.148

Barbieri, F., & Saggion, H. (2014). Automatic detection of irony and humour in twitter. *International conference on innovative computing and cloud computing* (pp. 155–162).

Bradley, M.M., & Lang, P.J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49–59, https://doi.org/10.1016/0005-7916(94)90063-9

Brown, B. (2016). *Cinematography: theory and practice: image making for cinematographers and directors.* New York: Routledge.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al. (2005). A database of german emotional speech. *Proceedings of interspeech 2005* (Vol. 5, pp. 1517–1520).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S.S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335–359, https://doi.org/10.1007/s10579-008-9076-6

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., Provost, E.M. (2016). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, *8*(1), 67–80, https://doi.org/10.1109/TAFFC.2016.2515617

Cabitza, F., Campagner, A., Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 6860–6868).

Canal, F.Z., Müller, T.R., Matias, J.C., Scotton, G.G., de Sa Junior, A.R., Pozzebon, E., Sobieranski, A.C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, *582*, 593–617, https://doi.org/10.1016/j.ins.2021.10.005

Chowdary, M.K., Nguyen, T.N., Hemanth, D.J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, *35*(32), 23311–23328, https://doi.org/10.1007/s00521-021-06012-8

De Bruyne, L., De Clercq, O., Hoste, V. (2019). Towards an empirically grounded framework for emotion analysis. *Proceedings of huso 2019, the fifth international conference on human and social analytics* (pp. 11–16). Retrieved from http://hdl.handle.net/1854/LU-8624200

De Bruyne, L., De Clercq, O., Hoste, V. (2021a). 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*, 1–29, https://doi.org/10.1007/s10579-020-09524-2

De Bruyne, L., De Clercq, O., Hoste, V. (2021b). Emotional RobBERT and insensitive BERTje: Combining transformers and affect lexica for dutch emotion detection. *Proceedings of the eleventh workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 257–263). Retrieved from https://aclanthology.org/2021.wassa-1.27

Denzin, N.K. (1984). *On understanding emotion.* San Francisco, CA: Jossey-Bass.

Du, Q., Labat, S., Demeester, T., Hoste, V. (2023). Unimodalities count as perspectives in multimodal emotion annotation. *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP.* Retrieved from https://ceur-ws.org/Vol-3494/paper14.pdf

D'Mello, S.K., & Westlund, J.K. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*, 1-36,

Ekman, P. (1973). Cross-cultural studies of facial expression. P. Ekman (Ed.), *Darwin and facial expression: A century of research in review* (pp. 169–222). New York: Academic Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200, https://doi.org/10.1080/02699939208411068

Ekman, P., & Friesen, W.V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues.* Los Altos, CA: ISHK.

Fischer, L., Brauns, D., Belschak, F. (2002). *Zur messung von emotionen in der angewandten forschung.* Lengerich: Pabst Science Publishers.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378, https://doi.org/10.1037/h0031619

Gao, J., Li, P., Chen, Z., Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*(5), 829–864, https://doi.org/10.1162/neco_a_01273

Ghafoor, Y., Jinping, S., Calderon, F.H., Huang, Y.-H., Chen, K.-T., Chen, Y.-S. (2023). TERMS: Textual emotion recognition in multidimensional space. *Applied Intelligence*, *53*(3), 2673–2693, https://doi.org/10.1007/s10489-022-03567-4

Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., ... Chen, K. (2023). *Multimodal-GPT: A vision and language model for dialogue with humans.* Retrieved from https://arxiv.org/abs/2305.04790

Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., ... others (2013). Challenges in representation learning: A report on three machine learning contests. *Neural information processing: 20th international conference, iconip 2013, daegu, korea, november 3-7, 2013. proceedings, part iii 20* (pp. 117–124).

Hajek, P., & Munk, M. (2023). Speech emotion recognition and text sentiment analysis for financial distress prediction. *Neural Computing and Applications*, 1–15, https://doi.org/10.1007/s00521-023-08470-8

Keltner, D., Sauter, D., Tracy, J., Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, *43*, 133–160, https://doi.org/10.1007/s10919-019-00293-3

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Los Angeles: Sage Publications.

Kumar, A., & Garg, G. (2020). Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia tools and Applications*, *79*,

15349–15380, https://doi.org/10.1007/s11042-019-7346-5

Kwon, H.-J., Ban, H.-J., Jun, J.-K., Kim, H.-S. (2021). Topic modeling and sentiment analysis of online review for airlines. *Information*, *12*(2), 78, https://doi.org/10.3390/info12020078

Labat, S., Demeester, T., Hoste, V. (2024). EmoTwiCS : A corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter. *Language Resources and Evaluation*, *58*, 505–546, Retrieved from https://doi.org/10.1007/s10579-023-09700-0

Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., Jia, J. (2018). MEC 2017: Multimodal emotion recognition challenge. *2018 first asian conference on affective computing and intelligent interaction (acii asia)* (pp. 1–5).

Liang, P.P., Salakhutdinov, R., Morency, L.-P. (2018). Computational modeling of human multimodal language: The MOSEI dataset and interpretable dynamic fusion. *First workshop and grand challenge on computational modeling of human multimodal language* (Vol. 1, p. 3).

Liliana, D.Y., Basaruddin, T., Widyanto, M.R., Oriza, I.I.D. (2019). Fuzzy emotion: A natural approach to automatic facial expression recognition from psychological perspective using fuzzy system. *Cognitive processing*, *20*, 391–403, https://doi.org/10.1007/s10339-019-00923-0

Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., ... Gao, K. (2022). Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. *Proceedings of the 2022 international conference on multimodal interaction* (p. 247–258). New York, NY, USA: Association for Computing Machinery.

Lu, Z., Cao, L., Zhang, Y., Chiu, C.-C., Fan, J. (2020). Speech sentiment analysis via pre-trained features from end-to-end ASR models. *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7149–7153).

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th berkeley symp. math. statist. probability* (pp. 281–297).

Maladry, A., Lefever, E., Van Hee, C., Hoste, V. (2022). Irony detection for dutch: A venture into the implicit. *Proceedings of the 12th workshop on computational approaches to subjectivity, sentiment & social media analysis* (pp. 172–181).

Mehrabian, A. (1971). *Silent messages*. Belmont, California: Wadsworth Belmont, CA.

Mehrabian, A., & Russell, J.A. (1974). *An approach to environmental psychology.* Cambridge: the MIT Press.

Ming, Y., Qian, H., Guangyuan, L., et al. (2022). CNN-LSTM facial expression recognition method fused with two-layer attention mechanism. *Computational Intelligence and Neuroscience*, *2022*, 1-9, https://doi.org/10.1155/2022/7450637

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D. (2020). M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 1359–1367).

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S. (2018, June). SemEval-2018 task 1: Affect in tweets. M. Apidianaki, S.M. Mohammad, J. May, E. Shutova, S. Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1–17). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S18-1001

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd international conference on knowledge capture* (pp. 70–77).

Nazir, A., Rao, Y., Wu, L., Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845–863, https://doi.org/10.1109/TAFFC.2020.2970399

Pan, T., Ye, Y., Cai, H., Huang, S., Yang, Y., Wang, G. (2023). Multimodal physiological signals fusion for online emotion recognition. *Proceedings of the 31st acm international conference on multimedia* (pp. 5879–5888). Ottawa, Canada.

Pfeifer, R. (1982). *Cognition and emotion: An information processing approach* (Tech. Rep.). Carnegic-Mellon University.

Picard, R.W. (1997). *Affective computing.* Cambridge: MIT press.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., ... others (2016). SemEval-2016 Task 5: Aspect based sentiment analysis. *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 19–30). San Diego, California.

Ray, B., Garain, A., Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, *98*, 106935, https://doi.org/10.1016/j.asoc.2020.106935

Rehman, A.U., Malik, A.K., Raza, B., Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, *78*, 26597–26613, https://doi.org/10.1007/s11042-019-07788-7

Reisenzein, R., Horstmann, G., Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in cognitive science*, *11*(1), 50–74, https://doi.org/10.1111/tops.12292

Rosenthal, S., Farra, N., Nakov, P. (2017, August). SemEval-2017 task 4: Sentiment analysis in Twitter. *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 502–518). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S17-2088

Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends in cognitive sciences*, *21*(3), 216–228, https://doi.org/10.1016/j.tics.2017.01.001

Shaver, P., Schwartz, J., Kirson, D., O'connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061, https://doi.org/10.1037//0022-3514.52.6.1061

Winters, T., Nys, V., De Schreye, D. (2018). Automatic joke generation: Learning humor from examples. *Distributed, ambient and pervasive interactions: Technologies and contexts: 6th international conference, dapi 2018, held as part of hci international 2018, las vegas, nv, usa, july 15–20, 2018, proceedings, part ii 6* (pp. 360–377).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the fifth international conference on language resources and evaluation.* Genoa, Italy: European Language Resources Association. Retrieved from https://aclanthology.org/L06-1082/

Xue, Z., & Marculescu, R. (2023). Dynamic multimodal fusion. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops* (pp. 2574–2583).

Yang, L., Li, Y., Wang, J., Sherratt, R.S. (2020). Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE access*, *8*, 23522–23530, https://doi.org/10.1109/ACCESS.2020.2969854

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., Wang, L. (2023). *The dawn of LMMs: Preliminary explorations with GPT-4V(ision)* (Vol. 9) (No. 1). Retrieved from https://arxiv.org/abs/2309.17421

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., ... Yang, K. (2020, July). CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3718–3727). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.343

Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2236–2246). Melbourne, Australia: Association for Computational Linguistics.

Zhang, H. (2020). Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. *IEEE Access*, *8*, 164130–164143, https://doi.org/10.1109/ACCESS.2020.3021994

Zhang, S., Zhao, X., Tian, Q. (2019). Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. *IEEE Transactions on Affective Computing*, *13*(2), 680–688, https://doi.org/10.1109/TAFFC.2019.2947464

Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., Li, H. (2022, May). M3ED: Multi-modal multi-scene multi-label emotional dialogue database. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5699–5710). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.391

Zhong, Q., Ding, L., Liu, J., Du, B., Jin, H., Tao, D. (2023). Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, , https://doi.org/10.1109/TKDE.2023.3250499