Multimodal Sentiment Analysis in Real-life Videos

Inaugural-Dissertation
zur Erlangung des Doktorgrades (Dr.-Ing.)
an der
Fakultät für Angewandte Informatik
der Universität Augsburg

vorgelegt von

Lukas Bernd Stappen

2021

Erstgutachter:	Prof. Dr. Björn Schuller
Zweitgutachter:	Prof. Dr. Elisabeth André
Drittgutachter:	PD Dr.- Ing. habil. Ronald Böck
Tag der mündlichen Prüfung:	19.10.2022

Acknowledgements

Throughout the journey from which this thesis emerged, I was accompanied by many extraordinary people to whom I am deeply grateful.

To my doctoral supervisor, Prof. Dr. habil. Björn Schuller, for the numerous conversations that sharpened my thinking and my research endeavours. His keen sense of the ravages of time in our research community, the pursuit of academic excellence, and the drive to never let the last mile slip away have been and always will be a role model for me.

My project partners at the BMW Group, especially the three Stefans, for their moral support and unbroken faith in me to tread new paths with uncertain outcomes. To all my colleagues at the EIHW and GLAM Chair whom I had the pleasure to work with over the years, for their intellectual stimulation and the energy to bring these ideas to life in collaborative work. Especially, I would like to emphasise my deepest gratitude to my colleagues and friends, Alice and George, for fantastic human and professional collaboration, without which I would write these lines with much less melancholy. Thank you for your time and thoughts on and off work, your unbroken will to reach impossible deadlines, and your tolerance to overlook my lack of politeness after working all-nighters. It has been an honour to digest defeats and celebrate victories with you on this journey. I would also like to sincerely thank Nadine for all the support to the people at the chair, especially to me, who got lost multiple times in the jungle of administration and convoluted formalities. Not forgetting Lea, Xinchen, Benjamin, Felix and all the many other young, brilliant minds for your support and cooperation. It has been a great pleasure to explore new directions with you and accompany you on your first academic steps.

As always, the best comes at the end. No words can ever suffice to express my wholehearted gratitude to my family and my better half, whom I love above all else and on whom I can always blindly rely. You are my backbone and the reason why I am here.

Lukas Stappen

November 2021

Abstract

This thesis extends the emerging field of multimodal sentiment analysis of real-life videos, taking two components into consideration: the emotion and the emotion's target.

The emotion component of media is traditionally represented as a segment-based intensity model of emotion classes. This representation is replaced here by a value- and time-continuous view. Adjacent research fields, such as affective computing, have largely neglected the linguistic information available from automatic transcripts of audio-video material. As is demonstrated here, this text modality is well-suited for time- and value-continuous prediction. Moreover, source-specific problems, such as trustworthiness, have been largely unexplored so far. This work examines perceived trustworthiness of the source, and its quantification, in user-generated video data and presents a possible modelling path. Furthermore, the transfer between the continuous and discrete emotion representations is explored in order to summarise the emotional context at a segment level.

The other component deals with the target of the emotion, for example, the topic the speaker is addressing. Emotion targets in a video dataset can, as is shown here, be coherently extracted based on automatic transcripts without limiting a priori parameters, such as the expected number of targets. Furthermore, alternatives to purely linguistic investigation in predicting targets, such as knowledge-bases and multimodal systems, are investigated.

A new dataset is designed for this investigation, and, in conjunction with proposed novel deep neural networks, extensive experiments are conducted to explore the components described above. The developed systems show robust prediction results and demonstrate strengths of the respective modalities, feature sets, and modelling techniques. Finally, foundations are laid for cross-modal information prediction systems with applications to the correction of corrupted in-the-wild signals from real-life videos.

Contents

List of Publications

As part of the research for this PhD thesis, the author (co-)authored several peer-reviewed publications in journals and conference proceedings, some of which are covered in more detail in this thesis. Below is a comprehensive list of these publications, ordered by publication category and relevance.

Journals

1. **L. Stappen**, A. Baird, L. Schumann, B.W. Schuller, "The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements," in *IEEE Transactions on Affective Computing*, July 2021, to appear, IEEE (IF: 10.506).

2. **L. Stappen**, A. Baird, E. Cambria and B. W. Schuller, "Sentiment Analysis and Topic Recognition in Video Transcriptions," in *IEEE Intelligent Systems*, March 2021, vol. 36, no. 2, IEEE (IF: 4.410).

3. **L. Stappen**, A. Baird, M. Lienhart, A. Bätz, B.W. Schuller, "An Estimation of Online Video User Engagement From Features of Time- and Value-Continuous, Dimensional Emotions," in *Frontiers in Computer Science*, 2022, to appear.

4. A. Baird, I. Lefter, **L. Stappen**, B.W. Schuller, "A Cross-Corpus Speech-based Analysis of Escalating Negative Interactions," in *Frontiers in Computer Science*, vol. 15, March 2022.

5. ZM. Ibrahim, H. Wu, A. Hamoud, **L. Stappen**, RJB. Dobson, A. Agarossi, "On Classifying Sepsis Heterogeneity in the ICU: Insight Using Machine Learning," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, March 2020, AMIA (IF: 4.497).

6. S. Amiriparian, M. Gerczuk, S. Ottl, **L. Stappen**, A. Baird, and others (2020), "Towards Cross-modal Pre-training and Learning Tempo-spatial Characteristics for Audio Recognition with Convolutional and Recurrent Neural Networks," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 19, October 2020, EURASIP (IF: 1.558).

7. A. Baird, A. Triantafyllopoulos, S. Zänkert, S. Ottl, L. Christ, **L. Stappen**, J. Konzok, S. Sturmbauer, E.M. Messner, B. Kudielka3, N., H. Baumeister, B. W. Schuller, "An

Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress", *Frontiers in Computer Science*, October 2021.

Conference Proceedings and Preprints

8. **L. Stappen**, E.M. Meßner, E. Cambria, G. Zhao, B.W. Schuller, "MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection," in *Proceedings of the 29th ACM International Conference on Multimedia (ACMMM)*, October 2021, Chengdu, China, ACM.

9. **L. Stappen**, B.W. Schuller, I. Lefter, E. Cambria, I. Kompatsiaris, "Summary of MuSe 2020: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media," in *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, October 2020, New York, USA, ACM.

10. **Best paper award:**
L. Stappen, V. Karas, N. Cummins, F. Ringeval, K. Scherer, B.W. Schuller, "From Speech to Facial Activity: Towards Cross-modal Sequence-to-sequence Attention Networks," in *Proceedings of the 21st IEEE International Workshop on Multimedia Signal Processing (MMSP)*, September 2019, Kuala Lumpur, Malaysia, IEEE.

11. **L. Stappen**[1], J. Thies[1], G. Hagerer, B.W. Schuller, G. Groh, "Unsupervised Graph-based Topic Modeling from Video Transcripts", *IEEE International Conference on Multimedia Big Data (BigMM)*, October 2021, Taichung, Taiwan, IEEE.

12. **L. Stappen**, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B.W. Schuller, I. Lefter, E. Cambria, "MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild," in *Proceedings of the 1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop 2020 (MuSe)*, October 2020, New York, USA, ACM.

13. **L. Stappen**, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.M. Messner, E. Cambria, G. Zhao, B.W. Schuller, "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress," in *Proceedings of the 2nd International Multimodal Sentiment Analysis Challenge 2021 (MuSe)*, October 2021, Chengdu, China, ACM.

[1] shared co-first authorship

14. **L. Stappen**, G. Rizos, B.W. Schuller, "X-AWARE: ConteXt-AWARE Human-environment Attention Fusion for Driver Gaze Prediction in the Wild," in *Proceedings of the 22nd International Conference on Multimodal Interaction (ICMI)*, October 2020, Utrecht, Netherlands, ACM.

15. **L. Stappen**, L. Schumann, B. Sertolli, A. Baird, B. Weigell, E. Cambria, B.W. Schuller, "MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox," in *Proceedings of the 2nd International Multimodal Sentiment Analysis Challenge 2021 (MuSe)*, October 2021, Chengdu, China, ACM.

16. **L. Stappen**, G. Rizos, M. Hasan, T. Hain, B.W. Schuller, "Uncertainty-Aware machine support for paper reviewing on the Interspeech 2019 Submission Corpus," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2020. Shanghai, China. ISCA.

17. **L. Stappen**, L. Schumann, A. Batliner, B.W. Schuller, "Embracing and Exploiting Annotator Emotional Subjectivity: An Affective Rater Ensemble Model," in *9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, September 2021, Virtual Event, AAAC.

18. **L. Stappen**, N. Cummins, EM. Meßner, H. Baumeister, J. Dineley, B.W. Schuller. "Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, Brighton, United Kingdom, IEEE.

19. **L. Stappen**, X. Du, V. Karas, S. Müller, B.W. Schuller, "Go-CaRD: Generic, Optical Car Part Recognition and Detection: Collection, Insights, and Applications," *arXiv preprint arXiv:2006.08521.*

20. **L. Stappen**, F. Brunn, B.W. Schuller, "Cross-lingual Zero-and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL", *arXiv preprint arXiv:2004.13850.*

21. A. Baird, **L. Stappen**, L. Christ, L. Schumann, B. Sertolli, B.W. Schuller, "A Physiologically-Adapted Gold Standard for Arousal during Stress," in *Proceedings of the 2nd International Multimodal Sentiment Analysis Challenge 2021 (MuSe)*, October 2021, Chengdu, China, ACM.

22. K. Friedl, G. Rizos, **L. Stappen**, M. Hasan, L. Specia, T. Hain, B. Schuller, "Uncertainty Aware Review Hallucination for Science Article Classification," in *Findings of the Association for Computational Linguistics: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, August 2021, Bangkok, Thailand, ACL.

23. A. Mallol-Ragolta, Z. Zhao, **L. Stappen**, N. Cummins, B.W. Schuller, "A Hierarchical Attention Network-based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2019, Graz, Austria, ISCA.

24. A. Baird, S. Mertes, M. Milling, **L. Stappen**, T. Wiest, E. Andre and B. Schuller, "A Prototypical Network Approach for Evaluating Generated Emotional Speech,", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2021, Brno, Czechia, ISCA.

25. S. Amiriparian, A. Awad, M. Gerczuk, **L. Stappen**, A. Baird, S. Ottl, B.W. Schuller, (2019). Audio-based Recognition of Bipolar Disorder Utilising Capsule Networks," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, July 2019, Budapest, Hungary, IEEE.

26. B.W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, **L. Stappen**, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. Rothkrantz, J. Zwerts, J. Treep, C. Kaandorp, "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, Escalation & Primates," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2021, Brno, Czechia, ISCA.

27. B.W. Schuller, A. Batliner, C. Bergler, EM. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, **L. Stappen**, H. Baumeister, A. MacIntyre, S. Hantke, "The Interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, October 2020, Shanghai, China. ISCA.

Introduction

1 Introduction

1.1 Motivation

The internet has become the port of call in almost every aspect of our lives. We communicate with family and friends on social networks, search for information, go shopping online, and entertain ourselves with music and films. The almost five billion people online[1] are not just passive consumers, but interact with each other and continually contribute new user-generated content. Along with the increasing number of participants, the duration and forms of interaction have led to rapid growth in the amount of data. The transition from text-based to multimodal content, especially in social networks such as YouTube, Instagram, and TikTok, plays a central role in this growth. This more complex content evokes a deeper willingness to engage and blurs the boundaries between the digital and the physical worlds. YouTube, for example, has become the second-largest social network, with nearly two billion active users and one billion hours of video watched each day[2].

Extracting, processing, analysing, and understanding relevant information from vast amounts of unstructured, user-generated data remains a challenge [1]. Sentiment analysis is a well-established method of managing and structuring this data. It allows opinions and sentiments (positive, neutral, and negative) on topics to be extracted and automatically measured on axes such as customer interest, satisfaction, and brand perception. Text-based sentiment analysis is now widely adopted in industry; however, the evident transition from text to video modalities also demands that available methods evolve.

A video is composed of three core modalities: a visual signal, an audio signal, and, derived from that, a textual transcription of the spoken word. Multimodality poses new challenges and, simultaneously, opens new avenues for processing and analysing this diverse information. The reasons for this are manifold: On the one hand, the individual modalities have specific strengths depending on the prediction target. For example, the visual component enables extraction of facial expressions and gestures to, for instance, recognise finger-pointing towards an object [2]. Voice is strongly associated with arousal-related emotions [3], and content can be extracted from the text [4]. On the other hand, robust intermodal dynamics can give a richer picture of a scene. The absence of one modality can be compensated by another.

[1] https://www.statista.com/statistics/617136/digital-population-worldwide/ accessed August 1, 2021.

[2] https://www.statista.com/statistics/272014/global-social-networks-/ranked-by-number-of-users/ accessed September 16, 2019.

For example, the person of interest's face might be obscured, but their voice may still be perceived [5, 6]. Fusing different modalities is often utilised to capture more fine-grained and complex aspects of emotion than just the sentiment, as in the field of Affective Computing (AC) [7, 8]. As a result, multimodal approaches often lead to superior prediction results than unimodal ones do [9].

From these insights emerges Multimodal Sentiment Analysis (MSA) [10]. Datasets are the cornerstones supporting the development of new methods for analysing multimodal interactions between emotions and topics in real-life, user-generated media ("in the wild") [11]. Despite the recent efforts that construct larger datasets [12], many in-the-wild paradigms remain unexplored to this day. These gaps lead to a lack of robustness and generalisation capabilities needed to develop and employ these techniques in real-world applications, which is still an ongoing challenge [13, 14]. Furthermore, two disciplines with differing computational backgrounds approach the topic from different angles. The sentiment (and opinion) mining community specialising in Natural Language Processing (NLP) methods for symbolic information analysis leverages the text modality and focuses on predicting discrete sentiment label categories [15]. At the same time, the field of affective (and behavioural) computing, specialising in intelligent signal processing, mainly focuses on one or both of the audio and visual modalities in order to predict the continuous-valued arousal and valence dimensions of emotion according to the Circumplex Model of Affect (CA) [16], while often disregarding the potential contribution of textual information [17–20]. Both communities have very similar objectives and highly influence each other, but nevertheless maintain separate points of view.

The aim of this work is to foster the desirable first steps towards unifying these communities [21–23], and to facilitate their convergence. On this path, the foundation is first laid by a newly collected and annotated multimodal dataset that integrates aspects of both worlds and goes beyond previous scopes. Then, through a combination of signal processing and the latest machine learning methods, new approaches are explored to extract meaningful representations from the vast amounts of data. These serve as the input to develop state-of-the-art prediction systems, expand previous knowledge, explore technological limits, and open new directions in the emerging research field of multimodal learning techniques.

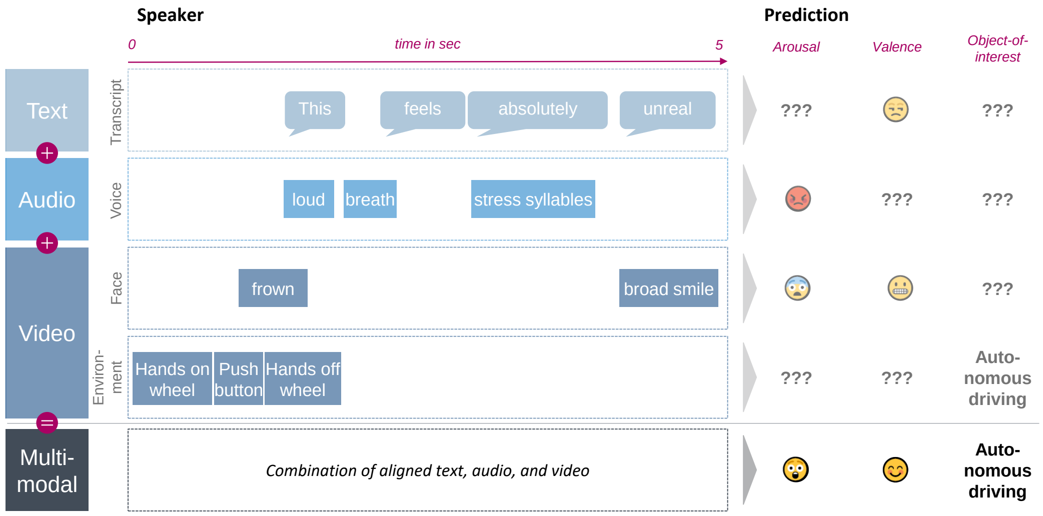


Figure 1.1: Illustration of a multimodal example to show how multiple features naturally complement each other for predicting arousal, valence, and the object of interest in multimodal sentiment analysis.

1.2 Problem Statement and Research Questions

The thesis's central task is to extend the frontiers of Multimodal Sentiment Analysis on user-generated video content. **Given a collection of videos, the aim is to develop machine learning methods to investigate the multimodal interaction between an opinion holder's communicated emotional response and the objective context in which it is triggered. While (perceived) emotions are inherently subjective, their target is rather objective.**

An example of Multimodal Sentiment Analysis is given in Figure 1.1. Here, the statement "This feels absolutely unreal" has no target because "this" is ambiguous. As shown by the multiple modality dimensions, the context can be inferred from hand gestures and objects that indicate that the "autonomous driving" feature has been turned on. The emotional perspective is also inconclusive. Given the textual information (of the spoken word) alone, it can indicate either an "expectation exceeded" (positive valence) or an "anxious uncertainty" (negative valence). In addition, the paralinguistic features of the vocal apparatus in this example indicate high arousal. Furthermore, the facial muscles, indicating attentive frowning and, at the end, a broad smile, further regulate the emotional outcome. All modalities together give a complete picture. For such a complex comprehension, however, the temporal appearance of the features must be robustly extracted and represented, temporally aligned and fused, and the intermodal context learned. This highlights the technical challenges that computational methods face when attempting to perform the task in an automated way.

As described, the focus of this thesis is to develop a set of novel computer-based methods for analysing user-generated videos from the real world in terms of **subjective and objective dimensions**. With this scenario in mind, the formulated Research Questions (RQ) investigated in this thesis are as follows:

RQ-1: To what extent and in what form can complex emotional states be effectively modelled by machine learning methodologies for Multimodal Sentiment Analysis? Furthermore, can these emotion representations be exploited for novel tasks specific to Multimodal Sentiment Analysis? To gain an understanding of this, the established primitives' arousal and valence, as well as the novel dimension of trustworthiness, are proposed and predicted in time- and value-continuous form. Additionally, a number of experiments are outlined to derive data-driven, discrete summary classes linked to the thematic boundaries of the objective dimension. Finally, the perceived emotion is utilised to estimate the popularity of user-generated videos.

RQ-2: How can the emotion's target be automatically extracted from vast amounts of user-generated videos without making a priori assumptions? Additionally, how can the coherent, annotated speaker topics be predicted? To explore this, an unsupervised graph-based machine learning method, particularly suited for automatic transcriptions of the spoken word in real-world settings, is proposed and experimentally evaluated. Furthermore, in a series of experiments, proposed supervised models utilising knowledge bases and multimodal representations are evaluated.

RQ-3: How can the subjective and objective dimensions of Multimodal Sentiment Analysis be predicted most effectively? How do audio, text, and video modalities perform as unimodal inputs and in multimodal fusion? To evaluate this, a multimodal dataset is designed including user-generated videos in complex settings. Furthermore, a variety of representations are extracted from several modalities, suitable deep neural network architectures are proposed, and, in combination, their robustness is experimentally evaluated.

RQ-4: Can cross-modal dynamics be useful to infer individual modalities? Given the often flawed representations in real-life recordings, understanding cross-modal, temporal dependencies is the cornerstone for more effective modality co-learning and robust multimodal representations. To investigate cross-modal interaction at a fundamental level, a series of experiments are conducted applying several proposed sequence-to-sequence networks that predict facial muscle activity from the voice alone.

1.3 Contributions

A broad spectrum of MSA topics is addressed in this work and in prior publications of the author. However, this thesis does not aim to cover all aspects at the same level of granularity. Instead, it focuses on specific challenges within these subtasks, making the following contributions:

Corpus and gold standard: The scope of the problem stated is specifically extended by a new collection of user-generated videos and novel gold standard methods.

- Introducing the selection, collection, and annotation process of the novel dataset Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) [24] — the most extensive annotated multimodal sentiment analysis dataset featuring continuous emotions and speaker topics.
- Presenting the novel gold standard annotator fusion method Rater Aligned Annotation Weighting (RAAW) as part of the Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox (MuSe-Toolbox) [25], alongside a developed procedure to extract time series features from these gold standards and create emotional classes for video segments of varying length.

Both were open-sourced to the research community through the international Multimodal Sentiment Analysis challenges (MuSe) [26, 27] and will continue to advance the field, while providing a standardised benchmark for novel methods. They are also at the core of this work's experiments.

The following new approaches and findings for insufficiently researched aspects of MSA contribute to answer RQ-1 and RQ-2, directly incorporating RQ-3, as well as addressing modality inference towards cross-modal systems (RQ-4).

Subjective dimensions:

- Proposing new Deep Learning (DL) architectures to demonstrate effective modelling of time-continuous emotion gold standards (e. g. , RAAW). In particular, providing new findings regarding modelling techniques; prediction strength of the individual audio, text, and video modalities and representations; the combination of these modalities for multimodal prediction; and benchmarking against other models found in the literature.
- Proposing and evaluating architectures for the classification of summary emotion classes on the same aspects as for continuous emotions. The main focus is on compar-

ing the intensity classes naïvely created from the annotations continuous in time and value with the emotion classes learned by the proposed new method.

- Demonstrating the value and predictability of the novel dimension of perceived trustworthiness proposed for quantifying trust in online videos. A particular effective model is proposed in this context and the individual functional elements are evaluated.

- Developing an experimental setting and methods to extract features from the subjective dimension, estimate relationships between these features and the popularity of a YouTube video, and predict popularity in terms of views, likes, and other user engagement criteria.

Objective dimensions:

- Proposing a novel unsupervised Graph-based Topic Modelling approach for Transcripts (GraphTMT) to model content and context understanding in videos without needing human-generated annotations. The semantic topic coherency and the performance are quantitatively and qualitatively compared to common benchmark models.

- Demonstrating Transformer and SenticNet-based Learning (SNL) approaches for predicting human-annotated topics of video segments and validating the multimodal use.

Cross-modal:

- Proposing stacked sequence-to-sequence and encoder-decoder DL architectures for cross-modal prediction of Facial Action Units (FAUs) from speech.

- Investigating these architectures in combination with context and local attention mechanisms to enhance the robustness of such systems.

1.4 Thesis Structure

This thesis is structured as follows:

- Chapter 2 lays the foundation for a common understanding of MSA by discussing its origins and current characteristics. Underlying principles and typical approaches are further broken down into the subjective and objective dimensions, providing important background information about state-of-the-art techniques and highlighting limitations.

- Chapter 3 introduces the extraction and representation methods employed on the raw data from the three modalities of audio, text, and video, along with the components necessary to develop Deep Neural Networks (DNNs).

- Chapter 4 describes the collection methodology, annotation process, and gold standard approaches for this thesis's central dataset.

- Chapter 5 introduces individual developed architectures and presents experimental results on the subjective dimensions for emotion regression and classification, trustworthiness recognition, and video popularity, as well as on the objective dimensions for target extraction and detection. The chapter further explores cross-modal prediction from speech to facial muscles as a potential future tool for correcting incomplete or corrupted visual data.

- Chapter 6 concludes the thesis with a summary, a review of the ethical implications of this research, a discussion of the limitations of the utilised methods, and directions for future work.

Background

2 Background

This section provides the reader with background and definitions of the key concepts underlying this work.

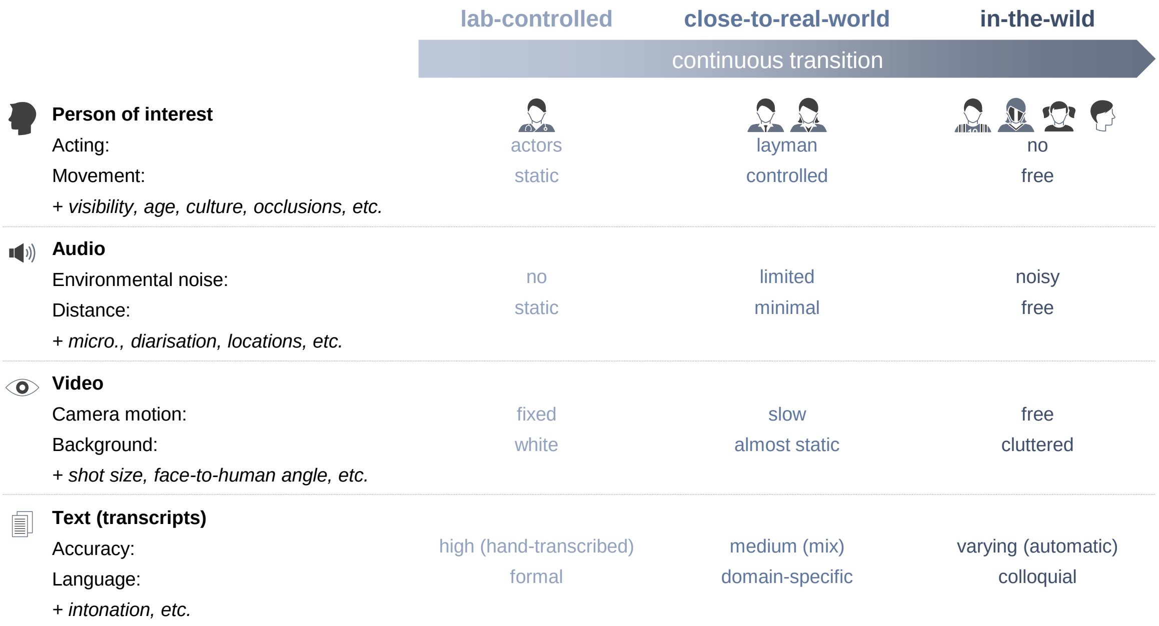
2.1 Characteristics of Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) is a rapidly growing research field with influences from various research communities, each of which offers distinct perspectives and research directions, leading to divergent understandings of relevant characteristics and terminology.

The first work claiming to have coined the term MSA [28] argues that its origins lie in the vast availability of unstructured information through the world wide web [29]. Derived from textual sentiment analysis, the goal of MSA is to automatically query multimodal content to gain insights into subjective states specifying sentiments, emotions, and opinions. The demonstrated proof of concept emphasises the challenges of natural language use. Similarly, other early work interprets the field as extending textual sentiment analysis, transforming its application to the multimedia web, as well as unlocking multimodal sources for a variety of research fields [9, 30]. The definition arising from these works encompasses an opinion holder, an object of interest, and the disposition of polarity, rather than just an opinion [31]. Service and product reviews [30] from social media platforms [32] are often mentioned as preferred domains.

In recent studies [12, 33], albeit partly controversial [10], automatic audio-visual emotion recognition from human-avatar and human-human interaction via video is also envisioned as a part of MSA. This also initiates the extension from simple positive, neutral, and negative sentiment polarities towards emotional classes and affective signals. In combination with automatic contextual inference, the broadened analysis can reveal deeper attitudes of a person towards entities [10]. The rationale behind this is that analyses of both polarity and emotion use linguistic, acoustic, and visual information extracted from videos, and the steps and methods for further processing and prediction are almost identical for both [29]. Lately, more MSA studies have explored the idea of predicting emotion intensities [11], emotion classes [13, 15, 34], and even emotionally complex emojis [35].

In this context, the estimation of continuous emotion primitives such as valence can be a valuable addition. Consequently, MSA and AC can be seen as overlapping. One example



continuous transition

	lab-controlled	close-to-real-world	in-the-wild
Person of interest			
Acting:	actors	layman	no
Movement:	static	controlled	free
+ visibility, age, culture, occlusions, etc.			
Audio			
Environmental noise:	no	limited	noisy
Distance:	static	minimal	free
+ micro., diarisation, locations, etc.			
Video			
Camera motion:	fixed	slow	free
Background:	white	almost static	cluttered
+ shot size, face-to-human angle, etc.			
Text (transcripts)			
Accuracy:	high (hand-transcribed)	medium (mix)	varying (automatic)
Language:	formal	domain-specific	colloquial
+ intonation, etc.			

Figure 2.1: Examples of the different disturbing influences on the individual modalities with increasing naturalness of the data source. An example of laboratory data are designed studies with defined speech dialogues and actors. At the other extreme are people in natural situations.

is distilling emotional responses to product promotional videos during human-to-human interaction via video communication [12]. The computational linguistics community [36], however, sees a distinction in that emotions tend to be short-lived and sentiments develop over a longer period of time. Regardless, both communities see the extraction and fusion of multimodal information as a core component of their work [9, 28–30, 32].

Given the lack of a common, mature definition, it is worthwhile to explain the understanding of MSA for the present work. The source domain provides review videos comprising user-generated, opinion-charged, multimodal content. This real-life media is at the end of the in-the-wild spectrum as shown in Figure 2.1. To exploit the modalities for modelling in a meaningful way, advanced approaches must be developed that deliver robust results despite the manifold interference.

The statements of an individual opinion holder are analysed for subjective and objective insights. This thesis follows the recent trend towards advanced emotional representations, reflected in prior publications, which attempt to bridge the gap between the communities. A terminological differentiation between short-term emotions and long-term sentiment appears reasonable from a psychological point of view. However, from a practical point of view, a model must pick up emotional cues from the same set and type of input data, regardless of the granularity and representation of the target. Following [12, 29], this thesis maps the subjective dimension as affective traces, e. g. , arousal and valence (see Section 2.2). The primary focus here is the fine-grained representation of both value and time continua. This seems to do better justice to the medium of video, itself sequential, fine-grained, and malleable. Moreover, it is (at least, technically) feasible to convert fine-grained to coarse-grained representations, unlike the other way around. Another essential element of MSA is an objective dimension wherein the topic of a given conversation is a core interest of this work.

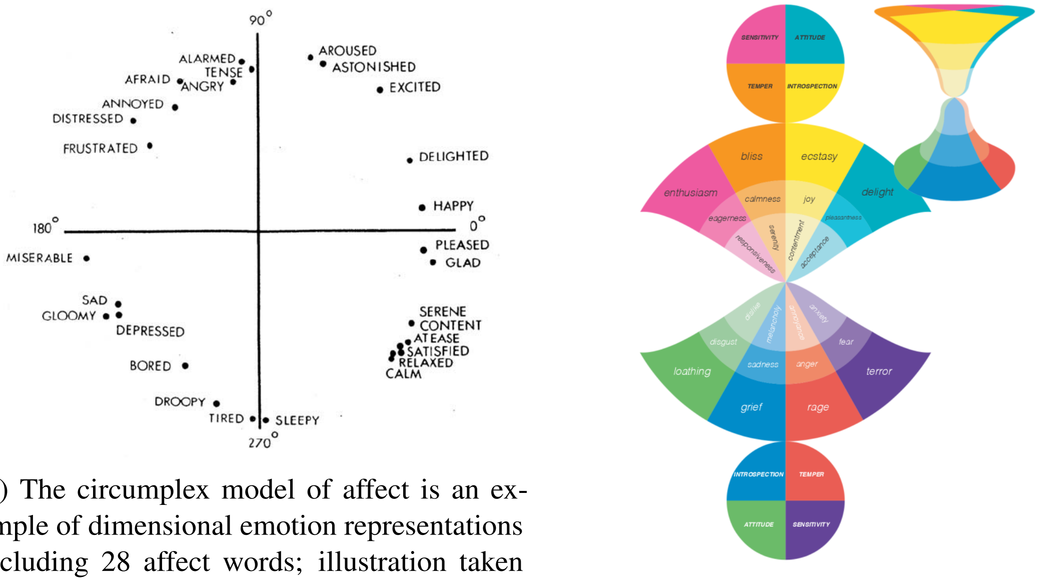
2.2 Subjective Emotion Dimensions

This section explains the subjective component of MSA in more detail, introducing fundamental emotional concepts, trade-offs between uni- and multi-modal input, and classification and regression approaches. The section rounds up with a brief background overview of one potential future subtask of the field: estimating popularity of social media videos.

As defined in the previous section, the subjective dimension encompasses more than simple positive-negative sentiments. The **conceptual mapping of complex emotions** is still the subject of ongoing research, with several research strands being pursued in parallel. Below are brief descriptions of the most prominent ones:

- **Discrete emotion theories:** Based on the understanding of innate, basic emotions hardwired in brain regions, these theories have several individual emotion expressions, each of which is represented by intensities. Ekmann and Friesen [37] conclude from their studies of facial expressions that there are six basic emotions, namely happiness, surprise, anger, disgust, fear, and sadness, which can be detected across cultures and are distinguishable from bio-psychological and physical reactions.

- **Dimensional emotion theories:** These theories describe the interplay of a few codependent primitive dimensional groupings reflecting an affective state. It is suggested that these are interconnected but have non-stationary locations in the nervous system. Therefore, they are seen rather as a reaction to the context and events of the environment, which are conceptualised and classified by our individual human understanding. Russell's **Circumplex Model of Affect (CA)** [16] is a popular advocate of this discipline. It is based on self-reported internal affective states in the form of emotion adjectives, which are represented in a two-dimensional circular order. Hereby, it maps a variety of affects on two axes of principal components, the arousal (vertical) and valence (horizontal) dimensions (see Figure 2.2a):



(a) The circumplex model of affect is an example of dimensional emotion representations including 28 affect words; illustration taken from Russell [16]. The vertical axis represents arousal and the horizontal, valence. The middle is seen as largely neutral. The intensity of both components corresponds to an affective state.

(b) The Hourglass of Emotions model is an example of a dyadic, categorisational emotion model typical for text sentiment analysis; illustration (revisited version) taken from [40].

Figure 2.2: Illustration of (a) dimensional and (b) dyadic emotion representation models.

- **Arousal** describes the degree of attention and alertness. It is activated by sensory impulses and can be measured by the central nervous system. For example, anger and astonishment show a very high degree of arousal, whereas sleep shows the opposite.
- **Valence** describes the spectrum between very negative and very positive. For example, frustrated is a highly negative state compared to satisfied, which is highly positive.

In theory, a specific emotion and the transition between emotions can be modelled through a combination of these primitives.

- **Dyadic theories:** Some attempts have been made to bring both of the above approaches together. The Hourglass of Emotions [38] is a mixture of discrete and dimensional approaches, inspired by the idea of emotional dyads [39]. Similar to CA, a circular representation of 24 complementary categories has been developed (see Figure 2.2b). Primitive emotions are on the inside of the circle, and emotions become more complex towards the outside. The representation reflects different emotional intensities.

These research directions are also reflected in computer-aided emotion and affect recognition. Ekman [37] is mostly applied in the field of sentiment analysis to predict several emotion classes on segments. This origin can be traced back to two factors: First, the original simple understanding of sentiment as a class with a positive, neutral, or negative expression. Second, the unavailability of other human signals beyond symbolic text representation of words and symbols, for example, prosodic features from the voice. Russell [16], on the other hand, is more prevalent in the AC and Signal Processing (SP) community, which has often been concerned with the analysis of signal-based data, such as acoustic or biological cues. Therefore, these emotions are often annotated as dimensional feel-traces, for instance arousal, in parallel to the (audio-video) signal and predicted as a sequence of continuous regression points [12, 41]. Although previous work has fused symbolic text and (quasi-)continuous audio-video signals to predict continuous affects, textual context information is still strongly underrepresented in the literature.

This thesis attempts to close this gap. In line with a large body of literature [13, 14, 34, 42], text is prioritised as an equal modality alongside audio-video signals — but with the difference that CA dimensions are continuously predicted instead of categorical basic emotion classes for predefined video segments.

Several attempts have been made to explore **additional focal dimensions** beyond arousal and valence, depending on the individual intention of the collected dataset and research direction [12, 41, 43, 44]. For example, value- and time-continuous likeability has been studied in human-human interaction [12, 18, 45]. Since the dawn of the internet age, the credibility of a trustee and the information provided from online sources has been an ongoing issue [46]. In the context of social media, trust is highly relevant when it comes to self-disclosure [47] and consumer decisions [48]. Previous studies have shown that social media protagonists who regularly convey content in a trustworthy way are more successful in attracting and interacting with online audiences [49, 50]. Understanding the mechanisms of how users gather, adopt, and trust information, and how this trust is reflected externally in a measurable way, is an open challenge [51]. A completely new approach could be to measure trustworthiness continuously as a new dimensional component.

Trustworthiness: Previous works have not yet settled on a final definition of trustworthiness [52–54] and sometimes found it non-trivial to quantify [53]. In Colquitt, Scott and LePine [55], analogous to the understanding in this thesis, trustworthiness is characterised as a trustee's capacity, benevolence, and honesty.

No previous study has attempted to quantify trust in videos from multimedia portals, such as YouTube, using a fine-grained, human-generated annotation of perceived trust. Through a continuous representation similar to arousal and valence, it is also possible to accurately identify relevant segments and build cross-domain recognition systems.

Table 2.1: Comparison of datasets focusing on at least one of three types of prediction targets: Sentiment (classes), Primitive (emotions), and Object of Interest (OoI) compared to the proposed dataset MuSe-CaR from Stappen et al. [56] (see Chapter 4 for a thorough introduction of the creation methodology). Modal(ities) available; Language: MULTI 1 = CN, DE, EN, GR, HU, SE & MULTI 2 = DE, ES, FR, PT; An(notation) Du(ration) (hh:mm)[1]; # (Number of minimum) Anno(tations per target). Subjectivity includes Sentiment: # (number of) sent(iment classes) (* intelligently derived from primitives); # (number of basic) Emo(tions); Cont(inuous), Primitive: Dim(ensions): V(alence), A(rousal), T(rustworthiness), L(ikability), I(ntensity), P(ower), E(xcitation), D(ominance); # (number of) Inc(rement) St(eps), T(race) annotations; O(bject) o(f) I(nterest): classes of topics or entities. Table adapted from Stappen et al. [24].

Name	**Modal**	**Language**	**AnDu**	**# Anno**	**Sentiment**		**Primitive**			**OoI**
					# Sent	# Emo	Class	# IncSt	Cont	
MuSe-CAR [24]	V,A,L	EN	40:12	5	3*	5*	V,A,T	✘	✔	✔
Text-centred										
UR-FUNNY [57]	V,A,L	EN	90:23	2	✘	1	✘	✘	✘	✘
MOSEAS [34]	V,A,L	MULTI 2	68:49	3	7	6	✘	V,A	✘	✘
MOSEI [58]	V,A,L	EN	65:53	3	7	6	✘	✘	✘	✘
ICT-MMMO [59]	V,A,L	EN	13:58	2	5	✘	✘	✘	✘	✘
Ext. POM [60]	V,A,L	EN	15:40	1	5	✘	✘	✘	✘	✘
CH-SIMS [61]	V,A,L	CN	2:20	5	5	✘	✘	✘	✘	✘
AMMER [62]	V,A,L	DE	1:18	1	✘	5	V,A	11	✘	✘
Youtubean [63]	V,A,L	EN	1:11	2	3	✘	✘	✘	✘	✘
MOUD [64]	V,A,L	ES	0:59	2	3	✘	✘	✘	✘	✘
YouTube [28]	V,A,L	EN	0:29	3	3	✘	✘	✘	✘	✘
Audio-video-centred										
SEWA [12]	V,A	MULTI 1	4:39	5	✘	✘	V,A,L	✘	✔	✘
HUMAINE [41]	V,A	EN	4:11	6	✘	✘	V,A,I	7	✔	✘
RECOLA [65]	V,A	FR	3:50	6	✘	✘	V,A	9	✔	✘
AFEW-VA [66]	V,A	EN	2:28	✘	✘	✘	V,A	21	✘	✘
VAM [67]	V,A	EN	12:00	6-8	✘	5	V,A	5	✘	✘
IEMOCAP [43]	V,A,L	EN	11:28	5	✘	9	V,A,D	5	✘	✘
SEMAINE [44]	V,A	EN	6:30	6	✘	7	V,A,I,P,E	✘	✔	✘
Belfast [68]	V,A	EN	3:57	6	✘	✘	V,A	3	✘	✘

Motivated by the importance of trust and emotion to engage users, this thesis proposes the novel continuous dimension of trustworthiness to study if and to what extent trust is another vital dimension in the emotional spectrum of user-generated content. The continuous dimensional representation is used as this is the only way to analyse the transition from content perceived as trustworthy to untrustworthy.

For the development of methods that automatically analyse content regarding subjective information, **data** is needed. According to Soleymani et al. [10], sentiment analysis using multiple modalities mostly emerges from audio-video content from social media platforms [69], e. g., video reviews [34, 59, 70] and human-machine and human-human interactions [12]. Table 2.1 gives an overview of common datasets, which are briefly explained below. First, datasets where the text modality plays a major role and that mostly follow the discrete theory are discussed, followed by audio-video datasets which ground their understanding in dimensional theory.

Text-centred datasets: Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) [58] contains 250 clips annotated with six basic emotion classes and seven sentiment classes expressing polarity. Some video material was removed due to alterations in the camera position during recording. The extension of the review database Persuasive Opinion Multimedia (POM) [71] by Garcia et al. [72] includes 600 audio-visual clips of people with visible faces talking about six facets of films. The average duration of the video snippets is one and a half minutes. The opinions are summarised in terms of segment-level annotation. Multi-Modal Movie Opinion (ICT-MMMO) [59] is similarly constructed with video material depicting reviewers looking into the camera. With this and the help of the subjects' voices, the aim is to be able to determine the sentiment of user-generated review videos. The Chinese MSA Dataset (CH-SIMS) [61] contains only Mandarin speakers. The 60 raw clips with fragments of 8–10 seconds in length were taken from TV shows, movies, and TV programmes. Again, only scenes where the voice and face are available at the same moment were included. Youtubean [63] includes mobile phone product reviews regarding seven typical product aspects which were made available in combination with sentiment annotations. Multimodal Opinion Utterances Dataset (MOUD) [64] also focuses on YouTube as a source and covers people facing the camera describing ideas in 30-second clips. Clips including background music were excluded. A broad spectrum of YouTube product reviews is offered by the YouTube corpus [28], which is labelled according to basic sentiment. The Automotive Multimodal Emotion Recognition (AMMER) [62] is a recently published dataset containing a simulation of a car ride with conversations in German. The object of observation here is the emotional relationships of the passengers.

Audio-Video-centred datasets: The following are datasets that focus more about the use of audio-video information than transcripts and provide some degree of dimensional annotation. Automatic Sentiment Estimation in the Wild (SEWA) [12] is one of the largest human-human interaction datasets. However, only four hours of the entire dataset have been fully annotated. In the interaction, participants converse via various webcams over advertisements presented to them beforehand. The recording is done statically, often in front of a white wall. In addition to the step-wise annotated arousal and valence, HUMAINE [41] provides continuous intensity annotations of various discrete emotions in a set of recordings of natural and played situations. Forty-six French participants interacting in a controlled laboratory environment and showing natural and spontaneous emotions are mapped in the REmote COLlaborative and Affective (RECOLA) dataset [65]. Arousal and valence are annotated via feel-traces. Besides visual input, the dataset also contains electro-dermal activity and electro-cardiogram traces. The Affect in-the-wild Valence-Arousal (AFEW-VA) dataset [66] covers a collection of video snippets extracted from feature films. Three raters annotated the clips with arousal and valence dimensional emotions on a scale from -10 to 10 with a one-step granularity. In addition to emotions, facial landmarks and seven facial gesture annotations are available. Vera am Mittag (VAM) [67] is composed of clips extracted from a German talk show. Parts of the recordings contain annotations of the six basic emotions. However, the main focus is on annotations for arousal, valence, and dominance, each on a 5-point scale. All three modalities are provided in the Interactive Emotional dyadic MOtion CaPture (IEMOCAP) dataset [43]. The record includes dyadic sessions in which 10 actors replicate controlled content. The SEMAINE [44] corpus includes 24 human-agent interaction sessions that were richly annotated. For the Belfast [68] recordings, emotions were actively stimulated and annotated in continuum.

Several observations can be drawn from this review of the latest datasets. Many datasets feature discrete emotions. The datasets with dimensional emotion theory aim to be generally applicable to many domains [73]. However, only recently have static experimental setups, e. g. , in laboratories, begun to be abandoned in order to collect noisy, large-scale, real-world data. In particular, this applies to the visual modality, which continued to be subject to defined specifications even though other parameters, such as recording equipment, have been made more flexible [12, 66]. User-generated videos from online sources that are particularly exposed to these challenges remain largely unexplored. Moreover, datasets with dimensional emotion annotations have so far shown little effort to provide content-related annotations, such as topic of conversation. This goes hand in hand with the fact that no attempt has yet been made to aggregate the dimensional annotations over the time period a specific topic is covered. With the improvements of models through deep learning and its efficient utilisation on large-scale data, giving up control of the data is the next plausible step to make models robust against various influences. Also, the properties of continuous traces might be helpful when dynamically breaking down a large sequence of audio-visual emotional annotations into shorter segments, e. g. , sentences, aspects, or noun-adjective pairs.

The mining of emotional information from **input modalities** has been an active research area within the Machine Learning (ML) and AC communities for more than a quarter century [74], and thus a comprehensive summary of this research area is beyond the scope of this thesis. The following discussion briefly introduces the unimodal perspective and then narrows the focus to the intramodal enhancement and the multimodal aspects. For further information, the interested reader is directed to the recent survey articles [75–77].

Unimodal: Using video data, three modalities are useful as a source to exploit emotional information. Each modality reflects unique aspects of human perception and is captured and stored in distinct ways. This also means that distinct expertise and tools are required for machine processing of each modality. For example, frequency (pitch range, contour slope), temporal (speech rate, stress), voice quality, and energy (loudness, breaks) characteristics are extracted from the audio signal, while changes in facial muscles and gestures are obtained from sequences of images. While in the past, filters to recognise these features were mainly developed by experts in these domains [78], potential inputs to recognition systems are now often learned by presenting large amounts of data to an Artificial Neural Network (ANN) [79]. For the text modality, which represents spoken language, there is no continuous signal; instead, there is a sequence of discrete string symbols. Due to computational limitations arising from the sparse representation of such information, those symbols need to be vectorised [80]. For this purpose, the context—the surrounding words and sentences in which a word appears—is considered [80, 81].

Recognition systems for prediction are based on the vectorised representations. While in the past, statistical Hidden Markov Models (HMMs) [82, 83] and Support Vector Machines (SVMs) [20, 26, 84] were the most common models trained to retrieve meaningful patterns, nowadays, ANNs are almost exclusively deployed. Networks that are able to feed in information in a sequential manner enjoy particular popularity due to their ability to learn temporal information [26, 85, 86]. However, Convolutional Neural Networks (CNNs) [87, 88] that lead to sparse extraction and end-to-end learning methods [88–90] by learning from raw image and audio material directly are also becoming increasingly popular.

Although these methods have led to considerable advancement in the predictive power and robustness of emotion models, the steadily increasing results on benchmark datasets suggest that the technological leeway is far from exhausted. Further progress is also needed for practical and in-the-wild data applications. Recently, a new ANN mechanism called attention [91] has greatly improved the learning of intramodality dynamics [13]. This often involves weighting information from a sequence step in relation to the surrounding information context. In terms of MSA, for example, this is very useful for spoken language, which is significantly disturbed by noise due to colloquial utterances (e. g. , “hmmm”, “it is like, you know, like”, “yeah”) and translation errors in the automatic speech-to-text pipeline ("bra" instead of "car") [92, 93]. This mechanism is widespread in ML tasks using text [82, 94–96], and its adoption in domains such as emotion recognition is accelerating [13, 97, 98]. In the text domain, more and more architectures that fundamentally rely on attention mechanism, namely Transformers [91, 99], are finding their way into mainstream research and beating previous benchmark results in unseen dimensions. Overall, paving the way for the exploitation of attention mechanisms at various levels would be beneficial to improve prediction quality. In particular, due to the strong intramodal properties, the representations of text from Transformer networks are predestined for a deep exploration.

In addition to unimodal intramodality, multimodal analysis focuses on facilitating intermodality dynamics.

Multimodal: The simultaneous use of multiple modalities, e. g. , trimodal, has been studied intensively in recent research [7, 100]. In this context, the linking of different modalities can be achieved through early fusion of input features [13, 98, 101] or fusion of the predictions at a later point [8, 20, 101]. The first approach, often called feature-level fusion, concatenates the inputs before they are fed into the ANN. This allows finding relationships from the earliest level, but leads to a high-dimensional feature space [42] that can contribute to overfitting. The second approach semantically models the independent modalities and fuses the result shortly before or after a prediction (decision-level).

Even though these techniques are gaining more and more traction, fusion of audio-video cues with the spoken word for predicting time-continuous emotions is still underrepresented in the field. This is caused primarily by the lack of suitable datasets (see above). In addition, small, hand-crafted feature sets are commonly preferred to much larger, ANN-learned representations to save computational resources [13]. As a result, the potential of combining information from multiple modalities is not fully exhausted and is an ongoing research challenge [42].

The conceptualisation of emotional information into discrete emotion classes and time-continuous primitive emotion dimensions almost naturally leads to the prediction tasks of **classification** and **sequences of value-continuous regression points**, respectively (for details, see Sections 3.2 and 4.3).

Target transformation: Even though a generalisable self-concept of emotion, e. g. , Russell's [16], may be better represented by dimensional axes, there is the challenge of interpretability of this abstract concept. People also often express themselves in concrete emotion classes, such as "I am disappointed" or "I was really happy". For better interpretability by humans, or to simplify the problem, continuous-time values are therefore transformed into classes. A naïve transformation would be to average the time-continuous values on fixed segment lengths and map them by hand to classes. A target transformation from discrete emotions to the dimensional emotion space [16] has been attempted by [102]. Similarly, clustering emotion tags was proposed [103] to find clusters corresponding to the four quadrants of the arousal-valence dimensions. In addition, transforming a continuous real-value annotation to time-continuous discrete annotation by using signal quantisation has been achieved [104]. However, this method fails to provide the possibility to summarise continuous emotion annotations to a certain class over a variable-length segment duration.

Little research has so far been undertaken to map more complex changes in the time-continuous signal to classes. For example, a rapid drop of arousal level within a short sentence is not reflected in an average value, as low and high values balance each other out, causing this characteristic to be lost. In addition, a flexible definition of segment length would render it possible to tailor the length rather than keeping to natural boundaries, such as sentences or self-contained thematic levels. This is especially relevant in the context of the objective dimension of MSA, which tries to achieve a thematic relation to the context of emotion (see Section 2.3).

The recognition of emotions in user-generated video content also opened up new avenues of subtasks and applications.

Popularity estimation: It is known that perceiving the transported emotional message from a video influences the viewer's feelings and thus their reactive behaviour [105]. User behaviour on (video) platforms, such as YouTube, is expressed in user engagement indicators such as the popularity of the video through likes, views, and comments. Social media network providers, in particular, could benefit from a deeper understanding of popularity through closer user engagement [106] and more meaningful recommendation systems [107]. Both still pose significant challenges today [108–110]. The dissemination of critically charged videos, such as fake news and hate speech, has also become a problem in both the virtual and real worlds. Finer control of distribution algorithms through an enhanced emotional understanding of a video could be a new path to a solution. Content creators may also be interested in applying this idea for marketing purposes to explicitly tailor communication content to customer groups.

Previous studies have shown that a portrayal of emotion [111], trust in the video protagonist [50], and positively reinforced content [112] influence viewers' engagement with the video [113, 114]. This has been investigated in relation to the consumption of traditional media, where it was found that emotional messages lead to consumers remembering content to a greater extent [115]. Emotional talk shows are also more popular than less emotional ones, a trait particularly recognisable through an analysis of audio characteristics [116]. In addition to the conventional understanding of primitive emotions, intricate emotions such as trustworthiness are also highly influential in this context. Influencers take advantage of this by portraying themselves as a close friend rather than a presenter in order to specifically build a parasocial relationship with the viewer [117].

The relationship between video platform metadata is the subject of a considerable body of research [118–123] e. g. , on user comments. It has been shown through correlation analysis [124] that the frequency, content, and sentiment of YouTube comments are indicators of user engagement with the video and viewer retention [125, 126]. Comments can also be used to predict the type and popularity of products in videos [127]. Furthermore, there have already been first attempts to transfer the reactions of individual users to comments into discrete emotions and to predict those emotions [128]. Such studies have also been conducted with regard to other platforms and social media reaction forms [129, 130]. For example, attempts have been made to map Facebook posts to the CA to predict the sentiment of future messages [131].

However, previous studies have not yet explored implicit emotional content and, hence, the relationships between time- and value-continuous arousal, valence, and trustworthiness annotations in combination with value-continuous popularity prediction (regression).

2.3 Objective Target Dimensions

The target dimension addresses understanding the content of a video in order to add context to an emotion [10, 132, 133]. Predicting an object of interest, such as a topic or aspect, is firmly rooted in conventional sentiment analysis. A topic represents a collection of semantically coherent aspects [134]. Aspects are a list of words that convey the topic's semantics. For example, given the aspect representatives {radio, screen, entertainment}, the topic could be "infotainment". Likewise, in the multimodal medium, the linguistic component takes a predominant role in thematic understanding. Spoken language influences accuracy and form, e. g. , through the intensive use of colloquialisms, automatic speech recognition errors, and long, convoluted statements [92, 93].

In the following, two approaches are discussed that focus on a) topic modelling for extracting aspect words without human guidance and b) targeted prediction of human-labelled relevant speaker topics.

Target extraction: Originating from the information retrieval community, target or topic modelling deals with automatically extracting topics from unstructured text data. The goal is to extract coherent topics and aspect representatives. Typical use cases are topic- and aspect-based sentiment analysis wherein, in addition to sentiment, representations (also called aspects or aspect terms) or name entities [135] are extracted from a text snippet. Improving topic modelling is therefore vital to enhance the fine-grained extraction of verbalised content and contextual nuances in sentiment analysis. From a multimodal perspective, the focus is on transcripts [92]. Here, the aim is to understand the significant themes discussed in one or more videos. The latent semantic structure thereby has to be inferred exclusively through the content. Potential applications include understanding of a video's content without watching, and finding suitable topics in preparation for human annotation for supervised prediction. In addition to understanding the ever-growing collection of video product reviews [136], related fields with a similar starting point, such as multimodal video indexing [137] and summarisation [138] are also increasingly dealing with this matter.

Due to the proximity to NLP and the early stage of work on multimodal corpora, user-generated online text data are included in the following discussion; however, this should be understood as a supplement and not as a comprehensive listing. For a deeper dive, the reader is directed to the surveys for text-based topic extraction [139], aspect mining [132, 133] and concept-level analysis [140]. In earlier approaches to topic modelling from transcripts, the subjectivity and high word-error rate of speech recognition systems posed a considerable challenge [93]. These problems can be partially mitigated by rigorous preprocessing, for example, by including only spell-checked, non-colloquial words [141]. Although semantic connections are lost, this approach has proven robust for extracting topic representatives [142, 143].

A methodological starting point for learning semantic latent structures without labels is clustering. Clustering on text aims to group semantically linked words together. While in the past, these were often based on a traditional text representation in the form of TF-Inverse Document Frequency (TF-IDF) matrices, current research is increasingly focusing on vector representations [144].

Word embeddings: Word-embedding topic models are based on vectorised text (see Section 3.1.2 for a detailed explanation). These embeddings are semantically related to each other, which is manifested in that related embeddings have a shorter distance to each other. Clustering methods exploit this property to group semantically similar words and discover the semantic structure of the underlying corpus [144–146]. This strategy outperforms traditional approaches for processing noisy online data from social media platforms, e. g., Twitter or Reddit [144]. Word-based topic models calculate an embedding per word of a given vocabulary corpus, resulting in meaningful, self-contained topics, thereby reducing overlap and increasing topic coherence [145]. In this setting, it has been shown that classical word embeddings like word2vec (one vector, one word) are more effective than dynamically computed contextual word embeddings like Bidirectional Encoder Representations from Transformers (BERT) (see Section 3.1.2) [145–147].

Clustering: The word-embedding representation of text opens up a high degree of flexibility, as it can be used by all sorts of clustering methods. K-means [148, 149], which assigns each point to a cluster centre, is widely utilised. It assumes consistent cluster sizes and the absence of outliers. This, in conjunction with limiting specifications needing to be done in advance by the user, such as the number of clusters, which would require a precise estimation of semantically related word groups, often leads to a lack of accuracy [150]. Density-based clustering algorithms are also becoming increasingly popular. They assume that good clusters have a high density, and between them lie areas of low point density [151]. The best known representative is the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [152, 153], which requires the definition of the minimum cluster size for clustering word embeddings [154]. Another widely used method is Latent Dirichlet Allocation (LDA) [141, 144, 155]. Here, a topic probability is assigned to each word in a document. By iteratively optimising the probabilities, robust topics of selected words are formed.

Several more methods have been proposed for this task in the past. However, due to their lack of topic coherence and the need for most clustering algorithms to specify the number of expected clusters beforehand [149, 139], no standard method has been established to date and new ways to approach this challenge are needed. A promising direction is graph-based methods [156, 157], which are becoming increasingly popular due to the intensive adoption of semantic network modelling of social media data [158].

Graph-based topic models: A graph consists of a set of nodes (vertices) and edges [159], wherein a graph node usually represents the text, and the edges represent the similarity between the connected nodes. Starting from a complete graph where all nodes are in the same cluster, connections are dropped, e. g., if the connectivity is low and the remaining nodes are clustered. Clustering algorithms form k-components by extracting subgraphs representing individual clusters from the graph [157, 156]. One advantage of this is that the number of cluster topics does not need to be defined in advance. An early evaluation of graph clustering on text [160] employed a fixed number of clusters and a threshold-logic-based TermCut algorithm to group short Chinese texts. For this purpose, TF-IDF representations were retrieved from Chinese short text snippets. Similarly, in another work, nouns were extracted to cluster crime reports [143]. Another use case involves patient incident records [161]. Here, the results of the graph connectivity algorithm proved superior to K-means for topic extraction [161].

In the context of multimodal sources and transcripts, graph methods have not yet been developed, and in light of recent findings, further research is needed.

Supervised prediction is another way to identify the target of a sentiment by using knowledge provided by a human via annotations.

Target detection: Target detection is a core task of text-based sentiment analysis [31, 162], wherein a reference target is detected [10, 36, 163] for later linkage to an emotional disposition. A target emerges in various granularities [36], such as an entity, topic, aspect, or physical object. In general, the goal is to train an algorithm to learn patterns indicative of a target from sample data in the domain of study to recognise them on unseen data. For this to be possible, topics have to be deemed relevant a priori to the automatic analysis [134]. Targets that humans have not annotated in advance cannot be predicted. The application is often particularly relevant in the case of customer reviews and social media data [164]. Typical supervised methods on text are rule mining [165] and lexicon methods [166, 167].

An explicit transfer from a textual perspective to a multimodal one in relation to target concepts has not yet been carried out [10].

Speaker topic: A natural target in the context of videos seems to be speaker topics, as the analysis in a video (segment) always refers to the perspective of an opinion leader. A speaker topic can be seen as an utterance with a defined start and end point within a discourse or a monologue [24, 26, 56]. In this understanding, the utterance on a topic goes beyond the conventional sense of an isolated visually similar scene or language separated by the means of sentences. It comprises a coherent section that enables the interplay of image, sound, and text from an opinion-centred perspective.

Similar to textual sentiment analysis, it is likely the targets will be refined over the years [163, 168], and more research has to be carried out.

Content and context understanding is gaining momentum in the analysis of video data.

Datasets: Using a variety of modalities, datasets are widely applied in human action and motion recognition [169–171], semantic scene segmentation [172], and video activity recognition [173]. One of the most extensive available datasets, MOSEI, provides videos covering 250 topics [58]. Each video covers only a single high-level topic; 16.2 % of the videos were labelled as reviews, 2.9 % as debate, and 1.8 % as advice. The humour dataset UR-FUNNY [57] covers more than 400 topics from TED Talk videos to examine humour and punchlines in context to topics. Here, the topic tags, such as technology, culture, and science, are very generic. Furthermore, the subjective and objective dimensions are contextualised for affect measurement in film scenes [174]. The dataset is very small, with only 14 video clips. Other datasets in the domain cover a large set of topics but do not provide annotations [64, 175].

In summary, current datasets do not support development of state-of-the-art MSA models. They focus primarily on detecting emotional characteristics, but lack suitable size and annotations. Many datasets centered on MSA cover a broad range of topics. Thematic understanding at a generic, high level of abstraction, e. g. , the theme of an entire review video is ”housing" or ”finance" [58], is desirable to increase the generalisability of models but does not serve an understanding of individual video segments in an opinion-topic structure, e. g. , a person is concerned about the leaking pipe in the loft. Moreover, the generalisation capabilities of general language models [176] have improved enormously in recent years and will continue to do so [177]; therefore, the research focus can and should now shift towards fine-grained, complex understanding. Furthermore, multimedia datasets that directly target speaker topics or aspects for prediction are clearly in the minority [63, 72]. This confines (supervised) context comprehension to language analysis nearly entirely. No approaches that rely solely on the spoken word will ever be able to provide a full understanding of interaction, for example, when verbal language is not accessible or is only partially so. Evidence shows that modalities such as the image play a crucial role in understanding aspects of an entity [178], and all complementary modalities should be integrated for an optimal outcome.

Without adequate data, only limited research has been done to develop suitable methods. The supervised recognition of targets in opinionised texts is a very intensively researched field [132] with diverse methodological approaches, for example, lexicon-based [179], graph-based [180], and neural learning [96, 181]. The added value of images has been demonstrated by developing an ANN on a bimodal, text-image sentiment analysis dataset [178, 182]. In relation to video-related textual information, initial research has been conducted. By creating a dataset consisting of seven videos for closed captions from YouTube, an attention Recurrent Neural Network (RNN) network was trained to extract and classify the sentiment and aspects of product videos [63]. Methods such as bag of features and LDA were elaborated using video representations to extract topics from promotional videos [134]. State-of-the-art ANN architectures such as Transformer [183] and end2end [184] multimodal learning methods have been neglected so far.

Although the linguistic component is given a high priority in previous work, the adoption of knowledge-base approaches [185–187], which have been successful in [188, 189], is also under researched in this context.

Knowledge-base approaches: As explained above, topics are superordinate to aspects. Likewise, aspects are not the smallest atom of human language. Knowledge-base approaches rely on this logic to construct a taxonomy of common-sense knowledge and part-of relationships [190, 191]. They also factor in social norms that play an essential role in contextualising words into knowledge [192]. This understanding has been created from the contexts of words in the language and stored symbolically. Since a manual specification of such would imply extensive domain knowledge and expenditure of time, this is often automated. Using multi-term keyword analysis of the content, context concepts can be extracted and further processed. The most comprehensive framework to date contains over 200 000 related concepts that map a word to subconcepts of life [179].

In summary, further investigations on video recordings using both unimodal and multimodal methods are warranted to specifically elicit the value of the individual modalities, the representations extracted from them, and the modelling methods for predicting human-prescribed targets.

Methodology

3 Methodology

This chapter familiarises the reader with the most fundamental methodology used in this work. To enable algorithms to learn from multimodal behaviour data, it has to be converted into a machine-readable form. The techniques and common frameworks for this are presented in the first section, and the second section presents different modelling methods to perform predictions from these representations.

3.1 Modality Representations

In the following, the types of hand-crafted and data-driven representations relevant for this work, and their corresponding extraction frameworks are introduced. To use raw audio, text, or video in Deep Learning (DL) (see Section 3.2), both hand-rafted and learnt representations are extracted and exploited [193]. However, they differ in their origin and composition:

- **Hand-crafted representations:** The majority of conventional machine learning algorithms require specifically crafted data representations that have been meticulously constructed [77, 78, 194]. Specialists obtain distinguishing traits, properties, and attributes, commonly referred to as expert-designed representations, from raw data while employing domain expertise in combination with mathematical methods, e. g. , from statistics or signal processing. The types of unstructured data with which this work is concerned include the following: Signal processing techniques can retrieve audio representations, such as the fundamental frequency (F_0), jitter, and shimmer of an audio signal [77]. In text analysis, the term frequency-inverse document frequency is often applied when it comes to analysing content [195]. As a numeric statistic, it counts the frequency with which a term appears in a document, weighted by the number of documents within the corpus that include the term. In computer vision, edges and corners are retrieved from pictures for training edge detectors [194]. More specifically, they find areas that where the picture's luminance changes abruptly.

- **Data-driven representations:** Various machine learning domains today employ Artificial Neural Networks (ANNs) as their primary method of problem-solving [196, 197], as explained in detail in Section 3.2. DL, in particular, is an umbrella term for techniques and Deep Neural Network (DNN) architectures with many layers (thus "deep") that have surpassed human capabilities in complex pattern recognition tasks [193]. Unlike traditional approaches, these robust representations are learned automatically from processing (almost) raw data. Instead of relying on engineered representations requiring human involvement, they find generalisable patterns from the data itself by passing it through DNN architectures. The neuronal activations of individual layers of these networks can subsequently be exported as a fixed vector and further exploited as a representation of the previously unstructured data, as with hand-crafted features.

The terms feature, feature set, and representation are often used interchangeably. The term "representation" was originally coined for a vector consisting of several features, where each represents a vector dimension, that are learned by a DNN. Hence, a feature is often equivalent to a single, activated neuron output (see Section 3.2). Various types of representation are often summarised as feature sets. For this reason, the terms feature and feature set are used in this work regardless of their origin, and representations are explicitly referred to as hand-crafted or data-driven to emphasise the type of creation.

3.1.1 Audio

Many tools have been proposed for making raw audio workable with machine learning methods. In the following, two methods are briefly presented each for hand-crafted [78, 198] and data-driven representation [199, 200] extraction that have a proven track record in audio processing and emotion recognition [199–201].

- **ComParE Low-Level Descriptors (ComParE LLDs)**: ComParE LLDs are the underlying base for calculating the INTERSPEECH ComParE functionals [202]. Both are widely used in computational paralinguistics [198]. Composing six voice-related, four prosodic- and energy-related, and 55 spectral LLDs and their first order derivatives results in a 130-dimensional vector, which can be extracted using the open-source framework Open-source Speech and Music Interpretation by Large-space Extraction Toolkit (openSMILE) [203]. For detailed descriptions, the interested reader is referred to Eyben et al. [203].

- **Geneva Minimalistic Acoustic Parameter Set (GeMAPS)**: The GeMAPS was developed for speech analysis research [78] and is specifically successful in the context of emotion recognition from speech [204]. The feature set is made up of a collection of statistical functions based on acoustic spectral, cepstral, and prosodic LLDs. Due to their theoretical relevance and simplistic computation, they are selectively picked by hand. The basic version has only 28 low-level descriptors, while the extended

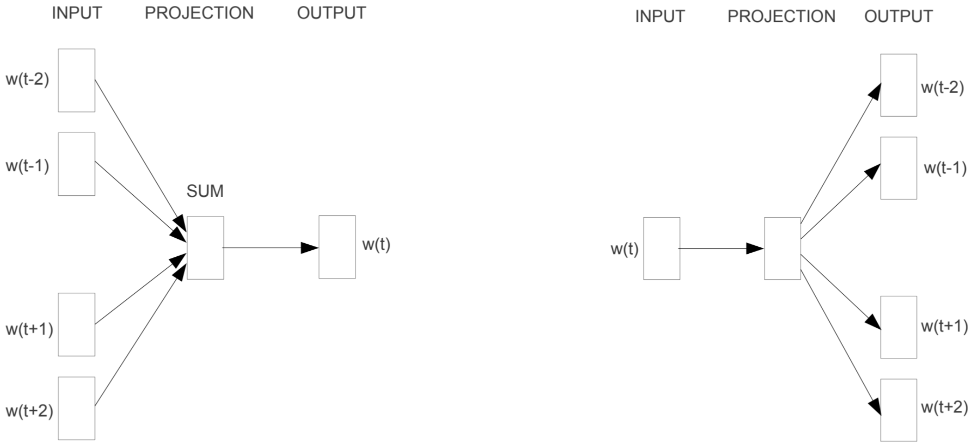
Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [78] has 88, which can be represented by a vector with the same number of dimensions. Both show robust prediction quality in various settings and tasks [8, 20, 84, 201]. As with ComParE LLDs, the open-source openSMILE toolkit [203] can assist in extracting these hand-crafted representations.

- **Spectrograms Feature Extraction from Audio Data with Pre-trained Convolutional Neural Networks (Deep Spectrum):** The Deep Spectrum representation extraction toolkit[1] provides deep representations from spectral images. After converting the audio input signal into mel spectrograms, pretrained Convolutional Neural Networks (CNNs) (see Section 3.2.3) can be fine-tuned with these images [199]. Alternatively, the images can also exclusively be processed through the network without fine-tuning. With either option, the final result is the network output of one of the final network layers, representing a compressed visual representation of an audio signal. These representations are effective for a wide range of audio-related tasks, including speech processing [79].

- **CNN Architectures for Large-Scale Audio Classification (VGGish):** As a deep acoustic representation extractor, VGGish also relies originally on a CNN architecture [200]. This VGGNet [205] is trained for auditory event recognition using log spectrograms from a large-scale audio dataset (70M training videos), crafted from 10-second YouTube snippets, called AudioSet [206]. Specifically, snippets used for pretraining were selectively picked through the probability of having acoustic events while making use of content analysis and metadata. Based on this, the more than two million identified audio recordings were categorised by human labelling into 632 audio event classes. This makes it a versatile training ground when it comes to robust generation of in-the-wild representations. The representation of the fully trained VGGNet has 128 dimensions. Although relatively small for DNN representations, e. g. , compared to the more than four thousand dimensions of Deep Spectrum, it has shown a high semantic expressiveness in numerous studies and often outperforms hand-crafted acoustic representations [207]. A framework which comes with pretrained models is available as open source.[2]

[1] https://github.com/DeepSpectrum/DeepSpectrum accessed January 15, 2021

[2] https://github.com/tensorflow/models/tree/master/research/audioset/ accessed July 5, 2021.



(a) The Continuous Bag of Words Model (CBOW) uses the context to predict the centre target word.

(b) The Continuous Skip-gram Word Model (CSG) uses the centre word to predict the context target words.

Figure 3.1: An abstraction of the underlying principles (a) CBOW and (b) CSG of Word to Vector (Word2Vec) training. Figure is taken from [80].

3.1.2 Text

Transcripts are the codependent, linguistical output of the spoken word and are reliant on the auditory signal. Due to the symbolic representation, efficient processing of raw text directly by ANNs is infeasible. Thus, it is a prerequisite to remould the word symbols to a vector. A straightforward way to represent words is to assign one position of a binary vector to each unique word of a corpus. Representing each word with a hot-encoding leads to a sparse vector of large dimensionality, which grows linearly to the number of unique words of the underlying text. This negatively influences memory and computational requirements. Furthermore, these equidistant vectors miss semantic information since the vector position does not have any deeper meaning.

These issues have motivated real-valued distribution representations within a semantic vector space of definite dimensions. There are two kinds of these so-called **(static) word embeddings** relevant to this work:

- **Word to Vector (Word2Vec):** Word2Vec [80] is a standard model that learns a vector for each term by employing a three-layered ANN in two ways: CBOW and CSG (see Figure 3.1). The CBOW model takes the context of each target word and learns to predict the encoding of the centre target word from the context window. A comparison is made from this encoding to the actual one-hot encoding of the target word to determine the output error. During the training process, this error is reduced step by step by tweaking the network weights. The CSG functions oppositely. It takes the

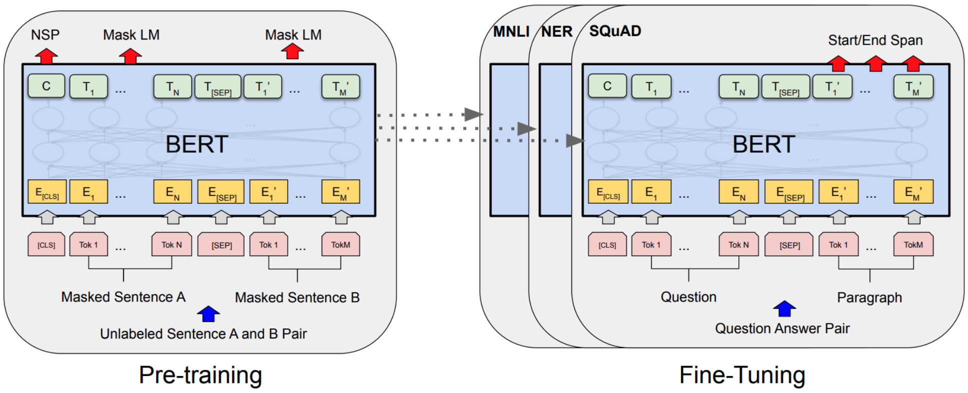


Figure 3.2: Training concept of Bidirectional Encoder Representations from Transformers (BERT), where the left side shows the general language pretraining approach using Next-Sentence Prediction (NSP). Two masked sentences (light red) are feed into the model, embedded (yellow), and processed through the Transformer network (blue) before they predict the masked words and the NSP target (arrows). The right side shows the subsequent downstream fine-tuning tasks. Figure is taken from [81].

target word and predicts the one-hot encoding of each word in the context window. The most effective and efficient version has a 300-dimensional embedding vector.

- **Fast Text Classifier (FastText):** FastText [208] is based on the CSG version of Word2Vec. In contrast, the constituents of a word, rather than the whole words, serve as input. A vector representation is associated to each character n-gram to enhance effectiveness. In addition, these subword pieces make word evaluation representation possible for out-of-vocabulary terms that did not appear in the initial training corpus. This appears advantageous when dealing with technical phrases and words from a domain-specific corpus. Like Word2Vec, this representation usually has also 300 dimensions. Several pretrained embeddings, for example, trained on the English Common Crawl corpus (600B tokens), are provided by Facebook's non-proprietary FastText Toolbox.[3]

The limitation of static word embeddings is that all senses of a polysemous word are trained to share a single embedding vector, discounting the contextual appearance of a word. Recent advances have resulted in models capable of generating **contextual word embeddings** based on Transformer architectures [209] (see Section 3.2.5):

- **Bidirectional Encoder Representations from Transformers (BERT):** Surpassing a set of Natural Language Processing (NLP) benchmarks [210], BERT became the

[3] https://fasttext.cc/docs/en/english-vectors.html accessed July 5, 2021

most commonly applied NLP Transformer base [81]. The main novelty of BERT is its context dependence enabled by considering the position and context of each word in a sequence. Besides the architecture, the training is different to previous embeddings. First, the words are segmented and tokenised using a WordPiece tokeniser. For this, the tokeniser initially starts with the individual characters of a corpus, iteratively fusing the most occurring character combinations to new tokens. Second, the language model is taught to predict around 15 % of the input tokens, which were masked beforehand, in a so-called masked language modelling task. The idea behind this is that the model has to use the tokens surrounding the input to predict the original token. From a high-level perspective, this is similar to static word embeddings; however, here the context is captured and incorporated in the network representations. Improving the context understanding further, BERT is indirectly instructed to learn the relation between the words and logically subsequent sentences by a NSP task as depicted in Figure 3.2. Next, it can be fine-tuned for typical NLP tasks, such as the Stanford Question Answering Dataset (SQuAD). Compared to static word embeddings, these word vectors are dependent on context, and therefore, the calculations are performed at run-time. Ultimately, final layer representations are extracted and concatenated, for example, the last token representations from the last four decoder layers. Pretrained BERT derivatives are available from Hugging Face.[4] The weights are pretrained on English Wikipedia (2.5B words) and BooksCorpus (800M words) [211] for representations extracted for this work.

- **A Lite BERT for Self-Supervised Learning of Language Representations (ALBERT):** An evolution of the BERT architecture is ALBERT [212], focusing on supervised fine-tuning ability. Using almost the same architecture as BERT for the Transformer encoder, it comes with two novel design choices to improve parameter efficiency: First, it proposes to separate the size of the WordPiece embedding layer from the hidden layer size. This coupling has the unwanted effect of blocking substantial memory, even though most parameters in the input layer are only updated rarely. By refactoring the embedding encoding from one large embedding matrix to two separate ones, it first projects the one-hot vectors into a lower dimensional embedding before propagating to the hidden space, making updates more frequently while having lower memory requirements. Second, it allows global weight-sharing, wherein all layers across the entire architecture share the parameters of the two Transformer layer components (Section 3.2.4.3). Sharing smooths interlayer transition and stabilises the network. Furthermore, it changes the intersequence modelling from the NSP loss (and

[4] https://huggingface.co/ accessed August 15, 2021

their associated tasks) to a sentence-order prediction loss, which forces the network to focus primarily on fine-grained coherence modelling. These changes lead to a parameter reduction of almost a factor of 20 while, at the same, time improving the capabilities of downstream supervised fine-tuning, leading to state-of-the-art results in several benchmarks. The final architecture is equipped with 12 blocks, each using 64 attention heads, and is also available from Hugging Face.

While word embeddings capture semantic and even contextual information of words and sentences using probabilistic, word-frequency models, natural language can provide even more conceptual insights [38, 40]. However, translating words into **high-level language concepts** requires common-sense knowledge, which is still difficult to learn, store, and access in its entirety by machines. One approach to this challenge is **SenticNet**, a knowledge base of 200 000 common-sense concepts, which offers a set of semantics, sentics, moodtags, and polarities associated with natural language representing fundamental human concepts [179, 187, 213]:

- **SenticNet:** SenticNet concepts capture evocative, indicative, and affective information associated with entities and phrases [140]. It employs two methods to establish its knowledge base — DL methods and symbolic approaches, such as ontologies — in a unified fashion [179]. The first identifies word and multiword expressions. The latter infers syntactic patterns to get reduced into primitives and superprimitives, as illustrated in Figure 3.3. Hence, it becomes less challenging to chart phrases and words into semantic representations linked to specific concepts [187]. Further following the hourglass model [38, 40], primary and secondary moodtags become existent. In addition, moodtags appear beneath the sentics and come with labels including delight, ecstasy, and bliss. Concept clusters associated with the segment are extracted as semantics and have an identical lexical function. All these high-level concepts can be exploited as representations for subsequent tasks [56]. For a more detailed elaboration on sentic computing, the reader is referred to the book by Cambria and Hussain [140].

This learnt conjunction of a given text to their common-sense concepts can be accessed by an application programming interface. When called, it returns a set of semantics, sentics, and polarity associated with the respective concept in a vectorised form. Several knowledge-base versions incorporating some evolution of this method have been made available. For example, in version five, the four sentics provided are introversion, temper, attitude, and sensitivity [187], while for version six, they are pleasantness, attention, sensitivity, and aptitude [179].

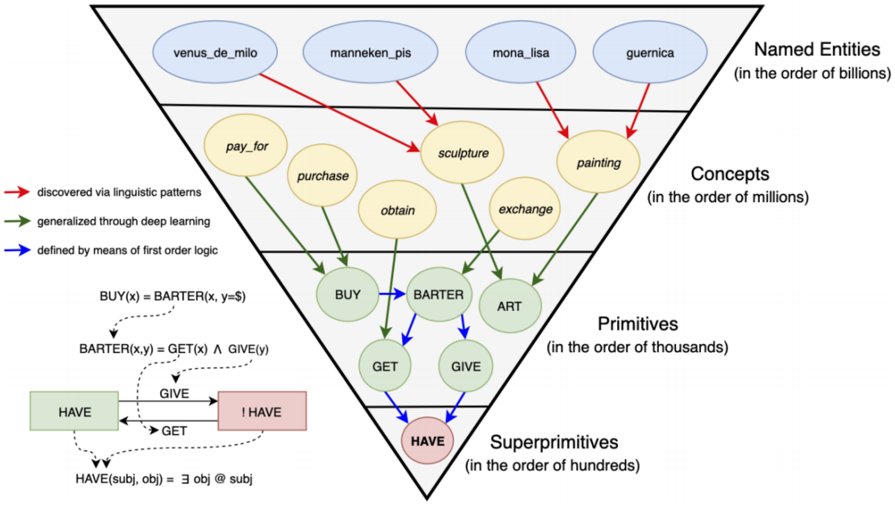

venus_de_milo
manneken_pis
mona_lisa
guernica
Named Entities
(in the order of billions)
pay_for
purchase
obtain
sculpture
exchange
painting
Concepts
(in the order of millions)
discovered via linguistic patterns
generalized through deep learning
defined by means of first order logic
BUY
BARTER
ART
GET
GIVE
Primitives
(in the order of thousands)
HAVE
Superprimitives
(in the order of hundreds)
BUY(x) = BARTER(x, y=$)
BARTER(x,y) = GET(x) ∧ GIVE(y)
GIVE
HAVE
! HAVE
GET
HAVE(subj, obj) = ∃ obj @ subj

Figure 3.3: SenticNet 6's reliance graph, showing the structure and source from name entities to primitives and superprimitives via concepts. Figure is taken from [179].

3.1.3 Video

Most visual extractors are designed to localise regions (e. g. , face), extract specific characteristics (e. g. , joint movement), or create discriminatory representations from an image. In the following, the representation frameworks relevant to this work are separated into human- and environment-focused (the visual context). Automatic extraction of visual information related to people is mostly practiced for emotion detection:

- **Multi-task Cascaded Convolutional Network Framework (MTCNN):** Before extracting facial representations, frameworks that locate the person of interest's face, such as MTCNN [214], come to play. It employs a three-phase cascaded structure in real-time for the automatic localisation of faces and facial landmarks. The datasets WIDER FACE [215] and CelebA [216] provide sufficient training material to learn the prediction from photos of real-faces.

- **Facial Action Units (FAUs):** FAUs are structured descriptions of perceived facial movements [37]. Derived from the Facial Action Coding System (FACS), they are considered an essential component of many emotion recognition systems that capture nonverbal cues using graphical perceptual methods. Facial cues are deconstructed into 17 specific modules related to muscle movement. The intensity of each is indicated on a scale from 0 to 5, with higher numbers reflecting higher intensities. The presence and the intensity can be automatically obtained from the OpenFace recognition toolkit [217], which aims to detect and analyse facial changes using cropped faces,

e. g., extracted by MTCNN as introduced before. This is exemplified by 2D (136 features), 3D (204 features), and other related features such as 288 gaze points and six head-pose positions.

- **Open Multi-person System to Jointly Detect Human Body, Hand, Facial, and Foot keypoints (OpenPose):** An effective representation extractor to recognise various sequences of people's movements by automatically identifying their joints[5] is provided by OpenPose [218]. As an automatic human-pose estimator, it is utilised in various fields, such as action recognition. It yielded the best results in several challenges such as the COCO 2016 Keypoints Challenge [219]. The underlying model consists of two branches of stacked CNNs, one of which identifies Part Affinity Fields that encode pairwise relationships between body parts. The other branch predicts 2D confidence maps for the vital points in question. On each layer, the outputs of each branch are concatenated and serve as input for the higher layer. Ultimately, the 2D coordinates for each of the 18 keypoints are available, as well as the corresponding confidence value for the presence of a keypoint.

- **Very Deep Convolutional Networks for Large-Scale Face Recognition Descriptor (VGGFace):** Originally intended for facial recognition tasks, VGGFace can be used for deep facial representations [220]. Its main advantage is its performance equivalence compared to other face recognition representations while using less data for training. The Oxford Visual Geometry Group developed the dataset, consisting of an excess of 2.6 million faces from 2 500 identities, used to train the deep CNN [205]. The top layer of the VGG16 network outputs a representation vector of 512 dimensions. Unlike other frameworks, for instance, OpenFace, the representations are not deterministic and do not model explicit intensity or presence scores.

Furthermore, for a better understanding of the environment and object interaction, the following information can be extracted:

- **Depthwise Separable Convolutions Network (Xception):** A state-of-the-art DNN, trained on a general image corpus, can extract general vision representations from an environment. The Xception [221] network uses residual blocks to stabilise its very deep network architecture. Given this greater depth, the architecture won 1st place on the ImageNet Large Scale Visual Recognition Challenge 2015, from which the well-known ImageNet benchmark dataset emerged. The dataset covers images

[5] The keypoints are: Nose, Neck, Right/Left Shoulder, Right/Left Elbow, Right/Left Wrist, Right/Left Hip, Right/Left Knee, Right/Left Ankle, Right/Left Eye, and Right/Left Ear.

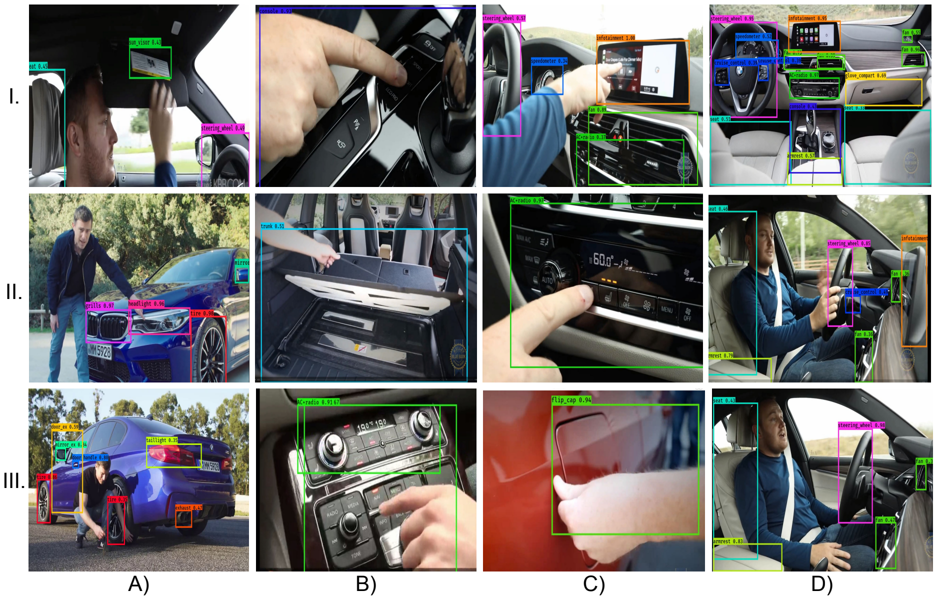


Figure 3.4: Examples of Generic, Optical Car Part Recogniser and Detector (GoCaRD)'s automatic detection of automotive parts with human interaction in vehicle environments from the video dataset Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) presented in Stappen et al. [24]. A), B), and C) show gestures and finger pointing. In A) the predicted bounding box around the part excludes the body parts, while in B) and C) the parts are fully recognised despite occlusion. D.I) shows the complete vehicle interior without people, while II) and III) compare the same situation with and without autonomous driving (hands on/off the steering wheel). Figure is taken from Stappen et al. [2].

structured in a hierarchical order of natural concepts. By relying on such a broad and deep dataset, a pretrained Xception network can be used as a representation extractor on images regardless of the domain. The last fully connected layer is extracted to obtain the deep representations, resulting in a 2048-dimensional vector for each image.

- **Generic, Optical Car Part Recogniser and Detector (GoCaRD)**: Stappen et al. [2] introduced the GoCaRD framework, specifically designed for the automotive domain and capable of localising 27 different car parts. By doing so, it aims to provide a deeper understanding of how a person interacts with the interior and exterior of a car, for instance, a hand localised by a third party framework (e. g. , OpenPose) overlapping the detected infotainment systems in video frames as depicted in Figure 3.4. The training material stemmed from combining a real-world photo database and video frames extracted from the MuSe-CaR database (see Chapter 4), wherein the former only depicts cars and the latter depicts human-vehicle interactions in YouTube car reviews [2, 24]. The material features several car makes and models. In total, more than 15 003 images were available for multi-label, multi-class labelling. The large scale and high variations within the dataset were deemed necessary to train a robust visual detector that can work with a variety of purposes in mind. In addition, the automotive domain has one of the highest variations in its products, such that a single model can achieve up to 10^{24} feature combinations [222], many of which affect the visual appearance. The underlying CNN, a Darknet-53 architecture, is trained in a two-step domain-adaptation procedure. First, the database without human involvement is used, followed by a specific fine-tuning using the videos depicting humans. As a result, a Mean Average Precision (mAP) of 67.57 % across all classes could be achieved; individual classes range from 14 % mAP for less distinctive car components, such as rooftop windows, to 94 % mAP for very distinctive ones, such as front grills. The prediction system uses the fully trained model as a backbone in a YoloV3 framework [223]. However, since the number of detected parts varies, the output's size had to be extended. In the implemented logic, the outputs are converted into a fixed-size vector, where only the ten objects with the highest confidence are included. By representing every object with a confidence value, the x and y coordinates, and the width and height, the representation vector is bounded to 134 dimensions (10 objects * (27 classes + 7 object-related data points).

3.1.4 Emotional

In previous publications introducing Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox (MuSe-Toolbox) [25, 224], Stappen et al. proposed extracting a 24-dimensional feature set reflecting the distribution and temporal changes of an emotionally human-annotated video segment, as discussed in Section 4.1, to make them more workable for further analysis. Generally, these hand-crafted representations can be separated in two groups: statistical measures and measures describing the changes in a time series of data.

- **Statistical measures:** Given the ability to summarise data quantitatively, numerical measures can provide easily understandable insights into the distribution of data. Typical measures are the standard deviation (*std*), and 5 %, 25 %, 50 %, 75 %, and 95 % *quantiles* ($q_5, q_{25}, q_{50}, q_{75}, q_{95}$), which have been shown useful in related works [116].

- **Time series measures:** To generate a deeper understanding of changes within a time series, time series statistics can be extracted following previous work in a similar field [225], such as speech emotion recognition where energy-related measures have been highly effective [226]. In the following, relevant measures are briefly described;

however, the reader is referred to the original sources for in-depth explanation and formalisation.

Using the adjusted Fisher-Pearson standardised moment coefficient, the dynamic sample skewness (skew) of a time series can be calculated to describe the strength and direction of the time series's asymmetry [227, 228]. Another measure of shape is the Kurtosis (kurt) as a descriptor of the "flatness", derived from the fourth standardised moment of a time series [229]. Energy-related descriptors are potent measures broadly applied in signal processing. A discrete-time signal's absolute Energy (absE) can be defined as the sum over the squared values [230]. The Sample Entropy (SaEn) is a measure of complexity. It is a variation of the approximate entropy that measures entropy independently of the time series length and has found widespread use in physiological applications [231, 232].

Several change-related features can be computed to assess a time series signal: The Relative Sum Of Changes (ASOC) can be obtained through the sum over the absolute value of subsequent changes. Similarly, the absolute difference between subsequent data points is called the Mean relative Absolute Change (MACh). The Mean Change (MCh) is defined as the general difference between consecutive points over a time series. The duration of a normalised consecutive subsequence is expressed by the measures strike above (LSAMe) and below (LSBMe) the mean. Lastly, higher-order changes of a time series signal can be calculated, such as the Mean value of a central approximation of the Second Derivatives (MSDC).

Taking the number of data points occurring more than once divided by the number of total points determines the normalised percentage of Percentage of Reoccurring Data points of non-unique single points (PreDa), a summary measure of a sequence of discrete-time distribution similarity. In addition, events relative to time can give deeper insights into the time series course. The first (F) and last (L) locations of the minimum (Mi) and maximum (Ma), *FLMi*, *LLMi*, *FLMa*, and *FLMa* can be specified relative to the length of the series. Furthermore, these events can also be counted in dependence to a constant m: The number of number of Crossings of a point (CrM), $m = 0$, tallies the number of crossings of m with two successive time series steps, the first of which is lower (higher) than m and the second is higher (lower) than m [230]. In the same way, the relative (to the length) number of *peaks* of a time series can be determined. Hereby, m is equal to the support, describing a subsequence of a series where a value larger than m occurs relative to its neighbours [230, 233].

3.2 Deep Learning

As a subset of Artificial Intelligence (AI), Machine Learning (ML) specialises in indirect problem-solving [193]. Its algorithms effectively instruct a system to generalise patterns seen in sample data (examples) and locate them again on unfamiliar data. These algorithms are used in a wide range of applications, including diagnosing medical disorders [234, 235], segmenting customers [236], and optimising robot movements [193]. ML employs two fundamental learning methods: unsupervised and supervised learning.

- **Supervised learning:** The goal of supervised learning is to model the relationship (mapping) between given input ($\mathbf{X}$) and output ($\mathbf{y}$) data. To do so, the approximation error, which is the sum of the differences between the ground truth class $y_i \in y$ and the estimated class $\hat{y}_i$ of each output, is iteratively reduced by a ML algorithm. The model improves when this disparity is reduced. The process of improving the algorithm's performance is called training, which uses a training set (X_{train}, y_{train}) for tweaking parameters θ of a model $\mathcal{M}$ and measures its generalisability capabilities on a development set (X_{devel}, y_{devel}), both subsets of (X,y). After training, the performance of the trained $\mathcal{M}$ is assessed using a test set (X_{test}, y_{test}), also a subset of (X, y). Based on the prediction task, the type of ground truth varies:

  - **Classification:** The ground truth y of a classification task is represented as k discrete values, each representing a category:

    $$f : \mathbb{R}^n \to \{1, ..., k\}. \tag{3.1}$$

    For example, in a binary (k = 2) sentiment classification scenario, *"negative"* = 0 and *"positive"* = 1.

  - **Regression:** The ground truth y of a regression is a continuous real value: [193]

    $$f : \mathbb{R}^n \to \mathbb{R} \tag{3.2}$$

    This can be a sentiment regression scenario, with an intensity value in $[-1, 1]$ where *"negative"*$= -1$ and *"positive"*$= 1$. In many tasks, classification is a coarse-grained formalisation of a regression task [237].

    Furthermore, this task can be extended to a sequence of regression points when not one but multiple regression targets are predicted through time. An example is a transition of the regression sentiment value every 10 seconds, starting with a rather negative state of -0.3, changing to a neutral of 0 (10 seconds later), and ending with a rather positive final state of 0.7 after 20 seconds.

- **Unsupervised learning:** The goal of unsupervised learning is to discover structures and regularities in X. For these algorithms, y is absent from the input data so that no target can be used, and model training relies solely on X. Clustering algorithms are the most common type of unsupervised learning.

Algorithms for the two supervised tasks can measure success directly. Classification often relies on a confusion matrix using the unique labels. Each data point is classified as to whether the prediction was true positive (TP), true negative (TN), false positive (FP), or false negative (FN). From these, further metrics can be derived, most commonly:

$$Accuracy(ACC) = \frac{TP+TN}{TP+TN+FP+FN}, \tag{3.3}$$

$$precision = \frac{TP}{TP+FP}, \tag{3.4}$$

$$recall = \frac{TP}{TP+FN}, \tag{3.5}$$

$$F1-score(F1) = 2*\frac{precision*recall}{precision+recall} = \frac{2*TP}{2*TP+FP+FN}, \tag{3.6}$$

where ACC reflects the proportion of correct classifications; precision, the proportion of positive predictions that were truly positive; recall, the proportion of all positive samples that were correctly predicted as positive; and the F1 score is the summary harmonic mean of precision and recall. An unweighted measure (also called a macro calculation) first collects the metric for each label and then calculates the mean. This avoids classes with larger data points having more weight in the final result, which is especially relevant for unbalanced classes. For regression, loss is usually taken as a measure (see loss function below).

With the shift to Deep Learning, the input X has also evolved. Traditional ML algorithms, such as Support Vector Machine (SVM), frequently employ hand-crafted features for X, while DL tends to either use or produce representations, as explained in Section 3.1. Nevertheless, ANNs are also employed to use hand-crafted representations, and a distinction between DL and shallow learning is not always apparent [197].

The following section provides an overview of the most fundamental ANN (layer) types, state-of-the-art mechanisms, and architectures, along with their mathematical equations. Note that in the following equations, the bias term is dropped for clarity, and capital letter variables represent trainable parameters. The first sections also include many underlying techniques, relevant to other sections, regardless of the network architecture. Naturally, this

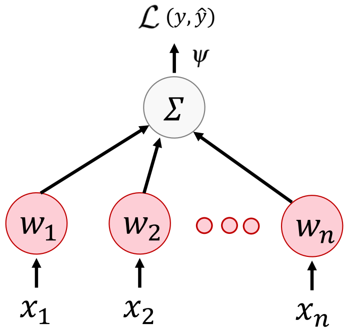


Figure 3.5: A network unit with inputs x which are weighted by w and summed to be squashed by an activation function ψ. For training, potential outputs $\hat{y}$ could be compared with ground truths y using a loss function $\mathcal{L}$.

should only be understood as a narrow insight into methods relevant for this work, given the broad spectrum that has been developed over the last few years. For a more in-depth description of the methodologies, the interested reader is referred to Goodfellow, Bengio and Courville [193].

3.2.1 Artificial Neural Networks

Neurons are the basic elementary units of an ANN. Loosely inspired by a human brain, a neuron receives many inputs $x_0,\dots,x_n$ and outputs a single value o. These inputs are weighed $w_0,\dots,w_n$ for each n inputs, added together, then passed through an activation function ψ:

$$o = \psi\left(\sum_{i=0}^{n} w_i x_i\right) \tag{3.7}$$

This concept is also called a perceptron. A Fully connected Feed-Forward Layer (FFL) is made up of many neurons which receive the same input simultaneously.

The simplest, rudimentary type of network is a Feed-Forward Neural Network (FNN), which transmits information from the input layer forward through an arbitrary number of hidden layers h [193] as depicted in Figure 3.5. Generally speaking, the FNN is an approximator of a differentiable non-linear function that defines the following mapping:

$$y = F(X;\theta), \tag{3.8}$$

with F being a ANN where the trainable parameters (weights) are indicated by θ.

To optimise these trainable parameters, the network is treated as a step-wise optimisation problem using an objective function. For this, the input has to first be propagated through the network with respect to the parameters θ in a forward pass until it reaches the last layer. There, the final outcome prediction is carried out. A **loss function** $\mathcal{L}$ calculates the error between the approximated output target $\hat{y}$ and the expected real value y. Next, backward propagation of errors uses gradient methods to compute the error gradient with respect to the parameters θ and adjusts them accordingly in the backward pass. The forward and backward pass of all data X is called an **epoch**. As explained in this section's introduction, there are several types of ground truth, and the loss function varies accordingly. Classification tasks target a set of finite categorical values predicted as the output of a pseudo-probability between 0 and 1, while regression tasks target a continuous value.

- **Cross-entropy (CE):** CE is the most prevalent **classification loss**. It increases when the predicted probability deviates from the actual label by multiplying actual probability with the log of the predicted probability for the ground truth class. Cross-entropy loss gives significant penalties when it comes to predictions that tend to be definite yet incorrect:
$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log\left(\hat{y}_i\right), \tag{3.9}$$
When it comes to multi-class classification, a distinct loss is determined for every class per observation and the result is summed up.

- **Mean Square Error (MSE):** A frequently applied **regression loss** is the MSE, measured as the average of the squared difference between predictions and real observations. It focuses on the average magnitude of the error values, notwithstanding the direction in which they occur. Predictions deviating from the exact values are clearly disfavoured owing to the square, in contrast to less deviating predictions:
$$\mathcal{L}_{MSE} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}. \tag{3.10}$$

- **Mean Absolute Error (MAE):** The MAE is similar, however, the average of sum of absolute differences is measured. In terms of mathematical characteristics, MAE differentiates in respect where gradients must be calculated using linear programming:
$$\mathcal{L}_{MAE} = \frac{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|}{n}. \tag{3.11}$$

- **Concordance Correlation Coefficient (CCC):** The CCC condenses both preciseness and accuracy of a sequence of regression points. It is often taken in the context of sequential emotion recognition as a metric for repeatability and efficiency. Further, it could come in handy as a loss function considering its resistance to changes in scale and location [238], whereas its theoretical features are comparable to those of other regression measures and losses [239]. Let o_j be a series of m ground truth labels and $\hat{o}_j$ a series of n corresponding prediction labels. The CCC is defined as follows:

$$\mathcal{L}_{CCC} = \frac{2 \times COV(\hat{o}_j, o_j)}{\sigma^2_{\hat{o}_j} + \sigma^2_{o_j} + (\mu_{\hat{o}_j} - \mu_{o_j})^2} = \frac{2 \times [(\hat{o}_j - \mu_{\hat{o}_j})(o_j - \mu_{o_j})]}{\sigma^2_{\hat{o}_j} + \sigma^2_{o_j} + (\mu_{\hat{o}_j} - \mu_{o_j})^2}, \tag{3.12}$$

where COV denotes the covariance, σ the standard deviation, and μ the mean.

The loss is reduced to obtain the global minimum of the function, irrespective of the ML task. This is accomplished through the use of gradient descent optimisation. To propagate the error of the last layer backwards through the network as gradients and adjust each parameter by computing the derivative of the fully differentiable network using the chain rule, **backpropagation** is used. These gradients are the error partial derivatives $\frac{\delta E}{\delta \theta}$ of the loss function $\mathcal{L}$ with respect to any trainable parameter θ (e. g., weights or bias) which update the parameters with regard to a learning rate (lr):

$$\theta^{i+1} = \theta^i - \alpha \frac{\delta L}{\delta \theta^i} \tag{3.13}$$

This process is performed iteratively rather than for the complete dataset. Therefore, instead of genuine gradients, a stochastic approximation is calculated on a subset of the data termed a **batch**, of a given number of randomly picked points of data, called the batch size (bs). **Adam** is a popular variant of this stochastic gradient descent approach [240]. It is computationally effective, resilient within noisy data domains that result in sparse gradients, and its initial parameters, such as lr, are autonomously modified depending on adaptive estimates, requiring minimum adjustment.

To approximate these highly complex, non-linear functions [241], the **activation function** ψ of the layers likewise has to be non-linear [242]. There are four widely adopted activation functions.

- **Rectified Linear Unit (ReLU):** Presently, the ReLU [243] is the most common activation function due to its effective computing properties, which merely involve comparison, multiplication, and maximisation (max):

$$ReLU(x) = \max(0, x). \tag{3.14}$$

It can be differentiated anywhere other than zero, where the derivative's number is picked at random.

- **Sigmoid:** For the last layer that predicts the target y, a softmax or a sigmoid function is frequently chosen. It is differentiable, saturates, and is bounded between 0 and 1 for every real-valued input. If employed individually for every output neuron, it can also be utilised for multi-label, multi-class classification:

$$sigmoid(x) = \frac{1}{1+\mathrm{e}^{-x}}. \tag{3.15}$$

- **Softmax:** Conversely, the softmax function generates a probability distribution over all neurons K throughout the final layer. The neuron with the highest value is set equivalent to the projected class:

$$softmax(x) = \frac{\mathrm{e}^{x}}{\sum_{k=1}^{K} \mathrm{e}^{x_k}} \tag{3.16}$$

- **Hyperbolic Tangent Function (*tanh*):** Another previously typical activation function is the *tanh*, which has very similar properties. Nonetheless, nowadays, it is not frequently utilised outside Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) (see Section 3.2.2) given its sluggish computation speed [242].

The ability to include additional layers and neurons in a deep and wide ANNs has led to issues with **generalisation**. Generalisation fails if the patterns observed in the finite set of training data cannot be applied to new data. This is caused by too many training epochs or when the complexity of the selected model is too high. Then the network covers the data by eventually lowering every bias and boosting the variance until it is too tightly fitted to the sample. A scenario in which non-generalisable noise and outliers are also taken into account in the decision boundary is called overfitting. The opposite, underfitting, occurs when the training error is excessively large due to a too-brief training period or a low degree of polynomial model complexity.

Since a shallow ANN is already capable of capturing an astounding amount of variance in the data, **overfitting** is generally a significant issue, and counterstrategies are usually applied during training. This procedure is called **regularisation** of the network, adjusting the training approach or model when updated to avoid overfitting.

- **Dropout:** Dropout is a simple, yet effective, and often utilised approach [244]. It prevents complex co-adaptations of neurons by performing model averaging that results in multiple independent internal representations. As a result, the network becomes less sensitive to the weights of individual neurons. To do so, at each step during training, input signals are randomly set to 0 and other inputs are scaled to keep the sum over all inputs unchanged.

- **Early stopping:** Another training strategy widely used is early stopping [245], in which the error within the validation set, e. g. , the loss or a measure, is tracked as a proxy for generalisation. Early stopping waits for a fixed amount of epochs before ending the whole training once this monitoring error grows, which would result in overfitting. The rationale for delaying multiple epochs is that occasionally, the model recovers because the optimiser slows the rate of learning, or discovers minor but novel and generalisable patterns. Alternatively, it could be because the gradient has accidentally gone into the wrong direction.

The peculiarities of the underlying data, for instance, the sequence pattern of text and audio or the grid-like architecture of pictures, are ignored by an FNN. This has prompted the development of more sophisticated network designs, which will be discussed next.

3.2.2 Recurrent Neural Network

An Recurrent Neural Network is a type of ANN that takes into account temporal dependencies, and, as a result, Recurrent Neural Networks (RNNs) are especially suitable for sequence modelling, e. g. , speech recognition and language modelling. Input sequences passed to the RNN are processed one at each time step t, while maintaining in their hidden units a “state vector” h that implicitly contains the passed-through information of the sequence’s previous steps, thus incorporating previous states at each step [193, 246]. This can be formalised as:

$$h_t = \psi(h_{t-1}, x_t; \theta), \tag{3.17}$$

where for a given time step t, an input x_t, the previous hidden state h_{t-1}, and non-linear function ψ are shown. Figure 3.6 depicts the internal dynamics of an RNN. It also maintains the input order while exchanging parameters from previous time steps via functioning in an auto-regressive way, absorbing information from the preceding step together with the current step.

When the causal structure of sequence data is of minor relevance, information from the beginning to end of the sequence, as well as from the end to the beginning, could be fed into such network layers. For this reason, **bidirectional** RNNs were invented to achieve this goal. One direction forwards information across the network in a chronological order, and the other utilises the information in a backwards fashion.

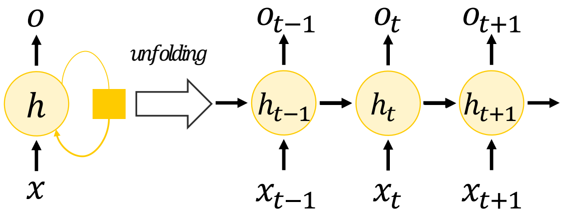


Figure 3.6: RNNs processing information from x through time by incorporating it into the state h, while o reflects the output, which can be squashed by a softmax function to unnormalised log probabilities. The loop indicates a delay of a single timestamp. The right side shows the unfolded computation graph. Figure adapted from [193].

As with FNNs, sequential models are trained with backpropagation. Due to the order dependency of the internal state on the previous inputs for each timestamp, an adapted version, **backpropagation through time**, has to be used. It differs insofar as the network has to be enrolled along the time axis for weight updating. Based on a pair of sequences for the current input and the corresponding output of a timestamp, the errors can be accumulated and the weights updated. This is repeated until all timestamps are updated. Since several iterations are required for a single update of the weights, the required computing power increases with increasing number of time steps.

In addition to the expensive computation, unwanted effects can occur while training RNNs when dealing with **long sequences**. The primary difficulties are vanishing and exploding gradients [247]. Exploding gradients refer to situations where the gradients grow excessively, causing unstable learning and highly fluctuating weights. On the other hand, vanishing gradients become extremely tiny, inhibiting the network from either learning completely or causing it to learn extremely slowly. To prevent the gradients growing out of control, a technique known as gradient clipping was developed, in which the gradients are rescaled when they reach a specified threshold.

Long Short-Term Memory Recurrent Neural Network (LSTM-RNN): To further overcome vanishing gradients, a new type of cell — the Long Short-Term Memory (LSTM) cells — advancing the previous hidden logic, are incorporated into an RNN layer to form an LSTM-RNN [85]. It has additional self-loops to create routes in which the gradient can be flowed over extended sequence lengths, extending the RNN's ability to preserve reoccurring patterns longer. The self-loops are implemented in a way that information can be removed or added to a cell state, regulated by structures known as gates. Gates consist of a pointwise multiplication operation together with a sigmoid bounding function. This logic is illustrated in Figure 3.7. On the right side, the inner structure of such an LSTM-RNN state is depicted. There are three types of gates:

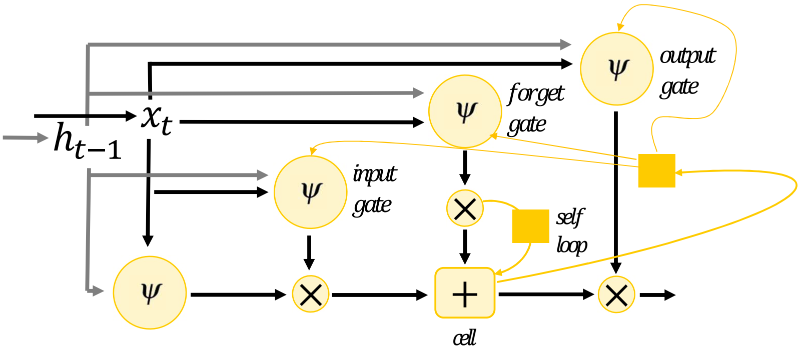


Figure 3.7: LSTM recurrent cell. The input feature is gated by a sigmoidal input gate. The state unit has a linear self-loop, where the value can be event-based "forgotten" through a weight controlled by a sigmoidal forget gate. Finally, the state gets squashed by a tanh function and the output gate regulates by a sigmoidal output gate. Figure adapted from [193].

the input gate, forget gate, and the output gate. The input gate specifies which of the new information needs to be added to the cell state, whereas the forget gate decides what information is deleted. The output gate, in the end, decides which information is valuable from the context and outputs it.

3.2.3 Convolutional Neural Network

In computer vision, the handling of extremely wide, correlated inputs is required, which FNN cannot manage efficiently [248]. For that reason, the visual context apparatus became the architectural muse of the CNN. A simple example is illustrated in Figure 3.8. Hierarchical topologies, employed by CNNs, extract spatial relationships efficiently from separated features, such as for object and facial recognition from images. Key components of this architecture are the pooling layer and convolution layer. They lead to sparse feature maps for detecting intricate patterns. In recent years, this powerful ability to efficiently condense large, correlated inputs into sparse feature maps has also found application in numerous other DL disciplines, e. g. , processing of raw audio [89].

A typical design can employ either one-, two-, or three-dimensional inputs. In the subsequent formalisation, an image source I is assumed, with the dimensions of height n_H, width n_W, and channels n_C. Normally, RGB pictures are used, leading to $n_C = 3$. The number of filters per layer is specified as l, while K filters are trained during the convolutional operation. Given an image which is convolved across the n_H and n_W dimensions, a convolutional product $\mathrm{CV}_{\mathrm{op}}$ is carried out between the receptive input field and the filters, where each matrix element is the sum of the element-wise multiplication:

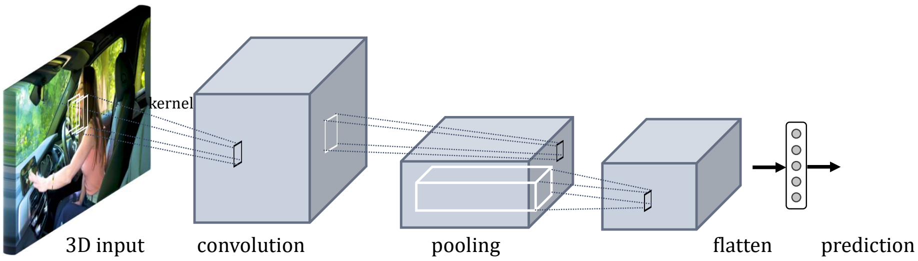


Figure 3.8: Basic functionality of a CNN including convolution and maximum pooling operations. Figure adapted from [193].

$$\mathrm{CV_{op}}(I,K)_{x,y} = \sum_{i=1}^{n_H}\sum_{j=1}^{n_W}\sum_{k=1}^{n_C} K_{i,j,k} I_{x+i-1,y+j-1,k}. \tag{3.18}$$

This results in the dimensions:

$$d = \left(\left\lfloor \frac{n_H + 2p - fs}{s} + 1 \right\rfloor, \left\lfloor \frac{n_W + 2p - fs}{s} + 1 \right\rfloor \right); \quad s > 0, \tag{3.19}$$

with the filter size (fs), strides (s) steps taken, and p, which specifies the padding type, so that a valid convolution that does not use padding leads to $p = 0$, and a same convolution pads the input matrix to ensure that the outputs have the same shape ($p = \frac{f-1}{2}$). In general, a small stride increases the size of the output, and conversely. A complete convolutional layer performs the convolutional process with several trainable filters $\left(f^{[l]} \times f^{[l]} \times n_C^{[l-1]}\right) \times n_C^{[l]}$ together with a broadcasted bias $b = (1 \times 1 \times 1) \times n_C^{[l]}$, accompanied by an activation function ψ that is frequently a ReLU function (see Section 3.2.1) or one of its derivatives:

$$\mathrm{CV_L}\left(a^{[l-1]}, K^{(n)}\right)_{x,y} = \psi^{[l]}\left(\mathrm{CV_{op}}(a^{[l-1]}, K^{(n)}) + b_n^{[l]}\right), \tag{3.20}$$

where $a^{[0]}$ indicates the input picture. The convolutional step is often followed by a pooling step, which downsamples the input by merging the outputs of neuron clusters from one CNN layer to a single neuron, most often by selecting either the maximum or mean value of the neuron cluster. These reductions also enhance shift invariance and can be expressed as:

$$\mathrm{POOL}\left(a^{[l-1]}\right)_{x,y,z} = \lambda^{[l]}\left(a^{[l-1]}_{x+i-1,y+j-1,z}\right); \quad (i,j) \in \left[1,2,\ldots,f^{[l]}\right]^2, \tag{3.21}$$

where λ is an average or maximum pooling operation.

These layers can be combined in a variety of ways to create blocks, which then get replicated identically and are arranged in a series for a deep feature extraction. When feature extraction by the CNN blocks is completed, the features are flattened to a data-driven representation and passed to a FFL as described before.

3.2.4 Attention Mechanisms

Attention can be seen as a building block capable of enhancing the internal network representations of each of the previously presented ANN architectures in different ways. Originally, it was introduced to support the memorialisation of long source sentences in Sequence to Sequence (S2S) models, common in language processing [249]. Similar to the LSTM-RNN cells, it strengthens the capabilities of modelling the context by focusing on one time step in respect to the surrounding ones. Intuitively, this is similar to a human focusing attention on relevant parts of a visual scene depending on the context [250]. This comparison shows the closeness to the computer vision domain [63, 251, 252]. Over the past years, several manifestations of this mechanism were found to generally improve representations, and hence it became a predominant concept in various domains [253, 254]. Furthermore, it is commonly used when aligning and fusing modalities in multimodal emotion recognition [101], as well as aligning them across modalities [254].

In the following, several concepts of attention are introduced that are relevant to this work and describe the slight differences between the various mechanisms. All have in common that the overall aim of the attention mechanism is to perform a linearly weighted sum of a sequence of input vectors. Hereby, the network should learn the optimal weights itself through a scoring function, and the resulting output derived from the input vectors should not be scaled in any manner.

3.2.4.1 Context Attention

Let us assume a typical scenario in neural machine translation where a sentence of a language A must be transformed into a sentence of a language B [255]. Formally speaking, the input sequence A must be encoded to a hidden representation, which is then decoded to the output sequence B. In this setting, the encoder and decoder are often RNNs [256]. Figure 3.9 depicts a comparison of an encoder-decoder architecture without (left) and with context- attention mechanism (right). Through compression, an information bottleneck naturally occurs in this transition between the input encoder and output decoder, causing the network to forget information of the input sequence when decoding. As a result, additive context attention [249] was created to get around the bottleneck through employing a weight distribution upon the input A that is primarily connected to a decoder output step of B. This can as well be thought of as a learned alignment function between two sequences. Due to this property,

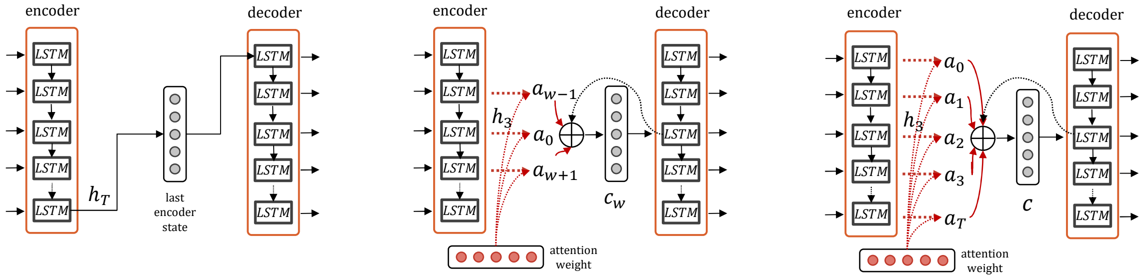


Figure 3.9: Encoder-decoder architecture with traditional information bottleneck after encoding (left), with local attention setting the window = 1 in each direction (middle), and, finally, context-attention mechanism, considering the entire encoding sequence in dependency to a decoding state (right). Figure adapted from [249].

the method is frequently applied to weight steps in sequence modelling to enhance the representation [193] as is done for textual question-answering [257]. For the formalisation of the additive attention, introduced in this context, let h_t represent the latent state of a network at time t with $h_t \in \mathbb{R}^F$ as the input to the attention layer. The input is transformed with one FFL sharing the weights W_a and b_a in the attention layer. Furthermore, a *tanh* activation introduces non-linearity:

$$u_t = \tanh\left(W_a h_t + b_a\right). \tag{3.22}$$

This score function can vary, as will be shown in the course of this section. To ensure the resulting score is still normalised, a softmax activation function is usually used, bounding the resulting values to a $[0,1]$ range so that the sum is equal to 1:

$$\alpha_t = \text{softmax}(u_t) = \frac{exp(u_t)}{\sum_{t=1}^{T} exp(u_t)}. \tag{3.23}$$

Finally, the input is weighted through the attention weights derived from Equation 3.23, allowing the network to attend certain steps and resulting in the context representation c:

$$c = \sum_{t=1}^{T_a} \alpha_t h_t. \tag{3.24}$$

3.2.4.2 Window Attention

Window or local attention is a version of context attention in which the dependencies are confined within a local attention window as depicted in Figure 3.9 (middle). Based on additive attention [258], an attention matrix L is introduced in the scoring operations [204] similar to [259]. It aims to limit the context of the decoder states g to the steps to a nonparametric window w. For this, it converts the encoder's hidden state representation h_t into a positioned hidden state g:

$$\alpha_{t,t'} = \text{softmax}((W_a g_{t,t'}) * L), \tag{3.25}$$

where the weight matrices of the hidden states are element-wise multiplied with L, a matrix of the shape [batch size, time steps, time steps]. The matrix comprises binary values of the chosen window size w across all time steps of a sequence. Thereby, w either takes into consideration the $w-1$ past steps or the previous and past $w/2$ steps. This step is followed by the usual normalisation procedure of the weighted summation to receive the hidden state representation of the current token.

3.2.4.3 Self-Attention

Vaswani et al. [91] introduced self-attention to simplify the attention concept for non-RNN encoder-decoder architectures. Self-attention makes attention more generic, in the sense that it aims to improve the representation of the input sequence independently of a potential decoder. As a result, it is also applicable for a larger set of task types. This is particularly relevant when dealing with extremely lengthy sequences. Rather than finding dependencies between the decoder and encoder, self-attention may be utilised to detect dependencies between various places in an identical sequence. Hereby, the energy score is not dependent on a previous decoder hidden state s_{j-1} (Equation 3.23). Furthermore, the additive attention is replaced by scaled dot-product attention which assigns the same vector to the query, key, and value. While looking at Equation 3.26, it can be seen that each input x_t appears thrice but is differently weighted:

- *Query (Q)*: In comparison with every other vector to calculate its own y_t,
- *Key (K)*: In comparison with every other vector to calculate the m-th output vector y_m,
- *Value (V)*: Once for each output vector in the weighted sum.

Thus, x_t is linearly transformed with trainable weights, each to allow the network to distinguish the inputs and learn individual linear transformations:

$$\begin{aligned} q_t &= W_q x_t, \\ k_t &= W_k x_t, \\ v_t &= W_v x_t. \end{aligned} \tag{3.26}$$

Equation 3.23 and Equation 3.26 can be rewritten to:

$$\alpha(\text{Q, K, V}) = \text{softmax}\left(QK^T/\sqrt{d_{key}}\right)V. \tag{3.27}$$

It is scaled by a factor of $\frac{1}{\sqrt{d_{key}}}$, where d_{key} is the dimension of the key. By doing so, vanishing gradients are prevented through limiting the expansion of the dot product.

3.2.4.4 Multihead Attention

Because self-attention merely estimates the weighted average of the tokens, it cannot distinguish the semantic difference of surrounding tokens. For this, multiple self-attention heads can be coupled, in a structure called the Multihead Attention Layer (MHAL), for learning the distinction between the surrounding tokens [91]. Specifically, multihead attention is used to obtain more meaningful sequence representations s_t, with T being the maximum number of steps. The multihead softmax dot-product attentions (Equation 3.27) can thus be computed in parallel. This allows for a block to jointly pay attention to information from several representation subspaces in each head. A head receives three linear projection inputs multiplied by an individual trainable query, key, and value weight matrix W_Q, W_K, W_V, wherein the division by $\sqrt{d_k}$ prevents the gradient from becoming minuscule. After scaling, the results of the individual heads are concatenated and fed into a subsequent linear layer W_S:

$$\text{MultiHead} = s_t = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_{at}\right) W_S, \tag{3.28}$$

$$\text{head}_j = \alpha\left(QW_Q, KW_K, VW_V\right), \tag{3.29}$$

where j is a single head with a maximum of at number of heads. Figure 3.10 depicts the parallel nature of multihead attention. Performing such attention functions in parallel is very efficient, with a time complexity of order $\mathcal{O}(T^2 \cdot F)$ compared to $\mathcal{O}(T \cdot F^2)$ for RNN (run-time $\mathcal{O}(1)$ vs $\mathcal{O}(T)$). This is highly relevant when dealing with large training corpora, vast representation vector sizes, and complicated sequences [91]. Multi-attention heads are the key building block of Transformer networks.

3.2.5 Transformer

As explained in Section 3.2.2, there are several issues with long sequences when training sequence architectures, even when the context-attention mechanism is employed. Since the information is propagated in a specific order, e. g. , the cell C_{t-1} has to be updated before cell

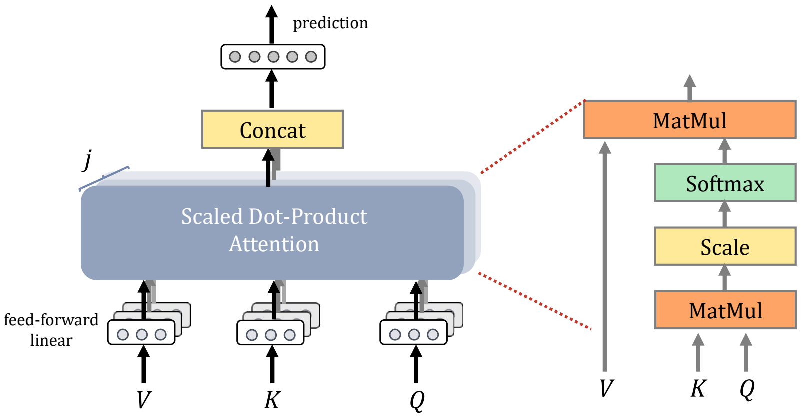


Figure 3.10: Multihead attention. Value, key, and query (V, K, Q) are processed for each head j for *at* attention heads in parallel. Figure adapted from [91].

C_t, parallel computation is not feasible. This prompted the development of the Transformer architecture, as depicted in Figure 3.11 [91], which can be trained in parallel even while dealing with sequences. It was originally used only for NLP, but it is now finding ever more applications in other domains, for example, computer vision [260] and multimodal research [101, 183]. Transformers are built on the foundation of two distinct layers:

- **Multihead self-attention:** As explained in Section 3.2.4.4, these layers capture long-range sequence patterns using a multihead attention mechanism [91].

- **Position-wise FFL:** This is a block of linear layers with dropout which applies itself to each position of the multihead output in parallel while sharing weights, so that the length of the source sequence does not matter.

Furthermore, since the learning signal is not propagated backwards through the time sequence, a spatial encoding is added to the input to indicate the position of every sequenced step.

In addition to the position-wise FFL, a Transformer block consists of layer normalisation and the multihead attention layer. N Transformer blocks are linked in sequence within the network, each having *at* attention heads. Furthermore, it can be built in an encoder and decoder design by stacking multiple blocks independently together.

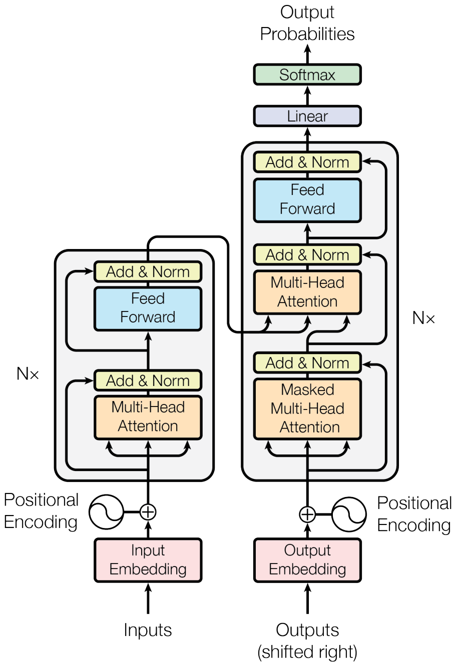


Figure 3.11: A detailed view of the plain Transformer encoder-decoder architecture. Figure taken from [91].

4 Data

In Section 2.1, it is pointed out that no previously existing dataset can be applied to answer the research questions **RQ-1, RQ-2**, and **RQ-3**. In Stappen et al. [24] the MuSe-CaR is introduced to facilitate research in Multimodal Sentiment Analysis (MSA) beyond discrete modelling, which is introduced in this chapter. It seeks to address the deficits of existing datasets identified in Sections 2.2 and 2.3 and can be summarised as follows:

- The covered videos must be unconstrained in respect to recording settings and emotions. This is also valid for the linguistic use; however, some overlapping throughout the topics and aspects is necessary to enable both unsupervised and supervised algorithm development.

- Instead of examining short sentences or segments in isolation, it should be possible to build up a continuous understanding of the context.

- The presenter's emotion-object interaction should take place in a variety of settings, such as inside and outside a car, to introduce more setting variety.

- Emotions should be inherently elicited based on the subject and situation, and the represented person should not be deliberately enacting them.

- The audio-video content should include both spoken and behavioural data. Furthermore, there should be times when one modality only delivers a finite amount of data, e. g. , the face appears while the conversion from speech to text fails.

Besides being the subject of numerous recent studies [261–263], MuSe-CaR is also the centrepiece database for this thesis. In the following section, the methodology of the collection is discussed first. Next, the annotation strategies are introduced, wherein human raters provide the prediction targets needed to train ML methods (see Section 3.2). The resulting individual annotations of the raters must be merged into a single target annotation. In Stappen et al. [25], a novel framework is presented that provides a wide range of state-of-the-art methods of the emotional gold standard and their transformation to discrete classes, which are described in Section 4.3.

4.1 Core Database

MuSe-CaR enables the development of MSA algorithms, including but not limited to the tasks of continuous emotion recognition (see Section 5.1), emotion-target engagement (see Section 5.1.2 and Section 5.2.2), and trustworthiness recognition (see Section 5.1.3), by means of comprehensively integrating the audio-visual and language modalities. It is more than three times larger than any other continuously annotated dataset in the field (see Section 2.1 and Table 2.1). Additionally, it offers novel annotations, specifically allowing modelling of the interaction between speaker topics, objects of interest, and emotions. The data is publicly available and served as the testing bed for the 1st and 2nd Multimodal Sentiment Analysis Challenge (MuSe 2020/2021) [26]. As in related studies [59, 126], the dataset videos are acquired in a semi-automatic fashion by feeding manually selected keywords into an automatic YouTube video crawler.

Acquisition: Contextual interaction with emotions, e. g., towards physical entities and topics, can have an infinite number of manifestations. The dataset aims to enable unsupervised as well as supervised learning, e. g., unsupervised extraction (see Section 5.2.1) and supervised classification (see Section 5.2.2) of topics. For this, it must be viable to label the videos and divide them into training, validation, and test sets, all covering a minimum (optimally balanced) number of classes. To ensure this and that every class occurs at least several times, a certain degree of content consistency is necessary. For this reason, the coverage of the content considered is limited to vehicle reviews.

In addition, many previous works neglect legal considerations or refer to the fair use principle for academic use [11, 34, 58] when crawling data from the internet. The legal situation for gathering data from internet crawling appears inconclusive in many jurisdictions[1].

Licence: Following the principles of an opt-in approach, the creators of videos with high user engagement (views, likes, etc.) were contacted to solicit their consent to use the videos for scientific purposes. This allows researchers — regardless of the legal sphere — to access the database without obtaining additional legal clearance. Over a period of three months, up to three (follow-up) emails were automatically sent to the creators

[1] Posting a video on YouTube immediately issues a YouTube licence. With regard to this licence, the data can only be used by YouTube directly or with the creator's permission. In the US, as an exception to intellectual property rights, research can refer to the principle of fair use; however, this does not seem to apply in the EU. In addition, both YouTube's general as well as application programming interface (API)'s terms and conditions have to be considered. Alternatively, videos under the Creative Commons licence only (CC-BY, full use if the creator is acknowledged) could be used, but would drastically reduce the amount of available data.

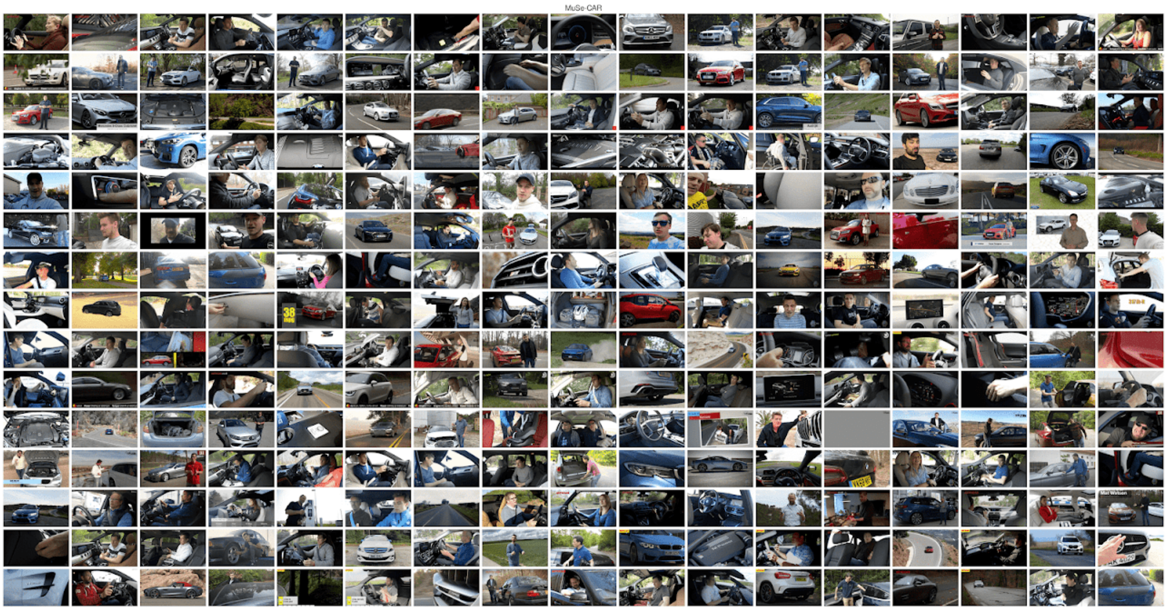


Figure 4.1: Thumbnails showing reviewers in various constellations to the camera and interacting with the object. Figure taken from Stappen et al. [24].

inviting them to agree to the End User Licence Agreement (EULA). In this process, an agreement was reached with almost half of the creators contacted, whose content were then be incorporated into the data set.

In absolute terms, this corresponds to 366 videos for which permission was granted. Example thumbnails are depicted in Figure 4.1. However, as is customary [11, 34, 59, 126], these acquired videos are subject to further assessment of their value in answering the research questions posed. To do this, three people inspected each video, looking at about 10 % of the content. The inspectors were asked to rate relevant properties (see Section 2.1), such as the degree of in-the-wild characteristics, emotionality, and video quality, in a survey.

Content characteristics: In summary, about 80 % of the videos were emotional or very emotional, and 85 % had good or very suitable video quality for the purpose of these experiments. The survey also supported the goal to quantify and balance uncontrollable in-the-wild influences and confine properties to enable generalisation for leveraging state-of-the-art DL methods. These in-the-wild features of MuSe-CaR include:

- **Video**: shot distance, face angle, occlusion, and camera angles in a wide range of scenes within a single video.
- **Audio**: ambient background sounds, dialects and accents, different microphone equipment, and speaker locations.

- **Text**: colloquialisms and terminology unique to the domain.

As an example, it was found that the shooting distances change multiple times in 80 % of the videos; that reviewers interact more often with parts inside than outside the car; that inside the car, the face is typically filmed from a slightly lower angle; and that on average less than 10 % of the duration of an average video is judged to be of poor sound quality.

Regarding the reviewers depicted in the videos, a predominant share appear to be (semi-)professional reviewers (e. g. , influencers). As a result, they possess more sophisticated equipment than casual amateur reviewers, which improves video quality and consistency of the topics covered. As the cohort of crawled videos was not actively preselected, the biographical characteristics of the reviewers must be estimated. All subjects in the dataset have a Caucasian appearance and fall roughly within an age range of about mid 20s to late 50s. Around a third of them are female, and approximately 30 % of the reviewers wear glasses. Most speakers are from the United Kingdom or the United States of America, and only a small number appear to be non-native but fluent English speakers. Although both countries have their accents, their prominence seems only faint in most of the recordings.

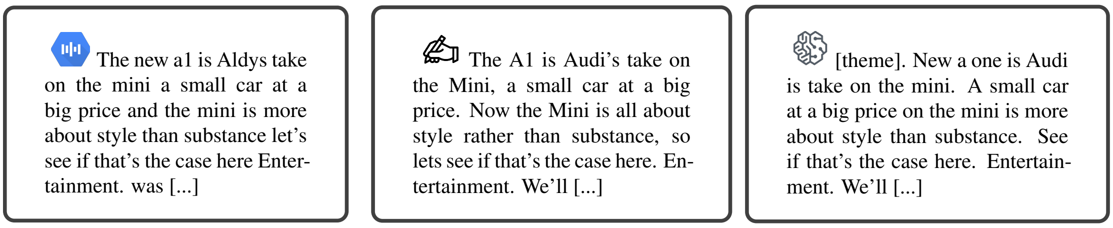
A comprehensive analysis of the characteristics described above can be found in Stappen et al. [24].

It is well known that spoken language information facilitates extracting content aspects and enhances understanding of emotions' meaning in a given context [86, 237, 264, 265]. Therefore, the spoken word in the videos is transcribed to support future studies on the interaction of speech, visual, and linguistic modalities. For this task, automatic speech-to-text systems are relied upon instead of human ones for two reasons. First, if linguistic representations are used for the real-time MSA in future in-the-wild situations, text capturing must work in the background without human intervention. Second, to develop methods for such environments, large, even very large, datasets are needed, making human transcripts prohibitive. This is reinforced by the rapid improvements of speech-to-text systems. In 2016, Microsoft researchers claimed to have reached human parity in this task on several English corpora [266].

Automatic transcripts: Two of the most reliable transcript services are Google Cloud's speech API[2] and Amazon Transcribe[3]. Both systems were applied to the videos, since both (individually or in combination) can be interesting for future research purposes.

[2] https://cloud.google.com/speech-to-text

[3] https://aws.amazon.com/transcribe/



The new a1 is Aldys take on the mini a small car at a big price and the mini is more about style than substance let's see if that's the case here Entertainment. was [...]

The A1 is Audi's take on the Mini, a small car at a big price. Now the Mini is all about style rather than substance, so lets see if that's the case here. Entertainment. We'll [...]

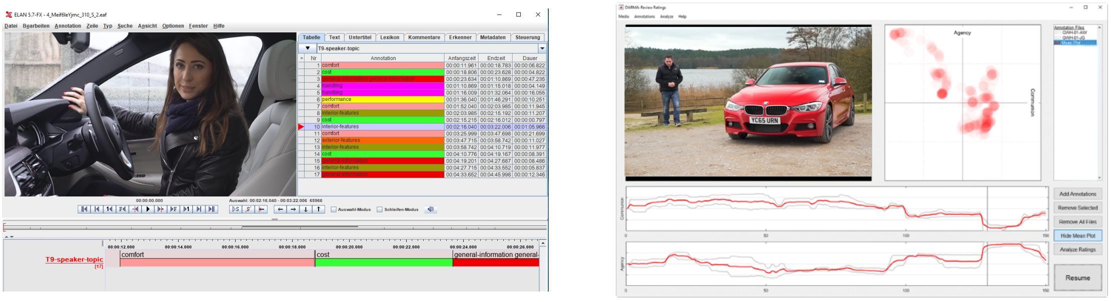
[theme]. New a one is Audi is take on the mini. A small car at a big price on the mini is more about style than substance. See if that's the case here. Entertainment. We'll [...]

(a) Google Speech-to-text (b) Human transcribed (c) Amazon Transcribe

Figure 4.2: Example snippet out of the selection of videos (id: 265) having the worst Word Error Rate (WER) compared to the hand-transcribed (b). (a) Google Speech-to-text reaches 37.85 % WER and (b) Amazon Transcribe 39.44 % WER. Figure taken from Stappen et al. [24].

By randomly selecting and transcribing 10 videos (10 576 words in total) by humans, the deviation between the automatic and manual transcripts was calculated. For Google's Speech-to-text a Word Error Rate (WER) of 25.04 % and for Amazon 28.39 % was found (lower is better). The worst videos had WER of up to 37.85 % for Google and 39.44 % for Amazon, displayed in Figure 4.2. Although this WER may seem high at first, many errors turn out to be small and not content-changing, for example "A1 is Audi's" (hand) vs "a1 is Aldys" (Google) vs "a one is Audi" (Amazon).

Google's Speech-to-text also incorporates non-verbal cues and auditory components, such as laughing and singing tags. However, the experiments in this work are based exclusively on the Amazon transcripts, as these can be fine-tuned with an individual dictionary. In random checks, the Amazon transcripts were found to produce superior results than Google's regarding the corpus's automobile domain-specific terminology. In addition, the service delivers punctuation and the beginning and ending timestamps for each word, allowing accurate alignment with extractions from other modalities that may be sampled at different rates. Previous work often had to use forced alignment [11, 13, 100] based on voice activity detectors. In forced alignment, word boundary detection can fail to recover after making an alignment error, and thus is more prone to errors for long sequences [267]. In addition, those methods usually rely on human-made transcripts, which would be an additional challenge in this setting. In total, the transcription yielded 28 295 sentences, surpassing the largest English-language MSA database to date, MOSEI (see Section 2.3 and Table 2.1), by almost 5 thousand records.



(a) Eudico Language Annotation Tool (ELAN) illustrating the segment annotations of the speaker topics. Figure taken from Stappen et al. [24].

(b) Dual Axis Rating and Media Annotation Software (DARMA) illustrating the annotation of time- and value-continuous arousal (lower) and valence (upper) traces.

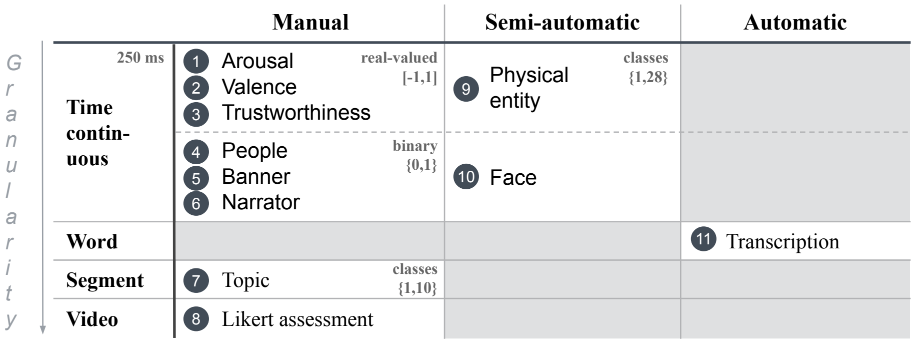
Figure 4.3: Screenshots of the (a) segment-level ELAN and (b) value- and time-continuous ELAN annotation tools utilised to annotate MuSe-CaR (id: 4).

4.2 Annotation

Annotation is a repetitive, time-consuming process that must be managed carefully to ensure a meaningful, high-quality, and ethical outcome [268]. This section first outlines the organisational structure, tasks, and software for annotating MuSe-CaR, adapted from comparable work on new databases [12, 43] to ensure an efficient process in terms of the defined criteria in the beginning of this section. Then, the focus moves to the definition of the annotations, derived from Section 2.1.

- **Functional roles:** The **annotator** labels data based on the understanding gained from the annotator protocol, the training sessions, and the audit feedback. The **auditor** reviews the annotators' qualitative and quantitative (e. g. , similarity compared to the mean of other annotators) performance and provides feedback. The **administrator** serves as an overseer for the entire annotation process.

- **Tasks:** Annotation tasks are **assigned** sequentially. Each focuses on one particular annotation type (e. g. , valence) comprising several videos equal to one hour of working including approximately 40 minutes of video content and 20 minutes of breaks in between tasks [43]. The bundle is sent to the auditor after completion. A new session is only assigned if the auditor confirms the old one. Every annotation has to be **reviewed, validated, and approved** by at least one auditor. Approximately 10 % of the annotated videos were found unsatisfactory and had to be annotated again. **Progress monitoring** is frequently performed by the administrator. By tracking the annotator's performance consistently, the **quality** of the output can be constantly improved.



Granularity	Manual	Semi-automatic	Automatic
250 ms Time contin-uous	1 Arousal 2 Valence 3 Trustworthiness real-valued [-1,1]	9 Physical entity classes {1,28}	
	4 People 5 Banner 6 Narrator binary {0,1}	10 Face	
Word			11 Transcription
Segment	7 Topic classes {1,10}		
Video	8 Likert assessment		

Figure 4.4: Overview of the annotation granularity (sampling rate) of all annotations: 1-3: continuous real-valued annotations; 4-6: continuous binary annotations; 7: speaker topic segments; 8: Likert assessment; 9: physical car entity classes; 10: faces; 11: automatic transcripts. Figure taken from Stappen et al. [24].

- **Software and hardware:** On the one hand, studies have found that emotions such as anger are conveyed more intensely via audio, whereas others, such as sadness, by visual signals [269]. On the other hand, context information is expressed by both modalities [270], making the need for multimodal annotation tools evident. Annotation of the entire video was sampled at 0.25 Hz with an axis magnitude between -1 000 and 1 000. Categorical annotations of video sections are done using the program ELAN 4.9.4 [271], which provides the audio wave and a video stream allowing the annotator to find exact start and endpoints (see Figure 4.3a). DARMA [272] in combination with a **Logitech Extreme 3D Pro Joystick** enabled the annotators to record perceived emotions in very intuitive way [12] as depicted in Figure 4.3b.

With almost 40 hours of user-generated video material, the continuous annotation of the three subjective dimensions (arousal, valence, trustworthiness) alone required more than 600 hours of human labour work.[4] The annotators consisted of 11 employees of the Chair for Embedded Intelligence for Health Care and Wellbeing of the University of Augsburg (six female and five male), all fluent in English. Each video dimension was annotated five times.

Next, the annotations that are of primary interest in this thesis' experiments (see Sections 5.1 and 5.2) are explained. An overview of the manual and semi-automatic annotation can be found in Figure 4.4. Whereas with manual annotations the entire data set is annotated exclusively by humans, this is only partially the case in semi-automatic procedures. Here, either an automatic annotation (e. g. , face extraction) is tested on a labelled subset of MuSe-CaR to estimate its quality, or the labelled subset is used to train or fine-tune an algorithm that subsequently annotates the data automatically (e. g. , physical entities). For a detailed

[4] Calculation: minimum of 5 independent annotators per video * 3 continuous annotation tiers * 40 hours of video = 600 hours, plus an additional 33 % of paid rests between annotations.

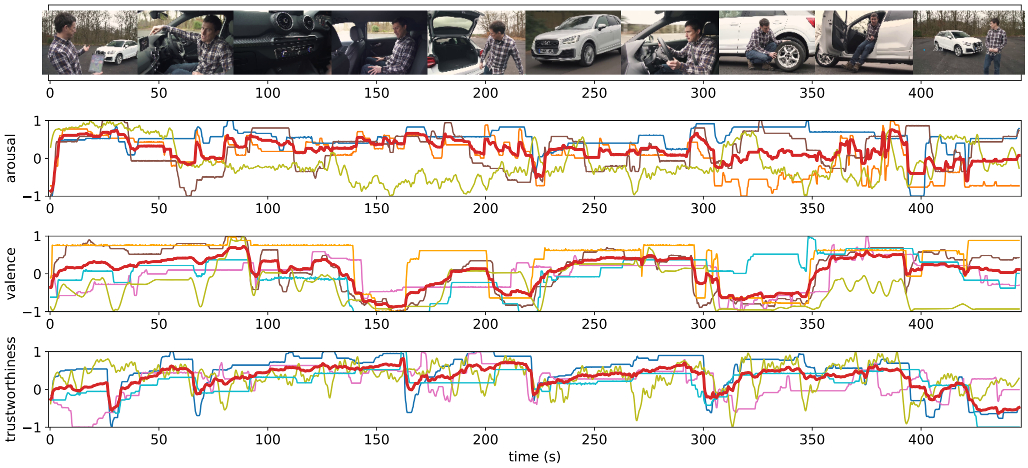


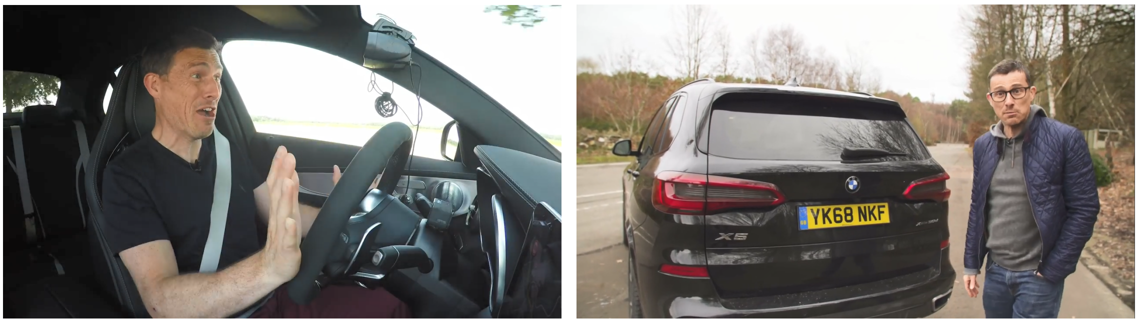
Figure 4.5: The three continuous, real-valued arousal, valence, and trustworthiness dimensions for an example video (id: 236), visualising the raw annotations and the gold standard Evaluator Weighted Estimator (EWE) (bold, red) with a sample rate of 250ms. Figure taken from Stappen et al. [24].

description of all the different layers, the interested reader is referred to the prior publication Stappen et al. [24].

The foundation for this work's derived understanding of MSA, as explained in Section 2.2, is the **manual** annotation of emotions in a **time- and value-continuous** fashion as depicted in Figure 4.5.

- **Arousal and valence:** The arousal and valence dimensions originated in the Circumplex Model of Affect (CA) theory [16] and are currently the most applied theoretical concept for time-continuous emotion recognition (see Section 2.2). Two example situations from the visual perspective can be found in Figure 4.6. As recommended in previous studies [12], the dimensional annotations are each annotated one at a time. In the MuSe-CaR data, examples of high arousal are stressful and elated (happy) situations; the first has a negative valence, while the latter is positively connoted.

- **Trustworthiness:** Given the social media setting of the dataset, it relies on the assumption that an individual (the video moderator) can objectively evaluate the matter and communicate their judgement with integrity, thus truthfully. Based on this conceptual definition (Section 2.2), the annotators rated it from a personal perspective, asking how truthful and informative the review is perceived over the duration of the video, for instance, covering specific product aspect addressed by the host. A negative impression could be that the host seeks to make a commercial profit rather than a genuine analysis of the product. Of course, similar to emotions, this may be a subjective standard for



(a) High arousal

(b) Positive valence

Figure 4.6: Excerpts of emotional scenes from the MuSe-CaR dataset. (a) High arousal triggered by the fast acceleration of the car, expressed by verbal and nonverbal cues. (b) Positive surprise at the beefy rear is expressed non-verbally. Figure taken from Stappen et al. [24].

annotation. Therefore, the annotators were given multiple video samples and cases to deepen their common understanding, and as with valence and arousal, five independent raters annotated this dimension.

To generally familiarise annotators with the concept, they first received instruction from a 15-minute explanatory video on the aspects and interpretations of CA dimensions. This was followed by an in-person training session. To learn to translate the theoretical understanding to high-quality annotations, training annotations were done individually and closely monitored. This first-hand experience is crucial to understand the functionality and reaction times of the joystick. The resulting annotations were compared and discussed in the training's group (max. five members) as well as benchmarked to a pre-recorded expert annotation. To maximise concentration, the annotation process after training was carried out alone in a quiet atmosphere and with headphones.
Categorical labels are also of particular interest for MSA.

- **Speaker topics:** These are concepts surrounding a generalised, high-level area of interest articulated by the moderator. These overarching themes under one definition can address several aspects within one labelled segment of a video. Therefore, they can consist of one or many sentences (see Figure 4.3a). Figure 4.7 provides a comprehensive overview of the subtopics and aspects covered by each subject and illustrates the distribution of topics across all sentences.

- **Physical entity:** The core modality of videos is vision in the form of an image stream. Automatically observing how a person of interest interacts with physical entities of the surrounding world can be seen as another source for deeper context understanding.

To manually label an appropriate amount of frames with bounding boxes would require an estimated time between 4 800 and 72 000 hours for a single annotator.[5] With this in mind, a **semi-automatic** annotation process was chosen. For this, a subset of MuSe-CaR was created containing frames across all channels and speaker topics. Multimodal Sentiment Analysis in Car Part Frames (MuSe-CaR-Part) is a selection of 1 000 frames extracted and labelled from the MuSe-CaR videos. It is annotated with more than 8 000 bounding boxes by three human annotators. Due to the image origin from video recordings, the in-the-wild characteristics are complex. For example, the camera is in motion and shot sizes are changing, resulting in a portion of the images being blurred. Slightly blurred frames were kept but others were removed, ending up with 1 124 frames depicting the 27 cars' interior and exterior parts mostly in use by humans, resulting in 6 146 labels in total with an average of 5.47 labels per frame. These images are used to train GoCaRD, as explained in depth in Section 3.1.3.

The fully trained model is applied to MuSe-CaR, reaching 41.07 % mAP on the labelled extract. The entire process is explained at length in Stappen et al. [2].

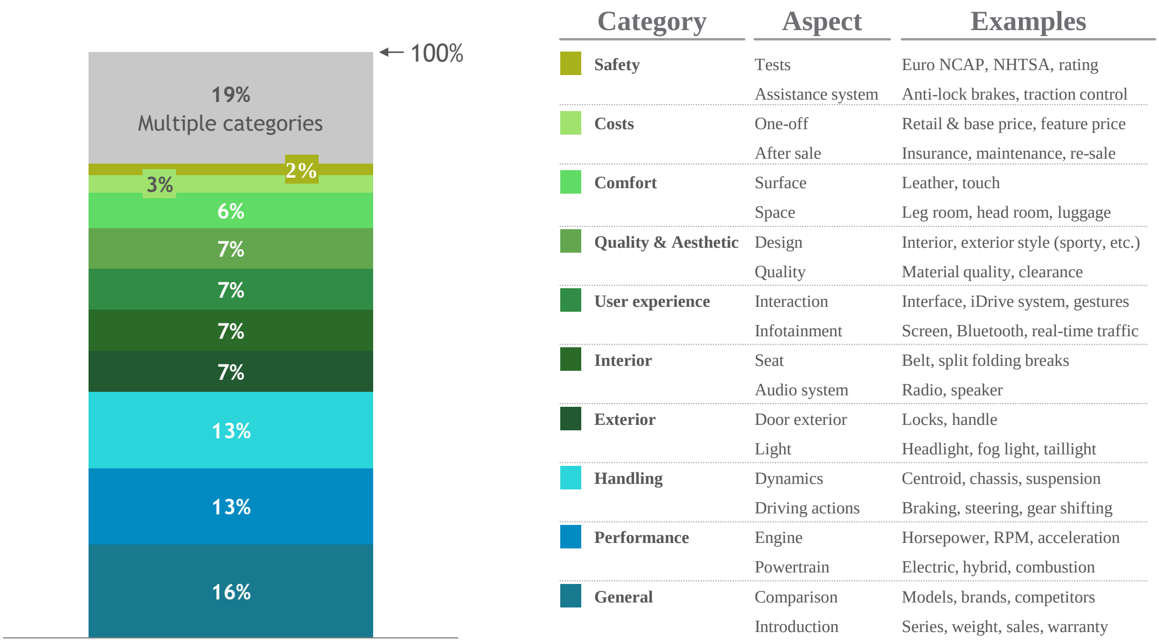
- **Face:** The human face plays a large role in recognising affect and emotions [216, 217]. These models rely on a robust extraction of the face in order to extract facial representations (see Section 3.1.3). Complex backgrounds and variations in scale make such a task non-trivial. To measure the robustness of the later approaches, a random set of up to a hundred frames per channel was selected and faces occurring in those frames were annotated by a human labeller.

  The annotated faces from Section 4.2 are used to investigate the quality of this approach. On the basis of the overlap of the prediction and human labelling, the detected bounding boxes were classified into true and false positives. The detector achieved an accuracy of 90 %, and an F1 of 86 % on the selection of MuSe-CaR. In addition, the bounding boxes are visually inspected to control the qualitative performance. Both performances underline the very good quality of MTCNN face extractions.

A detailed description of the annotations is omitted, as they are not the subject of this work but are available for future research. A full set is depicted in Figure 4.4

Other annotations: Besides the continuously real-valued emotional states, three annotations are made in a time-continuous, binary-valued fashion (see Figure 4.4): (4) the turns between the host (a visible person) and the narrator (a speaker from off-camera);

[5] Estimation: Around 4 frames per second are usually extracted in such a setting [12], resulting in over 576 000 frames for MuSe-CaR. The range is based on a minimum of 1 to a maximum of 15 bounding boxes for each frame, and the empirical value of an average of two boxes per minute from similar studies [273, 274].



	Category	Aspect	Examples
🟨	**Safety**	Tests	Euro NCAP, NHTSA, rating
		Assistance system	Anti-lock brakes, traction control
🟩	**Costs**	One-off	Retail & base price, feature price
		After sale	Insurance, maintenance, re-sale
🟩	**Comfort**	Surface	Leather, touch
		Space	Leg room, head room, luggage
🟩	**Quality & Aesthetic**	Design	Interior, exterior style (sporty, etc.)
		Quality	Material quality, clearance
🟩	**User experience**	Interaction	Interface, iDrive system, gestures
		Infotainment	Screen, Bluetooth, real-time traffic
🟩	**Interior**	Seat	Belt, split folding breaks
		Audio system	Radio, speaker
🟩	**Exterior**	Door exterior	Locks, handle
		Light	Headlight, fog light, taillight
🟦	**Handling**	Dynamics	Centroid, chassis, suspension
		Driving actions	Braking, steering, gear shifting
🟦	**Performance**	Engine	Horsepower, RPM, acceleration
		Powertrain	Electric, hybrid, combustion
🟦	**General**	Comparison	Models, brands, competitors
		Introduction	Series, weight, sales, warranty

Figure 4.7: Distribution of speaker topic annotations on sentence-level showing examples of subtopics and aspects. The percentage reflects the proportion of the sentences; almost 20 % of sentences have more than one topic.

(5) the appearance of banners; and (6) the appearance of more than one person (one person is speaking, and the others are there for demonstration reasons). Furthermore, there is a summary of several questions regarding appealingness to the annotator, host emotionality, content trustworthiness, and annotation confidence, wherein only one label for each video per annotator is given on a 10-point Likert scale (8).

4.3 Gold Standards

No framework has yet managed to unify the multitude of fusion methods for different time- and value-continuous and class-based discrete annotations. In this direction, Stappen et al. [25] proposes an open-source toolkit, MuSe-Toolbox, for creating a variety of time- and value-continuous and discrete-emotion gold standards, combining a wide range of fusion methods into a single framework central to the work at hand. Furthermore, these new methods are complemented by a novel annotation fusion method. In addition, for the first time, a procedure is proposed for meaningfully capturing dynamic emotional changes over time and aggregating them along the time dimension into a single, discrete label of a segmentation. An overview is displayed in Figure 4.8.

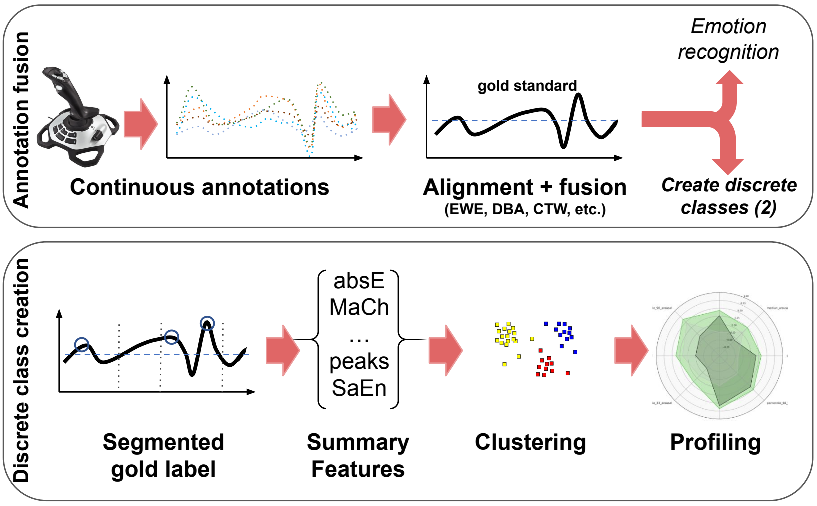


Figure 4.8: System overview of MuSe-Toolbox, showing the steps of the fusion process of value- and time-continuous annotations in the top panel, as well as the feature extraction and summary class creation in the bottom panel. Figure taken from Stappen et al. [25].

4.3.1 Dimensional Emotions

Personal emotion annotations are, by definition, hardly objective. Several approaches exist to merge individual subjective ratings of each annotator to form a collective emotion rating. For **dimensional emotions**, fused annotation is referred to as a gold standard [3, 184]. Naturally, one could assume that a mean over the signals could be applied to calculate the gold standard. However, this comes with two major disadvantages: first, it weights every annotator equally, independently of individual reliability and (dis-)agreement between annotators, and second, it ignores that each annotator can have different reaction times. Therefore, multiple gold standard creation methods were developed to calculate an even more objectively better fitting "mean". At its core, the approaches either aim to weight ratings [3, 45, 275] or try to align the continuous signals using advanced signal alignment methods [12]. Several are provided in the MuSe-Toolbox; however, due to relevance to this work, only two are focused on in the following (see Figure 4.9, left):

- **Evaluator Weighted Estimator (EWE):** Instead of exclusively including objective criteria in time series fusion [276, 277], Schuller et al. [3] proposed to directly model human diversity in emotion perception. For this, the reliability of an annotator is included in the fusion. In the EWE method, it is therefore assumed that the more similar a situation is perceived by several annotators, the better these opinions can be generalised and thus predicted. Mathematically, this is calculated through weighting each annotator or annotation. This is derived from a cross-correlation, hence the similarity of the continuous annotation of one rater with regard to the average of all others. Formally, this can be expressed as follows:

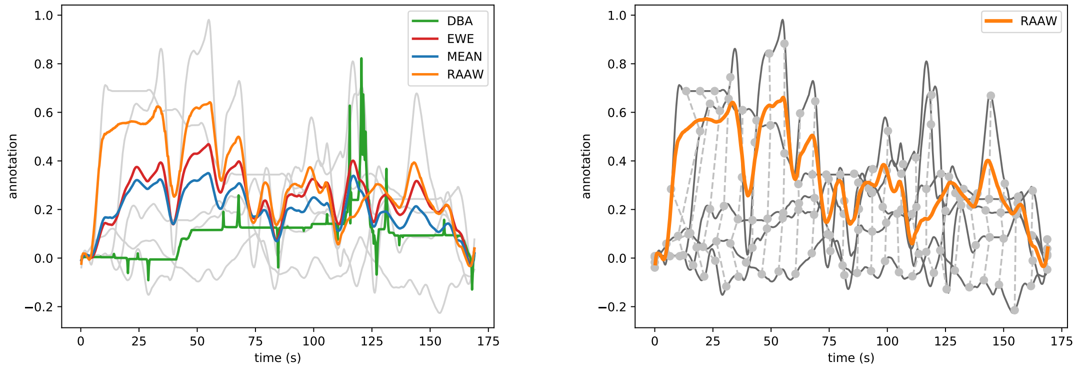
$$\hat{x}_n^{EWE} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k \hat{x}_{n,k}, \tag{4.1}$$

where r_k corresponds to the similarity of the kth annotator to the other continuous annotations. The similarity can be calculated with any time series similarity measure. EWE is the most prominent fusion approach for dimensional emotion annotation to date [65, 18, 12].

- **Rater Aligned Annotation Weighting (RAAW):** In Stappen et al. [25], an extension of the EWE method is proposed and provided by the MuSe-Toolbox. One problem with using human annotators is that they have varying delays in absorbing video content, assessing perceived emotion, and expressing this in joystick movement [278]. The phenomenon has been termed **reaction lag of evaluators** and found to be from one to six seconds. The problem is magnified when fusing annotations from different people with different reaction times, leading to unwanted smoothing of the fused annotation. For this reason, this alignment has so far been carried out manually or by brute force [279]. In the manual procedure, the individual annotations are shifted by sight. In the brute-force procedure, several models are trained on the basis of individual annotators, with the entire annotation shifted piece by piece by one to three seconds while monitoring the prediction results for improvements [18, 279]. The RAAW method eliminates this manual effort. For this purpose, a well-proven alignment method for time series, Dynamic Time Warping (DTW), is applied first (see Figure 4.9, right). It is known to condense a high degree of the original structure of the time series [276]. More specifically, since the complexity of the alignment computation increases with numerous annotators [280], a resource-efficient DTW variant, Generic-Canonical Time Warping (GCTW) is adopted from [277]. Then, the similarity for the individual aligned signals is calculated as done before with EWE to account for inter-rater agreement (subjectivity).

  Besides its use in MuSe-CaR introduced here, this method was also successfully applied in the creation of emotion-based gold standards that fuse physical-based arousal (electrodermal activity and beats per minute) with perceived arousal on a Trier Social Stress Test dataset. The interested reader is referred to Baird et al. [281].

On MuSe-CaR, the annotators achieved moderate correlations of .265 for arousal, .350 for valence, and .316 for trustworthiness before gold standard fusion, which is in line with



(a) EWE, DBA, Mean, and RAAW

(b) RAAWs warping paths

Figure 4.9: Fusion methods using an example video (id: 100) of MuSe-CaR. (a) shows the fusion methods including EWE (in red) and RAAW (yellow) applied to the arousal annotations; (b) illustrates the warping paths and the alignment of the RAAW method. Figure taken from Stappen et al. [25].

previous datasets in the field [12, 282, 283]. Both methods were applied to MuSe-CaR in earlier publications [26, 27] and are discussed in Section 5.1.1. If required, the framework also allows for upstream smoothing of the annotation as well as normalisation to the video or annotator level.

4.3.2 Summary Emotion Classes

A disadvantage for humans is the more intricate interpretation of continuous emotion signals compared to **categorical emotions**. Continuous emotions are also less estimable to determine a (sentiment) tendency of a segment [10]. An automatic transfer from dimensional annotations to classes would make precision and ease of interpretation possible as discussed in Section 2.2. Previous methods fail to provide the possibility to summarise continuous emotion annotations to a certain class over a variable-length segment duration. In [25], Stappen et al. proposed a new procedure towards this goal that does this largely automatically in three steps:

1. **Feature extraction:** A set of distribution and complex time series, hand-crafted representations are extracted from the emotion signals, precisely representing the temporal changes of the continuous emotions within a segment. This is seen as a feature set of emotional annotations, described in detail in Section 3.1.4. These representations can be used either collectively or separately for each emotion dimension. The extraction can be applied from annotation sections of varying lengths, often at sentence or segment

level. These are often based on transcribed speech or speaker turns. However, the absolute representations, in particular, must be normalised as a function of the segment length to reduce this influencing factor.

2. **Dimension reduction:** Clustering algorithms are often vulnerable and less robust when using broad feature sets with many dimensions due to effects such as the curse of dimensionality [284]. For this reason, the dimensionality of the created feature set is reduced by a Principal Component Analysis (PCA) [285]. In this process, the components are transformed by a projection procedure into new orthogonal axes that largely map the data variance. These components are then the starting point of the further procedure. Similarly, this is also possible with self-organising maps, which can be seen as a ANN without deeper layers. Here, a high-dimensional input is transformed into a low-dimensional output [286]. The output neurons correlate with patterns in the input, whereby the most important structures are represented even at low dimensions [287].

3. **Clustering:** The applied clustering methods intelligently construct more meaningful and homogeneous class clusters for these segments. Although a variety of methods are available in MuSe-Toolbox, only two are relevant to this thesis:

    - **K-means:** This is a crisp clustering method; therefore, each data point is always assigned to only one unique cluster during the course of the process [288, 149]. The number of clusters is specified a priori. Then, clusters are randomly placed centrally at this height, and the distance between the clusters and the data point is determined with the help of a distance measure, often the Euclidean distance. The nearest cluster centre corresponds to the cluster to which a point is assigned. Based on the Estimation-Maximisation Algorithm (EM), the centres and the cluster membership are optimised step by step until they converge after a few iterations.

    - **Gaussian Mixture Model (GMM):** In contrast, GMM is a fuzzy clustering method, so the membership of a data point in the resulting clusters is expressed with a certain probability [289]. This is based on a probabilistic algorithm that aims to describe the distribution of the points as well as possible with the help of Gaussian distributions. A method based on the EM algorithm is employed once again to find the appropriate covariance structure of the data and the centres of the latent Gaussian distributions.

This entire process is run several times for a set of hyperparameters and monitored by clustering and qualitative criteria until the user is satisfied with the results.

Experiments and Evaluation

5 Experiments and Evaluation

This chapter brings together the concepts (Chapter 2), underlying methodology (Chapter 3), and data (Chapter 4) to empirically answer the research questions described in Section 1.2. For this purpose, Section 5.1 and Section 5.2 exclusively use the Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) dataset. As the originators of the data set, suitable extraction frameworks were applied for the first time and the baseline models for the guidance of other researchers were introduced in Stappen et al. [26, 27]. However, the developed models go far beyond the simplistic baselines in comparable work [17, 18]. Their complexity, training, and fine-tuning are the results of extensive method development.

The experiments are executed on Graphics Grocessing Units (GPUs), which offer between 24 GB and 32 GB GPU RAM and serve as the hardware backbone. All experiments are implemented in Python. The design and training of the Deep Learning (DL) architectures are carried out with the Python packages PyTorch[1] and TensorFlow[2]. The Hugging Face package[3] is used for the NLP transformers. All other Machine Learning (ML) and clustering models are implemented with the ML library Scikit-Learn [290].

5.1 Subjective Dimensions

The focus in this section is on predicting and utilising the subjective dimensions of Multimodal Sentiment Analysis (MSA) (see Section 2.2). The presented MuSe-CaR dataset (see Chapter 4) is the basis for the experiments, which address **RQ-1** and **RQ-3** in the following manner:

- **RQ-1a:** Demonstrating the efficacy of predicting arousal and valence dimensions as a time- and value-continuous **regression** task, evaluating two gold standards.

- **RQ-1b:** Investigating the efficacy of predicting **arousal and valence** dimensions as a **classification** task, evaluating two "summary" class creation procedures for topic-specific segments.

- **RQ-1c:** Exploring the quantification of the perceived **trustworthiness** dimension as a continuous regression task (as in 1a).

[1] https://pytorch.org/ accessed July 11, 2021

[2] https://www.tensorflow.org/ accessed February 1, 2021

[3] https://github.com/huggingface/ accessed April 6, 2021

Table 5.1: Total amount of video footage of the MuSe-CaR dataset selected and speaker-independent partitioned for each task type. Reported are the number of unique videos and the duration for each task type in hh :mm :ss.

Partition	No.	Regression	Classification	Trustworthiness
Train	166	22 :16 :43	22 :35 :55	22 :45 :52
Devel.	62	06 :48 :58	06 :49 :46	06 :52 :22
Test	64	06 :02 :20	06 :14 :08	06 :12 :53
Σ	292	35 :08 :01	35 :39 :49	35 :51 :07

- **RQ-1d:** Exploring the relationships of the three dimensions of 1a and 1c on (predicting) the **popularity** of a video on the YouTube platform.

- **RQ-3:** Evaluating the predictive power of the three core **modalities** and their representations, individually and in combination.

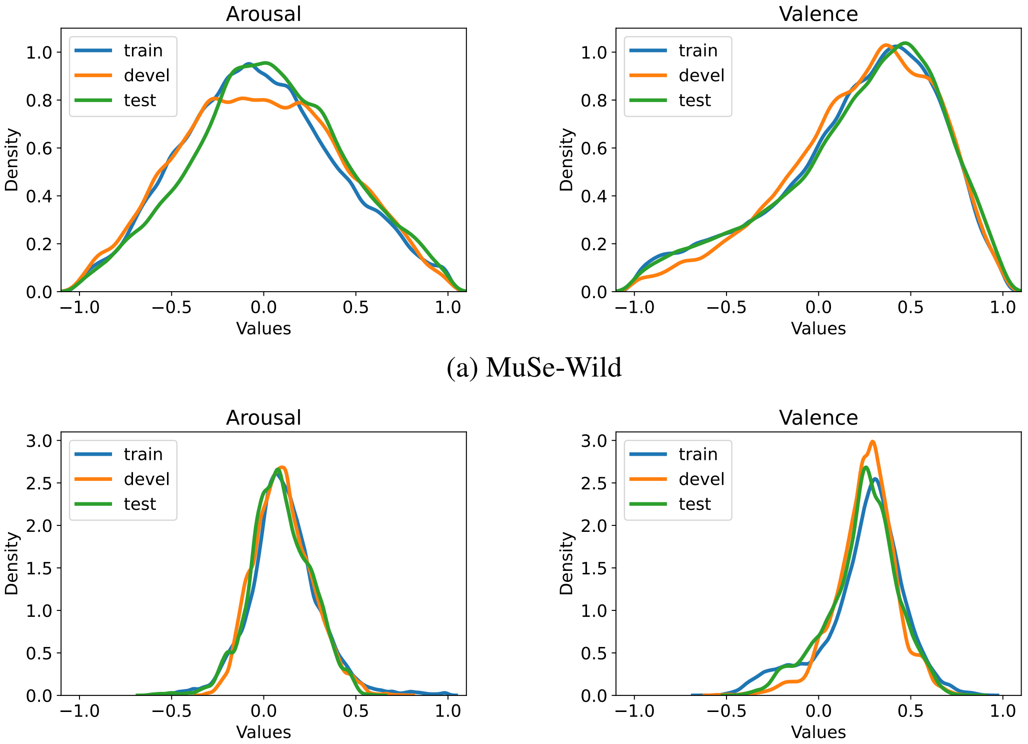
For each task, the characteristics (e. g. , data selection) are introduced at the beginning. The followed experimental setup explains details to the representation extraction and introduces the architectures, specifically developed for the task. After the results, the research questions raised here are revisited.

5.1.1 Emotion Recognition

5.1.1.1 Characteristics

Developing models robust against in-the-wild characteristics (Section 2.1) is a vital challenge in predicting arousal and valence levels. In general, the modalities extracted from the audio-visual recordings have varying amounts of presence and significance (e. g. , due to the influence of noise) and need to be systematically combined to fully exploit the models' potential. Methods involved include (dynamic) alignment and fusion of representations at appropriate linguistic levels (e. g. , word, sentence, segment).

For the following experiments, all people-focused video segments wherein a voice is present or a face is visible are included. At the same time, all non-product-related parts (e. g. , the introduction, social media mentions, and commercial remarks) are excluded. In total, more than 35 hours of video from MuSe-CaR fulfil these criteria, and are speaker-independent separated into train, development, and test sets as illustrated in Table 5.1. Stappen et al. proposed two slightly different versions of this task in previous publications [26, 27], differing only in the annotator fusion method (Section 4.2).



(a) MuSe-Wild

(b) MuSe-Wilder

Figure 5.1: Density distribution using 35 equal-width bins of the train, (devel)opment, and test partitions for continuous annotations of (a) MuSe-Wild and (b) MuSe-Wilder. The distributions between all partitions of one annotation are fairly similar, however, MuSe-Wilder is more sharply distributed than MuSe-Wild due to different standardisation and annotator fusion. Figure adapted from Stappen et al. [26, 27].

I. **Multimodal V-A Sentiments in-the-Wild Sub-challenge (MuSe-Wild):** The first version of the task [26] uses **Evaluator Weighted Estimator (EWE) fusion** on the ratings standardised on the video level.

II. **Multimodal Continuous Emotions in-the-Wild Sub-challenge (MuSe-Wilder):** The second version [27] uses **Rater Aligned Annotation Weighting (RAAW) fusion** and applies standardisation and min-max normalisation on the rater, not video, level as described in Section 4.2.

Modelling annotator subjectivity by these fusion methods has a vigorous impact on understanding the annotator agreement and generalisability of the models. Furthermore, the interested reader is referred to Stappen et al. [25] for an in-depth analysis of additional standardisation and fusion methods made available by Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox (MuSe-Toolbox) toolbox.

Analogous to other regression tasks and emotion recognition challenges [17, 18], the Concordance Correlation Coefficient (CCC), with an identical formulation as the CCC loss function (see Section 3.2.1), is used to assess and compare the models' performance.

5.1.1.2 Experimental Setup

For each task, the experiments are subdivided into establishing baseline results, as presented in the respective earlier publications, and evaluating the Multihead Attention Long Short-Term Memory Recurrent Neural Network (MHA-LSTM) architecture, which is derived from the lessons of the MuSe challenges. Even though the baselines differ for the respective tasks, all models are listed together in the following.

5.1.1.2.1 Feature Sets: In the following experiments, a wide variety of representations from the three modalities are utilised as introduced in Section 3.1. Extraction is performed over the entire length of a video, from which the relevant segments are selected. Most of the audio extractors rely on preprocessed audio data. The raw audio is normalised to -3dB and converted from stereo to mono, 16 kHz with 16 bit. The ComParE Low-Level Descriptors (ComParE LLDs), the voice-related hand-crafted representations, are extracted using 60 ms frames with a Gaussian window function, while all other Low-level Descriptors (LLDs) are based on 25 ms frames with a Hamming window function which is overlapped and sampled at 100Hz. In accordance with the specifications, a symmetric moving average window with a frame length of three and their first order delta regression coefficient with a window size of two frames is applied for smoothing. For Spectrograms Feature Extraction from Audio Data with Pre-trained Convolutional Neural Networks (Deep Spectrum) the audio signal is converted first into mel spectrogram plots using a Hanning window with 32 ms and an overlap of 16 ms. Next, 128 mel frequency bands are computed. Finally, these mel spectrograms are passed into a VGG-19 extraction network [205], resulting in a 4 096-dimensional vector. To be in line with the annotation sampling rate, the hop size is set to 250 millisecond steps for the remaining audio (e. g., CNN Architectures for Large-Scale Audio Classification (VGGish), extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)) and video (e. g., Depthwise Separable Convolutions Network (Xception), Very Deep Convolutional Networks for Large-Scale Face Recognition Descriptor (VGGFace), Facial Action Unit (FAU)) feature sets. For video representations, this corresponds to 4 evenly sampled frames per second. The only exceptions are the text inputs, wherein one feature vector for each word is generated, which can span across several hops.

Alignment: If feature extraction fails for a time step, for example, because no person is visible for a human-related feature such as a FAU, a zero feature vector is imputed. For the text representations, the extraction is done for each word. Since the word frequency is much lower than the annotation frequency, text representations are imputed for the time of vocalisation according to the timestamps of the transcripts. Substituting timestamps without articulation with a zero vector ensures a perfect alignment to representations of the other modalities.

5.1.1.2.2 Architectures: Four architectures with distinctive characteristics are proposed for this task. For MuSe-Wild, three models are used to create a baseline, while the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) model serves as a baseline for MuSe-Wilder only. The final architecture, incorporating all learnings, consists of a Multihead Attention Layer (MHAL) combined with an LSTM-RNN. In the experiments, the most promising feature sets introduced in Section 3.1 are used.

LSTM-RNN: Since the nature of the data provides a sequence of representations to predict a sequence of regression points, a many-to-many LSTM-RNN, as explained in Section 3.2.2, is inherently suited for this task. LSTM-RNNs are specifically designed to facilitate learning short- and medium-term patterns. As a stand-alone solution, the LSTM-RNN first performs a linear reduction of the feature input e to the hidden state dimensionality (h) $\in \{32, 64, 128\}$ of the first recurrent layer. This layer can be either uni- or bi-directional. The additional layers in the LSTM-RNN, number of layers (n) $\in \{1, 2, 4\}$, consist of the same type and dimensionality as the first while a Rectified Linear Unit (ReLU) activation function squashes the resulting hidden vectors of each. In the final layer, these states are transformed by a Fully connected Feed-Forward Layer (FFL) to a sequence of logits resulting in one prediction for each time step. As with comparable approaches [24, 26, 262], efficient training when using input sequences of different lengths is achieved by dividing them into shorter, equally long, and overlapping segments of the original sequence. For this purpose, a 200-step-sized window size (ws) (corresponding to 50 s at a sampling rate of 250 ms) is applied, shifting the window hop size (hs) by 100-step hops (25 s). Sequences that are too short are zero padded. The experiments are executed with a learning rate of learning rate (lr) $\in \{0.0001, 0.001, 0.005\}$ and a batch size (bs) $\in \{512, 1024, 2048\}$ running for up to 100 epochs (early stopping). For late fusion, an additional multimodal LSTM-RNN network can be trained on the stored predictions of the unimodal networks as done in [262]. Due to the 2- or 3-dimensional input, parameters are set as $h = 32$, $n = 1$ and the network is fine-tuned until the loss does not decrease further for 15 epochs.

Long Short-Term Memory Recurrent Neural Network with Self-Attention (LSTM-SA): Similar to LSTM-RNN, LSTM-SA copes with the sequential nature of the input representations by feeding them into a bidirectional LSTM-RNN with $h \in \{30, 40, 50, 100\}$ neurons in each of the $n \in \{1, 2, 3\}$ hidden layers. As described in Section 3.2.4, the encoded representation is then used as a query to apply self-attention based on the entire sequence. The resulting attention-encoded sequence is concatenated to the sequence of original query representations. At each time step, the vector is processed by a FFL for continuous-time regression. Furthermore, the input length is restricted to 50 time-step segments with no overlap and 0.1 Gaussian noise is added to the input for better generalisation. The initial $lr \in \{0{,}00001, 0{,}0001, 0.001\}$ and $bs \in \{50, 100\}$ are used while the training is running for up to 50 epochs, applying early stopping when the loss on the development set does not improve for 5 epochs. As with the LSTM-RNN, the network has an early fusion mechanism.

End-to-End Learning (End2You): The End2You Framework [89], is frequently applied for end-to-end DL in unimodal audio [90] and multimodal settings [26, 291, 292]. It is based on a Recurrent Convolution Neural Network (RCNN), combining a Convolutional Neural Network (CNN) mechanism (see Section 3.2.3), for low-level, spatial audio feature extraction from the raw (audio) signals, with an LSTM-RNN to learn short-term temporal dynamics. The configuration comprises three CNN building blocks with a ReLU activation function for audio feature extraction, followed by concatenating representations from other modalities and an LSTM-RNN layer. One CNN block consists of a one-dimensional CNN layer with a filter size (fs) and kernel size (ks), followed by a maximum pooling layer with a pool size (ps) and strides strides (s), as well as 50 % random dropout. After preliminary experiments, the following architecture is selected: The first block has filter size (fs) = 50, $ks = 5$, $ps = 10$, and $s = 10$; the second block increases fs to 125 and reduces ks to 8, while ps and s are set to 8; the final block sets $fs = 250$, $ks = 6$, $ps = 6$, and $s = 6$. The LSTM-RNN layer has $h = 256$. Due to the high degree of noise, further modalities are necessary to achieve stable training. In addition to audio, VGGFace is used for vision — since they are low-level face representations (see Section 3.1.3) — and Fast Text Classifier (FastText) representations for text. While the VGGFace representations are imputed with a zero vector if missing, the text feature vector is repeated until the next recognised word, equivalent to a successfully extracted word vector. Based on previous findings [291, 292], to learn underlying low-level audio representations, the lr has to be set very low, starting from 0.00001, and is increased in equal steps to 0.00009 for up to 40 epochs at a $bs = 10$.

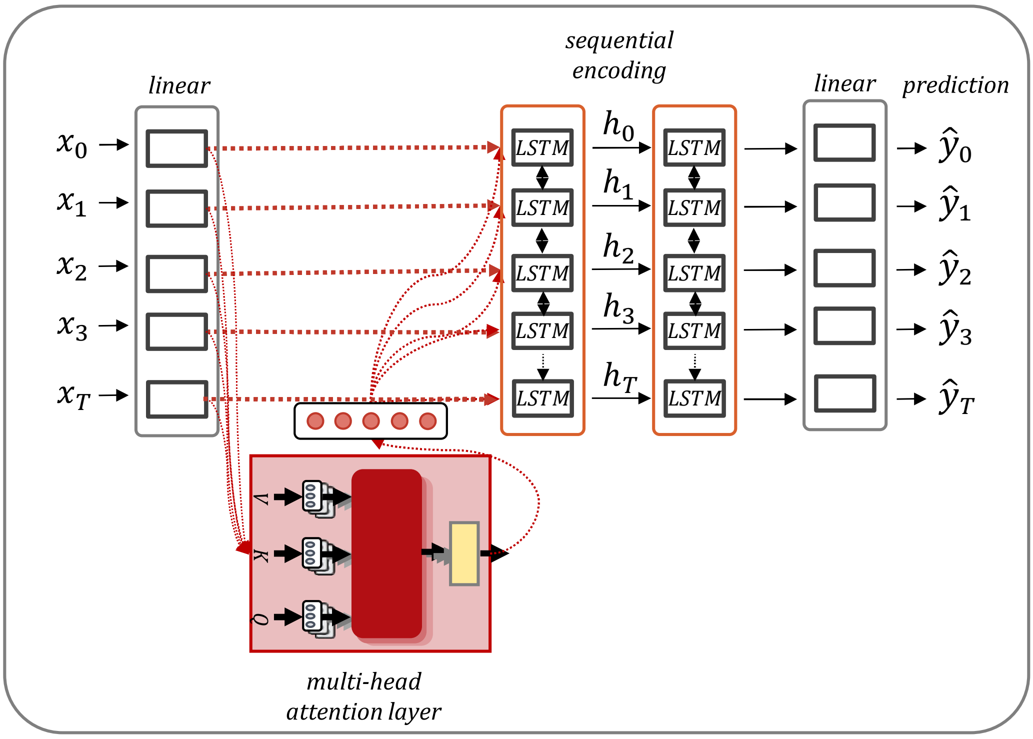


Figure 5.2: Illustration of the developed MHA-LSTM architecture utilising multihead attention ($at = 3$ in layer) for intramodality enhancement and two bidirectional LSTM-RNNs to capture the sequential context.

MHA-LSTM: To model the short- and long-term dependencies, two DL mechanisms are leveraged, as depicted in Figure 5.2. These improve the representation of the input state(s), e. g. , early fused audio, text, and video feature sets, by one or multiple MHAL (see Section 3.2.4) and again encode the spatial patterns of state transitions with one or multiple (bidirectional) LSTM-RNNs. The attention heads (*at*), reinforce the robust representations of the extracted local representation and can preserve the long-term (global) dynamics of a sequence. However, this mechanism is not sufficient to develop a deeper understanding of the positional encoding (see Section 3.2.2) [91]. For this purpose, the properties of n LSTM-RNN layers are required. Similar combinations have been proposed previously [27, 262, 293]. Depending on the task, the top layer is a FFL with one output for each encoding step (Sequence to Sequence (S2S)) for regression. For optimal performance, the following architectural hyperparameters are searched: $n \in \{1,2,4\}$, $h \in \{32,64,128\}$, number of attention heads $(at) \in \{2,4,8\}$. In addition, the training behaviour is optimised for up to 100 epochs (early stopping) using $lr \in \{0,0001,0.001,0.005\}$ and $bs \in \{512,1024,2048\}$, and the sliding windows values are set to $ws = 200$ and hop size to $hs = 100$. For MuSe-Wilder only, the number of MHALs is evaluated in the same range as *n*. The architecture enables early and late modality fusion. For this, multiple modalities can be incorporated by an early concatenation of aligned representations, or the extracted sequences of predictions utilised by an additional LSTM-RNN.

5.1.1.3 Results

In the following, results of MuSe-Wild and MuSe-Wilder are first discussed individually. The conclusion brings the findings of the two tasks together to give an overall picture.

5.1.1.3.1 MuSe-Wild: An overview of the results for predicting continuous, EWE fused emotions is given in Table 5.2. The upper half of the table shows the baseline models, while the lower half highlights the architectures developed as challengers.

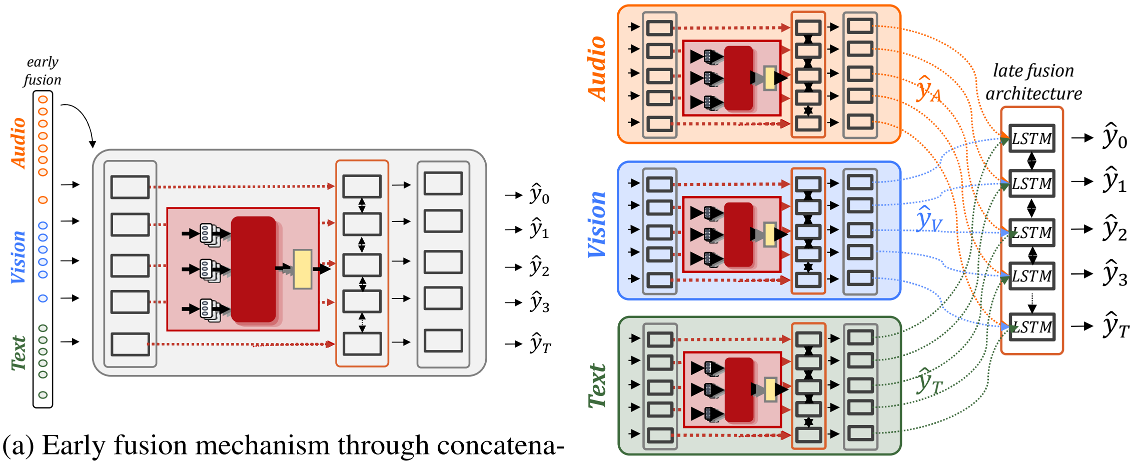
Baseline: For valence prediction, the best results are yielded by the end-to-end architecture using a range of representations consisting of FastText, VGGISH, and audio representations learned from the raw audio signal. Here, the model shows a CCC on valence of .1506 on development set and .2431 on the test set. This architecture also achieves robust results for arousal prediction, with .2587 CCC on the development set and .2706 on the test set. However, the LSTM-SA performs considerably better for predicting arousal with a CCC of .3088 and .2884, respectively, on the development and test sets. In this case, the audio representations, especially the LLD and Deep Spectrum, show the most robust performance of all input representations. In comparison, results for valence are poor, with the best feature set, FastText, only resulting in a CCC of .1816 on the test set.

Challenge: Compared to the baseline models, the two best teams in the MuSe-Wild challenge achieved substantial performance gains. Ruichen et al. [294] achieved test results of .4346 for arousal and .4513 for valence with a LSTM-RNN attention architecture, and Sun et al. [262] developed a LSTM-RNN incorporating multihead attention that achieved .4726 and .5996 on arousal and valence, respectively, on the challenge test set fusing up to six different feature sets.

Post-challenge models: Inspired by the challenge improvements and to find the crucial success mechanisms, further experiments (see lower panel in Table 5.2) are conducted using two architecture with (MHA-LSTM) and without (LSTM-RNN) multihead attention. Again, the best performance for predicting valence is achieved with the text

Table 5.2: Reporting arousal and valence for **MuSe-Wild** (using EWE annotation fusion) in Concordance Correlation Coefficient (CCC) on the devel(opment) and test partitions. Audio feature sets: Low-level Descriptors (LDD), eGeMAPS (Ge), Deep Spectrum (DS), and VGGish (VG); vision features sets: GoCaRD (Go), VGGFace (VF), and Xception (X); and text feature sets FastText (FT) and BERT (BT) are fed into the models. Furthermore, the raw audio signal (RA) is utilised by End2You. The features are aligned to the label timestamps.

Approach	**Modality**	**Feature(s)**	**Valence**		**Arousal**	
			devel	test	devel	test
		Official Baselines [26]				
		Unimodal				
LSTM-SA	A	LLD	.0711	.0349	**.3078**	**.2834**
		DS	.0165	.0024	.1585	.1723
		Ge	.0435	-.0097	.1090	.0827
	V	X	.0499	.0426	.0776	.0683
		aV	.0098	.0272	.1598	.1227
	T	FT	.1273	.1816	.0959	.1074
		Multimodal				
LSTM-SA	A+T	Ge + FT	.0520	.0361	.1375	.1018
	T+A+V	FT + Ge + aV	.0393	.0654	.1809	.0865
End2You	T+A+V	FT + VG + RA	**.1506**	**.2431**	.2587	.2706
		Post-Challenge Models				
		Unimodal				
LSTM-RNN	T	FT	.2398	.3202	.2038	.1629
		BT	.4522	.5216	.2497	.1615
	A	VG	.1843	-.0349	.4249	.1822
		Ge	.1809	.0867	.3711	.1810
	V	VF	.1083	.0455	.3497	.2593
		AU	.0646	.0542	.2713	.0662
MHAL-LSTM	T	FT	.2742	.3806	.1669	.1004
		BT	**.4569**	**.5987**	.3527	.2954
	A	VG	.1481	.0652	**.4909**	**.4027**
		Ge	.1271	.0728	.3984	.4024
	V	VF	.0921	.0686	.3733	.3652
		AU	.1109	.0392	.2840	.0961
		Multimodal				
MHAL-LSTM (early fusion)	best A + V		.1174	.0634	**.5327**	.3592
	best A + T		.4284	.5577	.4235	.1757
	best V + T		.4451	.5547	.4127	.2777
	best V + A + T		.4302	.5476	.4683	.3073
MHAL-LSTM (late fusion)	best A + V		.1918	.1185	.4622	.3891
	best A + T		.4711	.6216	.4622	.4104
	best V + T		.4645	**.6237**	.4107	.3814
	best V + A + T		**.4737**	.6204	.4748	**.4271**



(a) Early fusion mechanism through concatenation of modality representations.

(b) Late fusion by sequential encoding via LSTM-RNN.

Figure 5.3: Illustration of the multimodal (a) early and (b) late fusion techniques used to improve MHA-LSTM.

modality on both models. All other modalities achieve less than .1 CCC on the test set and are of limited relevance. For arousal, audio representations show the strongest results. Using MHA-LSTM, VGGish yielded the best results with .4027 CCC on the test set. Both vision representations show solid results on the development set for the prediction of arousal, however, only VGGFace generalises on a similar level on the test set. Text, e. g. , Bidirectional Encoder Representations from Transformers (BERT) achieves up to .2954 CCC on the test set with MHA-LSTM. Compared to the participants' results, the performance of the proposed models are comparable and in most cases slightly outperform them. The MHA-LSTM architecture achieves **new state-of-the-art results** with .6237 CCC on the test set predicting valence fusing BERT and FAU. The best result for the prediction of arousal falls slightly behind the best challenge result with .4271 CCC, however, both utilise only two, three representations, compared to the six of the model that won the challenge.

In comparison to the baseline models, several changes have been made that seem to have positive impact on the results. Three underlying mechanisms are highlighted here as contributing, at least partially, to the performance improvements:

- **Shorter window size and overlapping segments:** Compared to the baseline, the input window length and overlapping is more sensible crafted for the post-baseline models. Experimental results are shown in Table 5.3 wherein the window *ws* and hop size *hs* in the data segmentation process are varied. For predicting arousal and valence, the best performing representations (VGGish and BERT) in combination with the best architecture (MHA-LSTM) and most suited hyperparameters are applied. Looking at the average results on the test set, the results clearly improve with increasing length, for example from .4142 CCC (using *ws* = 100 and *hs* = 50) to .5223 CCC (using *ws* = 750 and *hs* = 500). For VGGish, the results with overlap (*ws* > *hs*) are in all cases better than without (*ws* = *hs*) when comparing the performance on a *ws*-level, for example .4027 CCC at *ws* = 200 and *hs* = 100 compared to .3752 CCC at *ws* = 200 and *hs* = 200. For BERT, this is only the case for the *ws* = 750.

The disadvantage of longer sequences is the vastly increased computation time for architectures employing LSTM-RNN, which increases computational cost and GPU memory quadratically with each sequence step (see Section 3.2.2), making it impractical for broad hyperparameter search (*ws* > 750 exceeds 32 GB GPU memory in the standard setting used here).

- **Stronger representations:** A summary of these experiments in tabular form is shown in the lower half of Table 5.2. BERT outperforms FastText by a large margin independently of the architecture and prediction target. Although the acoustical representations still dominate in the arousal prediction, the gap to the best text (BERT) and best vision (VGGFace) feature sets is substantially reduced (cf. MHA-LSTM vs LSTM-RNN). On audio, which previously showed better results on arousal, the new VGGish clearly outperforms eGeMAPS on the development set. While the results using the LSTM-RNN are also much better on the test set, the results using the MHA-LSTM architecture are almost identical on the test set.

- **Late fusion:** Finally, in a multimodal setting, late fusion is superior in combining various modalities (see lower half of Table 5.2). In these experiments, the representation which reached the best unimodal results for each modality is chosen. These results confirm the participants' results, so that here too, a later fusion is more beneficial to performance than an early one. The results of early fusion are always worse than the best result of the best unimodal approach. For example, MHA-LSTM achieves .5987 CCC with BERT representations for valence prediction of but only .5577 CCC when including VGGish representations (alone: .0652 CCC) in the early fused mode. In contrast, late fusion of the same combination achieves .6216 CCC for the same target, greater than both unimodal results. The primary reason for this may be that overfitting is avoided. Early

Table 5.3: Results of **MuSe-Wild** varying the data segmentation parameters: the window size (*ws*) and the overlap hop size (*hs*) showing the results of the best model MHA-LSTM with the strongest feature sets, VGGish for arousal and BERT for valence, as well as the average of these two Ø. The other hyperparameters remain static. The results are given in Concordance Correlation Coefficient (CCC) for the devel(opment) and test sets.

steps		**BERT**		**VGGish**		**Ø**	
ws	*hs*	**devel**	**test**	**devel**	**test**	**devel**	**test**
750	750	.5641	.5540	.4274	.4344	.4957	.4942
750	500	**.5739**	.5747	**.5604**	**.4699**	**.5671**	**.5223**
750	250	.5189	.5693	.5386	.4686	.5288	.5190
200	200	.4512	.5245	.4566	.3752	.4539	.4499
200	100	.4569	**.5987**	.4909	.4027	.4739	.5007
200	50	.4282	.5160	.4440	.4081	.4361	.4621
100	100	.4167	.5128	.4782	.2815	.4475	.3971
100	50	.4312	.5068	.4958	.3215	.4635	.4142
100	25	.4233	.5216	.5016	.3233	.4624	.4225

fusion of the representations might lead to modelling very complex interacting representations, which are not necessarily found in the test set.

5.1.1.3.2 MuSe-Wilder: The proposed baseline models for the MuSe2021 challenge incorporate the lessons learnt and findings of MuSe2020, aiming for better results while using a less complex architecture.

Baseline: To predict the RAAW-fused arousal and valence annotations, a competitive baseline with a variety of feature sets was first established in Stappen et al. [27]. In terms of the predictive strength of the modalities, the picture is consistent with the results of the EWE fused annotations: valence is best predicted by text, whereas audio has a strong predictive power for arousal. Looking more closely at the representations, the BERT representations combined with the hyperparameters of $n = 4$, $lr = 0.005$, and $h = 128$ achieve the best results by far with .4613 CCC on the development set and .5671 CCC on the test set. Visual representations still achieve results of up to .1637 CCC on test for Xception. Even though the performance of audio is marginally superior on the development set, the models generalise poorly on the test set, so that these results are not transferable, and Deep Spectrum just exceeds .1 CCC. This performance gap is analogous to the EWE annotation results. The result is different for arousal prediction. Here, Deep Spectrum demonstrates the strongest performance in combination with the hyperparameters $n = 2$, $lr = 0.001$, and $h = 64$: .4831 CCC on the development and .3386 CCC on the test set. VGGish and eGeMAPS are still

Table 5.4: Reporting arousal and valence for **MuSe-Wilder** (using RAAW annotation fusion) in Concordance Correlation Coefficient (CCC) on the devel(opment) and test partitions. The feature sets tested are Deep Spectrum, VGGish, and eGeMAPS for audio; Xception, VGGFace, and FAU for video; and BERT for text. All utilised features are aligned to the label timestamps by imputing missing values or repeating the word embeddings.

	LSTM-RNN Baseline [27]				**MHAL + LSTM**			
Features	**Valence**		**Arousal**		**Valence**		**Arousal**	
	devel	test	devel	test	devel	test	devel	test
Unimodal								
Deep Spectrum	.1901	.1019	**.4841**	**.3386**	.2109	.1217	**.4688**	.3670
VGGish	.1500	.0054	.4027	.2545	.1898	.1111	.4541	**.3768**
eGeMAPS	.1916	.0019	.3877	.2428	.1867	.0827	.4042	.3471
Xception	.1872	.1637	.2870	.1793	.1730	.1530	.3622	.2123
VGGFace	.1203	.1197	.3201	.2970	.1654	.1291	.3633	.3735
FAU	.0682	.1275	.3045	.1165	.0993	.1456	.3166	.2983
BERT	**.4613**	**.5671**	.2716	.1873	**.4660**	**.6132**	.3741	.2374
Multimodal Late Fusion								
best A + V	.2362	.1220	.4821	.2822	.2193	.1715	.4442	**.4242**
best A + T	.4782	.5950	.4754	.3046	**.4677**	.6147	.4538	.3713
best V + T	.4641	.5874	.3111	.1767	.4647	.6035	.4131	.3138
best V + A + T	**.4863**	**.5974**	**.4929**	**.3257**	.4762	**.6150**	**.4834**	.4148

competitive on the development set, but perform slightly worse on the test set with .2545 CCC and .2428 CCC, respectively. Of the visual representations, VGGFace shows a strong and balanced performance across both partitions with .3201 CCC and .2970 CCC on the development and test set, respectively.

Bimodal late fusion produces mixed results with isolated increases in performance. Multimodal late fusion from the three best performing representations of each modality (Deep Spectrum, VGGFace, and BERT) leads to improved outcomes for valence by around .03 to .5974 CCC on the test set. For arousal, the results increase only slightly on the development set, but are slightly worse than Deep Spectrum.

MHA-LSTM: As seen with MuSe-Wild, the attention-enhanced architecture outperforms the unimodal baseline on almost every feature set. In line with the baseline results, text is the best predictor of valence, while the audio representations are most effective in predicting arousal. For the valence prediction, BERT rose by almost .05 CCC, from .5671 to .6132 CCC, on the test set using h = 128, one LSTM-RNN, and a MHAL with four blocks trained with a $bs = 512$ and a $lr = 0.005$. The visual representations show only isolated marginal improvements for valence prediction. For the audio representations, the generalisation ability of the models increases, hence closing the gap between development and test set results for all audio feature sets on both targets. For example, it almost halved the margin from .14 to .08 CCC and .15 to .08 CCC on the test set for eGeMAPS on valence and VGGish on arousal respectively. In predicting arousal, Deep Spectrum gains almost .03 CCC on the test set and remains the best feature set. VGGish jumps from .2545 to .3768 CCC with hyperparameters of $n = 4$, $h = 128$, $at = 2$, and training parameters of $bs = 512$ and $lr = 0.005$, thus, as with MuSe-Wild, it shows the best performance for unimodal prediction of arousal on the test set. For visual representations, FAU in particular greatly increases to .2983 CCC on the test set, but remains just behind VGGFace ($n = 2$, $h = 64$, $at = 4$, $bs = 1024$, and $lr = 0.001$) with .3735 CCC on the test set, putting the best vision feature only slightly behind VGGish. Text also improves significantly in predicting arousal on the development set, but these results can only be partially generalised on the test set.

The multimodal experiments behave almost identically to the baselines with top results of .6150 CCC on the test set by fusing Deep Spectrum, Xception, and BERT on the prediction of valence and .4242 CCC on the test set by fusing Deep Spectrum and VGGFace on the prediction of arousal. In general, the bi- and tri-modality results only slightly improved compared to the baseline results when predicting valence. For arousal, the largest gain is achieved by the fusion of VGGFace and BERT. Otherwise, the results on the development set perform slightly worse than the baseline models, however, vast improvements are shown on the test set, e. g. , the best A + V increases from .2822 to .4242, the best A + T from .1767 to .3138, and the best V + A + T from .3257 to .4148 CCC. The improvement of development and test set results indicates an increased generalisation robustness of the attention-enhanced architectures.

5.1.1.4 Conclusions

Overall, these findings show that both annotation gold standards of the novel MuSe-CaR dataset can be effectively predicted through a carefully selected combination of representations and models (**RQ-1a**). Besides several extracted representations, a variety of models for the tasks of MuSe-Wild (EWE) and MuSe-Wilder (RAAW) were developed and successfully tested. In particular, performance improvements can be attributed to intra modality feature enhancement through the attention mechanism prior to sequential coding. This finding links to literature that associates the robustness of representation with attention encoding [91]. In addition, segmentation through fine-tuned window selection strategies has demonstrated its effectiveness. Both have been united in the proposed MHA-LSTM architecture. Due to the ongoing development of attention mechanisms, further advances can be expected in the future.

Furthermore, it was found that the distributions of the prediction points of the widely used EWE gold standard differ vastly from the proposed configuration of the RAAW method (see Figure 5.1). However, the experiments presented could verify that the new proposed gold standard has no negative impact on predictability by emotion recognition models. For example, BERT, as the best text representation, achieves .5987 CCC using EWE and .6132 CCC using RAAW. Comparable results can also be seen for arousal prediction. In fact, it could be shown that the structure, hence the combination of successful modalities or representation types to prediction targets (arousal, valence), is identical. These findings indicate that RAAW allows comparable results to EWE, while the influence of the rater lag is reduced fully automatically.

From a uni-modal perspective (**RQ-3**), the textual representations proved to be particularly suitable in this novel setting, especially in the prediction of valence. Audio was most effective in the prediction of arousal. Regardless of the modelling technique, the novel contextual Transformer word embeddings (BERT) specifically turned out to be much more predictive than the classical word embeddings. In the case of audio representations, data-driven representations (VGGish, Deep Spectrum) demonstrated their suitability, often outperform hand-crafted representations (eGeMAPS). The vision representation, e. g. , VGGFace, also led to promising results in predicting arousal. Looking at these results, it is worth noting that although MuSe-CaR brings a higher complexity in terms of in-the-wild characteristics, the results are in the same range as on other, much less complex datasets [12]. However, the performance of vision on valence is below expectations. This can probably be attributed to the negative influence of the noisy visual traits. To counteract noisy influences, further investigation is needed, as are improved extraction methods, which can conveniently be conducted on this novel dataset. Among the fusion techniques, late fusion (MHA-LSTM) outperformed early fusion (End2You, LSTM-SA, and MHA-LSTM). In the future, improving intermodality fusion using attention mechanisms in a hybrid integration could be a next step.

5.1.2 Emotion Classification

5.1.2.1 Characteristics

Multimodal Sentiment Analysis aims to investigate the engagement between emotions (see Section 2.2) and the domain-specific context (see Section 2.3). The latter is contextually self-contained; for example, a topic has a clearly defined beginning and endpoint and engages with a group of heterogeneous topic aspects in between. While the recognition task in the previous chapter focuses on detecting small emotional changes (sequence of regression points), the classification task implies two subtasks in this setting:

(a) **creating higher-level summary emotion labels** from the continuous annotations of a coherent content segment, and

(b) **predicting** the artificially created summary emotion labels.

In this light, the emotion component is considered in isolation while using an identical data basis as in Section 5.2.1 for speaker topic prediction. Therefore, the applied methods are distinct in their viewpoints and training but share a common ground, and the individual results can be connected to an EMOTION-SPEAKER TOPIC pair for each segment.

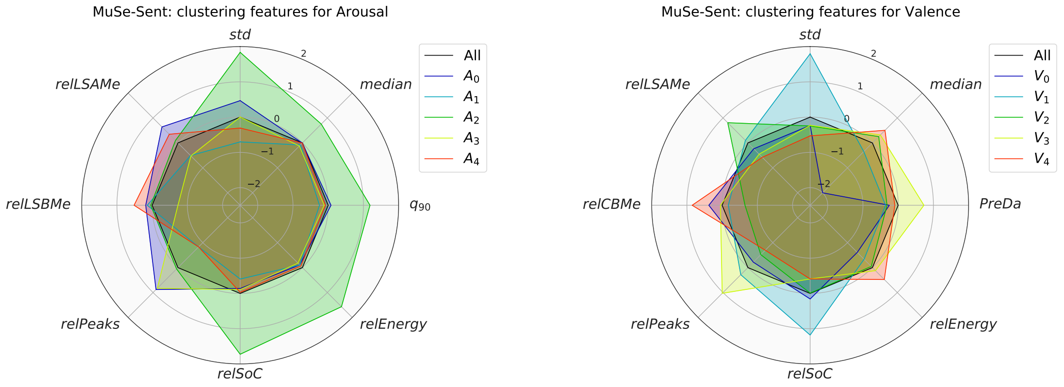
As for emotion recognition, MuSe-CaR serves as the testing bed. However, the focus is only on segments with an active voice, because it is a well-known fact (see Section 2.3) that a contextual understanding can better be understood from linguistics, such as sentence transcripts in the case of videos, than from other modalities. In addition, to avoid highly fragmented short segments, adjacent segments covering the same target are concatenated if the gap is smaller than 2 seconds. Therefore, one segment can comprise one or more sentences, and slightly more data is available, as for the previous task (see Table 5.1). The data partitioning follows the same logic as before.

5.1.2.1.1 Class Creation: Due to the lack of human labelling of emotions for the speaker-topic-based segments, a transfer from continuous emotion annotations to discrete summary emotion classes is needed before models can be trained to predict them automatically. Stappen et al. present the MuSe-Toolbox [25], which provides multiple options to accomplish this task.

The following sections summarise one possible naïve method for the subsequent prediction as well as a more complex approach. The first is the base for Multimodal Emotion-Target Sub-challenge (MuSe-Topic) and the second for Multimodal Sentiment Sub-challenge (MuSe-Sent) tasks:

I. **MuSe-Topic Naïve Intensity Classes:** In [26], Stappen et al. craft classes by averaging the annotations across a segment and transforming the summary score into classes. This can be interpreted as intensity classes within the emotion distribution. The classes are calculated by first taking the mean value of the gold standards to aggregate them to a regression estimate on the temporal axis. As in Section 5.1.2 I., the EWE gold standard is used for annotation fusion. The regression estimate is transformed into a class label depending on the class boundaries set. The class boundaries are naïvely specified based on an equal distribution of the average values in three low-, medium-, and high-intensity classes. For this, two class thresholds are chosen, leading to a balanced number of segments that fall into each class (33.3 % each), as depicted in Table 5.5.

II. **MuSe-Sent Learnt Emotion Classes:** A major shortcoming of naïve classes is the loss of information regarding the temporal changes occurring within a segment, as only the mean value of the annotation segment is used. To address this, Stappen et al. propose an evolution of MuSe-Topic in [27]. Instead of the emotion classes being statically selected based on the class distribution, they are dynamically constructed by extracting time series, hand-crafted representations (see Section 4.3) expressing temporal changes and forming cluster classes using unsupervised clustering methods. An in-depth explanation of this concept is given in Section 4.3. By running preliminary experiments using the MuSe-Toolbox [25], the following final settings are chosen by monitoring the provided qualitative and quantitative measures (as described in Section 4.3.2), resulting in five representative classes for each dimension: 1) as time series representations (median, standard deviation, percentile {10, 90}, relative energy, relative sum of changes, relative number of peaks, relative longest strike {below, above} mean, and relative count below mean) for arousal, and the same representations for valence plus the mean, percentile {5, 25, 33, 66, 75, 95}, and the percentage of reoccurring data points to all data points. They are calculated relative to the segment length from the RAAW gold standards of each segment. 2) By using the provided dimension reduction function, the representations are reduced to five components by a Principal Component Analysis (PCA). Based on these, the K-means algorithm [149] is selected for valence and 3) a Gaussian Mixture Model (GMM) clustering [289] for arousal based on exhaustive evaluation runs by the toolbox. The clustering is only applied to the training partition to ensure it is applicable to new data from the same domain. Using the distance to the resulting cluster centres, the segment representations of the development and test partitions are labelled. The silhouette coefficients [295] (ranging from -1 to 1) for the arousal and valence dimensions are



(a) Arousal-derived

(b) Valence-derived

Figure 5.4: MuSe-Sent classes for (a) arousal and (b) valence. The figures show the standardised characteristics of the selected time series representations (“relative” means normalised by segment length) for each of the five classes. The abbreviations are as following: standard deviation (*std*), *median*, the 90th percentile (q_{90}), percentage of reoccurring data points to all data points (*PreDa*), relative energy (*relEnergy*), relative sum of changes (*relSoC*), relative number of peaks (*relPeaks*), relative count below mean (*relCBMe*), relative longest strike below mean (*relLSBMe*), and relative longest strike above mean (*relLSAMe*). Figure taken from Stappen et al. [27].

0.19 and 0.10, respectively. It may be difficult to obtain a notably higher coefficient given the sensitivity of the measure to varying cluster densities [296], which are naturally caused by the application of PCA as it structures the data along orthogonal axes. Hence, the created classes expressing the emotional changes of a segment cannot simply categorised as low, medium, and high. Instead, the provided visualisations and statistical measures (correlations, distribution statistics) give guidance for interpretation. Figure 5.4 shows the most distinctive relative representations for valence ($V_{\#}$) and arousal ($A_{\#}$). The number (#) is meaningless and just given for structural reasons. A small amount of corrupted data points are removed, so that the classes cover slightly fewer segments in the training and development partitions compared to Table 5.5.

5.1.2.1.2 Prediction: As a standard measure of ML classification tasks [19], (macro) F1-score (F1) (see Section 3.2) is reported and the primary focus of the performance analysis. For MuSe-Topic, the Unweighted Average Recall (UAR) is further reported as it is also widely used in speech recognition tasks [8].

Table 5.5: Distribution of the naïve intensity classes of valence and arousal on the train(ing), devel(opment), and test set. Note: For technical reasons, some segments had to be excluded after segmentation from the development set of Arousal. Table adapted from Stappen et al. [26].

	Valence				**Arousal**			
	train	devel	test	total	train	devel	test	total
low	1432	434	384	2250	1433	453	343	2229
medium	1483	486	435	2404	1460	416	432	2308
high	1398	466	441	2305	1420	415	485	2320
$\sum$	4313	1386	1260	6959	4313	1284	1260	6857

Table 5.6: Distribution of the five valence and arousal classes across train(ing), devel(opment) and test partitions. Table adapted from Stappen et al. [27].

Valence					**Arousal**				
#	train	devel	test	total	#	train	devel	test	total
V_0	528	71	89	688	A_0	612	249	178	1039
V_1	552	159	277	988	A_1	534	135	194	863
V_2	1178	458	378	2014	A_2	312	96	53	461
V_3	1112	405	271	1788	A_3	1255	388	448	2091
V_4	837	242	245	1324	A_4	1494	467	387	2348
$\sum$	4207	1335	1260	6802	$\sum$	4042	1335	1260	6802

5.1.2.2 Experimental Setup

While the experiments are conducted from two viewpoints, the experimental settings for both are described together because of the common focus on the same type of prediction and, thus, theoretical suitability.

5.1.2.2.1 Feature Sets: Even though a high sampling rate of representations is not necessarily required for the classification task with one prediction per segment, this fine granular input enables time-sensitive modelling. For the extraction of the audio representations (Deep Spectrum, eGeMAPS, VGGish), the audio track is converted to mono (16 kHz with 16 bit) and normalised to -3dB. To calculate the 128 mel frequency bands of Deep Spectrum, the mel spectrograms are extracted using a Hanning window with 32 ms with an overlap of 16 ms. The other representations (Xception,VGGFace) are assumed to have one extraction every 250 ms as before. The only exception are the text-based representations (FastText, BERT), which are imputed and aligned per word. To align the representations meaningfully with each other and across modalities, the same logic is used in this chapter as is described in detail in Section 5.1.1.2. Due to the unimodal use of SenticNet representations, the sampling rate is the same regardless of the type.

5.1.2.2.2 Architectures: Except for LSTM-RNN (see Paragraph 5.1.1.2.2), three new architectures are introduced for this task.

LSTM-RNN-based: The architecture and hyperparameter search are almost identical to the LSTM-RNN emotion recognition models in Paragraph 5.1.1.2.2. The primary difference is the output of the prediction layer, which, instead of predicting a sequence of regression points, predicts only one label to classify the entire sequence. For this, the prediction layer of **LSTM-SA** is replaced by a global max-pooling layer. The pooling operation is applied over the concatenated sequence of the top attention layer, providing the logits of each class prediction. Furthermore, the set of each sequence fed into the LSTM is padded or cropped to 500 time steps. For the **LSTM-RNN** and **MHA-LSTM**, the sequence of hidden vectors from the final LSTM-RNN layer here only outputs a single value per prediction target. The range of learning rates is changed to $lr \in \{0.001, 0.005, 0.01\}$ based on first experiments. In addition, the CCC loss function for regression is replaced by a cross-entropy loss function for all models.

Multimodal Transformer (MMT): Another suitable architecture for emotion classification is the MMT architecture [183]. As explained in Section 3.2.5, the Transformer units allow for a deep, cross-modal fusion of representations. On similar tasks [297, 183], such architecture outperformed other advanced network architectures in predicting emotion classes. Similar to Yao et al. [183], the implementation evaluates a deep representation fusion method using cross-modal attention heads between bimodal feature sets (text to audio, text to video, audio to video) which are temporally fused to a trimodal representation. The chosen configuration utilises five cross-modal attention heads. The network is optimised using a *bs* of 16 and an Adam optimiser. To avoid overfitting, dropout is set on multiple levels: 0.25 dropout on the embedding, 0.1 dropout on each attention, and 0.1 dropout on the ReLU layers. Preliminary results found a learning rate of 10^{-3} optimal, leading the network to converge after 20 epochs.

Support Vector Machine (SVM): A Support Vector Machine (SVM) is a supervised ML algorithm for classification and regression, frequently used as a baseline model for emotion recognition [8, 84]. For this, the data points are represented in a vector space in which a hyperplane is projected to separate the binary-labelled data points. By using quadratic programming solvers, the largest possible margin between points of different classes is computationally determined. The calculation of this decision boundary utilises (non-)linear kernels and transfers the space to higher dimensions until a separation is feasible. Support vectors then flank the optimal hyperplane in the higher dimensional space on both sides of the decision boundary. This process is repeated until an optimal separation is achieved. The method is relatively memory efficient as the support vectors only require a portion of the training data and is particularly effective in high-dimensional input data, even if the number of dimensions is higher than the number of samples. However, the computational and memory requirements increase rapidly with the amount of data and, hence, the training vectors required. For computational optimisation and to avoid overfitting, a complexity parameter C is set. With increasing influences of noise, this parameter is lowered and serves for regularisation. One-against-one or one-against-all strategies, in which classes are combined, or several SVMs are trained, make it necessary to use them for multi-class problems. Since SVMs cannot work with continuous-time feature extractions, these are condensed by averaging.

A linear SVM is employed for the MuSe-Topic task, optimising $C \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ up to 1 000 iterations on the training set and validating the results on the development set. For the best C value, the model is then retrained on a concatenated version of the training and development set to measure the performance on the data points of the test set. This training procedure is used for SVM training only.

SENtic Sentiment Analysis Learner (SenSA): There are several ways to represent text in the form of high-contextual representations, for example SenticNet-based Learning (SNL), as discussed in Sections 2.3 and 3.1.2. However, only SenticNet represents a natural hierarchy of logical concepts of the real world through the automatic extraction of subsymbolic terms and their ordering and arrangement. In prior work, Stappen et al. [56] introduce the method SenSA, which extracts sentics, primary and secondary mood tags, and semantics from the transcription, enhances them into sentence- or segment-level representations, and enables the prediction of classes.

In the data cleaning process, stop words, for example, sentence conjunctions, personal pronouns, and articles, are removed from the text snippets. The system dynamically separates the cleaned text into n-grams, applies the SenticNet application programming interface (versions 5 (SenSA5) and 6 (SenSA6)) to extract n-gram-level representations and fuses them back together to receive sentence and segment representations. The fusion process has to take the different sequence lengths n of the n-grams $\bar{n}_s = [n_1, ..., n_n]$ as well as the segment s into account. The resulting sequence m of concepts $\bar{c}_s = [c_1, ..., c_m]$ is also of varying length and has to be compressed into a single embedding vector representation h_s to incorporate the entire context. In its simplest form, h_s is equal to a concatenated n-hot encoded vector of discrete semantics and sentic concepts. For continuous-valued extractions of intensities (e. g. , mood tags, polarity), n and m are step-wise averaged considering the normalised values based on the respective length. The resulting representations can then be used for emotion classification by an SVM.

The linear SVMs, as explained above, are trained, tuning the C value from 10^{-5} to 1 on the development set over up to 10 000 iterations. As before, the configuration yielding the best results on development is applied to test.

5.1.2.3 Results

First, suitable architectures are described for the naïve intensity classes MuSe-Topic, beginning with the baseline model selection from the earlier publication Stappen et al. [26], and the challenger text-only SenSA models, whose concepts are explained in Sections 2.3 and 3.1.2, from Stappen et al. [56]. This is followed by the results of the LSTM-RNN models predicting the learnt emotion classes MuSe-Sent.

5.1.2.3.1 MuSe-Topic: For the classification tasks MuSe-Topic, an overview of the results is given in Table 5.7. As can be seen, some of the prediction results of the naïvely created classes fall below the random level of 33 %.

Baselines: Both unimodal baseline approaches achieve the best results on the test set using vision representations (Xception and VGGFace). The SVM outperforms LSTM-SA, yielding 37.94 % for both F1 and UAR on the test set for valence prediction. Using the VGGFace representations, a F1 of 42.46 % and an UAR of 43.07 % on the test set is reached for arousal prediction. None of the models give compelling results when solely trained on text representations.

Comparing the two multimodal fusion models, LSTM-SA and MMT, the MMT reaches stronger results. Here, the MMT, with 39.93 % F1 and 40.52 % UAR on the test set, achieves the best overall results taking FastText, eGeMAPS, and Xception as inputs. For arousal prediction, the MMT, with the same configuration, demonstrates the best results among all fusion combinations, but is slightly outperformed by the SVMs trained with vision representations.

Table 5.7: Reporting arousal and valence for **MuSe-Topic** (using EWE annotation fusion) in Unweighted Average Recall (UAR) and F1 on the devel(opment) and test partitions. Audio feature sets: Low-level Descriptors (LDD), eGeMAPS (Ge), Deep Spectrum (DS), and VGGish (VG); vision features sets: GoCARD (Go), VGGFace (VF), and Xception (X); and text feature set FastText (FT) are fed into the models. Furthermore, high-level text concepts: sentics (SE), mood tags (MO), and polarity (PO) are evaluated. Furthermore, all vision features (aV) are utilised by LSTM-SA. The features are aligned to the label timestamps. The by-chance level is 33 %.

Approach	**Modality**	**Feature(s)**	**Valence**				**Arousal**			
			F1		**UAR**		**F1**		**UAR**	
			devel	test	devel	test	devel	test	devel	test
Official Baselines [26]										
Unimodal										
LSTM-SA	A	DS	34.17	34.60	34.07	35.00	38.03	**37.54**	38.43	36.78
		eG	33.26	34.44	32.16	33.94	34.39	33.33	34.44	32.87
	V	X	36.21	**36.83**	35.75	**36.61**	40.38	35.16	**40.51**	34.87
		aV	35.61	34.92	35.10	34.41	**38.11**	34.21	38.26	35.39
	T	FT	**38.41**	36.19	**37.75**	36.22	35.15	34.92	35.78	**37.10**
SVM	A	DS	34.08	34.29	33.21	34.07	41.35	42.30	40.18	40.18
		eG	36.33	33.10	34.79	34.13	43.52	34.37	42.27	33.43
	V	X	**38.28**	**37.94**	**37.09**	**37.94**	46.22	41.35	**45.25**	40.52
		VG	37.08	32.94	37.01	32.63	46.44	**42.46**	45.21	**43.07**
	T	FT	37.90	36.43	36.00	35.37	45.17	38.25	44.53	39.67
Multimodal										
MMT	A+V+T	FT + eG + X	38.28	**39.92**	37.62	**40.52**	41.87	37.30	40.83	37.87
		FT + eG + VG	37.38	32.78	38.19	32.53	**47.12**	**41.19**	**45.55**	**39.01**
		FT + eG + AU	36.93	39.92	37.35	39.57	43.15	34.76	41.88	34.87
		FT + eG + OP	**39.48**	38.81	**39.17**	38.64	38.88	37.70	38.95	38.10
LSTM-SA	A+V+T	eG + FT + aV	36.06	37.14	35.20	37.14	39.92	35.16	40.44	34.76
Post-Challenge Models [56]										
SenSA5	T	SE	35.66	35.16	34.65	35.17	33.56	33.73	34.51	33.69
		MO	33.78	37.86	33.53	38.04	**36.78**	35.79	**36.30**	33.50
		SE + MO	34.38	37.78	34.22	37.89	35.96	35.63	35.49	33.60
		SE + MO + PO	34.53	38.41	34.40	38.55	35.51	36.51	35.15	34.54
SenSA6	T	SE	35.51	36.59	34.77	36.72	32.43	35.16	33.45	35.91
		MO	**38.65**	36.83	**38.28**	37.57	35.66	38.02	35.33	35.86
		SE + MO	38.13	38.57	37.89	**38.90**	36.18	38.02	35.68	35.90
		SE + MO + PO	37.68	**38.65**	37.54	38.88	35.96	**38.33**	35.50	**36.12**

SenSA: Combining all the constructed SenticNet representations (sentics, mood tags, polarity), the results achieve 38.65 % F1 on the test set predicting valence and 38.22 % F1 on the test set predicting arousal utilising text-only representations. In terms of F1, they outperform the text baseline. The concepts from SenticNet-6 show slightly better results to those from SenticNet-5. By fusing sentics, mood tags, and polarity, the model exceeds the LSTM-SA and SVM using FastText for valence prediction by an average of 3 % for UAR and F1. Furthermore, the mood tags representations perform

Table 5.8: Reporting arousal and valence for **MuSe-Sent** in F1 score across five classes. Feature sets tested are Deep Spectrum, VGGish, and eGeMAPS for audio; Xception, VGGFace and FAU for video; and BERT for text. All utilised features are aligned to the label timestamps by imputing missing values or repeating the word embeddings. Table is taken from [27]. The by-chance level is 20 %.

	LSTM-RNN			
Features	**Valence**		**Arousal**	
	devel	test	devel	test
Audio				
Deep Spectrum	30.23	27.26	33.52	33.16
VGGish	30.76	25.08	36.05	31.66
eGeMAPS	32.93	25.80	36.04	31.97
Video				
Xception	30.40	28.74	35.16	31.14
VGGFace	32.29	28.86	34.57	31.32
FAU	31.37	27.38	35.21	31.43
Text				
BERT	32.68	31.90	**38.27**	30.63
Late Fusion				
best A + V	**32.96**	27.92	37.72	**35.12**
best A + T	30.15	30.29	37.63	32.87
best V + T	30.17	**32.91**	37.51	32.82
best V + A + T	30.37	31.01	36.72	33.20

better than the conceptual sentics, with all models benefiting from fusion. Conversely, in predicting arousal, the text-embedding baseline outperforms them.

In summary, the models emphasise the importance of inter- and intra-modality fusion. In the unimodal setting, they contradict the common view that valence is more predictable by textual representations and arousal correlates more strongly with audio representations (see also Section 5.1.1). The latter can be attributed to the generally low level of the results caused by the difficulty of the task and the trivial generation of labels. The contextual high-level SenSA results are only of limited value, although the results for valence are considerably more plausible. Their usefulness, e. g. , as supplements in fusion with other low-level representations in a multimodal context, could be given, but remains to be investigated in more depth.

5.1.2.3.2 MuSe-Sent: The results for predicting the learned class labels (MuSe-Sent) can be found in Table 5.8. It can be seen that the level of results obtained are relatively higher than before, compared to the by-chance level of 20 % (5 classes).

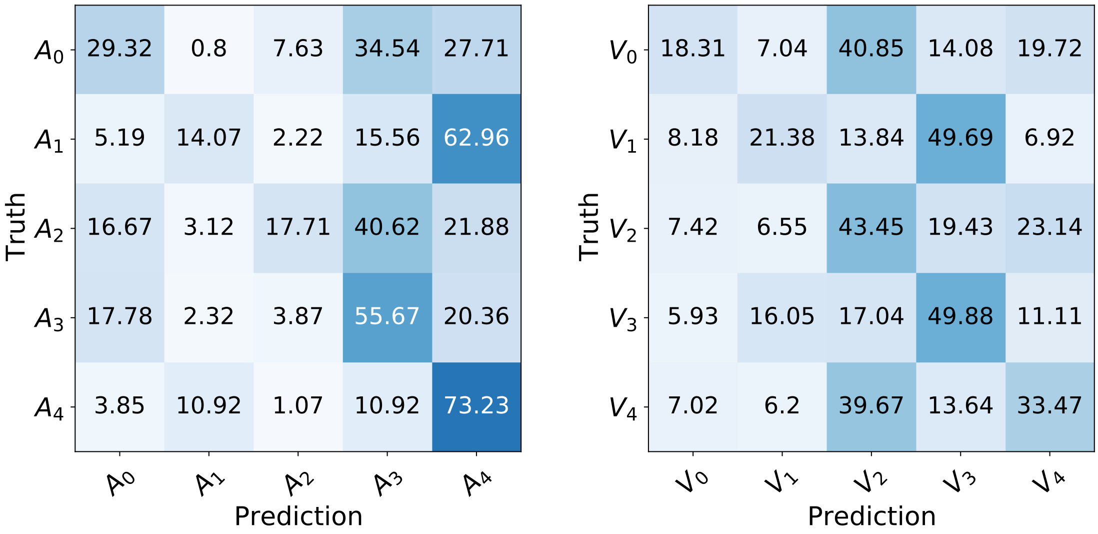
Unimodal: With regard to the baseline model, LSTM-RNN, on unimodal representations, the textual representations for valence and the audio representations for arousal show the most compelling results. For valence prediction, the BERT representations score 32.68 % F1 on the development set and 31.90 % F1 on the test set. This is followed by the visual representations VGGFace with 28.86 % and Xception with 28.74 % F1 on the test set. Leading to the highest F1 for arousal with 33.52 % on the development and 33.16 % on test set are the audio-based DEEP SPECTRUM representations, closely followed by VGGish with 31.66 % on the test set.

The best unimodal setups for each prediction target are mapped in the relative confusion matrices. Figure 5.5 shows that valence class V_2 is frequently predicted for data points from classes V_0 and V_4, and class V_3 is frequently misclassified for data points from class V_1. With arousal, there is some confusion with classes A_1 and A_2. Classes A_3 and A_4 also classify many data points from other classes. A connection with the particularly salient representations (deviating from the average behaviour) from Figure 5.4 is not directly evident. One implication of this could be that the time series of representations are only reflected in the modalities to a limited extent. Without a direct link, no generalisable patterns for recognition can be obtained. The best performing hyperparameters in the experiments vary greatly depending on the representation and target. In general, bidirectional layers and at least two layers have shown to be slightly advantageous.

Multimodal: The bimodal late fusion performs slightly better on both prediction targets, but can only occasionally reach new peaks. For example, 32.91 % F1 on the test set shows a marginal improvement of 1 percentage point when VGGFace and BERT are combined for predicting valence. For arousal prediction, the combination of audio (VGGish) and video (FAU) achieves slightly improved scores of almost 2 percentage points to 35.12 % F1 on the test set.

5.1.2.4 Conclusions

The main aspect of the evaluation was the nature of the target itself, being transformed from the time- and value-continuous annotation into a discrete class per segment (**RQ-1b**). Overall, for both directions, effective systems could be developed with results above chance-level. The peak results of 39.92 % F1 for valence and 42.26 % for arousal at a 33% chance-level indicates a severe loss of information in the case of the naïvely created classes. A logical assumption is that this loss is caused by the heavily reduced granularity of the target (discrete classes instead of continuous points) and the temporal compression (one target instead of a sequence of targets). The proposed approach of elaborately constructed classes aimed to mitigate these weaknesses by expanding the granularity in terms of more classes and by



(a) Arousal (b) Valence

Figure 5.5: Relative confusion matrices over the 5 (a) arousal and (b) valence classes for the MuSe-Sent sub-challenge. The LSTM baseline model with hyperparameters of $n = 4$ (bidirectional), $h = 128$, and $\alpha = 0.001$ using the eGeMAPS feature set for arousal, and for valence, the BERT representations with a unidirectional model setting of $n = 2$, $h = 64$ and, $\alpha = 0.01$ is used. Figure taken from Stappen et al. [27].

including features that describe the temporal course of an annotation segment. Furthermore, very little human involvement is necessary since the classes are proposed automatically, driven by unsupervised machine learning methods. Classes generated this way exhibit plausible characteristics. However, robust classification proved to be difficult. The best results showed 32.91 % F1 for valence and 35.12 % F1 for arousal at a 20 % chance-level. In terms of absolute F1, this represents a slight improvement over chance level, for example, 15 % instead of 9 % for valence. A robust, categorical representation of a summarised sequence of emotion targets would be very convenient for human interpretation, and another reason to build on continuous emotion annotations. Given the consistencies in the findings across both experiments, further steps are required to refine the procedure, making it more generally applicable. One idea could be human annotation of both forms to learn a mapping directly, or cross-corpus class formation on large datasets to find more robust, generalisable characteristics in this data-driven process. Additionally, further efforts are also necessary to make data-driven classes more interpretable.

A large set of uni- and multi-modal architectures was proposed (**RQ-3**). For the naïve created classes, the SenSA architecture, developed based on knowledge graphs, showed improvements compared to conventional word embeddings. This may relate to the known ability of the underlying knowledge-base features to extract high-level emotional and thematic concepts from thematically closed segments. The SVM architecture using only vision representations (VGGFace) demonstrated the best results for arousal with 42.26 % F1. This was only beaten by the very complex MMT achieving 39.92 % F1 for valence, but fell slightly behind on arousal with 41.19 % F1. The elaborately constructed classes show peak values of 31.90 % F1 for valence using BERT and 33.16 % F1 for arousal using Deep Spectrum on the test set of the 5-class problem. The multimodal fusion was slightly superior, with peak values of 32.91 % F1 for valence and 35.12 % F1 for arousal. Regarding the effectivity of the feature sets and modalities, the results of the second series of experiments are well in line with the outcome of the value- and time-continuous prediction in Section 5.1.1. For the first series, it can be speculated that the isolated superior result of the unimodal SVM was due to generalisation achieved by chance on the grounds of the less efficient class formation procedure and closely spaced results.

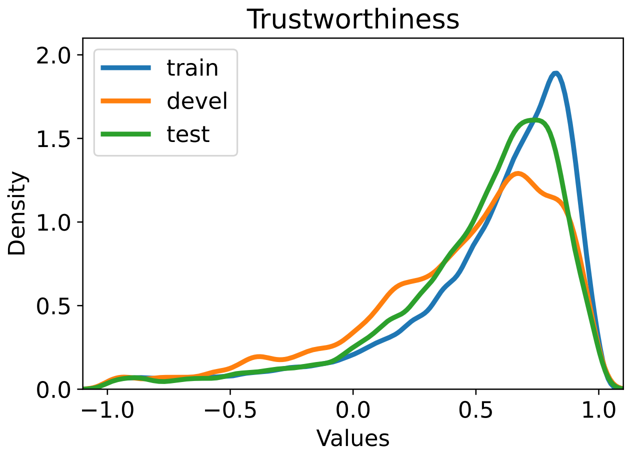


Figure 5.6: Density distribution using 35 equal-width bins of the partitions train(ing), (devel)opment, and test for the continuous annotations of MuSe-Trust. The distributions between the individual partitions are very similar, however, the distribution is skewed towards the positive end. Figure adapted from Stappen et al. [26].

5.1.3 Trustworthiness Recognition

5.1.3.1 Characteristics

In Stappen et al. [26] a completely novel, continuously annotated dimension, trustworthiness, was proposed that attempts to capture the subjective feeling about how objective information (see Section 2.1) is conveyed and perceived. For this dimension, the data selection from MuSe-CaR differs modestly from the emotion regression and classification tasks (see Sections 5.1.1 and 5.1.2) and was captured under the name Multimodal Trustworthiness Sub-challenge (MuSe-Trust). The non-product segments of a video, previously excluded to prevent introducing bias on the task targets, are now included. Those elements, such as an advertisement, might be an essential factor of perceived trustworthiness. The annotation covers a total of 35 :51 hours of relevant video material (see Table 5.1), is strongly left skewed as depicted in Figure 5.6, and EWE-fused by the MuSe-Toolbox [25] as before with MuSe-Wild.

5.1.3.2 Experimental Setup

The below depiction of the experiments is separated into setting up a baseline, as introduced in Stappen et al. [26], and reporting observations from the deeper exploration of the task, as described in Stappen et al. [24]. Since there has never before been an attempt to quantify trustworthiness in a sequential manner alongside a video stream, these in-depth experiments cover various aspects of modelling in greater detail. For example, the influence of the sequence length in data augmentation, the choice of the loss function, and individual network configurations are all evaluated. Furthermore, arousal and valence annotations alongside trustworthiness are learnt in a multitask fashion to explore the interaction between the established emotion dimensions and trustworthiness at a large scale. The same feature alignment approach is used as in Section 5.1.1. Like the other regression tasks, the CCC is evaluated on the development set after each epoch, and after the training, the best configurations are subsequently evaluated on the test set. The following architectures are used:

Baselines: As in the previous regression chapter (Section 5.1.1), the **LSTM-RNN**, the **LSTM-SA**, and **End2You** are utilised as baseline models using the same configuration and hyperparameter range.

Deep Trust Multihead Attention Network (DeepTrust): To show the impact of the various experimental aspects, one architecture is put in the centre of the analysis. Stappen et al. [24] named this architecture DeepTrust when it was first proposed and that name will be used here as well; however, it is in its core identical to the MHA-LSTM architecture. It has proven to be very robust and has achieved the best results in the two previous emotion tasks. Since the hyperparameter space allows endless combinations, only the most relevant hyperparameters are changed. Step by step evaluations are performed for the effects of, for instance, augmentation, the number of heads, the choice of loss functions, the type and number of layers, fusion, and multitask learning (trustworthiness together with arousal and valence). Thereby, one set of hyperparameters is searched within a specific range to consider cross-interactions, and one set remains static. The static initial parameters are $ws = 200$, $hs = 100$ for augmentation and $n = 64$ hidden neurons of the bidirectional LSTM. For each experiment, the hyperparameters are optimised: $at \in \{2, 4, 8\}$ heads, $lr \in \{0.0001, 0.001, 0.005\}$, and $bs \in \{512, 1024, 2048\}$. All experiments optimise the network for 100 epochs using Adam as the optimisation algorithm of choice. The model reduces the learning rate by 0.1 if it reaches a plateau for more than ten epochs, and stops training if no further loss reduction is reached.

5.1.3.3 Results

First, different experimental settings are evaluated using the DeepTrust architecture. The best configuration is then compared to the other baseline approaches.

Table 5.9: Results of unimodal DeepTrust (**MuSe-Trust**) experiments, reported in Concordance Correlation Coefficient (CCC) on the devel(opment) and test set using feature sets for text (T): FASTTEXT (FT) and BERT; audio (A): VGGISH and EGEMAPS; vision (V): VGGFACE and FAU. Table taken from Stappen et al. [24].

Modality	**Feature sets**	**devel**	**test**
T	FASTTEXT	.4559	.4782
	BERT	.5624	**.5539**
	BERT+ FASTTEXT	**.5648**	.5478
A	EGEMAPS	.3921	.1220
	VGGISH	**.5376**	**.4035**
	VGGISH+ EGEMAPS	.4751	.2402
V	FAU	.3675	**.3623**
	VGGFACE	**.4000**	.2802
	VGGFACE+ FAU	.3936	.3298

5.1.3.3.1 Unimodal results: The first experiment evaluates the capabilities of the individual audio, text, and video feature sets. Table 5.9 gives an overview of the most effective representations. In particular, the textual BERT representations using $at = 4$, $bs = 512$, and $lr = 0.005$ and the deep acoustic VGGish representations (same hyperparameters) stand out for their strong performance, while the vision representations tend to underperform. In addition, the extracted FAU (same hyperparameters) score very weakly in predicting trustworthiness, whereas the VGGFace using $at = 4$, $bs = 1024$, and $lr = 0.005$ does better on the development set but generalises poorly to the test set. Early fusion of the representations within the modality before feeding them into the network yields a slight advantage only for text (test set). The following experiments use only the most convincing feature set per modality (BERT, VGGish, and FAU).

5.1.3.3.2 Segmenting: Increasing the amount of data by segmenting and adding slightly overlapping segments has proven to be an effective technique for improving results [17]. To determine an optimal ratio, the number of sequence steps *ws* and hop sizes *hs* is varied. Table 5.10 illustrates that a higher value of *ws*, hence a longer context (*ws* = 750), leads to more effective training of the task. The experimental results on the overlap *hs* are rather ambiguous and produce both better and worse results depending on *ws*. The absence of any overlap (*ws* = *hs*) yields a better generalisation only for very short $ws = 100$. A strong interdependence of both factors seems likely. From $ws > 200$ onwards, a larger overlap with increasing sequence length seems useful, from which a sound reference value of approximately *hs* = 0.3–0.5 *ws* can be derived. As the duration of the sequence increases, both the memory requirements and the training time expand. The maximum supported

Table 5.10: Results of several augmentation hyperparameter combinations of DeepTrust to predict MuSe-Trust on the devel(opment) and test set, reported in Concordance Correlation Coefficient (CCC). Table taken from Stappen et al. [24].

steps		**BERT**		**VGGish**		**FAU**		**Ø**	
ws	*hs*	**devel**	**test**	**devel**	**test**	**devel**	**test**	**devel**	**test**
750	750	.5641	.5540	.4274	.4344	.3775	.3719	.4563	.4534
750	500	.5739	**.5747**	.5604	**.4699**	.3671	.3705	.5005	.4717
750	250	.5889	.5693	.5386	.4686	.4305	**.4843**	.5193	**.5074**
200	200	.5512	.5245	.5566	.4752	.3614	.2710	.4897	.4236
200	150	.5500	.5533	.5517	.3034	.3558	.3508	.4858	.4025
200	100	.5624	.5539	.5376	.4035	.3675	.3623	.4892	**.4399**
200	50	.5282	.5160	.5440	.4081	.3319	.1820	.4680	.3687
100	100	.5167	.5128	.5064	.4445	.3711	.3709	.4647	**.4427**
100	50	.5312	.5068	.5369	.2918	.3816	.3294	.4832	.3760
100	25	.5233	.5216	.5264	.2517	.3642	.2911	.4713	.3548

sequence length on the 32 GB GPU machine is $ws = 750$, which means performance has to be sacrificed for usability.

5.1.3.3.3 Heads: Table 5.11 shows that the number of heads relates to the modality or feature set dimension size. Four heads achieve the best results for text (BERT: 768 dimensions), .5539 CCC; 16 heads for audio (VGGish: 512 dimensions), .4592 CCC; and two heads for vision (FAU: 28 dimensions), .3774 CCC (all on the test set). Even on average, no consistent pattern emerges. Two, four, and sixteen heads produce comparably good on the development set, with a slight advantage for more heads. However, as with the sequence length, more heads require considerably more computing resources.

Table 5.11: Evaluation of the number of heads using DeepTrust on the devel(opment) and test set reported in Concordance Correlation Coefficient (CCC). Table taken from Stappen et al. [24].

heads	**BERT**		**VGGish**		**FAU**		**Ø**	
	devel	**test**	**devel**	**test**	**devel**	**test**	**devel**	**test**
2	.5698	.4745	.5375	.4368	.3561	**.3774**	.4878	.4296
4	.5624	**.5539**	.5376	.4035	.3675	.3591	.4892	**.4388**
8	.5539	.5454	.4035	.2671	.3623	.3280	.4399	.3802
16	.5693	.5112	.5619	**.4592**	.3548	.3352	.4953	.4352

5.1.3.3.4 Loss: To fairly evaluate the best loss function (the loss and metric of the previous experiments was CCC), the Pearson Correlation Coefficient (PCC) and the Root Mean Square Error (RMSE) are also reported in Table 5.12. When BERT and VGGish representations are used, both correlation-based metrics (CCC and PCC) show better results with the CCC than with the L1 [193] and Mean Square Error (MSE) loss. However, the VGGish representations show a contrary picture, with the RMSE leading to particularly compelling results.

Table 5.12: Evaluating CCC, L1, and MSE loss functions of DeepTrust and reporting Concordance Correlation Coefficient (CCC), Pearson Correlation Coefficient (PCC), and Root Mean Square Error (RMSE) as metrics on the devel(opment) and test set. Table taken from Stappen et al. [24].

loss	metric	BERT		VGGish		FAU		Ø	
		devel	**test**	**devel**	**test**	**devel**	**test**	**devel**	**test**
CCC	CCC	.5624	.5539	.5376	.4035	.3675	.3623	**.4892**	**.4399**
	PCC	.5684	.5998	.5384	.4421	.3770	.4301	**.4946**	**.4907**
	RMSE	.3652	.3485	.3867	.4199	.4693	.4780	.4071	.4155
L1	CCC	.5076	.5211	.3650	.2408	.3678	.3407	.4135	.3675
	PCC	.5432	.5712	.3877	.3270	.3724	.3728	.4344	.4237
	RMSE	.3595	.3409	.4031	.3978	.4350	.3690	.3992	**.3692**
MSE	CCC	.5215	.5433	.3932	.4094	.3537	.3243	.4228	.4257
	PCC	.5455	.5570	.3932	.410	.3584	.3498	.4324	.4392
	RMSE	.3584	.3470	.3932	.4160	.4407	.3796	**.3974**	.3809

5.1.3.3.5 Model: The performance of a neural network is strongly influenced by the depth (number of layers) and width of the layers. Table 5.13 shows the results of experiments with different configurations of the two main mechanisms (LSTM-RNN and MHAL). Overall, an architecture combining an MHAL and a bidirectional LSTM seems to lead to solid results. For BERT, this can be further improved with a second MHAL.

Table 5.13: Experiments with different DeepTrust model configuration. Results of MuSe-Trust reporting Concordance Correlation Coefficient (CCC) on the devel(opment) and test set. Table taken from Stappen et al. [24].

network	BERT		VGGish		FAU		Ø	
	devel	**test**	**devel**	**test**	**devel**	**test**	**devel**	**test**
MHAL	.3117	.3248	.4230	.3150	.3677	.3351	.3675	.3250
LSTM	.5165	.5170	.5441	.3771	.3270	.2513	.4625	.3818
MHAL+LSTM	.5423	.5526	.5368	.2248	.3609	.3047	.4800	.3607
MHAL+2 Bi-LSTM	.5456	.5504	.5259	.3688	.3642	.2973	.4786	.4055
MHAL+Bi-LSTM	.5624	.5539	.5376	**.4035**	.3675	**.3623**	.4892	**.4399**
2 MHAL+Bi-LSTM	.5548	**.5762**	.4918	.3818	.3645	.3447	.4704	.4342
2 MHAL+2 Bi-LSTM	.5410	.5344	.4942	.3233	.3553	.3523	.4635	.4033
3 MHAL+3 Bi-LSTM	.5437	.5089	.4977	.3376	.3455	.3104	.4623	.3856

5.1.3.3.6 Multimodal fusion: Modality fusion displays a mixed picture in terms of performance (see Table 5.15). Having a positive effect on predicting the development set in almost all combinations, it generalises poorly for some models (see the fusion of BERT and VGGish). Others, such as the fusion of text and image representations, show improved results on the test set and outperform (.5880 CCC) all others, including all unimodal models.

5.1.3.3.7 Multitask learning: Besides merging multiple input representations, one can also simultaneously predict multiple outputs (targets) to assess whether a jointly learned representation enhances the ability to predict a single task. Multitask training, in which the model simultaneously learns to predict arousal, valence, and trustworthiness, outperforms the baseline models (see Table 5.14) by 0.2 on the development and almost 0.15 on the test set in terms of CCC. This shows that trustworthiness benefits from a jointly learned representation that is equally compelling for arousal and valence. This becomes even more noticeable when the importance of the loss of trustworthiness is increased (II.), and the other two modes are assigned a weaker learning signal.

Table 5.14: Evaluating multitask learning on trustworthiness (T), arousal (A), and valence (V) of MuSe-Trust task using the DeepTrust and reporting results on devel(opment) and test set in Concordance Correlation Coefficient (CCC). Configurations: (I.) equal loss weight of 0.33 (II.) 0.5 x trustworthiness, 0.25 x {arousal,valence}. Table adapted from Stappen et al. [24].

Configuration		**T**		**A**	**V**
Model	**Features**	**devel**	**test**	**devel**	**devel**
End2You-Multitask [26]	FASTTEXT+ VGGFACE+ A	3264	.4119	–	–
MHAL+LSTM-Multi (I.)	BERT+ VGGISH+ FAU	.5428	.5456	.4102	.4442
MHAL+LSTM-Multi (II.)	BERT+ VGGISH+ FAU	.5497	.5518	.4132	.4215

5.1.3.3.8 Best model: An overview of all results are given in Table 5.15. The baselines were drastically improved upon with an in-depth hyperparameter search of DeepTrust. The unimodal LSTM-SA falls far behind, with the best result achieved using the textual representations, leading to .2549 CCC on the test set. The multimodal version of the same architecture results in further deterioration of the test set results with a CCC of .2054. Here, the end-to-end baseline system using the raw audio signals alongside FastText and VGGish yield the best results with .4128 CCC on test. This setting is identical to the prediction of MuSe-Wild (cf. Table 5.2). However, all are outperformed by a large margin by the unimodal (as above), including at least one MHAL, DeepTrust architecture. Comparing those models with the baselines while using the same representations (eGeMAPS and FastText), one

Table 5.15: Reporting final results for the prediction of **MuSe-Trust** using the developed DeepTrust and the baseline models in Concordance Correlation Coefficient (CCC) on the devel(opment) and test partitions. Audio feature sets: Low-level Descriptors (LDD), eGeMAPS (Ge), Deep Spectrum (DS), and VGGish (VG); Vision features sets: GoCaRD (Go), VGGFace (VF), and Xception (X); and Text feature sets: FastText (FT) and BERT (BT) are fed into the models. Furthermore, the raw audio signal (RA) is used by End2You. The features are aligned to the label timestamps.

Approach	**Modality**	**Feature(s)**	**Trustworthiness**	
			devel	test
Unimodal Baselines				
LSTM-SA [26]	A	LLD	.2560	.1343
		DS	.0875	.0874
		Ge	.1576	.1385
	V	X	.1167	.1378
	T	FT	.2278	.2549
Multimodal Baselines				
LSTM-SA [26]	A+T	Ge + FT	.2296	.2054
	T+A+V	FT + Ge + aV	.1245	.1695
End2You [26]	T+A+V	FT + VG + RA	**.3198**	**.4128**
MultiFusion [263]	T+A+V	FastText+ DS + 2D	.3426	.3259
Post-Challenge: Best DeepTrust [24]				
early fusion	best V + A + T	BT + VGG + AU	.6241	.5073
	2xT + 2xA + V	BT + FT + VGG + eG + AU	.5445	.4998
late fusion	best V + A + T	BT + VGG + AU	.6075	.5796
	2xT + 2xA + V	BT + FT + VGG + eG + AU	**.6507**	**.6105**

notices a considerable improvement (e. g. , FastText doubles the result on test), which can therefore be attributed to the architecture as a whole. The best versions use late fusion. This increases results by almost 100 % to .6507 CCC on the development and by more than 50 % to .6105 CCC on the test set.

5.1.3.4 Conclusions

In this section, the experiments demonstrated that the novel dimension trustworthiness can be predicted and opened the door to a new way to automatically understand the trust mechanism in online videos (**RQ-1c**). In doing so, the baseline models achieved results up to .4128 CCC on the test set, which were gradually improved with careful tuning. On a DL architectural level, the integrated modules (loss, heads in the attention layer) of the proposed DeepTrust architecture exhibit similar improvements as for emotions, suggesting that this form of temporal modelling might helpful for other, related sequence-to-sequence tasks. Furthermore, longer sequences in the segmentation increased the results even more than previously with arousal and valence emotions. This may indicate that trustworthiness fluctuates less than emotions, or that a longer temporal context is more relevant to estimate credibility later in the course. This rigorous evaluation provides valuable insights for future model design for trustworthiness prediction.

The evaluation has led to numerous findings regarding the effectiveness of the individual modalities (**RQ-3**). Text achieves the strongest prediction performance, regardless of the representation. Thus, content seems to be the most important indicator of perceived trustworthiness. This is followed by audio representations, suggesting that the acoustic environment and how the content is communicated (e. g. , prosodically) are of value. The feature sets based on facial expressions (video), which are known to be another non-verbal communication layer, performed the worst but still achieved solid results. The best result was equal to arousal and valence prediction achieved by the late-fusion with trimodal fusion. By extending this with an additional text and audio representation each, hence unimodal and multimodal fusion, this result increased even further to .6105 CCC on the test set. Overall, the results suggest that even for trustworthiness, the multimodal perspective has advantages.

One phenomenon that should be investigated further is the high level of trustworthiness that appeared in MuSe-CaR (see Figure 5.6). This could be caused by selection bias due to the data selection of popular videos with mainly semi-professional actors, the domain, or the YouTube platform, which may not reflect the general collection of videos available on the internet. More data sets with such annotations are needed to be able to more precisely assess the applicability and the full potential of the new dimension.

5.1.4 Video Popularity

The previous two chapters have explored the influencing factors in modelling the subjective dimensions, namely arousal, valence, and trustworthiness, to achieve accurate prediction results. The following chapter shows that these subjective dimensions can estimate how popular a video will become. This research direction is inspired by the high relevance of video portals for collecting videos, in this case, review videos from YouTube, which are particularly suited for training MSA systems. Improving recommendation engines for platforms like YouTube is seen as a potential application of such systems, as discussed in Section 2.1.

5.1.4.1 Characteristics

Many methods have been developed to predict user engagement indicators. It is known that portraying certain emotion patterns by the actor, for instance, high fluctuation in arousal and trustworthy content, improves video attractiveness. Stappen et al. [224] are the first to deeply investigate the relationship between YouTube video popularity and trustworthiness time series, hand-crafted representations and predict cross-modal user popularity indicators as a regression task. Compared to previous research (see Section 2.1), both tasks are purely based on feature representations extracted from the subjective emotion annotations without using audio, text, or video modalities. The MuSe-CaR dataset (see Section 4.1), with its 600 hours of continuously annotated emotions and collected metadata, is the optimal testing bed for such a study. To also include the viewers' opinions expressed in comments and to combine video and text modalities in the empirical investigation, it is necessary to extend the MuSe-CaR collection, as described in Section 5.1.4.3.

With the linkage of the time series emotion representations with various user engagement indicators, including views, like/dislike ratio, and the sentiment of comments, this section aims to:

(I) **identify interpretable patterns** that correlate to user engagement and determine positive and negative effects of the individual representations, and

(II) **estimate** the value of user engagement indicators by predicting them with an interpretable model.

5.1.4.2 Experimental Setup

As can be seen from Figure 5.7, the proposed approach aims to reveal relationships between emotions (arousal, valence, trustworthiness), user-engagement-related metadata, and user

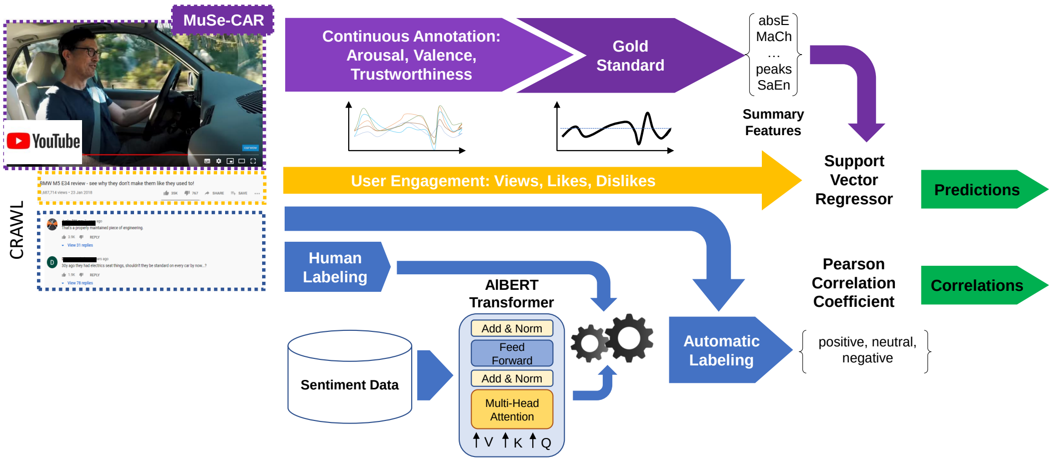


Figure 5.7: For the proposed approach to examine relationships and estimate the engagement, three core components are necessary: a) representations derived from continuous emotion and trustworthiness (purple), b) (semi-)automatic annotations of YouTube comments regarding the sentiment (blue), and c) the collected user engagement data (orange). Figure taken from Stappen et al. [224].

comments, as well as to predict these engagement indicators from emotion representations. As a starting point, predictive representations from the EWE fused gold standard annotation (see Sections 4.3 and 5.1.1) in **purple** are extracted using the procedures described in Section 3.1. Furthermore, the MuSe-CaR dataset needs to be extended (see Section 5.1.4.3) to incorporate the user engagement data (**yellow**) and the comments of each video (**blue**). To enable a comparison of the representations with the comments originally represented in unstructured textual form, it is necessary to transform them into structured data. The target structure first captures the sentiment of the comments, either positive, neutral, or negative, which can then be summarised. This is achieved by manually labelling some of the comments and automating the process of labelling for the remaining ones using a developed Transformer sentiment classifier (see Section 3.2.5). Finally, all three streams are brought together to conduct the defined experiments.

Support Vector Regressor (SVR): To predict the engagement regressors, an SVR is used. This is the regression version of an SVM as explained in Paragraph 5.1.2.2.2 and used in Paragraph 5.1.2.3.1. The SVR employs a linear kernel for interpretability. The experiments build upon the sensible crafted MuSe-Wild data partition (see Section 5.1.1), guaranteeing speaker-independence among other splitting criteria. The emotional input representations are standardised, but the targets are left in their original form to allow the interpretation of the measured Mean Absolute Error (MAE).

As before in Paragraph 5.1.2.2.2, the training procedure for each input-target combination has two steps: first, hyperparameter optimisation is performed for the C value $\in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ for up to 10 000 iterations, scoring on the development set. Second, with the best C value identified in step one, the model is retrained from scratch using a concatenated training and development set and scored on the hold-out test set.

Feature selection: Selection is done semi-automatically (cross-task) and automatically (task-specific), and combinations of representations that predict user engagement indicators are automatically identified.

For semi-automatic selection, correlations between the feature and the target variables are applied. These correlations are independent of the predictive target (selected across tasks). Only representations whose mean across all prediction tasks is between $-0.2 > r_{mean} > +0.2$ (minimum low positive/negative correlation) are selected.

Automatic selection is done by a brute-force methodology. This probes through all combinations of representations with $5 < k < k_{max-1}$ using a statistical, univariate regression test (f) whose linear F-test estimate is converted to a p-value. The test is used as a scoring function in which all input representations are evaluated task-specifically in various combinations:

$$score(f,y) = \frac{X_{k_i} - \bar{X_{k_i}} \cdot (y - \bar{y})}{\sigma_{X_{k_i}} \cdot \sigma_y}, \tag{5.1}$$

where k_i represents the index of the feature.

The representations with the highest k number are determined based on the p-value.

5.1.4.3 Data

The MuSe-CaR dataset is extended in two directions:

1. **Video Comments**

   a. **Collection:** The functionality of the YouTube crawler is broadened to be able to also collect YouTube comments based on the pre-collected video identifier. The selection was limited exclusively to parent comments while ignoring sub-comments. In total, 79 000 comments are retrieved. The number of likes and dislikes of these comments were also harvested. These are designed to signify the supposed agreement (“likes”) or disagreement (“dislikes”) of other users with the opinion expressed. By random sampling, 1 100 comments are chosen for human labelling. These comments allow quantifying the success of the automatic labelling model, as explained in the next section. Three Amazon Turk workers independently categorised the comments as positive, neutral, negative, and not applicable, reaching an average agreement of 0.47 inter-rater joint probability. A single ground truth was established by majority voting; less than a tenth of the comments were removed because no majority could be reached.

b. **Sentiment extraction:** To maximise the study ground of the cross-modal study, data collection must go beyond hand-categorised comments; hence, a system is developed that accurately and automatically assigns a sentiment label to every comment collected. To date, context-learning language Transformer networks (see Section 3.1.2) provide the best results in classifying text but require a large amount of data to produce robust results. Since sufficiently large datasets of YouTube comments are not available, multiple datasets from various domains are selected, modified, and merged to generally pretrain the system: *a)* 70 000 text snippets equally positive and negative from the Sentiment140 dataset [298], *b)* a mix of all sentiments consisting of more than 14 000 opinions of the US Airline Sentiment dataset [299] and *c)* the datasets utilised from 2013 to 2017 in the Semantic Evaluation challenge, a series of challenges for computationally classifying text into sentiments including but not limited to the domains of Twitter, SMS, and sarcasm [300] (more than 75 000 texts). The combined dataset has almost 150 k text snippets (60 k positive, 32 k neutral, and 56 k negative). After training with these data, the model is fine-tuned on the Amazon Turk-labelled MuSe-CaR sample set to reflect the writing styles and expressed opinions in the domain more closely.

As a DL architecture, A Lite BERT for Self-Supervised Learning of Language Representations (ALBERT) (see Section 3.1.2) is chosen (see further explanation in Section 2.2 and Section 5.2.2). For all experiments, the system operates on half-precision numbers (FP16), limiting the sequence length to 300, with shorter sequences padded and longer ones truncated, and the number of samples for each *bs* is set to 12 due to GPU memory limitations (32 GB). Data cleansing is applied to all texts, removing words starting with a “#”, “@”, or “http” as well as transforming emoticons to emoticon names. For pretraining, the data is stratified based on the class distribution and partitioned across training, development, and test sets (80-10-10). For both settings, the class weight is additionally allocated to each data point to mitigate the detrimental effects of class imbalance. The

Table 5.16: Relative sentiment distribution of the YouTube comments predicted by the developed sentiment labelling classifier along with examples. # indicates number. Table taken from Stappen et al. [224].

sentiment	**# comments**	**predicted [%]**	**example**
positive	26 032	33	"the metaphors are just flying like the raindrops in this video." #47620
neutral	28 518	36	"Are engines for F30 made in Germany?" #4
negative	24 494	31	"Poor review unfortunately, the microphone quality was very muffled..." #31

network is optimised using an Adam optimiser converged after three epochs using a lr of 10^{-5} and a warm-up ratio of *0.06* with an $\varepsilon = 10^{-8}$, and gradient clipping is applied at a value of 1.0 to avoid exploding gradients. The fully trained model yields an F1 of 81.13 % on development and 78.09 % on the test set. Further, when fine-tuned for one epoch with a reduced learning rate of 10^{-6} using the crawled YouTube comments, the model achieves 75.41 % F1 on the test set of the YouTube comment dataset. Table 5.16 illustrates the relative distribution across all classes and examples using this classifier to label the remaining, not manually labelled snippets. The distribution is quite balanced, with slightly more neutral predicted snippets.

2. **Video popularity indicators:** The video identifiers from the first crawl (see Section 4.1) also enable collecting additional information from each video. The number of views **(Vp/d)**, likes **(Lp/d)**, dislikes **(Dp/d)**, comments **(Cp/d)**, and likes of comments **(LCp/d)** are collected from the videos, where p/d indicates that each indicator is calculated on a per-day basis, given the crawling and the individual video upload date. Averaged over all videos, the statistical distributions are as follows (μ mean, σ standard deviation): Vp/d: $\mu = 863.88, \sigma = 2048.43$; Lp/d: $\mu = 9.73, \sigma = 28.75$, Dp/d: $\mu = 0.4125, \sigma = 1.11$; Cp/d: $\mu = 0.91, \sigma = 3.00$; and LCp/d: $\mu = 5.28, \sigma = 16.84$.

5.1.4.4 Results

5.1.4.4.1 Interpretable patterns: First, the correlations within the user engagement indicators themselves are examined. As shown in Figure 5.8, the four indicators (Vp/d, Lp/d, Dp/d, and Cp/d) are correlated. Also correlations are not necessarily transitive in the Euclidean plan, it can be an indicator that a correlation to any of these variables might

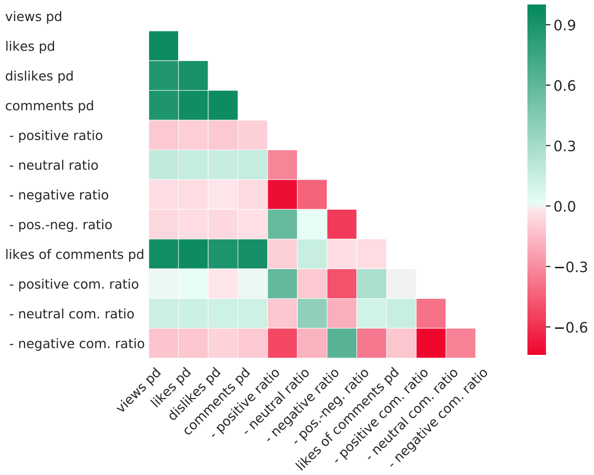


Figure 5.8: Pearson correlation matrix of indicators. All results are significant at a 0.01 level. Abbreviations: comment (com.); positive (pos.); negative (neg.); per day (pd).

imply a correlation to the others. It is also evident that individual scores cannot represent complex interactions. For example, Lp/d and Dp/d have a complementary instead of a contrary relationship (if one increases, the other increases), which is relativised, if one looks at the ratios (pos.-neg. ratio vs pos. ratio, neg. ratio). Therefore, observing the indicators needs to be considered in isolation, as otherwise, it can easily lead to a wrong conclusion.

In the following, the correlation results for each emotional dimension are discussed individually based on the obtained Pearson correlation coefficient Figure 5.9:

Valence: Almost all of the conventional distribution statistics extracted from the valence gold-standard tend to have a very weak linear correlation with the engagement indicators, typically with r values slightly below .2. Notable exceptions are that higher values around the middle of the distribution (kurt – $r = -.313$) result in more likes per comment. A reduction in the standard deviation ($r = -.276$) causes a shift towards more positive comments. Meanwhile, the more complex representations reveal strong positive correlations for absE e. g. , $r_{Vp/d} = .467, r_{Lp/d} = .422, r_{dislikes} = .355, r_{Cp/d} = .350$, followed by the peaks, CBMe and LSBMe, suggesting that user engagement increases with the value of this feature. In contrast, MACh and the *SaEn* are negatively correlated, so as the complexity of the valence gold-standard within a video increases, the number of user interactions with the video decreases.

Arousal: As can be seen from Figure 5.8, there are various correlations among the arousal representations that differ in intensity. For example, Vp/d, Lp/d, Cp/d and CLp/d decrease slightly (e. g. , $r_{(v_{p/d}, std)} = -.293, r_{(V_{p/d}, q_{95})} = -.212$) with increasing standard deviation and the 95% *quantile*. This effect is coherently mirrored to the opposite

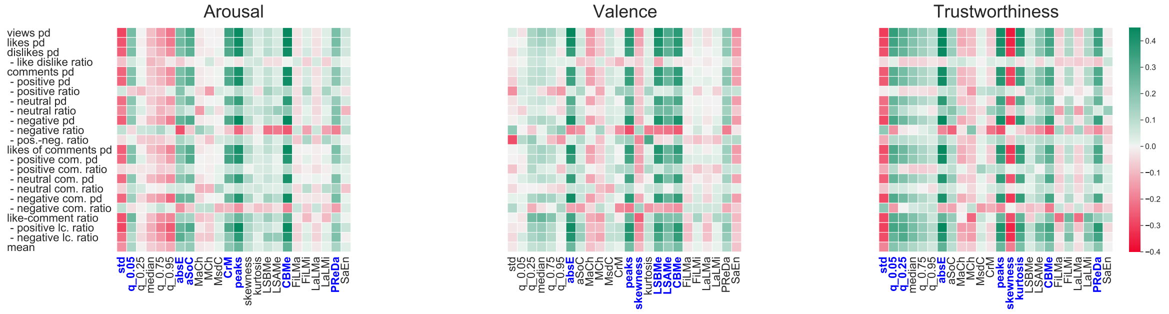


Figure 5.9: Pearson correlation matrix of user engagement indicators and the statistics/ representations extracted from each dimension. The latter are standard deviation (*std*), quantile (q_x), absolute Energy (absE), Mean relative Absolute Change (MACh), Mean Change (MCh), Mean value of a central approximation of the Second Derivatives (MSDC), number of Crossings of a point (CrM) *m*, peaks, dynamic sample skewness (skew), Kurtosis (kurt), Last Strike Above the Mean (LSAMe), Last Strike Below the Mean (LSBMe), Count Below Mean (CBMe), Relative Sum Of Changes (ASOC), first and last location of the minimum and maximum (*FLMi*, *LLMi*, *FLMa*, *FLMa*), Percentage of Reoccurring Data points of non-unique single points (PreDa), and Sample Entropy (SaEn). Representations in blue are utilised as cross-task, semi-automatic representations for user engagement prediction.

quantile, including the comment-like ratio (clr) (e. g. , $r_{(V_{p/d},q.05)} = 0.231, r_{(clr,q.05)} = -.248$).

Overall, peaks and *CBM* show the strongest positive correlations, for example, $r_{(V_{p/d},peaks)} = .440, r_{(L_{p/d},CBMe)} = .456$ and $r_{C_{p/d},peaks} = .409$. Furthermore, the share of neutral comments changes much less than the share of positive and negative comments, as these representations trend upwards. Also, CrM, *aSoc*, absE and PreDa significantly correlate to the user engagement criteria. No conclusion can be drawn with regard to the ratios (e. g. , like-dislike and positive-negative comments), as none of the correlations to any feature is sufficiently significant.

Trustworthiness: The skew of the distribution of the trustworthiness gold-standard appears to be of inherent importance. This is reflected in two feature categories: On the one hand, the skew itself, which shows a significant negative correlation to most indicators over $r < -.3$ (interpretation: negative skew equals to a shift of the distribution to the right (left skewed) leads to a stronger, positive effect). However, on the other hand, this pattern is also recognisable in the quantiles, which correlate positively with decreasing relevance, for example $r_{(views,q.05)} = .356, r_{(likes,q.75)} = .175$. In this context, the level of the trustworthiness increases with decreasing standard deviation increases, showing that a strongly concentrated, left skew is beneficial (e. g. , $r_{(views,std)} = -.304, r_{(likes,std)} = -.287, and r_{dislikes,std} = -.274$). In addition, a number of other representations show clear correlations, such as absE and the number of peaks. Regarding the ratios, no conclusion could be drawn due to the lack of significance.

Discussion: In the previous sections, a variety of correlations between emotion signal statistics (including trust) and user engagement were shown. Representations of arousal and trustworthiness exhibited distinct patterns. A less broadly distributed level of arousal (higher low and lower high quantiles) and an overall high level of trustworthiness showed more user interaction with the video. In contrast to the findings from [116], valence appears to have less apparent linear correlations in the analysis.

The number of peaks is defined (according to the parameter set: ten consecutive ascending followed by ten descending time steps) as the strongest correlated and broadly applicable indicator. Likewise, the signal-energy related feature for trustworthiness and valence have proven to be helpful. Regardless of the representations, the number of negative comments tend to be stronger correlated, followed by likes and positive comments, often showing weak to moderate correlations.

Even though these results show many significant correlations, they are based exclusively on simple correlation analyses. The real impact and predictive strength for the prediction are therefore still to be established. Similarly, non-linear correlations, even spanning several representations, may exist.

5.1.4.4.2 Estimation results: All four prediction tasks are presented in Table 5.17 and addressed in the following paragraphs. As explained in Section 3.2.1, the MAE allows rendering the results under consideration of the underlying distribution of individual target variables (cf. Section 5.1.4.3). Each SVR version is trained with all semi-automatic and automatically selected representations. The semi-automatically chosen representations of each dimension are highlighted in blue in Figure 5.9 and kept identical across targets.

On average, the fewest number of representations were selected by the automatic process for trustworthiness (6.0), followed by arousal (9.3) and valence (9.5). Broken down by targets, these are 7.6 Vp/d, 23.3 Cp/d, 29.3 Lp/d, and 20.1 LCp/d.

A comparison of the feature selection per method is provided for CLp/d in Figure 5.10. It depicts the relevance of each feature for the prediction by the corresponding weight of the SVM. In addition, the p-values of the automatic (univariate) selection are displayed, which are almost equivalent to the auto-selection. This also holds for the hand-selected representations in this example, with automatic representations slightly outperforming all others (cf. Table 5.17), however, indicating sensitivity to the inclusion and exclusion of representations.

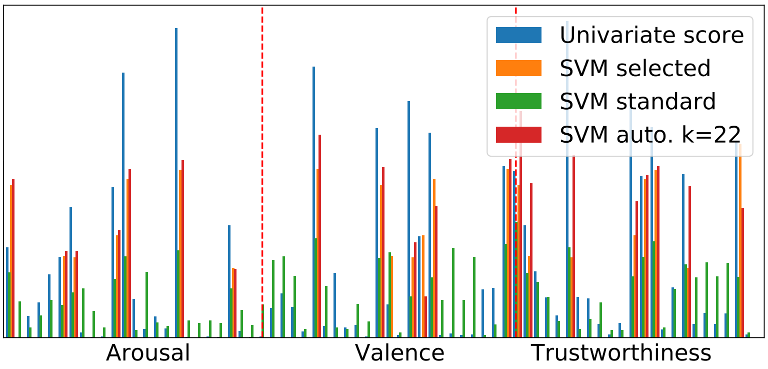


Figure 5.10: Comparing different feature selection methods based on the SVM weights of arousal, valence, and trustworthiness representations. It shows the p-values of all, manually selected (24), and automatically selected $k = 22$ representations. For the sake of clarity, the p-values of the automatic selection are scaled by applying a base 10 logarithm and dividing the result by 10 ($-Log(p_{value})/10$).

Views per day: Representations from all gold-standards show potential to predict Vp/d. When valence and trustworthiness representations are fused, the results improve. When arousal is added to them, the result decreases slightly. Without any feature selection, trustworthiness is the strongest. However, the results deteriorate if only individual representations are selected, suggesting that only a broad range of representations allows robust prediction without causing generalisation problems. When considering the feature selection for the other uni-modal results, the semi-automatic, cross-task version clearly increases these to 198.5 and 184.8 MAE respectively. With automatic feature selection, the result for valence can be increased even further to 169.5 MAE. Selecting representations from the fused modalities, however, does not seem to be clearly advantageous.

Likes per day: Comparable to Vp/d, the use of all unimodal arousal and valence representations appears to be more sensitive than their fusion. This only seems worthwhile with the cross-task selection method, which improves all results except for tri-modal fusion. With the automatic selection approach, the results in the single and bi-gold-standard models reach their peak, with valence having the most robust prediction performance with 1.23 MAE Vp/d. Only the tri-modal fusion seems to make sense with the cross-task selection approach. The tri-modal fusion slightly degrades the selection methods. For the prediction of Lp/d, trustworthiness exhibits weaker performance than the other feature types on the test set while being superior on the development set. Although

Table 5.17: Prediction of views, likes, comments, and likes of comments utilising extracted and selected representations from **A(rousal)**, **V(alence)**, and **T(rustworthiness)** annotations. The *C* parameter of the SVR is specified, which is tuned from 0.00001 to 1. Using the best *M* mean absolute error on the devel(opment) set the best *C* is selected for the prediction on the test set. (%) indicates the relative change of the automatic (auto.) and semi-automatically selected (sel.) in % to the unchanged representations, while "+" indicates an improvement, hence a decrease of the Mean Absolute Error (MAE) compared to the original feature sets.

Type	**Views**						**Likes**						**Comments**						**Likes of Comments**					
	devel			**test**			**devel**			**test**			**devel**			**test**			**devel**			**test**		
	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.	all	sel.	auto.
	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%	MAE	rel.%	rel.%
A	231.8	+6.8	+5.0	220.3	+9.9	+3.1	2.30	-0.3	+2.6	1.55	+5.9	+3.0	.288	-0.1	+3.7	.154	+2.5	+0.6	1.19	+5.7	+5.9	.50	-19.1	-22.7
V	253.1	+8.7	+7.2	223.8	+17.4	+24.3	2.29	+0.6	+1.0	1.61	+17.6	+24.0	.288	+3.1	+3.9	.154	+5.1	+2.4	1.17	-1.4	+3.6	.51	-2.8	-18.4
T	237.4	+11.8	+16.3	207.9	-5.2	-9.7	2.21	+5.3	+14.4	1.92	+13.3	+3.6	.262	+5.8	+6.4	.225	+2.1	-5.3	1.11	-0.1	+9.5	.75	+8.8	+6.7
A+V	237.6	-1.0	+2.1	210.7	+4.1	+18.3	2.27	-11.4	+0.3	1.79	+24.2	-0.7	.277	-4.3	+3.4	.161	+9.9	+0.1	1.16	+0.1	+2.0	.54	+16.8	-27.3
A+T	240.3	+9.2	+15.7	207.9	-6.7	-3.9	2.26	+4.8	+10.6	2.02	+11.8	+10.3	.268	+1.6	+7.2	.182	-34.9	-1.1	1.11	-0.2	+3.7	.59	-17.9	-14.7
V+T	249.1	+15.5	+20.0	205.8	-3.1	-2.6	2.07	-11.8	-0.2	1.99	+17.2	+0.1	.262	-2.7	+5.5	.188	+10.9	-24.7	1.04	-6.2	+0.3	.78	+11.4	-0.1
A+V+T	228.9	-1.2	+8.7	205.9	-8.4	+0.2	2.06	-12.6	+0.6	2.08	-22.9	+0.3	.264	-0.0	+2.7	.192	-7.5	+0.5	1.10	+0.9	+4.3	.60	-16.6	+8.4

fusion and automatic feature selection yield robust results on the development set, they are not transferable.

Comments per day: Without selection, the representations of the two conventional emotion dimensions demonstrate the most substantial results in predicting Cp/d with .154 MAE each. For valence, in particular, the findings from the previous tasks are confirmed and highlight it is a meaningful gold-standard which prediction results can even be improved by the auto-selection and cross-task method. However, the combination of the two shows little added value across the board, even though it achieves the best result with the cross-task selection of .145 MAE. As in the previous tasks, trustworthiness is convincing on the development set but is incapable of validating the results on the test set.

Likes of comments per day: Using all representations, arousal achieves the best result in the prediction of CLp/d. Unlike other prediction goals, the feature selection procedures lead to a deterioration of the results except for the fusion of arousal and valence in combination with the task-specific procedure. It is apparent once again that trustworthiness does not meet the results of the development on the test set.

Discussion: Examining the results obtained across all prediction targets, it seems likely that feature selection leads to consistently improved results. In particular, the proposed cross-task feature selection seems to have good generalisation capabilities, which makes it slightly superior to automatic selection. All user-engagement criteria have an inherent sentiment relationship, which probably explains valence's consistently superior results for almost all prediction targets and makes it the most predictive signal overall. This is not necessarily reflected in the simple correlations. Furthermore, arousal without any selection as well as in fusion with valence (see Vp/d) proves to be valuable. In general, the early fusion of representations seems to be of limited applicability. However, it cannot be ruled out that it still leads to better results, since late fusion of the predictions and deeper fusion by more complex models was beyond the scope and has not been tested so far.

The additional dimension trustworthiness fails to reach the same predictive power as valence and arousal. In particular, this is due to the generalisability problem on the test set, as very good results are achieved on the development set, for example, uni-modal and without feature selection on Vp/d performs better than the others. This weakness can be attributed to the lack of consistency of the trustworthiness gold-standard across the derived partitions and the lack of normalisation. Since the targets remain unnormalised to allow for interpretability, this leads to much stronger effect of outliers in a tightly and skewed distribution across partitions. In general, these initial results are promising, indicating that trust also appears to be generally valuable to viewers and supports the formation of a parasocial relationship [301].

5.1.4.5 Conclusions

Through the extensive series of experiments, it could be shown that each of the emotional dimensions studied (arousal, valence, and trustworthiness) have relationships with key criteria for user engagement and can be exploited to predict them to a certain extent (**RQ-1d**). Such automatic prediction appears to facilitate potential benefits, for example, the patterns discovered in the emotional understanding of the video content (e. g. , short-term fluctuations in arousal) could improve the parasocial relationship and thereby drive the user engagement.

This effect may be reinforced by YouTube's own algorithms, which favour content that goes viral (thus having a high reach or increased user engagement). The use of these representations may also open up financial improvement of income sources for the creators, e. g. , through more views of the video per day. In order to control these effects in the future, however, it may also be sensible for the platform operators to gain a deeper understanding of emotional mechanisms and, if necessary, integrate them into suggestion algorithms. For this, further research on the indicators themselves are necessary, e. g. , a comparison between the real increase and the calculated average per day.

In view of the above-mentioned parasocial relationship theory, it could be expected that trustworthiness will be very predictive in estimating user interaction. Although the results are promising in some cases, conventional emotional dimensions seem to be more effective. However, the selected data domain only considers car review videos, which means that applying the novel dimension to other domains (e. g. , comedy or entertainment) may reveal itself to be more important in exploring the relationship between trust and user engagement.

5.2 Objective Dimensions

Besides the emotional information in audio-video data, the other essential dimension of MSA is the context, specifically, the target the emotion is directed towards. In the following, experiments are presented dealing with the target dimension in two different fashions:

- **RQ-2a:** Improving contextual understanding by proposing a target **extraction** method (alongside others) that works without labels, as well as assumptions about the number of expected topics.
- **RQ-2b:** Predicting the proposed speaker topic dimension using **detection** models, which are trained on manually labelled segments.
- **RQ-3:** Evaluating the uni- and multi-modal dynamics and the strengths of the three core **modalities**.

The characteristics of each task are explained in detail in the introduction, followed by the proposed methods and the experimental results. The RQs covered are discussed at the end of each section.

5.2.1 Target Extraction

5.2.1.1 Characteristics

It is well known that the most meaningful modality for contextual understanding is the linguistic component of language. This section aims to discover latent semantic structures (the topics) algorithmically. As explained in Section 2.3, this is usually done by specifying the number of expected topic clusters as an initial parameter for the model. For each topic cluster, the model returns a set of words, called topic representatives, that portrays the topic's semantics. Also, the capabilities of speech-to-text are improving, the field is rather unexplored for spoken language from videos. Here, the main differences in extracting context from written text lie in colloquial language, expressed by longer, rapidly produced text sequences.

Stappen et al. [302] propose a novel approach, Graph-based Topic Modelling approach for Transcripts (GraphTMT), to these challenges, aiming to extract meaningful and semantically coherent topics from transcribed video reviews. To avoid making any (wrong) assumption regarding the data distribution, the algorithm aims to find the number of relevant topic clusters automatically. The approach transforms the corpus into a single graph, which is split into subgraphs utilising edge connectivity to find prominent topics in an unsupervised way.

Furthermore, the differences between transcribed language of MuSe-CaR and another corpus, Citysearch New York corpus (Citysearch), of written short-sentenced reviews, are analysed, as well as the generalizability of this novel approach.

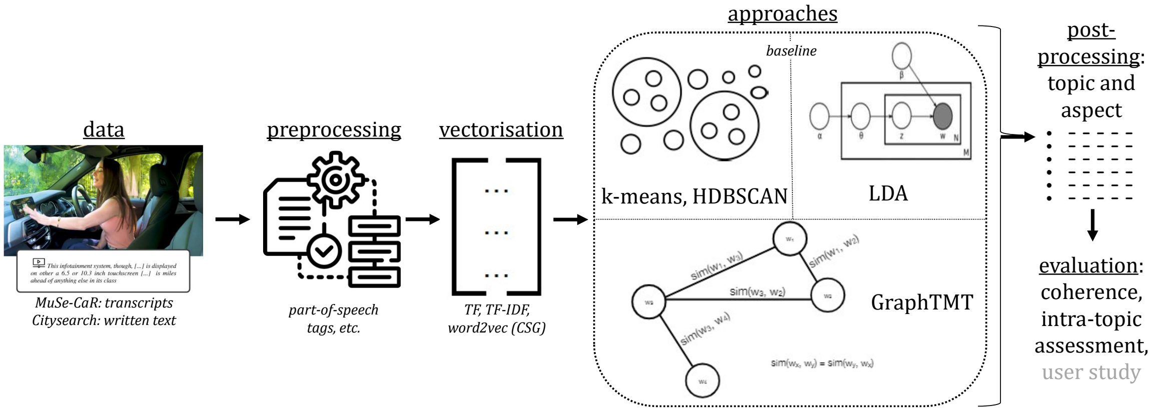

data
This infotainment system, though, [...] is displayed on other a 6.5 or 10.3 inch touchscreen [...] is miles ahead of anything else in its class
MuSe-CaR: transcripts
Citysearch: written text
preprocessing
part-of-speech tags, etc.
vectorisation
TF, TF-IDF, word2vec (CSG)
approaches
baseline
k-means, HDBSCAN
LDA
α
θ
z
w
β
N
M
GraphTMT
w_1
w_2
w_3
w_4
sim(w_1, w_3)
sim(w_1, w_2)
sim(w_3, w_2)
sim(w_3, w_4)
sim(w_x, w_y) = sim(w_y, w_x)
post-processing: topic and aspect
evaluation: coherence, intra-topic assessment, user study

Figure 5.11: Overview of the topic extraction process: First, the transcripts are preprocessed, vectorised, and grouped by different methods (baseline: K-means, HDBScan, LDA vs. GraphTMT) to extract relevant topics and aspects. The results are considered in terms of the topic coherence and intratopic structure. In addition, a user study is conducted for MUSE-CAR.

5.2.1.2 Experimental Setup

The entire workflow using the transcripts is illustrated in Figure 5.11, and a detailed explanation is provided in the following paragraphs.

Datasets: The primary focus of the evaluation is on the MuSe-CaR covering 5 467 transcribed segments, each assigned to one of the ten topics, featuring more than 20k sentences as described in Chapter 4. Furthermore, the popular Citysearch [303–305] corpus is utilised.[4] It covers over 50 000 restaurants reviews from 30 000 distinct users. However, only a subset of 3 400 sentences was labelled by Ganu et al. [306] using five core topics: *Ambience*, *Anecdotes*, *Food*, *Price*, and *Staff*.

Preprocessing: Motivated by Schofield and Mimno [307], a part-of-speech system extracts tags for the words of the transcripts [308]. Using the unique words, a Continuous Skip-gram Word Model (CSG) model is trained for 400 epochs using a window size of 15 (see Section 3.1.2), as suggested by other studies [145, 154, 309], to deliver a stable outcome.

[4] Download Citysearch: http://www.cs.cmu.edu/~mehrbod/RR/, accessed on 29 April 2021

Baseline approaches: The Latent Dirichlet Allocation (LDA) [155], K-means [310], and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [153] algorithms are selected as baseline models. The first one is broadly applied in topic modelling [144] to gain an understanding of semantic clusters in documents. By using statistical Dirichlet-priors and bag-of-words methods, a distribution of topics as well as their representatives can be determined. The sentence structure is left aside, and the focus is on pure word co-occurrences and their cardinality. In the context of NLP, it is often initialised by the Term Frequency (TF) and the TF-Inverse Document Frequency (TF-IDF) [144]. These as well as the document-topic density (α), word-topic density (β), and the number of topics (t_n) are tuned.

K-means uses the same initialisation techniques, where clusters are formed based on the closest distances between the words [146, 161, 309, 311]. As the distance measure, the Euclidean distance is commonly applied. As explained in the context of MuSe-Toolbox (see Section 5.1.2), K-means divides the text data into a predefined number of k clusters by executing an Estimation-Maximisation Algorithm. In the assignment step, each point is assigned to the closest cluster centroid, and in an update step, the new positions of the centroids are calculated based on the assigned data points. For the experiments, a hyperparameter search for $k = \{4, 20\}$ using both types of initialisation (TF, TF-IDF) are used.

The latter algorithm, HDBSCAN, is very efficient in finding hierarchical and density-based clusters [153], wherein the number of clusters k does not have to be specified a priori. Creating a minimum spanning tree that is reduced into smaller trees creates clusters until the min_{size} is reached and converges. In the configuration used here, the automatically detected outliers are ignored.

GraphTMT approach steps:

1. **Graph construction:** A graph G of N nodes and E edges is constructed, where $|N| \leq n$ nodes represents the embeddings of all vocabulary words, and each edge $e \in E$ reflects the semantic closeness between words relying on the cosine similarity, which is calculated on the word embeddings [80, 312]. After construction, the graph is complete so that every edge is adjacent to every other edge.
2. **Edge removal:** Based on the construction, a higher similarity between nodes is expressed by a higher weight value of the edge. This property can be used to disconnect nodes. By dropping edges of low similarity, subgraphs of highly similar nodes, representing topics, are revealed. To isolate the nodes, a Percentile Similarity Threshold (PTH) is proposed.

3. **Subgraph topic clustering:** The incomplete graph is further clustered to obtain maximal connected subgraphs, each expressing a concise topic. This is achieved using the K-Components algorithm, which is highly efficient in finding locally strongly connected subgraphs for various definitions of edges and nodes [156, 313, 314]. An optimal subgraph of G is defined by having at least a maximal node connectivity K, while at least K nodes are removed. The nested nature of these subgraphs is hierarchical. This means that a 1-component graph can enfold one or more 2-components, where each of them comprises several 3-components, each consisting of several 4-components, and so forth.

Post-processing: The minimum number of representative words for a cluster is defined as six, meaning that clusters have to meet this minimum to be considered a relevant cluster topic. The selection of assigned words is chosen according to the proximity of the average embedding vector of a cluster. Words can be ranked either by the Topic Vector Similarity (TVS) or Node Degree Connectivity (NDC) logic. Finally, as in comparable studies [305, 304], each derived topic is assigned by hand to one of the gold standard topics based on the representative words.

Evaluation: Three measures are used in order to assess success of the proposed method.

- **Coherence:** The semantic similarity between topic clusters can automatically be assessed using a coherence score. Röthe et al. [315] found that c_v better mapped human understanding of coherence within a cluster than alternatives such as the widely applied u_{mass} [145, 305]. The c_v is intrinsically calculated by a sliding window, calculating the cosine similarity on the base of the normalised pointwise information. It can be interpreted as an intrinsic evaluation of how the modelled topics reflect the data set [316].
- **Intratopic assessment:** Although the coherence score provides information on semantic consistency, this does not necessarily correspond to human understanding [315, 317]. For this reason, intratopic assessment is used to compare unsupervised created clusters labelled by a human to human-crafted annotations (gold topics). For this purpose, the representation words of a class are interpreted by a human and a topic is inferred. By comparing annotated classes and assigned classes, the topic coverage (t_c) as well as the topic overlap (t_o) can be calculated. The first is the proportion of inferred labels that are also present in the annotations. The second is the ratio of repeated inferred topics.

Parameter	**Values**
Number of topics (t_n)	[4; 20]
Document-topic density (α)	[0.1, 0.4, 0.7, 1.0, 1/t_n]
Word-topic density (β)	[0.1, 0.4, 0.7, 1.0]
Weighting strategy	[TF, TF-IDF]
Minimum cluster size(C_{min})	[5; 30]
Edge connectivity (K)	[1, 2, 3]
Edge weight threshold ($percentile_{rank}$)	[0.50, 0.60, 0.70, 0.80, 0.90, 0.95]

Table 5.18: Overview of all hyperparameter settings evaluated in the experiments. Table is taken from [302].

- **User study:** Computational methods are easy to use and therefore often preferred. However, such indicators do not replace the human ability to abstract concepts and sort them into their understanding of the world. For this reason, semantic validity is assessed through a user study similar to [142, 317, 318]. In this study, 31 participants with at least an upper-intermediate English level (minimum of B2 in the Common European Framework of Language Reference) performed word intrusion tasks for MuSe-CaR. In this task, six words are suggested to the user, one of which is a representative of another cluster. The participants have to locate the intruder. Alternatively, they can select "not sure" to express high uncertainty. Given the following intrusion task: {acceleration, steering, stability, voice, chassis, anticipation, *not sure*}, all words besides the intruder word "voice" represent the topic of "handling". The precision of finding the intruder can be expressed by the Word Intrusion Precision (WIP):

$$WIP_k^m = \sum_s \mathbb{1}(i_{k,s}^m = w_k^m)/S, \tag{5.2}$$

where S the number of all participants, $i_{k,s}^m$ is the intruder selected by the s^{th} participant on the k^{th} topic, and w_k^m is the intruder from the k^{th} topic inferred by model m. Similarly, the rate of "not sure" (NSR) selections is relatively measured.

5.2.1.3 Results

Here, the pre-experiments in terms of the preprocessing procedures are first discussed. Details of the experiments of MuSe-CaR and Citysearch follow. An overview of the hyperparameter settings is given in Table 5.18.

Table 5.19: Overview of the best results for the baseline (LDA, HDBSCAN, K-means) and GraphTMT approach reporting the intrinsic coherence score (c_v), the number of topics (t_n), topic coverage (t_c), and topic overlap (t_o) on MuSe-CaR. The best hyperparameters (HP) for LDA are α = 0.10, β = 0.70; HDBSCAN with $min_{size} = 6$; and K-means are $k = 8$ and $k = 10$ while using the TF-weighted initialisation. The graph approaches utilise PTH = 80 for $K = 1$ and $K = 2$. Table adapted from Stappen et al. [302].

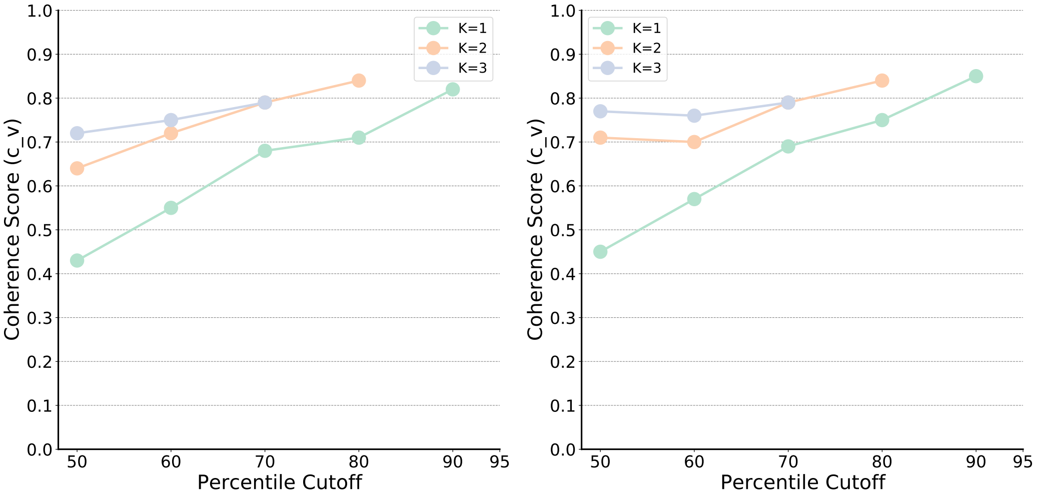
Approaches	**HP**	c_v	t_n	t_c	t_o
LDA	α = 0.10, β = 0.70	.51	8	**.60**	.25
HDBSCAN	min_{size} = 6	.63	**11**	**.60**	.40
K-means	k = 8	.73	8	**.60**	.25
	k = 10	.69	10	**.60**	.25
GraphTMT	K = 1	.76	6	.50	**.17**
	$K = 2$	**.85**	5	.40	.20
Ø		.70	8	.45	.25

5.2.1.3.1 Preprocessing: In preliminary experiments, variations of typical preprocessing procedures are evaluated [307, 318]. No positive effect arises for MuSe-CaR from the removal of stop words, or the normalisation of nouns through lemmatisation. However, a positive effect appears when the words are not stemmed but limited to all types of nouns provided by the tagger. This is in line with multiple studies [141–143], which have shown that generalisation is superior when dealing with noisy textual data as given by automatic transcripts, e. g. , words that are implausible in context, incorrect grammar due to failures in the speech-to-text system, and colloquialisms in spoken language. Moreover, the preprocessing has shown to be effective in related tasks [143, 154, 319]. For Citysearch, using all tags achieved the best results. For conciseness, the following detailed results are limited to these preprocessing procedures.

5.2.1.3.2 MuSe-CaR evaluation: Table 5.19 depicts the best results of the baseline and graph-based methods.

Coherence score: Following are results of the strong word-embedding clustering baseline: K-means achieves a c_v = 0.73 for $k = 8$. Higher values of k, for example, k = 10, show more incoherent (miscellaneous) topics also reflected by c_v = 0.69. HDBSCAN with a min_{size} = 10 has its best result with 11 topics resulting in a c_v = 0.63. LDA creates the most coherent clusters at α = 0.1, β = 0.7, and t_n = 8 with c_v = 0.51. The best configurations for GraphTMT outperform the baseline models, reaching at PTH = 80 with K=1 and c_v = 0.75 and for K=2 and c_v = 0.84.

Furthermore, for the proposed GraphTMT, the c_v is exhibited in detail for different levels of K, NDC, and TVS in Figure 5.12. With increasing K, c_v also increases,



(a) Node Degree Connectivity (NDC)

(b) Topic Vector Similarity (TVS)

Figure 5.12: GraphTMT for (K = {1, 2, 3} and using (a) NDC and (b) TVS (right) reporting the intrinsic coherence score on MuSe-CaR. For all K's, the c_v is increasing with higher percentile cutoff points, while TVS shows better performance in the lower range. Missing dots indicate that the experiment outcome did not fall in line with the experimental guideline.

Table 5.20: Exemplary results for GraphTMT at 80th PTH with $K = 1$ on the MuSe-CaR corpus showing inferred and gold topics as well as their representative words. Table adapted from Stappen et al. [302].

Inferred Topic	Topic Representatives	Gold Topic
Handling	suspension, changes, sport, steering, response	Handling
Infotainment	hand, screen, pop, information, entertainment	User Experience
Infotainment	touch, ways, climate, buttons, controls	User Experience
Passenger Space	area, head, roof, room, headroom	Interior Features
Performance	seconds, turbo, twin, acceleration, cylinder	Performance
YouTube	channel, dot, please, thanks, share	General Information

suggesting that higher edge connectivity leads to more representative topics. In addition, TVS boosts coherence scores; however, for K={1, 2} at 80th PTH level, experiments do not reach the minimum threshold of six subgraphs. The NDC variation does not have this issue but generally performs worse than TVS.

Intratopic assessment: Looking at the qualitative side, the baseline approaches give more fine-grained clusters than the human-annotated ground truth and achieve a slightly higher topic coverage (t_c) than GraphTMT. As shown in detail for K-means in Table 5.21, the mixed golden label *General Information* [24] is further separated into *YouTube* and *Storage*. This means that from six unique gold topics that can be matched (t_c = 6/10) they included two duplicate topics (t_o = 2/8). HDBSCAN leads to almost the

Table 5.21: List of topics extracted from the MuSe-CaR corpus by using weighted K-means ($k = 8$, TF-weighted). Table adapted from Stappen et al. [302].

Inferred Topic	**Topic Representatives**	**Gold Topic**
Handling	suspension, handling, dampers, corners, chassis	Handling
Infotainment	menus, satnav, swivel, commands, entertainment	User Experience
Interior Features	dash, design, events, wood, plastic	Interior Features
Performance	engine, turbo, litre, cylinder, engines	Performance
Safety	detection, assist, safety, collision, airbags	Safety
Storage	storage, items, space, boot, hooks	General Information
YouTube	please, enjoy, click, share, wow	General Information
Miscellaneous	cars, guys, opportunity, brand, tomorrow	General Information

Table 5.22: User study of the most successful topic models on MuSe-CaR showing the Word Intrusion Precision (WIP), lowest and highest hit rates, and the "not sure" ratio (NSR). Table adapted from Stappen et al. [302].

Approaches	**WIP**	**Lowest Hit**	**Highest Hit**	**NSR**
LDA	43 %	15 %	**78 %**	13 %
K-means	61 %	47 %	75 %	15 %
GraphTMT	**63 %**	**56 %**	72 %	**8 %**
Ø	56 %	39 %	75 %	36 %

same topics, but also adds two *Miscellaneous* ones as well as a *Passenger Space* topic with the respective words: *{area, head, roof, room, headroom}* compared to K-means. Example cluster topics are shown for $K = 1$ in Table 5.20. The topic overlap t_o is lower for GraphTMT than that of the baseline models, indicating high precision. The consistent representative words and the absence of a miscellaneous topic show the high quality of this approach. For $K = 2$, the structure is almost identical while *Performance* is further divided and *Handling* and *Infotainment* disappear. Therefore and due to the hierarchical structure of this approach, it can be assumed that the overlapped topics *Performance*, *Infotainment*, *Passenger Space*, and *YouTube* reflect the most stable, coherent structures.

User study: Results of the user study are given in Table 5.22. WIP measures the number of successfully identified randomly added intruders by a participant. If the respondent selects the wrong option or "not sure", this counts as missed. For each method, a number of randomly selected topic clusters are suggested to the user. The average WIP rate across all users of a model is shown for all topic clusters as well as the lowest and highest hit rate across all topic clusters. GraphTMT performed best with an average WIP rate of 63 %, followed by K-means. Even the lowest hit rate of 56 % shows that the majority of users are still very confident even with the fuzziest topic of GraphTMT

Table 5.23: Overview of the best results for the baseline (LDA, HDBSCAN, K-means) and GraphTMT approach reporting the intrinsic coherence score (c_v), the number of topics (t_n), the topic coverage (t_c), and topic overlap (t_o) on Citysearch. The best hyperparameters (HP) for LDA have α = 1/t_n, β = 0.40; HDBSCAN has min_{size} set to 5; and K-means has $k = 8$ while using the TF-weighted initialisation. The graph approaches utilise PTH = 80 for $K = \{1–3\}$. Table adapted from Stappen et al. [302].

Approach	**HP**	$\boldsymbol{c_v}$	$\boldsymbol{t_n}$	$\boldsymbol{t_c}$	$\boldsymbol{t_o}$
LDA	α = 1/t_n, β = 0.40	.48	8	.67	.50
K-means	$K = 8$	**.64**	8	**.83**	.38
HDBSCAN	$min_{size} = 5$	.61	3	.33	.33
GraphTMT	$K = 1$	.40	9	.67	.56
	$K = 2$	.60	6	.67	.33
	$K = 3$	**.64**	5	.67	**.20**
Ø		.56	7	.64	.38

and that the inferred representation words are the most straightforward to interpret. It also achieves the lowest NSR. LDA has the least interpretable topics with only 15 %, but the highest hit rate with 78 %, closely followed by K-means and GraphTMT.

The results can be summarised by stating that both the baseline and the GraphTMT approaches found meaningful topics. Hereby, clustering-based approaches performed better than the statistical LDA approach. The most topics were inferred by the GraphTMT, while by increasing K, the topic coherence increases but the number of clusters decreases. In this case, the user study corresponds well with the coherence scores, so high scores also have the highest topic hit rates and seem to be the more interpretable topics on average.

5.2.1.3.3 Citysearch evaluation: Finally, the **generalisability** of the proposed approach is evaluated using the Citysearch corpus (see Table 5.23). In general, stronger results are achieved using the full preprocessing capability.

Coherence score: The best configuration of each model type along with the associated results are listed in Table 5.23. With a c_v of .64 each, both K-means (with $k = 8$ and TF-initialised) and GraphTMT (with K = 3 and PTH = 0.8) achieve the best c_v result. As K of GraphTMT decreases, so does the coherence score, however, the number of topics increases. HDBSCAN scores c_v = .61 followed by LDA with c_v = .42 and applying α = 1/t_n, β = 0.4. Furthermore, Citysearch has six manual annotated topics. K-means infers eight and GraphTMT (K=3) five. At a lower K = 1, GraphTMT even finds nine topics. HDBSCAN, however, infers only three topics.

Table 5.24: Inferred topics and their representatives from the GraphTMT (K = 1) approach on the Citysearch corpus, including positive (pos.) and negative (neg.) *Service* topics as well as Miscellaneous (Misc.) topics as gold topic.

Inferred Topic	Topic Representatives	Gold Topic
Adjectives	incredible, fantastic, outstanding, fabulous, amazing	Miscellaneous
Ambience	painted, baguette, mirror, coloured, wood, leather	Ambience
Anecdotes	celebrate, celebrated, celebrating, wedding, anniversary	Anecdotes
Food	fennel, puree, polenta, jalapeno, pate	Food
Food	poivre, hanger, sirloin, ribeye, frites	Food
Location	high, hill, bronx, murray, queen	Miscellaneous
Location	chelsea, downtown, soho, midtown, district, uptown	Miscellaneous
Music	piano, playing, jazz, band, played, background	Miscellaneous
Service (pos.)	personable, gracious, polite, knowledgeable, professional	Staff
Service (neg.)	arrogant, unfriendly, incompetent, unattentive, unprofessional	Staff
Weekdays	tuesday, wednesday, monday, friday, thursday	Miscellaneous

Table 5.25: Inferred topics and their representatives from the K-means (k = 8, TF) approach on Citysearch.

Inferred Topic	Topic Representatives	Gold Topic
Handling	suspension, handling, dampers, corners, chassis	Handling
Infotainment	menus, satnav, swivel, commands, entertainment	User Experience
Interior Features	dash, design, events, wood, plastic	Interior Features
Performance	engine, turbo, litre, cylinder, engines	Performance
Safety	detection, assist, safety, collision, airbags	Safety
Storage	storage, items, space, boot, hooks	General Information
YouTube	please, enjoy, click, share, wow	General Information
Miscellaneous	cars, guys, opportunity, brand, tomorrow	General Information

Intratopic assessment: K-means produces convincing results, finding all six golden themes ($T_c = .83$) and only resulting in a single miscellaneous topic, as depicted in Table 5.24. For GraphTMT, only location and food are duplicated (see Table 5.24), resulting in t_c $= .67$ (see Table 5.23). Many of the classes found seem very plausible and intrinsically coherent, such as, e. g. , *Music* with *{piano, playing, jazz, band, played, background}*, although these were not necessarily mapped by a human as a gold topic. Overall, the proposed method inferred many relevant topics. Regarding the golden labels, the degree of coverage for GraphTMT is slightly lower than for K-means and LDA with t_c = 83.3 as well as the overlap with t_o = 38 %.

5.2.1.4 Conclusions

From these experiments, it can be concluded that the proposed GraphTMT is able to extract meaningful target clusters without being dependent on human annotations and a priori assumptions regarding the number of topics (**RQ-2a**). This finding is promising as it supports the development of graph-based machine learning methods to tackle the issue of unsupervised exploration of long, noisy transcript snippets. It can also be useful as an exploration step before domain-specific labelling. In addition, the results underline the robustness of this method with very little fine-tuning. On transcripts, GraphTMT outperformed the baseline topic modelling approaches in terms of coherence, uniqueness, and interpretability of the clusters. The simplicity when fine-tuning is, in comparison to other approaches, an additional benefit. For example, a high K led to highly encapsulated clusters, which suggest that semantically inconsistent topics and words can further be excluded with a single parameter. In contrast, by manually setting k in K-means, it is possible to extract a larger number of topics. However, this is not so much a fine-tuning parameter, but rather a strict setting that leads to the exact number of output topics k, requiring assumptions about the data and repeated runs to optimise the random initialisation.

Generally speaking, the GraphTMT properties help in every scenario where no assumptions can be made and no labels are available. The experiments on the Citysearch corpus showed that the method is generally transferable to other domains and datasets. Although the coherence results were just behind the best model, K-means, it achieved the highest WIP rate and a higher uniqueness.

Overall, the proposed approach of graph construction, edge removal, and subgraph clustering showed promising results. However, this needs further investigation on a wider selection of datasets. For the future, a multimodal view is necessary for a full understanding of the content topics factoring in human actions. The development of such extension can be seen as an intriguing next step.

5.2.2 Target Detection

5.2.2.1 Characteristics

After extracting the speaker topics in an unsupervised fashion, this chapter focuses on modelling the human-annotated speaker topics as presented in Section 4.2. Understanding the domain-specific context relies strongly on the textual representation of the spoken language (transcripts). To this end, uni- and multi-modal approaches from Stappen et al. [26] are evaluated on various modalities. This is followed by the work on subsymbolic, high-level representation from Stappen et al. [56]. All experiments are conducted on the data selection of MuSe-Topic from MuSe-CaR introduced in Stappen et al. [27] (see Section 5.1). This enables a linkage of the two components, emotion and topic, as explained before. The distribution and examples across the ten speaker topics can be found in Section 4.2.

5.2.2.2 Experimental Setup

The experimental setup closely follows what is described in Section 5.1.2.2.

5.2.2.2.1 Feature Sets: As with the prediction of emotion classes, the sampling rate is maintained for all representations. As a result, the audio-video feature sets (Deep Spectrum, eGeMAPS, VGGish, Xception, VGGFace, and Generic, Optical Car Part Recogniser and Detector (GoCaRD)) produce a vector every 250ms. The preprocessing of the audio track and the imputing of the text-generated feature vectors (FastText and Sentic) also remains equivalent. The extracted representations and feature alignment are the same as the ones used in Section 5.1.2.2.

5.2.2.2.2 Architectures:

ALBERT: ALBERT [81] is employed as a state-of-the-art NLP Transformer for fine-tuning, as explained in Section 3.1.2. The network core and pretrained weights are utilised from Hugging Face[5] package. To optimise computational resources for such a parameter-intensive architecture, the training procedure is conducted in half-precision numbers. This purely textual model is fine-tuned for three epochs with a learning rate of 10^{-5} (ε to 10^{-8}) using an Adam optimiser with a *bs* of 12. The sequence length is limited to 300, applying padding and truncating where appropriate.

SenSA: SenticNet provides a unique way to integrate high-level language semantic concepts $\bar{c}_s$ from knowledge-based representations (see Section 3.1.2) into speaker topics

[5] https://github.com/huggingface accessed June 25, 2021.

classification. On the same base as in Paragraph 5.1.2.2.2, one-hot embedded sentics are utilised for training classifiers. Inspired from word-embedding training (see Section 3.1.2), the stronger learning impulse from this language-centric task can be channelled to neurally learn a domain-specific projection $h_s = \sigma(\bar{c}_s)$, where σ is a sigmoid FFL, from these one-hot encoded vectors. The compression leads to a strongly reduced input dimension from > 5k valid concepts to only 100, benefiting computation of the subsequent SVM. While training the embeddings, embedding dropout on single representations and time steps is applied to improve generalisation by avoiding learning the identical matrix. As before, the final prediction is made by a linear SVM, tuning the C value from 10^{-5} to 1 on the development set over up to 10 000 iterations (see Paragraph 5.1.2.2.2).

Others: The training configurations for the **LSTM-SA**, **MMT**, and **End2You** models are identical to Section 5.1.2.

5.2.2.2.3 Measure: The evaluation metric is inspired by classification tasks of similar challenges [8, 19] and is aimed at balancing the UAR and F1 (micro) measures in a combined score of $0.66 \cdot F1 + 0.34 \cdot UAR$.

5.2.2.3 Results

Illustrated in Table 5.26 are the results of the proposed approaches for the speaker topics classification task (by-chance 10 %).

Baselines: The unimodal LSTM-SA leads to around 35 % combined score on test. The best result in this combination is achieved by the textual representations FastText with 36.20 % combined score. The Transformer-based ALBERT architecture proved to be especially competitive, reaching 76.79 % combined score.

In a multimodal setting of the LSTM-SA network, the results improve slightly to 37.14 % combined score on the test set. However, these results fall behind those achieved by the MMT, which is equipped with a more advanced modality fusion. Here, all feature combinations result in better scores, with the best of FastText, eGeMAPS, and FAU improving results by over 15 percentage points to 52.98 % combined score.

Comparing the performance of MMT and ALBERT at the level of individual classes (see Figure 5.13), both show the strongest results for the classes Performance, Comfort and General Information. However, MMT performs at lower levels across all classes. For example, many segments from the Interior are falsely predicted as the Exterior

Table 5.26: Reporting arousal and valence for **MuSe-Topic** (using EWE annotation fusion) in a score combining Unweighted Average Recall (UAR) and F1 ($0.66 \cdot \text{F1} + 0.34 \cdot \text{UAR}$) on the devel(opment) and test partitions. Audio feature sets: eGeMAPS (eG), Deep Spectrum (DS), and VGGish (VG); Vision features sets: GoCaRD (Go), VGGFace (VF), and Xception (X); and Text feature set FastText (FT), are fed into the models. Furthermore, high-level text concepts based on contextual sentics embeddings, either encoded through n-hot vector embedding or the neural network (NN), are evaluated. Furthermore, all vision features (aV) are utilised by LSTM-SA. The features are aligned to the label timestamps. The by-chance level is 10 %.

Approach	**Modality**	**Feature(s)**	**Combined**	
			devel	test
Official Baselines [26]				
Unimodal				
LSTM-SA	A	DS	17.50	34.74
		eG	16.75	34.27
	V	X	24.14	36.75
		aV	25.43	34.75
	T	FT	21.44	36.20
ALBERT	T		**70.96**	**76.79**
Multimodal				
MMT	T+A+V	FT + eG + X	44.86	51.81
		FT + eG + VG	41.62	48.84
		FT + eG + AU	44.33	52.98
		FT + eG + OP	42.67	51.05
LSTM-SA	T+A+V	FT + eG + aV	25.03	37.14
Post-Challenge Models [56]				
SenSA5	T	n-hot	56.18	66.15
		NN	47.08	56.71
SenSA6	T	n-hot	46.22	57.09
		NN	40.67	49.01

class. In addition, there is a high level of confusion between Interior and Comfort. The prediction of Safety also fails almost completely for MMT, while ALBERT shows an almost error-free classification. This highly divergent behaviour suggests that language may be a dominant factor in learning this class. Both also show weaknesses in predicting User Experience and often confuse this with the close proximate class Interior. Unclear delineation in the definition (see Table Figure 4.7) and discrimination problems in human annotation may have played a role here.

These results demonstrate that text is exceptionally well-suited for speaker topics prediction. Multimodal approaches showed strong results, but were not able to keep up with the ALBERT. This can be an indication that a Transformer architecture with more profound modalities, such as ALBERT, could outperform these architectures in the future.

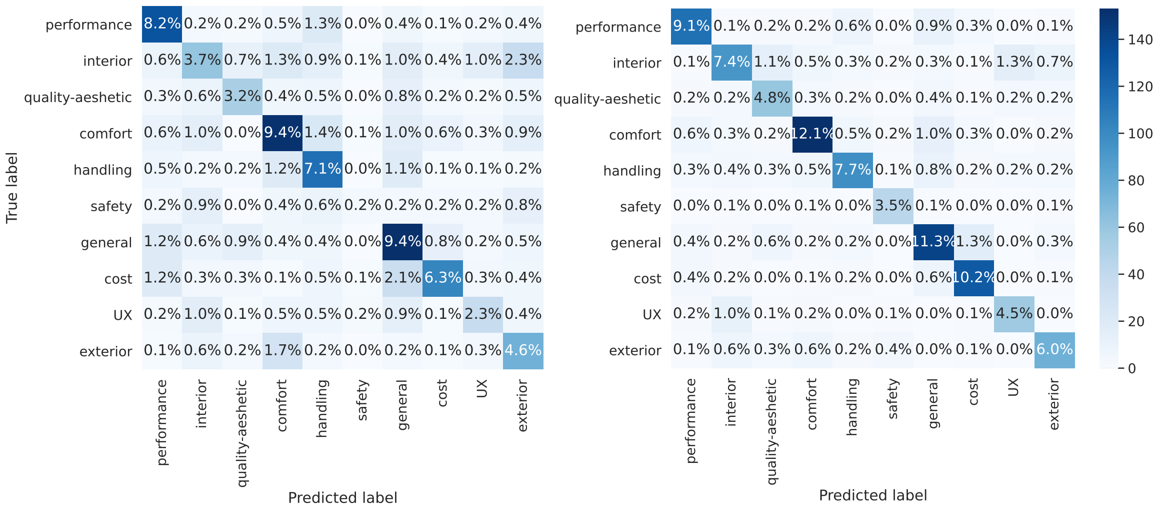
SenSA: The learnt embeddings and models building on SenticNet-5 extractions seem more predictive than SenticNet-6 with an almost 15 percentage point improvement for the task. The domain-specific representations using the Artificial Neural Network (ANN) architecture achieved solid results with 56.71 % combined score on the test set (see Table 5.26) using version 5 extraction. The naïve n-hot encoded representations in combination with a SVM performed even better, achieving 56.18 % combined score on the development set and 66.16 % combined score on the test set.

As shown in Figure 5.14, the prediction errors are relatively evenly distributed, except for the Interior and Aesthetics and Cost and General Information classes, each of which the algorithm confuses more frequently. Compared to the baseline models, the prediction behaviour shown by the confusion matrix is very close to that of the other language-driven model ALBERT, but at a slightly lower level. For example, a confusion between the class Cost and the classes General Information and Performance is evident as in MMT. The result for the class Interior is comparable to ALBERT and superior to MMT, which shows the potency of Transformer textual representations in drawing a fine line between them and similar classes.

In summary, the semantic concepts SenSA prove to be very predictive, which indicates a good contextual understanding, and are only beaten by the ALBERT. However, one should bear in mind that the state-of-the-art Transformer was pretrained with considerably more data, both unsupervised and supervised, and possesses substantially more parameters, which leads to higher computational costs. As a result, the other models such as the LSTM-RNN with self-attention are more than 30 percentage points and the multimodal Transformer [26] is almost 15 percentage points behind the sentic learning results.

5.2.2.4 Conclusions

From these experiments, it is evident that the text modality is the most effective one in predicting the targeted speaker topics (**RQ-3**). This can mostly be attributed to the content-dependency of the prediction target. The deeply integrated multimodal fusion of the MMT experimentally revealed small differences in the results depending on the video feature selected. Here, both the environmental representation Xception and FAU proved most predictive, suggesting that other modalities may be beneficial for in-depth contextual understanding (e. g. , the visual representation captures a person driving, indicating an indoor topic). However, the unimodal encoding of the linguistic cue is clearly decisive.



(a) MMT

(b) ALBERT

Figure 5.13: Relative confusion matrix predicting 10 speaker topics (MuSe-Topic) of fine-tuned MMT and ALBERT on the test set.

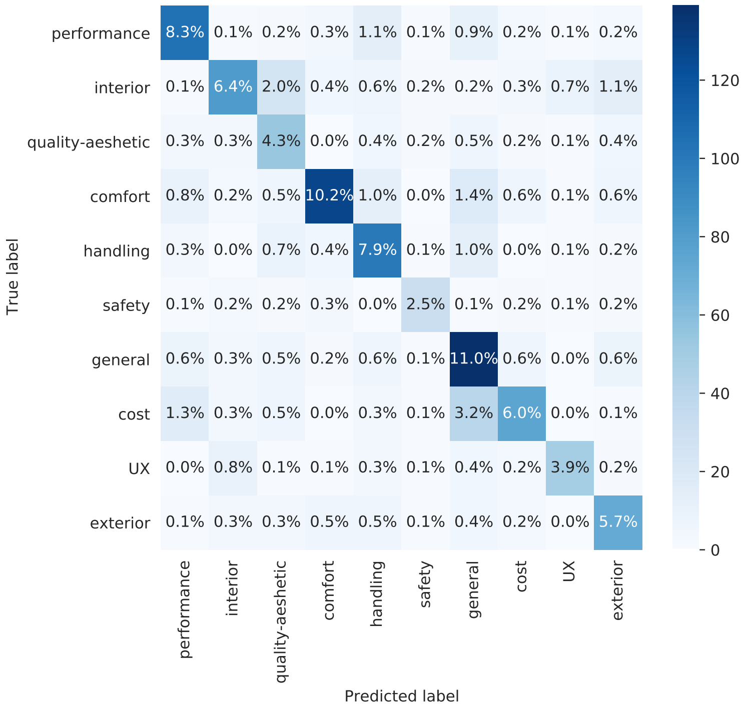


Figure 5.14: Relative confusion matrix predicting 10 speaker topics (MuSe-Topic) of SenSA5 on the test set. Figure taken from Stappen et al. [56].

Regardless of the model design, text-focused models naturally use the semantics of video transcripts, achieving the most promising results. However, the results of the text-based models vary greatly (**RQ-2b**). This demonstrates the importance of dedicated architectures for this particular subtask of MSA. The parameter-heavy NLP Transformer ALBERT dominates the benchmark results. Although this form of BERT Transformer has almost 85 % fewer parameters than other BERT variants [212], computational limits were encountered during the network training, leading to a restriction of the segment window length and, thus, an information loss. This led to developing the second-best, parameter-light, subsymbolic SenSA architecture. The results generally highlight great strengths in the inclusion of sentic concepts in the modelling of this task. Furthermore, the integration of high-level human concepts requires drastically less representational encoding, and the architecture is simpler.

5.3 Towards Modality Inference for Real-life Videos

Video data acquired from in-the-wild environments poses several challenges for the development of reliable prediction systems (see Section 2.1). Available information is often of low quality or feature extraction fails when the object of interest is missing or (partially) occluded at a certain time point in a data stream [9]. In the context of this work, speech and facial expressions are used in Section 5.1 and Section 5.2, for example, predicting emotion-related targets. Some of these systems occasionally performed worse than expected. One issue is that a face can be shaded or unfavourably positioned or a person stops talking for a few moments [320], so that representations are missing or are of low quality. This issue is aimed to be addressed in the following manner:

- **RQ-4:** Investigating capabilities for modelling the cross-modal dynamics of facial muscle activity based on voice in a sequence-to-sequence prediction scenario.

5.3.1 Cross-modal Recognition

5.3.1.1 Characteristics

In the award-winning publication Stappen et al. [204], explored several ways of how facial muscle movements can be estimated from speech signals. A first attempt of this task has been made by Ringeval et al. [320], which simplified this highly complex task of fundamental research into subtasks. The subtasks of onset, apex, offset, and occurrence of a particular FAU from the Facial Action Coding System (FACS) system (see Section 3.1.3) are individually estimated by a classifier. In the approach taken in this work, robust sequence-to-sequence ANN architectures utilising several attention mechanisms are first developed on a single well-suited FAU (chin raiser) to understand effective network mechanisms. Then, the efficiency is compared to the results of Ringeval et al. [320]. Finally, the results will be linked back to the posed RQ.

5.3.1.2 Experimental Setup

Architectures: This work proposes an encoder-decoder architecture (see Figure 5.15) and a stacked, bidirectional LSTM-RNN(s) architecture to model the sequence-to-sequence problem of predicting FACS from speech signals. The encoder distils the input sequence to an abstract representation. The decoder encodes the representation to a converted output sequence of another modality. To improve the encoding-decoding process, attention is commonly applied to put more weight on relevant input steps

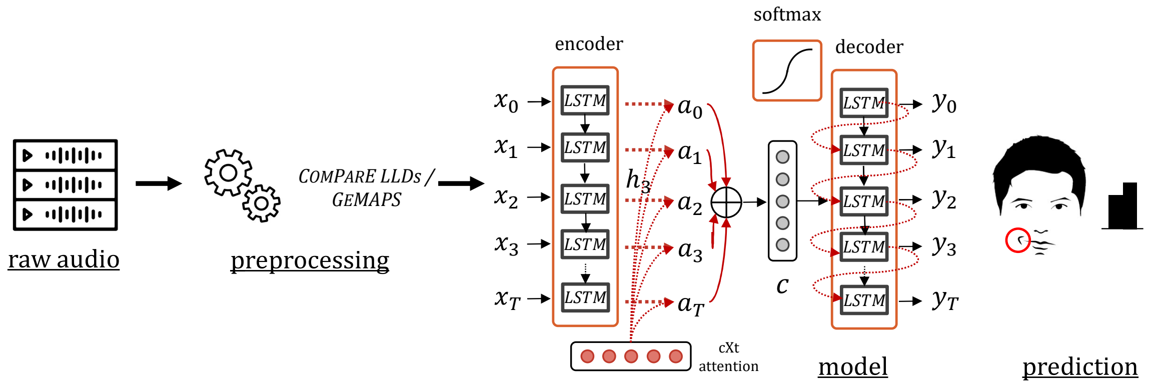


Figure 5.15: Pipeline for cross-modal prediction from an audio signal to FAU through an encoder-decoder architecture with a *cXt* attention module. First, to extract the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and ComParE LLDs representations, the raw audio original is preprocessed. The input sequence x is encoded, context vectors x computed to enhance decoding, and, finally, combined with $\hat{y}_{t-1}$, $\hat{y}_t$ is sequentially predicted.

and align the sequences dynamically [256, 249, 98]. These experiments distinguish between two attention types: context attention (*cXt*) and window attention (*win*), as described in Section 3.2.4. In the encoder-decoder architecture, the latter can attend the previous time steps (*pa*). An attention regulariser weight of 10^{-4} is added to the context attention. The bilayered stacked architecture passes the prediction through time without encoding the whole sequence at first. The LSTM-RNN layers are bidirectional and fused using summation. Between the stacked layers and before the prediction layer, 0.2 dropout is applied. The output transformation is done by a softmax prediction layer. Also here, *win* attention is feasible but since there are no sequence constraints of the decoder as for the encoder-decoder model, the attention windows can capture the context from both directions *bi*. The number of neurons for all hidden layers of both architectures is set to 20 in accordance to the rather low-dimensional input audio feature sets. Furthermore, both apply L1 bias regulariser of 10^{-4} and a L2 kernel regulariser of 10^{-1}.

Baseline: There is an internal and an external baseline to which the advanced architectures are compared. The external baseline is the result of the 3-layered stacked LSTM-RNN architecture of [320]. The following experimental settings correspond to this baseline, e. g. , in terms of data, so that a fair comparison is possible. Furthermore, to evaluate the positive effect of attention, the internal baseline is the architectures without any applied attention (*No-Att*).

Dataset: As a high-quality data source, the *General Multimodal Emotion Representations* (GEMEP) database, which was specifically designed to enable fundamental multimodal research, is chosen [321]. Comprising 7 000 audio-video emotional portrayals, it depicts 18 emotions acted by 10, equally balanced female and male, actors. The portrayals show vocal interactions in French, captured by a static camera setting with a frame rate of 25Hz. One of the three cameras is pointed at the actor's face and additional microphones recording at 44.1kHz are positioned next to the left ear of a playing actor [320]. For comparability, the focus of this work is on a subset of 158 portrayal segments, provided by [320] for tasks aiming at cross-modelling the speech and vision modalities. It was carefully designed to avoid any modelling imbalance, using only segments which had a high label agreement rate between annotators, a minimum of 5 % occurrence of each FAU, and a balanced amount of recordings from each actor.

Preprocessing: Audio-video synchronisation and segmentation is based on the manual effort of [320], which ensures highly accurate alignment between modalities. Furthermore, the loudness of the audio is normalised to 0 db peak amplitude. This study is designed with emotion recognition in mind, which is why the use of domain-typical representations and targets for modelling is a natural choice. From the audio data, two low-level descriptor acoustic representations, ComParE LLDs and GeMAPS, are extracted at a sample rate of 100Hz. Speaker-wise normalisation is performed on the LLDs in respect to the variance of representations over the different actors in relation to the facial action units, while all other parameters are kept as described in Section 3.1.1. In addition, to smooth the GeMAPS, a symmetric moving average window of three frame lengths and a context window of three is utilised. To receive a systematic coding of different facial expressions, the FAUs [37] based on the FACS are utilised as **labels**. While the systematic coding is identical to the automatic extraction as explained in Section 3.1.3, the focus is on a limited set of eight FAUs connected to voice, which are manually annotated by two independent FACS encoders: Inner Brow Raiser (FAU-1), Outer Brow Raiser (FAU-2), Brow Lowerer (FAU-4), Cheek Raiser (FAU-6), Lid Tightener (FAU-7), Nose Wrinkler (FAU-10), Lip Corner Puller (FAU-12), and Chin Raiser (FAU-17). A sequence of continuous intensity values for each point in time is transformed from a regression to a classification problem by changing the intensity to sequences of binary activation as preprocessed in [65]:

1. The **onset** corresponds to the start of muscle activation, thus, the period when the intensity of an FAU increases for longer than 2 consecutive frames.

2. The **apex** indicates the peak of muscle activation after an onset and before an offset.
3. In contrast to the onset, the **offset** is the part of a sequence in which the muscles are released. The start point is the end of an apex and the end point is either an intensity of 0 or the start of another onset.
4. The **occurrence** covers any kind of facial muscle activation, independent of onset, apex, or offset.

Measure: The evaluation is done following a speaker-independent leave-one-speaker-out (LOSO) cross-validation on each combination. As with this work, this approach is often chosen for small datasets and is also applied by the reference work [320]. In a binary case, the by-chance level of the calculated UAR is 50 % independent of the class distribution. The average across all partitions is reported.

Training: The LOSO validation method limits the options of hyperparameter search [204]. For this reason, a few state-of-the-art techniques are applied to achieve a smooth training behaviour of the networks but refrain from comprehensive parameter tuning. Batch normalisation reduces the training time and leads to stable information passed to downstream layers [322]. The length of the sequences is set to 100, so that shorter sequences are zero padded and network parts, related to missing timestamps, are not updated (masked). Since the size of the dataset is rather small, the maximum number of epochs is set to 100 and early stopping with patience of five is applied to avoid overfitting. Adam is chosen as the gradient descent optimiser, setting the *lr* to 10^{-4}, with a *bs* to 32. Finally, the data points of each class are weighted in the loss function to prevent class imbalance effects.

5.3.1.3 Results

In the first experiments, the focus is to evaluate various architectural configurations using ComParE LLDs representations to identify components which are particular suitable for this task.

Architecture comparison: As shown in Table 5.27, the **stacked** architecture (*Stacked-winAtt*) performs better with larger context windows ($bi = 5$ vs $bi = 15$). Adding more stacked layers (2*x*) also improves the results, e. g. , for occurrence from 70.2 % UAR of the $bi = 15$ single-stacked version to 76.6 % UAR of the double-stacked one. This is also the overall most successful configuration for occurrence; however, the onset, apex, and offset results increase when only the past (*pa*) time steps are

Table 5.27: Results are reported on the exemplary FAU 17 using the stacked (*Stacked*) and encoder-decoder architectures (*EncDec*) in three configuration types: no attention (*NoAtt*), either bidirectional (*bi*) or only past steps (*pa*) local attention (*WinAtt*), and context attention (*cXtAtt*). The COMPARE LLDS features are used to report the unweighted average recall as a percentage for onset, apex, offset, and occurrence (occur.) labels. Table taken from Stappen et al. [204].

Architectures		**FAU17**			
Name	**Window**	**Onset**	**Apex**	**Offset**	**Occur.**
Stacked-NoAtt	-	75.8	73.1	80.2	75.4
Stacked-WinAtt	bi = 5	62.9	62.1	66.8	61.9
Stacked-WinAtt	bi = 15	79.1	70.3	74.9	70.2
2xStacked-WinAtt	pa = 15	76.9	76.3	80.7	65.7
2xStacked-WinAtt	bi = 15	75.7	72.8	75.4	**76.6**
Enc-Dec-NoAtt	-	77.0	76.8	83.3	73.8
Enc-cXtAtt-Dec	-	**80.5**	**77.6**	**89.2**	73.3
Enc-cXtAtt-Dec-WinAtt	pa = 15	76.2	76.4	83.9	71.1

considered. The **encoder-decoder** architecture using only the *cXtAtt* module performs strongly, achieving the best results on onset (80.5 %), apex (77.6 %), and offset (89.2 %) subclasses. For both architectures, it can be seen that just because one of the subclasses improves, it does not necessarily lead to an improvement in the occurrence class. It may be that the more frequently changing binary labels of occurrence are more challenging to learn from the immediate neighbours, which is underlined by the *NoAtt* results for this class, which either did not change or only slightly improved in contrast to the subclasses.

The next series of experiments extends the view from a single FAU-17 to all FAUs and compares the results to the LSTM-RNN baseline architecture. Including the four subclass experiments for each of the eight FAUs, ten separately trained models for each partition, and two feature sets (GeMAPS, ComParE LLDs) results in more than 640 models. This high computational effort makes it necessary to limit the experiments to only the encoder-decoder with context attention.

Baseline comparison: Compared to the baseline models [320], all subclasses show increased average results except for the combination of occurrence and GeMAPS representations. Specifically, using COMPARE, the average results improved on a percentage point basis over those in [320] by 1.9 to 63.9 % UAR on onset, 3.1 to 66.2 % UAR on apex, and 19.4 to 79.7 % UAR on offset. A similar picture is obtained when looking at the GeMAPS use, achieving even better results for onset with 65.9 % and apex with 67.8 %. One explanation is that both the enhancement by attention and the addition

Table 5.28: Full results on all FAUs based on the training of the proposed encoder-decoder architecture (*Enc-Dec-cXtAtt*) with the GEMAPS and COMPARE features and an absolute comparison to [320] shown in brackets. The FAUs are as defined in Section 3.1.3: Inner Brow Raiser (FAU1), Outer Brow Raiser (FAU2), Brow Lowerer (FAU4), Cheek Raiser (FAU6), Lid Tightener (FAU7), Nose Wrinkler (FAU10), Lip Corner Puller (FAU12), and Chin Raiser (FAU17). Results are given in Unweighted Average Recall (UAR) as a percentage. Table taken from Stappen et al. [204].

	COMPARE				**EGEMAPS**			
FAU	**Onset**	**Apex**	**Offset**	**Occurrence**	**Onset**	**Apex**	**Offset**	**Occurrence**
1	64.2 (+2.3)	68.1 (+3.6)	80.0 (+20.5)	69.7 (+2.6)	69.7 (+7.6)	67.1 (-0.8)	81.0 (+19.8)	63.8 (-3.8)
2	60.9 (-3.4)	73.3 (+5.9)	70.9 (+8.6)	68.1 (-1.2)	64.8 (+0.2)	63.2 (-7.8)	66.5 (+3.5)	65.2 (-5.7)
4	68.9 (+7.5)	65.8 (-0.4)	75.6 (+12.3)	67.6 (+3.0)	64.7 (+2.8)	72.9 (+6.2)	73.3 (+10.1)	61.5 (-6.4)
6	65.1 (+1.1)	64.2 (-1.0)	90.3 (+26.9)	64.0 (+0.2)	65.7 (+1.4)	72.9 (+5.2)	81.4 (+18.2)	62.2 (-1.0)
7	62.0 (-0.7)	65.3 (+8.7)	85.6 (+27.3)	60.3 (+5.9)	62.8 (+0.0)	65.7 (+6.0)	69.7 (+8.2)	57.8 (+5.1)
10	55.7 (-6.1)	56.2 (-4.3)	73.8 (+17.4)	60.1 (-0.3)	59.5 (-2.5)	62.6 (+1.5)	91.1 (+33.1)	57.8 (-2.5)
12	53.6 (-5.7)	59.5 (-0.4)	72.0 (+15.3)	59.0 (+0.4)	60.5 (+2.5)	67.2 (+6.3)	79.6 (+21.8)	62.0 (+3.3)
17	80.5 (+20)	77.6 (+12.8)	89.2 (+26.9)	73.3 (+8.7)	79.6 (+18.0)	70.6 (+5.6)	79.8 (+16.7)	67.1 (+1.3)
Avg.	63.9 (+1.9)	66.2 (+3.1)	**79.7** (+19.4)	**65.3** (+2.4)	**65.9** (+3.7)	**67.8** (+2.8)	77.8 (+16.4)	62.2 (-1.2)

of state-of-the-art network properties, such as batch normalisation, early stopping, and class weights, are the main contributors to this change. Interestingly, the largest increases in performance are seen in the offset subclass results regardless of FAU, while the results on occurrence are mixed across FAUs. Compared to the occurrence prediction in [320], the results improve by 2.4 to 65.3 % UAR using COMPARE, however, they fall behind by -1.2 using GeMAPS. On an individual level, some FAUs stand out. FAU-7 (Lid Tightener), FAU-12 (Lip Corner Puller), and FAU-17 (Chin Raiser) seem especially suitable for this task, reaching new state-of-the-art results on all subclasses and feature sets. In contrast, the occurrence results on the FAUs related to the brow area of the face (inner = 1, outer = 2, lower = 4) seem especially hard to predict from the GeMAPS voice representations.

5.3.1.4 Conclusions

The experiments demonstrated that there are human-centred relationships which can be inferred across the audio and video modalities (**RQ-4**). For this, FAU-17, Chin Raiser, which is strongly connected to mouth movements, was first evaluated. Second, the encoder-decoder architecture was used to show generalisability of the designed network on all FAUs. The architectures seem to profit from the attention mechanism, with an advantage for context over local attention, and other state-of-the-art settings which improve sequence modelling. The best average results predicting FAU were achieved using GeMAPS with 65.9 % UAR on the onset and 67.8 % UAR on the apex, while using COMPARE was more effective on offset with 78.7 % UAR and occurrence with 65.3 % UAR.

The results suggest that the reasons for the improvement are manifold. Improved mechanisms such as batch normalisation and class weights lead to more robust predictions. The clear advantages of the attention mechanism can presumably be attributed to improved temporal modelling, clearly evident in the case of offset. It can be suspected that the mixing of onset, apex, and offset in the occurrence subtask does not allow these to be modelled equivalently, as the prediction target is exclusively binary and therefore in-between states cannot explicitly be modelled by the advanced architecture.

This improved understanding of cross-modal interactions paves the way for applications to in-the-wild data. Specifically, similar architectures can be applied to the problem of imputation, so that parts of a disturbed sequence are replaced by estimations. However, since inference from one modality to another is naturally limited, thus cannot now and probably never will be perfect, it seems best that the context of the target modality should also be integrated in such an approach. The additional information from a (past) context of the target modality ought to make inference easier. The experiments demonstrated a way towards this idea by only considering a limited context window. In this regard, it might be worthwhile to look for stronger alternatives to local attention windows, which work well in this setting, but do not fully reach the level of contextual attention where the whole (source) sequence is considered. Furthermore, it remains to be thoroughly evaluated whether large datasets, as with in-the-wild data, can stabilise the problem to the extent that regression points can be directly estimated rather than the simplified binary target subclasses employed in this experimental evaluation.

Concluding Remarks

6 Concluding Remarks

This chapter first provides a summary of the findings in Section 6.1, followed by the ethical and social considerations, which have been a continuous companion to this work, in Section 6.2, before concluding by outlining limitations in Section 6.3, and future directions of this thesis's research in Section 6.4.

6.1 Summary and Discussion

Representing an emerging field of Deep Learning research, Multimodal Sentiment Analysis seeks to structure the fastest-growing human-made information sources of the 21st century — unstructured user-generated data — into emotional and thematic contexts through Machine Learning (see Chapter 1). For this purpose, approaches from the fields of signal processing, Affective Computing, and Natural Language Processing are united to develop methods that can decompose video, a multimodal medium, into its three components to automatically learn and recognise unimodal as well as overarching dependencies between multimodal combinations. Starting from current research trends (see Chapter 2), this work has identified shortcomings in the latest research and defined new directions to establish a closer link between the MSA and AC communities. Methodologically, these challenges can be addressed by proposing new methods, largely grounded in key concepts of modality representation and deep learning (see Chapter 3). However, the subjective and objective dimensions of MSA could not sufficiently be addressed with existing training material. This led to the creation of a new dataset, MuSe-CaR, collected specifically with our research objectives in mind (see Chapter 4). It provides the ability to fully integrate the audio-video and linguistic components, as well as to predict continuous emotions and targets in context. All three modalities exhibit challenging in-the-wild characteristics, for example, domain terminology in automatically transcribed spoken word, free-floating video perspectives, and ambient audio soundscapes. Furthermore, source-specific annotations were proposed, such as perceived trustworthiness in user-generated content and speaker topics in multimodal scenarios. In addition, the MuSe-Toolbox was introduced to facilitate the generation of emotional gold standards for MSA. It provides a new gold standard RAAW and data-driven emotion class creation. RAAW is motivated by the manual effort when aligning several annotators due to the rater reaction delay. It extends the idea of agreement-based weighted annotators, EWE [3], while counteracting holistic individual lagging rater reaction [278] through GCTW alignment [277].

Equipped with new data and tools, novel architectures were proposed and experimental frameworks set up to investigate the initially posed research questions (see Chapter 5). The main focus of this work was on the emotion dimension of MSA (**RQ-1**). First, the thesis addressed to what extent and how time- and value-continuous arousal and valence dimensions can be effectively predicted from real-world, user-generated content (see Section 5.1.1). For this purpose, a variety of audio (DEEP SPECTRUM, eGeMAPS, LLD, and VGGish), vision (Xception, VGGFace, and FAU), and text (FastText and BERT) representations were extracted. Targeting two different gold-standards (EWE and RAAW), several sequential models were proposed for this sequential regression task, both without attention mechanisms (LSTM-RNN and End2You) and with them (LSTM-SA and MHA-LSTM). It is clear from the experiments that the audio features are the most effective at modelling arousal, and the text-based ones outperform all others on valence. This is found to be consistent with other work [8]. Learnt representations from ANN mostly performed stronger than hand-crafted ones. Having already been successfully used in the context of audio-video emotion recognition on smaller, less realistic datasets [79], the results clearly underline their advantages in the context of exceptionally large-scale, in-the-wild data. In particular, BERT, which calculates the embeddings at run-time to integrate context, proved to be highly effective on automatically transcribed spoken language. The merits of such text embeddings for continuous-time prediction targets are also valuable insight for the interdisciplinary field of MSA. Many researchers from the field of MSA have their origins in linguistic analysis with segment-based discrete targets [28, 34] and are unfamiliar with the possibilities of the continuous-time form, while the AC community, for its part, has largely neglected text [8]. Generally, it can be expected that data-driven representations will fully replace hand-crafted representations in the long run as interest in analysing real-life videos increases. In addition, equivalent results using EWE and RAAW suggest that the proposed gold-standard method is similarly suitable for prediction while providing improved theoretical properties. Other work with RAAW also suggests that the integrated alignment has advantages when it comes to fusing human-made annotations with machine-recorded biological signals, such as arousal and electrodermal activity for stress recognition [281]. Among other findings on useful architectural features (see RQ-3 below), the integration of the multihead attention mechanism before sequential coding in the MHA-LSTM leads to a robust internal encoding of a representation. This is promising as it shows that modelling accuracy in the in-the-wild domain can be further improved by more advanced ANN attention mechanisms, suggesting that pure-attention networks will be even more successful in the future (see Section 6.4).

Second, this work experimentally addressed whether these time- and value-continuous arousal and valence annotations can be transformed reasonably into segment summary classes to allow a combined emotional and thematic summary understanding on the same granularity (see Section 5.1.2). A naive idea towards a more intelligent unsupervised learned method was also explored. Among the proposed models for prediction, the knowledge-based SenSA architecture yielded the strongest text-based performance, while the MMT was the most successful overall on the naive classes. However, the results remain at a low level. The prediction results on the learned classes were found to be quantitatively improved. Again, audio representations for arousal classes and text representations for valence classes, especially BERT, exhibit the best performance. This consistency suggests that fundamental traits are preserved in the class creation process. Qualitatively, the visualisations provide insights into typical (temporal) characteristics of the classes, but this is not comparable to the ease of interpretation of conventional emotion classes [37]. Considering the novelty of time-and-value compression compared to previous value-only quantisation [104] and extrapolation [102] approaches, these first experimental results bear further potential for improvement towards dynamic granularities in MSA (see Section 6.4).

Third, motivated by the success of time-continuous annotations and the respective origin of the data from online sources, it was investigated for the first time whether models can be developed to quantify perceived trustworthiness (see Section 5.1.3). Due to the similarity of the tasks, the experimental design, representations, and architectures could benefit from the findings of the arousal and valence dimensions. The resulting DeepTrust architecture employing multihead attention outperformed all baselines by a large margin of 50 %. The content conveyed (text) seems to be of the greatest effectiveness for prediction, followed by audio and video (e. g. , face) signals. With regard to the architectural design, it has been shown that multiple attention heads and a large segmentation window are beneficial. This suggests that long-term context is more influential for perceived trustworthiness than, for example, for arousal and valence. For many use cases, such as consumer decision-making [48], a fine-grained assessment of credibility offers a new entry point for research that can profit from the scale of social media.

Finally, a potential new field of utilising time-continuous annotations in the context of MSA to predict video popularity (e. g. , view, likes, comments) was outlined (see Section 5.1.4). This way differs substantially from other attempts that have used metadata and text such as comments [125, 126]. Interpretable patterns of up to medium-strong feature-target correlations were found. For example, the absolute energy of the annotation course of all three dimensions is positively correlated to views per day. Considering the findings in the context of previous research, certain emotional characteristics found could support content creators' abilities to develop a parasocial relationship with their viewers in a more purposeful way [117]. To explore the prediction of a video's popularity, a semi-automatic and an automatic feature selection method combined with an interpretable SVR was proposed. Given this exclusively one-dimensional view, the automatic feature selection method yields strong results based on features extracted from the valence annotations. This is likely caused by the inherent sentiment relationship of the predicted user-engagement criteria.

The second elementary dimension of MSA, the emotion's target (**RQ-2**), was attempted to be extracted (almost) without human intervention and predicted using human annotations in the form of objective speaker topics (see Section 5.2). First, experiments were conducted to determine whether coherent speaker topics of a (video) corpus can be extracted without a priori assumptions (see Section 5.2.1). For this, a graph-based method using only the transcripts, GraphTMT, was proposed. It consists of specifically designed edge removal mechanisms (e. g., PTH) and representative-word prioritisation mechanisms, (TVS and NDC). As the results have shown, these allow for a high robustness of outcome with very little to no need for fine-tuning, which previously caused low-quality outcomes when using other methods [150]. The success was measured using three different evaluation approaches. The proposed method outperformed all baselines (LDA, HDBSCAN, and K-means) in terms of measured coherence of the topics. Additionally, intratopic assessment found fine-grained clustering. Finally, this method resulted in the most interpretable clusters with the highest average hit rate in an intruder task conducted as a user study from a human-centred evaluation perspective. The high error rate in transcripts is considered a severe difficulty of this task [93]. The experiments suggest that the iterative approach to reduce edges based on linkage distances seems to be particularly robust for large-scale automatic transcripts. As a side experiment, the generalisability of the approach was investigated using a text evaluation dataset (Citysearch), where GraphTMT also showed competitive results in terms of coherence and intratopic evaluation. These promising findings demonstrate that graph-based machine learning models are a genuine alternative in topic modelling.

Secondly, within the context of RQ-2, it was explored whether and how human-annotated speaker topics can be effectively predicted (see Section 5.2.2). The multimodally annotated speaker topics differ from the typical purely textual perspective [10]. In a rigorous set of experiments, a large collection of representations and classifiers suitable for detecting ten speaker topics from video segments was adopted. Besides the architectures already used in the previous emotion-related tasks (LSTM-SA, MMT, and End2You), networks optimised for an advanced linguistic understanding (ALBERT and SenSA) were introduced. The results clearly support the understanding that text is the dominant modality for learning content-related topics. Combined input from text and other modality representations exhibited slight improvements, predominantly for face-related and environmental visual features. It can be speculated that this relates to the domain context, where very strong visually discernible reactions might serve as an additional indicator for certain classes (e. g. , broad smile under strong acceleration). Training a state-of-the-art Transformer (ALBERT) obtains the strongest performance. However, the proposed parameter-light and second-best subsymbolic architecture SenSA integrating common knowledge concepts needs only a fraction of the parameters. In the sense of MSA, this opens up an efficient alternative, for example with regard to real-time MSA.

As given by the name of the field, exploring inter- and intra-modal dynamics of the three core modalities was a steady companion (**RQ-3**). From the unimodal perspective, most notably text, in its original representation of symbolic and irregularly occurring strings of words in time, has been proven to be extremely useful, despite the difficulty of aligning it with regularly sampled audio and video signals. Similar to other work [98], the proposed models exploiting cross-temporal attention mechanisms were able to cope with the poor quality of automatic transcription of colloquial utterances and misfit to continuous-time prediction targets more efficiently than models without. While there are clear strengths of one modality depending on the prediction goal (e. g. , audio for arousal), multimodal fusion almost invariably showed more effectiveness in almost all subtasks of MSA. This proves to be very pertinent for the emotional dimension. In this work, early (e. g. , LSTM-SA), late (e. g. , DeepTrust), and hybrid fusion (e. g. , MMT) were explored. Early fusion often achieved very good results on the development set, but failed to replicate on the test set. It is known that a large input dimension in networks lead to bad generalisation capabilities [42]. Hybrid fusion holds great promise for the future, but is currently still subject to many limitations (see Section 6.3). Therefore, in this work, intermodal, temporal late-fusion ANN proved to be a good compromise between strong unimodal predictors, reasonable training duration, and learning temporal, intermodality dynamics.

Finally, to address the need for a more profound understanding of cross-modal relationships, it was explored whether attention networks are suitable to explicitly model facial muscles from the spoken word (**RQ-4**). The mechanisms were investigated in a fundamental study predicting facial muscles from the voice on the multimodal dataset GEMEP. The developed double-stacked LSTM-RNN with bidirectional local attention windows effectively predicted the occurrence of chin raises and the encoder-decoder architecture with context attention showed remarkably robust onset, apex, and offset prediction. This architecture was applied on all FAU, and achieved several new state-of-the-art results, e. g. , predicting lip corner puller and lid tightener, compared to previous benchmarks from neural networks without attention mechanisms. It is likely that the reason for the improvement lies in the refined temporal modelling through attention weights applied after the input encoding. By weighting the individual sequence steps depending on each output decoding step, the coding bottleneck is bypassed. In doing so, no information is lost compared to first compressing the entire sequence into a single representation for the subsequent decoding. In addition, state-of-the-art mechanisms such as batch normalisation and class weighting stabilise the training procedures on this reasonably small dataset.

6.2 Social and Ethical Considerations

Throughout this work, the data collection and annotation process has been intensively explored and, above all, models for emotion and object recognition for real-life environments have been developed. In recent years, emotion recognition systems, especially those making decisions about humans, have faced hefty criticism [323–325]. While initial ethical discussions on the theory of emotional machines themselves date back a long time [326], many practical implications are only slowly surfacing as technology continues to advance and new sources of data emerge (e. g. , the internet, smartwatches, CCTV surveillance). For example, in the context of this work, ethicists and privacy advocates have intensified efforts to elaborate whether working with data from the public domain (e. g. , the internet) can pose risks to individual privacy [327–329]. Constant companions of all the research and implementation decisions faced here have been ethical, legal, and social considerations. This section examines these aspects collectively and presents them to the reader in a structured way. For each section, the basic theoretical principles are presented first, followed by the use case specific policies derived for the crafted dataset, organised challenges, and designed models. As in previous work [7], this is approached from the data collection and the potential application side.

6.2.1 Data Collection

Subject of investigation: The people displayed and the data's origin has a strong influence on the sensitivity of the data. As with data characteristics (see Section 2.2), these can be roughly divided into three groups, though the boundaries are fluid and different characteristics can occur in each case. In the first group fall data collections from the beginning of the emotion recognition field. These often include posed situations that actors had to imitate [321]. These performances are, therefore, in a thoroughly professional environment, whose participants are ordinarily compensated. These data are worthy of protection, but since neither the artificial context, the generated emotion, nor the environment allows conclusions drawn about the person, personality rights are scarcely affected. A side effect is that the recorded emotion does not necessarily correspond to a natural person's authentic reaction, which are often internalised because most people associate a specific behaviour such as gestures and facial expressions with an emotion. In this setting, aspects such as spontaneity or naturalness of an affect cannot be examined at all [10]. The second group consists of non-professional performers, such as students. In these, recordings capture the exchange of personal views on a topic, only led by a rough conversation guideline [12]. The monetary compensation of the subjects is presumably lower than that of actors, and the impressions conveyed correspond to their nature. This data is particularly worthy of protection. For example,, in therapeutic settings [330], personal experiences are shared, and participants in the study will always remain identifiable from their personal environment [7]. Informed consent and thorough explanation are necessary [331]. The trend goes towards the last group, collection of data whose use for a study was not apparent to the participants at the time of observation [11, 27, 58]. From a research perspective, this is highly desirable and an accepted solution, as the consciousness of a study can influence the subject's behaviour being studied. Databases whose consent is obtained after recording are very rare [332]. More often, however, data recorded for another purpose is used [58, 67] with the particularity that the data is freely available on YouTube. From an ethical point of view, this raises several questions: The participants whose private states are mined have no awareness of the study, so they are unlikely to be compensated for it, and, in the context of internet sources, publicly visible content, free of charge, does not comply without rights [327].

Dealing ethically with data that is freely available on the internet is relevant for this work since the dataset also falls into this category. For this, a preliminary review describing the study's aims in a structured manner was submitted to the Ethical Board and the Data Protection Commissioner of the university [24]. As in other works [327], both parties concluded that no in-depth review or specific approval is necessary from the Boards and the participants. The reasons given are a) the data (car reviews) per se do not touch any personal rights, b) are already recorded and publicly available, therefore researchers do not interact with the subjects and creators should have a general understanding that the data might be analysed when openly accessible, and finally, c) no harm to the health of the database users can result from watching the videos.

However, copyright concerns have been raised because the videos have to be downloaded to be annotated and distributed to future challenge participants — without the explicit permission of the creators [327, 333]. For this matter, the legal doctrine of the Fair Use Principle exists in the US legal sphere [334]. It states that copyrighted material may be reused under certain circumstances without obtaining permission from the copyright holder. Research is particularly affected by this. As a result, research-relevant data from the public domain lose copyright protection for non-commercial use. A comparable principle does not exist in the European Union, entailing a risk for academic bodies [327]. As concluded in Stappen et al. [24], to legally protect researchers in other countries outside the US, it is necessary to contact creators individually. This process is a very time-consuming and hardly practicable task, as creators cannot be contacted via an in-messenger platform and private individuals tend not to provide an email address. In the case of this study, this led to relying more on (semi-)professional reviewers for data collection, losing more than 50% of the targeted selection size due to missing responses and contact information.

Environment: Not only the subjects themselves but also their surroundings and public spaces display privacy-sensitive information. The camera might unknowingly catch other people during an in-the-wild recording [12], the immediate environment displays private premises [58], or an object enters unexpectedly, for example, capturing the licence plates of passing vehicles [24]. Legally, the definition of private information varies between jurisdictions; however, research is done globally. For MuSe-CaR, the focus was, therefore, on videos that only depict one person. Even though people and elements in public areas can be camouflaged manually, this way is not scaleable in practice with increasing data volumes. At the same time, no automatic system is or will ever be perfect, so risk can only be reduced, not eliminated. Until almost perfect disguise systems are developed for large-scale use and are safe from unmasking attempts [328, 335], the responsible handling lies not only in the creation process but also in the hands of (academic) receivers of in-the-wild datasets.

Sharing and storage: One aspect of this handling is the storage. Numerous papers emphasise the importance of making collected data available for research [7, 327, 336], especially in the field of MSA, a tremendous challenge [7, 337]. Open data allows other researchers to build upon previous work, both from a scientific and technical viewpoint. Scientifically, reproducibility is ensured, and technically, data sharing allows building upon laborious work such as data preparation. Releasing the data comes with obstacles. While some aspects, such as protected storage, are well investigated, others are open to research. In order to legally pass on rights acquired from creators or, in other cases, participants to others, a licence agreement is needed. This contract, between the licence holder and licensee, is known as an End User Licence Agreement (EULA). In theory, it is legally binding and is intended to ensure ethical exploitation. However, a right without the possibility to enforce it is not a right. This problem becomes evident in the following example: Let us assume a German university grants an EULA of a depression dataset to a professor in a foreign country, who subsequently sells the data for an application to a start-up. This is a clear violation of a standard non-commercialise clause [8, 26]. A potential product would capture models trained with this data but would not be publicised in the product's advertising or in its use by the customer. The infringement is therefore difficult if not impossible to detect. For the sake of this thought experiment, let us assume that this violation was discovered by chance. A university is exclusively regulated within a national legal sphere and, apart from international research projects, operates within that jurisdiction. It will have little legal means at its disposal to enforce rights across national borders. Besides capacity and know-how, another issue is that there are few regulative frameworks [327] that make such a contract legally binding in every country of the world. Furthermore, the EULA might cover elements such as, e. g. , copyright, which are not covered, not indictable (e. g. , fair use), or differently interpreted by the foreign legal system. At least in the case of academic EULA holders, there is the option of contacting other officials at the university where the signee is employed to request support in this matter. Another option would be to make such violations public so that other institutions no longer share data with the contract infringer. However, the effectiveness of this enforcement is questionable and the concept as such offers no protection against such a violation occurring in the first place. This illustrates that data sharing must always be seen as challenging to control and a high risk for ethical violations.

Zenodo [338][1] is a data-sharing platform mandated by the European Commission to make it easier for all researchers to share, curate, and publish data and software. The research project is at the forefront of the Open Access and Open Data movement in Europe and provides a secure platform run by an established academic institution, CERN, on EU servers. Even though the problems mentioned largely remain, this technical support increases control over data accessibility. To access MUSE-CAR, a participant first needs to sign the EULA and submit it digitally. The form checks that the email address is a valid academic one. A human controller also checks whether the person holds a permanent academic position. For the organised challenges and other research, more than 100 teams were given access in this way for a fixed period of three months. Further, statistics are provided (e.g., number of downloads), which helps monitor access on an account level. To date, the repositories featuring different data selection of MuSe-CaR have recorded more than 900 absolute (650 user unique) downloads.

For data of higher sensitivity, e. g. , health care [8, 339], this standard might still not be sufficient and new ways are needed. The latest ideas, such as OpenMinded[2] seek to tighten this concept further to resolve the issues around EULAs. Owners retain complete control over their data while enabling scientists to train models on (private) datasets without ever accessing them directly. A future challenge in the EU is posed by introduction of the General Data Protection Regulations (GDPR). To be highlighted are the principles around the right to

[1] www.zenodo.org accessed July 15, 2021.

[2] www.openmined.org accessed August 8, 2021.

withdraw consent and the right to be forgotten will present researchers with further challenges in data sharing, both in terms of collection and technical feasibility [340].

6.2.2 Application

The focus of this work on publicly available in-the-wild data illustrates how far the groundwork has advanced and that a measure of sentiment on specific topics in large-scale datasets is on the verge of widespread application.

Reliability: Even though systems are constantly improving, perfect reliability in recognition has not yet been achieved and probably never will be [341]. There are several reasons for this. As is the case today, future algorithms will also have the challenge that emotions can by definition never be completely objective, or thus generalisable [326], nor do they need a perfect model of the world including all person-related information to be understood. This is especially the case for very complex emotions [7]. This work has looked at the fusion of multiple subjective emotion annotations and seen room for improvement in these approaches. Sources of error, such as human error or distractions during annotation, are not unlikely. There is also the haptic challenge of transferring a rapidly changing emotion to a recording device (joystick) [278] and the challenge of explaining a complex emotion definition across many annotators, social groups, and cultures. In the latter case, it is evident from Section 2.2 how difficult it is to uniquely define a complex emotion such as trustworthiness.

The first prerequisite for emotional reasoning or even the synthetisation of emotions in applications is their recognition. Faulty recognition, therefore, permeates an emotionalised system [325]. With this awareness in mind, one also has to imagine an average user. Users are used to computers making objective, clearly comprehensible decisions. Regardless of its accuracy limitations, the output of a computer will be perceived as factual [7, 325]. With emotions, this is not the case. In addition, it is fundamentally questionable whether a simulated understanding of emotion can lead to more objective decision-making by a machine.

The error in recognisability and uncertainty has minor effects in use cases studied here. Usually, two applications are seen as closely connected to the analysis of online reviews [10]. On the one hand is an automated understanding of trends through aggregation of different topics, their aspects, and the associated domains of videos uploaded. Therefore, it serves exclusively for an aggregated understanding of a situation or change and thus has an indirect influence on decisions. On the other hand is the improvement of search and recommendation algorithms through a better object-emotion understanding of video content. Here, the focus is on optimisation, so humans do not expect a perfect result, and suggestions instead serve as a recommendation for a decision made by the user.

Nevertheless, the dataset and the findings of these experiments can generally also be leveraged for other applications. In particular, applications in the medical field are increasingly using real customer in-the-wild data. This may include medical decisions that can save or endanger human lives [339]. The necessity and proportionality of such an invasive application must be continuously examined for each use case, as this is where a person's most private data is collected and processed. Faulty profiling or inferences based solely on the association with a certain group of people showing the same emotions cannot be ruled out either.

Collection bias: Another frequently addressed problem is the different treatment of people based on external characteristics such as ethnicity [342]. Although this creates critical problems, human discrimination due to a collection bias [341] towards skin colour may be a more temporary problem. Providers of such applications will naturally try to achieve improved recognition rates [343]. Nevertheless, this problem should continue to be closely monitored by the research community, as other biases exist or will arise due to social inequalities inherent in human nature.

Limiting freedom of speech: Through automated analysis of online data, multimodal sentiment analysis can detect problematic content such as hate speech or fake news more accurately and quickly [10]. However, the automatic detection of these fuzzily defined concepts is difficult, and even simple forms of hate speech detection present social media platforms and algorithms with to-date insurmountable obstacles [344]. Furthermore, automatic blocking can also capture ordinary content, which might become blocked as collateral damage. These unsystematic errors can also have an impact on important opinion-forming processes. Another risk is oppressive states that engage in large-scale censorship [345].

Although this topic is receiving increased attention from the public, research community, and press, clear conclusive regulation is still in its infancy. The EU has taken the first steps by the regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).[3] Here, emotion recognition applications (outside of research) are classified as biometric technology. The act has made clear that even with the primary goal not being the identification of individual people, risks in terms of ethics and data protection need to be examined more closely. Other geographical regions such as the United States are adopting similar plans but, in some cases, have a fundamentally different cultural understanding of data privacy. This will continue to be a significant challenge in implementing global frameworks in the future. Involving the

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=DE/ accessed August 4, 2021.

research-intensive Asian and North American regions will be vital to ensure worldwide uniform standards for ethical and sensible applications.

6.3 Limitations

Despite the successful creation of the MuSe-CaR dataset and several state-of-the-art methods, three limitations of the methodology are highlighted in the following.

Data generalisation, specialisation, and trade-offs: The breakthrough of MSA has emerged through the exponential growth of data and novel DL methods. Unlike classical ML methods, saturation effects do not occur. Instead, more complex patterns can be learned from the available data variance [193]. Limiting factors, however, are the available data sources and the need for human-generated annotations [283]. User data from public data sources harbour new dependencies that did not exist with lab-generated data. For example, as in the case of MuSe-CaR, demographic distributions in training data are only estimable (see Section 4.1). Furthermore, some groups of people do not use the internet at all or only to a limited extent. This can lead to an unintended selection bias (see Section 6.2). Using a sample for training that does not reflect the population leads to the risk of erroneous predictions for groups of gender, age [346], and ethnicity. A partial solution to create a more accurate understanding of the collected online data and possible implications is to estimate demographics at a large scale using biometric facial recognition techniques [347].

However, this does not solve the contextual challenge, which also forces a trade-off between deep and broad data due to financially expensive human annotations. Wide data enables generalisation, which means that, for example, emotion recognition can be applied regardless of the domain. Complex emotions like sarcasm are ambiguous and situation-dependent [348]. However, recognising them often requires deep domain understanding, and thus deep data. Modelling a complex thematic understanding only allows the humorous twist to become recognisable [57]. An understanding of domain-specific entities and other contextual objects of interest (in the dataset presented here for car parts and topics in the automotive environment) as well as their relationship to each other and to humans (e. g. , human-object interaction [349]) is only possible through fine-grained modelling. This applies as well as to the rapid development of generic understanding of text modality were the arguments to go deep instead of broad. However, this only leads to an investigation of mechanisms in a narrow domain and cannot necessarily be generalised to complexities in other domains. One potential future solution could be artificial general language and multimodal intelligence models with zero-shot capabilities [350, 351]; however, those are not in sight yet.

Model interpretability: Even though DL systems showed tremendous success in this and many other works, they present the developer with a lack of interpretability being a broadly noted structural challenge [352]. Modelling a complex world leads to incomplete information, which in turn will always lead to erroneous predictions. Without a qualitative understanding of what is learned and how it is learned, it is difficult to understand errors, debug them, and develop custom solutions. Typically, this leads to technically generic rather than domain hypothesis-driven approaches and objectives, wherein fundamentally more robust learning mechanisms are developed and proven by higher scores on benchmark datasets [353], such as the context, local, or MHAL attention mechanisms used in this work, or by indirectly exerting control over the training data to be used. In the context of this work, the issue is specially relevant because multimodal fusion has co-dependencies on all upstream recognition systems [10]. In specific application fields of MSA for sentiment detection or recommendation engines, this can lead to a poor user experience and, in the adjacent AC or precision medicine applications, even to human-damaging decisions. Explainable AI research approaches attempt to offer a solution by combining the proposed methods with attention decoding or gradient mapping to achieve better transparency while maintaining the prediction level of deep learning methods [354].

Multimodal fusion: As described earlier, this work mainly used early and late fusion. Early fusion of representations led to overfitting, while in later fusion, training improved the representation only in terms of intermodal dynamics. Until the final fusion step, much information that might be precious in the intermodal context is already lost. Hybrid fusion has emerged to fuse modalities after initial encoding and solve these problems [42, 13]. However, efficient mechanisms that are able to cope with large ANN extracted representations have not been found yet. For example, fusion by multihead attention, as experimented with in this work, requires equal dimensions from all modalities due to mathematical constraints [91]. Reducing all representations to the dimensions of the smallest leads to heavy information loss. Increasing them to the largest is technically impractical given current GPU memory limitations.

6.4 Outlook

Establishing MuSe-CaR into the research community through two challenges and with more than 120 academic groups successfully having requested access, the publications, and accompanying source codes, this thesis has laid a multitude of foundations for future research. In the following, the limitations and other ideas are taken up and an outlook on future work is given.

Semantic vision: MSA aims to explore the context alongside emotions in videos. Currently, semantic contexts are still mostly derived from the spoken word, such as speaker topics. However, the opinion holder uses the full spectrum of communication in videos and incorporates other modalities [355], such as pointing and gestures. Predictive vision representations were used in this work, but only as additional input to speaker topic recognition. Interaction with the environment can be modelled directly for MSA and extracted in the form of object-human structures [349]. The object and hand do not necessarily overlap or may not in line with a fixed camera in the wild, so that, for example, for finger-pointing, both the finger and the object in a room have to be calibrated spatially by anchor points. Stappen et al. has shown similar first approaches for gaze recognition [253]. The explicit derivation of human-object relationships represents another exciting extension of MSA to open up a new semantic level.

Dynamic granularities: Sentiment analysis is carried out at different levels of granularity, for example, on aggregated topics [356] and subdivided aspects [304]. There are also methods of thematic breakdown into hierarchical levels and concepts [38, 357]. Such hierarchies are hardly established in MSA so far, especially with regard to the emotional component. Affective Computing is based on fine-grained, continuous affect annotations, while language-influence MSA relies on sentiment and emotion classes for individual topics, aspects, or subaspects. However, an emotion does not necessarily have to arise in a direct spatial context ("food was excellent"), but can be a logical conclusion from complex, long utterances that requires a larger emotion context. A hierarchical context and different granularity would allow a seamless zoom in and out on the level of targets.

Pure-attention networks: This work focuses mostly on method development within conventional ANN architectures (e. g. , LSTM-RNN) that embed attention mechanisms. While pure-attention models, such as Transformers, are the new state-of-the-art in the textual domain [81, 212] and have also achieved very strong results in this work (see Paragraph 5.1.2.1.2), pure-attention-based architectural variants are only slowly gaining acceptance in the video [358] and audio domain. In the MuSe 2021 challenge, which had just ended at the time this work was finalised, models using audio transformer representations such as wav2vec [359] yielded the strongest performance gains. Further modality-specific adaptations are necessary for a fully integrated approach [350, 351]. In particular, multimodal Transformers, hybrid fusion networks or spiking neural networks could lead to a much deeper dovetailing of the different modalities than early or late fusion [253]. However, initial approaches still lack targeted mechanisms, especially for modelling continuous emotions and multimodal target extraction, but present an exciting new research direction.

Given the technological leaps in recent years, the research field of MSA can look forward to a prosperous future. We can soon expect that machines will help us to open up the vast amounts of multimodal, unstructured knowledge and take us to a new level of the information age.

Acronyms

Symbols | A | B | C | D | E | F | G | H | K | L | M | N | O | P | R | S | T | U | V | W | X

Symbols

at number of attention heads.

bs batch size.

fs filter size.

h hidden state dimensionality.

hs hop size.

ks kernel size.

lr learning rate.

n number of layers.

ps pool size.

s strides.

t_c topic coverage.

t_o topic overlap.

$tanh$ Hyperbolic Tangent Function.

ws window size.

A

absE absolute Energy.

AC Affective Computing.

ACC Accuracy.

AI Artificial Intelligence.

ALBERT A Lite BERT for Self-Supervised Learning of Language Representations.

ANN Artificial Neural Network.

ASOC Relative Sum Of Changes.

B

BERT Bidirectional Encoder Representations from Transformers.

C

CA Circumplex Model of Affect.

CBMe Count Below Mean.

CBOW Continuous Bag of Words Model.

CCC Concordance Correlation Coefficient.

CE Cross-entropy.

Citysearch Citysearch New York corpus.

CNN Convolutional Neural Network.

ComParE LLDs ComParE Low-Level Descriptors.

CrM number of Crossings of a point.

CSG Continuous Skip-gram Word Model.

D

DARMA Dual Axis Rating and Media Annotation Software.

Deep Spectrum Spectrograms Feature Extraction from Audio Data with Pre-trained Convolutional Neural Networks.

DeepTrust Deep Trust Multihead Attention Network.

DL Deep Learning.

DNN Deep Neural Network.

DTW Dynamic Time Warping.

E

eGeMAPS extended Geneva Minimalistic Acoustic Parameter Set.

ELAN Eudico Language Annotation Tool.

EM Estimation-Maximisation Algorithm.

End2You End-to-End Learning.

EULA End User Licence Agreement.

EWE Evaluator Weighted Estimator.

F

F1 F1-score.

FACS Facial Action Coding System.

FastText Fast Text Classifier.

FAU Facial Action Unit.

FFL Fully connected Feed-Forward Layer.

FNN Feed-Forward Neural Network.

G

GCTW Generic-Canonical Time Warping.

GDPR General Data Protection Regulations.

GeMAPS Geneva Minimalistic Acoustic Parameter Set.

GEMEP *General Multimodal Emotion Representations*.

GMM Gaussian Mixture Model.

GoCaRD Generic, Optical Car Part Recogniser and Detector.

GPU Graphics Grocessing Unit.

GraphTMT Graph-based Topic Modelling approach for Transcripts.

H

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise.

HMM Hidden Markov Model.

K

kurt Kurtosis.

L

LDA Latent Dirichlet Allocation.

LLD Low-level Descriptor.

LSAMe Last Strike Above the Mean.

LSBMe Last Strike Below the Mean.

LSTM-RNN Long Short-Term Memory Recurrent Neural Network.

LSTM-SA Long Short-Term Memory Recurrent Neural Network with Self-Attention.

M

MACh Mean relative Absolute Change.

MAE Mean Absolute Error.

mAP Mean Average Precision.

MCh Mean Change.

MHA-LSTM Multihead Attention Long Short-Term Memory Recurrent Neural Network.

MHAL Multihead Attention Layer.

ML Machine Learning.

MMT Multimodal Transformer.

MSA Multimodal Sentiment Analysis.

MSDC Mean value of a central approximation of the Second Derivatives.

MSE Mean Square Error.

MTCNN Multi-task Cascaded Convolutional Network Framework.

MuSe-CaR Multimodal Sentiment Analysis in Car Reviews.

MuSe-CaR-Part Multimodal Sentiment Analysis in Car Part Frames.

MuSe-Sent Multimodal Sentiment Sub-challenge.

MuSe-Toolbox Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox.

MuSe-Topic Multimodal Emotion-Target Sub-challenge.

MuSe-Trust Multimodal Trustworthiness Sub-challenge.

MuSe-Wild Multimodal V-A Sentiments in-the-Wild Sub-challenge.

MuSe-Wilder Multimodal Continuous Emotions in-the-Wild Sub-challenge.

N

NDC Node Degree Connectivity.

NLP Natural Language Processing.

NSP Next-Sentence Prediction.

O

OpenPose Open Multi-person System to Jointly Detect Human Body, Hand, Facial, and Foot keypoints.

openSMILE Open-source Speech and Music Interpretation by Large-space Extraction Toolkit.

P

PCA Principal Component Analysis.

PCC Pearson Correlation Coefficient.

PreDa Percentage of Reoccurring Data points of non-unique single points.

PTH Percentile Similarity Threshold.

R

RAAW Rater Aligned Annotation Weighting.

RCNN Recurrent Convolution Neural Network.

ReLU Rectified Linear Unit.

RMSE Root Mean Square Error.

RNN Recurrent Neural Network.

S

S2S Sequence to Sequence.

SaEn Sample Entropy.

SenSA SENtic Sentiment Analysis Learner.

skew dynamic sample skewness.

SNL SenticNet-based Learning.

SP Signal Processing.

SVM Support Vector Machine.

SVR Support Vector Regressor.

T

TF Term Frequency.

TF-IDF TF-Inverse Document Frequency.

TVS Topic Vector Similarity.

U

UAR Unweighted Average Recall.

V

VGGFace Very Deep Convolutional Networks for Large-Scale Face Recognition Descriptor.

VGGish CNN Architectures for Large-Scale Audio Classification.

W

WER Word Error Rate.

Word2Vec Word to Vector.

X

Xception Depthwise Separable Convolutions Network.

Bibliography

[1] B. Bhardwaj, “Text mining, its utilities, challenges and clustering techniques,” *International Journal of Computer Applications*, vol. 135, no. 7, pp. 22–25, 2016.

[2] L. Stappen, X. Du, V. Karas, S. Müller, and B. W. Schuller, “Go-card–generic, optical car part recognition and detection: Collection, insights, and applications,” *arXiv preprint arXiv:2006.08521*, 2020.

[3] B. W. Schuller, *Intelligent audio analysis*. Springer, 2013.

[4] V. Karas and B. W. Schuller, “Deep learning for sentiment analysis: an overview and perspectives,” in *Natural Language Processing for Global and Local Business*, F. Pinarbasi and M. N. Taskiran, Eds. IGI Global, 2020, pp. 97–132.

[5] X. Chen, Y. Wang, and Q. Liu, “Visual and textual sentiment analysis using deep fusion convolutional neural networks,” in *2017 IEEE International Conference on Image Processing (ICIP)*. Beijing, China: IEEE, 2017, pp. 1557–1561.

[6] M. P. Fortin and B. Chaib-draa, “Multimodal sentiment analysis: A multitask learning approach,” in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Prague, Czech Republic, 2019, p. 368–376.

[7] B. Schuller, J.-G. Ganascia, and L. Devillers, “Multimodal sentiment analysis in the wild: Ethical considerations on data collection, annotation, and exploitation,” in *Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHI-CA2)*, Portoroz, Slovénie, 2016, pp. 29–34.

[8] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, “The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks,” in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[10] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[11] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *arXiv preprint arXiv:1606.06259*, 2016.

[12] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, “Sewa db: A rich database for audio-visual emotion and sentiment research in the wild,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2021.

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark: ACL, 2017, pp. 1103–1114.

[14] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, “Multi-attention recurrent network for human communication comprehension,” in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*. New Orleans, USA: AAAI, 2018.

[15] A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria, and S. Scherer, “Proceedings of grand challenge and workshop on human multimodal language (challenge-hml),” in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Seattle, USA: ACL, 2018, pp. 1–83.

[16] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[17] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*. Barcelona, Spain: ACM, 2013, pp. 3–10.

[18] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proceedings of the 7th Annual Workshop on*

Audio/Visual Emotion Challenge (AVEC). Mountain View, USA: ACM, 2017, pp. 3–9.

[19] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, “Analysing affective behavior in the first abaw 2020 competition,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020).* Buenos Aires, Argentina: IEEE, 2020, pp. 637–643.

[20] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, “The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats.” in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH).* Hyderabad, India: ISCA, 2018, pp. 122–126.

[21] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, “Gated multimodal networks,” *Neural Computing and Applications*, vol. 32, no. 14, pp. 10 209–10 228, Jan. 2020.

[22] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).* Snowmass Village, USA: IEEE, 2020, pp. 1470–1478.

[23] X. Qiu, Z. Feng, X. Yang, and J. Tian, “Multimodal fusion of speech and gesture recognition based on deep learning,” in *Journal of Physics: Conference Series*, vol. 1453, 2020, pp. 2092–2098.

[24] L. Stappen, A. Baird, L. Schumann, and B. Schuller, “The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements,” *IEEE Transactions on Affective Computing (Early Access)*, June 2021.

[25] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigel, E. Cambria, and B. W. Schuller, “Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox,” in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM).* Changu, China: ACM, 2021.

[26] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, “Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and

trustworthiness detection in real-life media," in *Proceedings of the 1st International Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop (MuSe), co-located with the 28th ACM International Conference on Multimedia (ACMMM).* Seattle, USA: ACM, 2020, p. 35–44.

[27] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM).* Changu, China: ACM, 2021.

[28] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th International Conference on Multimodal Interfaces (ICIMI).* Alicante, Spain: ACM, 2011, pp. 169–176.

[29] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

[30] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Lisbon, Portugal: ACL, 2015, pp. 2539–2544.

[31] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, A. Charu and X. Z. Cheng, Eds. Springer, 2012, pp. 415–463.

[32] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[33] C. Clavel and Z. Callejas, "Sentiment analysis: from opinion mining to human-agent interaction," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 74–93, 2015.

[34] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "Moseas: A multimodal language dataset for spanish, portuguese, german and french," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Virtual: ACL, 2020, pp. 1801–1812.

[35] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, "Multimodal emoji prediction," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, New Orleans, USA, 2018, pp. 679–686.

[36] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014.

[37] P. Ekman and W. V. Friesen, *Manual for the facial action coding system.* Consulting Psychologists Press, 1978.

[38] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioural Systems,* , A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds., 2012, pp. 144–157.

[39] R. Plutchik and H. Kellerman, *Theories of emotion.* Academic Press, 2013, vol. 1.

[40] Y. Susanto, A. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[41] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *International Conference on Affective Computing and Intelligent Interaction (ACII).* Lisbon, Portugal: AAAC, 2007, pp. 488–500.

[42] S. Poria, A. Hussain, and E. Cambria, *Multimodal sentiment analysis*, ser. Socio-Affective Computing. Springer, 2018.

[43] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2011.

[45] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proceedings of*

the Tenth International Conference on Language Resources and Evaluation (LREC). Portoroz, Slovénie: ELRA, 2016, pp. 2156–2161.

[46] J. Tang and H. Liu, "Trust in social media," *Synthesis Lectures on Information Security, Privacy, & Trust*, vol. 10, no. 1, pp. 1–129, 2015.

[47] W.-Y. Lin, X. Zhang, H. Song, and K. Omori, "Health information seeking in the web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure," *Computers in Human Behavior*, vol. 56, pp. 289–294, 2016.

[48] M. Irshad, M. S. Ahmad, and O. F. Malik, "Understanding consumers' trust in social media marketing environment," *International Journal of Retail & Distribution Management*, 2020.

[49] C. Schwemmer and S. Ziewiecki, "Social media sellout: The increasing role of product promotion on youtube," *Social Media+ Society*, vol. 4, no. 3, pp. 1–20, 2018.

[50] A. Nikolinakou and K. W. King, "Viral video ads: Emotional triggers and social media virality," *Psychology & Marketing*, vol. 35, no. 10, pp. 715–726, 2018.

[51] J. Tang and H. Liu, "Trust in social computing," in *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, Seoul, Korea, 2014, pp. 207–208.

[52] H. Horsburgh, "Trust and social objectives," *Ethics*, vol. 72, no. 1, pp. 28–40, 1961.

[53] S. T. Moturu and H. Liu, "Quantifying the trustworthiness of social media content," *Distributed and Parallel Databases*, vol. 29, no. 3, pp. 239–260, 2011.

[54] J. C. Cox, R. Kerschbamer, and D. Neururer, "What is trustworthiness and what drives it?" *Games and Economic Behavior*, vol. 98, pp. 197–218, 2016.

[55] J. A. Colquitt, B. A. Scott, and J. A. LePine, "Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance." *Journal of Applied Psychology*, vol. 92, no. 4, pp. 909–927, 2007.

[56] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 88–95, March 2021.

[57] M. K. Hasan, W. Rahman, A. B. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, and M. E. Hoque, "Ur-funny: A multimodal language dataset for understanding humor," in *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019, pp. 2046–2056.

[58] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, 2018, pp. 2236–2246.

[59] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[60] E. Marrese-Taylor, C. Rodriguez, J. Balazs, S. Gould, and Y. Matsuo, "A multimodal approach to fine-grained opinion mining on video reviews," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Seattle, USA: ACL, 2020, pp. 8–18.

[61] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Virtual: ACL, 2020, pp. 3718–3727.

[62] D. Cevher, S. Zepf, and R. Klinger, "Towards multimodal emotion recognition in german speech events in cars using transfer learning," in *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*. Erlangen, Germany: GSCL, 2019, pp. 79–90.

[63] E. Marrese-Taylor, J. Balazs, and Y. Matsuo, "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark: ACL, 2017, pp. 102–111.

[64] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sofia, Bulgaria: ACL, 2013, pp. 973–982.

[65] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Shanghai, China: IEEE, 2013, pp. 1–8.

[66] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, “Afew-va database for valence and arousal estimation in-the-wild,” *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[67] M. Grimm, K. Kroschel, and S. Narayanan, “The vera am mittag german audio-visual emotional speech database,” in *2008 IEEE International Conference on Multimedia & Expo (ICME).* Hanover, Germany: IEEE, 2008, pp. 865–868.

[68] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, “The belfast induced natural emotion database,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2011.

[69] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proceedings of the 21st ACM International Conference on Multimedia (ACMMM).* Barcelona, Spain: ACM, 2013, pp. 223–232.

[70] V. P. Rosas, R. Mihalcea, and L.-P. Morency, “Multimodal sentiment analysis of spanish online videos,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.

[71] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, “Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach,” in *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI).* Istanbul, Turkey: ACM, 2014, pp. 50–57.

[72] A. Garcia, S. Essid, F. d’Alché Buc, and C. Clavel, “A multimodal movie review corpus for fine-grained opinion mining,” *arXiv preprint arXiv:1902.10102*, 2019.

[73] A. Luneski, E. Konstantinidis, and P. Bamidis, “Affective medicine: a review of affective computing efforts in medical informatics,” *Methods of Information in Medicine*, vol. 49, no. 3, pp. 207–218, 2010.

[74] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[75] S. Li and W. Deng, “Deep facial expression recognition: A survey,” *IEEE Transactions on Affective Computing (Early Access)*, 2020.

[76] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, 2017.

[77] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[78] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[79] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, “An image-based deep spectrum feature representation for the recognition of emotional speech,” in *Proceedings of the 25th ACM International Conference on Multimedia*. Mountain View, USA: ACM, 2017, pp. 478–484.

[80] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv: 1301.3781*, 2013.

[81] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.

[82] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Lujiazui, China: IEEE, 2016, pp. 4960–4964.

[83] B. Schuller, G. Rigoll, and M. Lang, “Hidden markov model-based speech emotion recognition,” in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Hong Kong, China: IEEE, 2003, pp. 401–404.

[84] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, “The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language,” in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH).* San Francisco, USA.: ISCA, 2016, pp. 2001–2005.

[85] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[86] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, “Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH).* Makuhari, Japan: ISCA, 2010, pp. 2362–2365.

[87] S. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, “Recurrent neural networks for emotion recognition in video,” in *Proceedings of the 2015 ACM International Conference on Multimodal Interaction (ICMI).* Seattle, USA: ACM, 2015, pp. 467–474.

[88] P. Tzirakis, J. Zhang, and B. W. Schuller, “End-to-end speech emotion recognition using deep neural networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Alberta, Cananda: IEEE, 2018, pp. 5089–5093.

[89] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, “End2you–the imperial toolkit for multimodal profiling by end-to-end learning,” *arXiv preprint arXiv:1802.01115*, 2018.

[90] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, “End-to-end multimodal affect recognition in real-world environments,” *Information Fusion*, vol. 68, pp. 46–53, 2021.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS).* Long Beach, California: NeurIPS, 2017, pp. 5998–6008.

[92] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, “Multilingual sentiment analysis: from formal to informal and scarce resource languages,” *Artificial Intelligence Review*, vol. 48, no. 4, pp. 499–527, 2017.

[93] M. Morchid, G. Linares, M. El-Beze, and R. De Mori, “Theme identification in telephone service conversations using quaternions of speech features.” in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH).* Lyon, France: ISCA, 2013, pp. 1394–1398.

[94] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH).* San Francisco, USA: ISCA, 2016, pp. 3047–3051.

[95] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI).* San Francisco, USA: AAAI, 2017.

[96] M. Yang, W. Tu, J. Wang, F. Xu, and X. Chen, "Attention based lstm for target dependent sentiment classification," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI).* San Francisco, USA: AAAI, 2017, pp. 5013–5014.

[97] C. Zhou, J. Bai, J. Song, X. Liu, Z. Zhao, X. Chen, and J. Gao, "Atrank: An attention-based user behavior modeling framework for recommendation," in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI).* New Orleans, USA: AAAI, 2018.

[98] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR).* Miami, USA: IEEE, April 2018, pp. 196–201.

[99] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, Ł. Kaiser, N. Kalchbrenner, N. Parmar *et al.*, "Tensor2tensor for neural machine translation," *arXiv preprint arXiv:1803.07416*, 2018.

[100] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[101] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Barcelona, Spain: IEEE, 2020, pp. 3507–3511.

[102] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. Traue, and H. Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC).* Seoul, South Korea: IEEE, 2012, pp. 3316–3320.

[103] C. Laurier, M. Sordo, J. Serra, and P. Herrera, “Music mood representations from social tags,” in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009, pp. 381–386.

[104] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, “Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies,” in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH) and the 12th Australasian International Conference on Speech Science and Technology (SST)*. Brisbane, Australia: ISCA, 2008, pp. 597–600.

[105] R. W. Picard, “Affective computing for hci.” in *Human Computer Interaction*. Citeseer, 1999, pp. 829–833.

[106] P. Lebreton and K. Yamagishi, “Predicting user quitting ratio in adaptive bitrate video streaming,” *IEEE Transactions on Multimedia (Early Access)*, 2020.

[107] P. Zhou, Y. Zhou, D. Wu, and H. Jin, “Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks,” *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1217–1229, 2016.

[108] Y. Liu, X. Shi, L. Pierce, and X. Ren, “Characterizing and forecasting user engagement with in-app action graph: A case study of snapchat,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, Anchorage, USA, 2019, pp. 2023–2031.

[109] C. Yang, X. Shi, L. Jie, and J. Han, “I know you’ll be back: Interpretable new user clustering and churn prediction on a mobile social application,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, London, United Kingdom, 2018, pp. 914–922.

[110] Z. Lin, T. Althoff, and J. Leskovec, “I’ll be back: On the multiple lives of users of a mobile activity tracking application,” in *Proceedings of the 2018 World Wide Web Conference (WWW)*, Lyon, France, 2018, pp. 1501–1511.

[111] K. English, K. D. Sweetser, and M. Ancu, “Youtube-ification of political talk: An examination of persuasion appeals in viral video,” *American Behavioral Scientist*, vol. 55, no. 6, pp. 733–748, 2011.

[112] J. Berger and K. Milkman, “What makes online content viral?” *Journal of Marketing Research*, vol. 49, no. 2, pp. 192–205, 2012.

[113] E. Shehu, T. H. Bijmolt, and M. Clement, “Effects of likeability dynamics on consumers’ intention to share online video advertisements,” *Journal of Interactive Marketing*, vol. 35, pp. 27–43, 2016.

[114] F. Kujur and S. Singh, “Emotions as predictor for consumer engagement in youtube advertisement,” *Journal of Advances in Management Research*, vol. 15, no. 2, pp. 184–197, May 2018.

[115] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher, “Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes,” *Journal of Vision*, vol. 14, no. 3, pp. 31–49, 2014.

[116] H. Sagha, M. Schmitt, F. Povolny, A. Giefer, and B. Schuller, “Predicting the popularity of a talk-show based on its emotional speech content before publication,” in *Proceedings 3rd International Workshop on Affective Social Multimedia Computing, 18th Annual Conference of the International Speech Communication Association (INTERSPEECH) Satellite Workshop*. Stockholm, Sweden: ISCA, 2017.

[117] C. Chapple and F. Cownie, “An investigation into viewers’ trust in and response towards disclosed paid-for-endorsements by youtube lifestyle vloggers,” *Journal of Promotional Communications*, vol. 5, no. 2, pp. 19–28, 2017.

[118] M. Yan, J. Sang, C. Xu, and M. S. Hossain, “Youtube video promotion by cross-network association:@ britney to advertise gangnam style,” *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1248–1261, june 2015.

[119] S. D. Roy, T. Mei, W. Zeng, and S. Li, “Towards cross-domain learning for social video popularity prediction,” *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1255–1267, October 2013.

[120] Z. Tan and Y. Zhang, “Predicting the top-n popular videos via a cross-domain hybrid model,” *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 147–156, June 2019.

[121] J. Biel and D. Gatica-Perez, “The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs,” *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, January 2013.

[122] X. Cheng, J. Liu, and C. Dale, “Understanding the characteristics of internet short video sharing: A youtube-based measurement study,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184–1194, August 2013.

[123] R. G. Garroppo, M. Ahmed, S. Niccolini, and M. Dusi, “A vocabulary for growth: Topic modeling of content popularity evolution,” *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2683–2692, October 2018.

[124] Z. Wu and E. Ito, “Correlation analysis between user’s emotional comments and popularity measures,” in *Proceedings of the 3rd International Conference on Advanced Applied Informatics (IIAIAAI)*. Kokura Kita-ku, Japan: IEEE, 2014, pp. 280–283.

[125] R. Yang, S. Singh, P. Cao, E. Chi, and B. Fu, “Video watch time and comment sentiment: Experiences from youtube,” in *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*. Washington, USA: IEEE, 2016, pp. 26–28.

[126] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam, “Retrieving youtube video by sentiment analysis on user comment,” in *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. Kuching, Malaysia: IEEE, 2017, pp. 474–478.

[127] A. Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, “Multi-lingual opinion mining on youtube,” *Information Processing & Management*, vol. 52, no. 1, pp. 46–60, January 2016.

[128] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, “How useful are your comments? analyzing and predicting youtube comments and comment ratings,” in *Proceedings of the 19th International Conference on World Wide Web (WWW)*, Raleigh, USA, 2010, pp. 891–900.

[129] C. Gilbert and E. Hutto, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International Conference on Weblogs and Social Media (ICWSM)*, Ann Arbor, USA, 2014, pp. 216–225.

[130] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to italy and france,” *New Media & Society*, vol. 16, no. 2, pp. 340–358, April 2014.

[131] D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman, “Modelling valence and arousal in facebook posts,” in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. San Diego, USA: ACL, 2016, pp. 9–15.

[132] A. Nazir, Y. Rao, L. Wu, and L. Sun, “Issues and challenges of aspect-based sentiment analysis: A comprehensive survey,” *IEEE Transactions on Affective Computing (Early Access)*, 2020.

[133] T. A. Rana and Y.-N. Cheah, “Aspect extraction in sentiment analysis: comparative analysis and survey,” *Artificial Intelligence Review*, vol. 46, no. 4, pp. 459–483, Feburary 2016.

[134] S. Hou, L. Chen, D. Tao, S. Zhou, W. Liu, and Y. Zheng, “Multi-layer multi-view topic model for classifying advertising video,” *Pattern Recognition*, vol. 68, pp. 66–81, Aug. 2017.

[135] S. Arora, A. May, J. Zhang, and C. Ré, “Contextual embeddings: When are they worth it?” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Virtual: ACL, Jul. 2020, pp. 2650–2663.

[136] K. Park, S. Lee, and Y. Tan, “What makes online review videos helpful? evidence from product review videos on youtube,” *SSRN Electronic Journal*, pp. 1–38, September 2020.

[137] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, “Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 5, pp. 996–1009, May 2019.

[138] M. Husain and S. M. Meena, “Multimodal fusion of speech and text using semi-supervised lda for indexing lecture videos,” in *National Conference on Communications (NCC)*. Bangalore, India: IEEE, 2019, pp. 1–6.

[139] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey,” *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, November 2019.

[140] E. Cambria and A. Hussain, *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*, E. Cambria and A. Hussain, Eds. Springer, 2015.

[141] S. Basu, Y. Yu, and R. Zimmermann, “Fuzzy clustering of lecture videos based on topic modeling,” in *14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. Bucharest, Romania: IEEE, 2016, pp. 1–6.

[142] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in *Proceedings of the Australasian Language Technology Association Workshop (ALTA).* Sydney, Australia: ATLA, 2015, pp. 111–115.

[143] P. Das, A. K. Das, J. Nayak, D. Pelusi, and W. Ding, "A graph based clustering approach for relation extraction from crime data," *IEEE Access*, vol. 7, pp. 101 269–101 282, July 2019.

[144] S. Curiskis, B. Drake, T. Osborn, and P. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Information Processing & Management*, vol. 57, no. 2, p. 102034, March 2019.

[145] M. Sahlgren, "Rethinking topic modelling: From document-space to term-space," in *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP).* Virtual: ACL, 2020, pp. 2250–2259.

[146] L. Wang, C. Gao, J. Wei, W. Ma, R. Liu, and S. Vosoughi, "An empirical survey of unsupervised text representation methods on twitter data," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT).* Virtual: ACL, 2020, pp. 209–214.

[147] V. K. R. Sridhar, "Unsupervised topic modeling for short texts using distributed representations of words," in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing (NAACL).* Denver, USA: ACL, 2015, pp. 192–200.

[148] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classification," *Biometrics*, vol. 21, no. 3, pp. 768–780, 1965.

[149] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, March 1982.

[150] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: A simple yet principled alternative algorithm," *PLOS ONE*, vol. 11, no. 9, pp. 1–28, September 2016.

[151] J. Sander, *Density-Based Clustering.* Springer, 2010, pp. 270–273.

[152] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *24th Pacific-Asia Conference Advances in Knowledge Discovery and Data Mining (PAKDD).* Singapore, Singapore: Springer, 2013, pp. 160–172.

[153] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," in *IEEE International Conference on Data Mining Workshops (ICDMW).* New Orleans, USA: IEEE, 2017, pp. 33–42.

[154] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020.

[155] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[156] J. Torrents and F. Ferraro, "Structural cohesion: Visualization and heuristics for fast computation," *Journal of Social Structure (JoSS)*, vol. 15, no. 1, pp. 1–35, December 2015.

[157] J. Moody and D. R. White, "Structural cohesion and embeddedness: A hierarchical concept of social groups," *American Sociological Review (ASR)*, vol. 68, no. 1, pp. 103–127, February 2003.

[158] R.-H. Li, L. Qin, J. X. Yu, and R. Mao, "Influential community search in large networks," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 8, no. 5, pp. 509–520, January 2015.

[159] D. B. West, *Introduction to Graph Theory.* Prentice Hall, 2000.

[160] X. Ni, X. Quan, Z. Lu, L. Wenyin, and B. Hua, "Short text clustering by finding core terms," *Knowledge and Information Systems (KAIS)*, vol. 27, no. 3, pp. 345–365, June 2011.

[161] M. T. Altuncu, S. N. Yaliraki, and M. Barahona, "Graph-based topic extraction from vector embeddings of text documents: Application to a corpus of news articles," in *Complex Networks & Their Applications IX.* Springer, 2021, pp. 154–166.

[162] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Seattle, USA: ACL, 2013, pp. 1631–1642.

[163] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *International conference on machine learning (ICML).* Beijing, China: PMLR, 2014, pp. 703–711.

[164] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, September 2016.

[165] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Seattle, USA: ACM, 2004, pp. 168–177.

[166] Q. Su, K. Xiang, H. Wang, B. Sun, and S. Yu, "Using pointwise mutual information to identify implicit features in customer reviews," in *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, Y. Matsumoto, R. W. Sproat, K.-F. Wong, and M. Zhang, Eds. Springer, 2006, pp. 22–30.

[167] S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," in *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*. Dublin, Ireland: ACL, 2014, pp. 28–37.

[168] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *2011 IEEE 11th International Conference on Data Mining Workshops (ICDM)*. Vancouver, Canada: IEEE, 2011, pp. 81–88.

[169] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware modeling and recognition of activities in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Portland, USA: IEEE, 2013, pp. 2491–2498.

[170] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Marseille, France: Springer, 2008, pp. 817–829.

[171] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE, 2019, pp. 244–253.

[172] S. Gupta and R. J. Mooney, "Using closed captions to train activity recognizers that improve video retrieval," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. Miami, USA: IEEE, 2009, pp. 30–37.

[173] S. Gupta and R. Mooney, "Using closed captions as supervision for video activity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Atlanta, USA: AAAI, 2010.

[174] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2085–2098, December 2014.

[175] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic, "The mahnob mimicry database: A database of naturalistic human interactions," *Pattern recognition letters*, vol. 66, pp. 52–61, Nov. 2015.

[176] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, November 2020.

[177] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, 2019.

[178] Q.-T. Truong and H. W. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, July 2019, vol. 33, no. 01, pp. 305–312.

[179] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual: ACM, 2020, pp. 105–114.

[180] H. Sayyadi and L. Raschid, "A graph analytical approach for topic detection," *ACM Transactions on Internet Technology*, vol. 13, no. 2, pp. 1–23, December 2013.

[181] S. Wang, S. Mazumder, B. Liu, M. Zhou, and Y. Chang, "Target-sensitive memory networks for aspect sentiment classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia: ACL, 2018, pp. 957–967.

[182] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Honolulu, USA: AAAI, 2019, pp. 371–378.

[183] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy: ACL, 2019, pp. 6558–6569.

[184] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1665–1678, August 2016.

[185] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, September 2013.

[186] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan: ACL, 2016, pp. 2666–2677.

[187] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. New Orleans, USA: AAAI, 2018, pp. 1795–1802.

[188] S. Poria, I. Chaturvedi, E. Cambria, and F. Bisio, "Sentic lda: Improving on lda with semantic similarity for aspect-based sentiment analysis," in *2016 International Joint Conference on Neural Networks (IJCNN)*. Vancouver, Canada: IEEE, 2016, pp. 4465–4473.

[189] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain, "Sentic lstm: a hybrid network for targeted aspect-based sentiment analysis," *Cognitive Computation*, vol. 10, no. 4, pp. 639–650, March 2018.

[190] C. Brun, D. N. Popa, and C. Roux, "Xrce: Hybrid classification for aspect-based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*. Dublin, Ireland: ACL, 2014, pp. 838–842.

[191] H. Bai, F. Z. Xing, E. Cambria, and W.-B. Huang, "Business taxonomy construction using concept-level hierarchical clustering," in *Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Macao, China: ACL, 2019, pp. 1–7.

[192] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The general inquirer: A computer approach to content analysis*. University of Chicago, 1966.

[193] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, T. Dietterich, Ed. MIT Press, 2016.

[194] R. R. Atallah, A. Kamsin, M. A. Ismail, S. A. Abdelrahman, and S. Zerdoumi, "Face recognition and age estimation implications of changes in facial features: A critical review study," *IEEE Access*, vol. 6, pp. 28 290–28 304, 2018.

[195] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," *Introduction to Information Retrieval*, vol. 100, pp. 2–4, 2008.

[196] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. The MIT Press, 2020.

[197] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[198] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[199] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features." in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 3512–3516.

[200] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, USA: IEEE, 2017, pp. 131–135.

[201] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 148–152.

[202] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 148–152.

[203] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[204] L. Stappen, V. Karas, N. Cummins, F. Ringeval, K. Scherer, and B. Schuller, "From speech to facial activity: towards cross-modal sequence-to-sequence attention networks," in *Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP).* Kuala Lumpur, Malaysia: IEEE, 2019, pp. 1–6.

[205] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[206] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, USA: IEEE, 2017, pp. 776–780.

[207] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC).* Seoul, Republic of Korea: ACM, 2018, pp. 65–72.

[208] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[209] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* New Orleans, USA: ACL, 2018, pp. 2227–2237.

[210] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP).* Minneapolis, USA: ACL, Jun. 2019, pp. 72–78.

[211] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMLP): System Demonstrations. Virtual: ACL, 2020.

[212] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” in *2020 International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[213] M. Dragoni, S. Poria, and E. Cambria, “Ontosenticnet: A commonsense ontology for sentiment analysis,” *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 77–85, 2018.

[214] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, april 2016.

[215] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” *arXiv preprint arXiv:1511.06523*, 2015.

[216] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, pp. 3730–3738.

[217] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Placid, USA: IEEE, 2016, pp. 1–10.

[218] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[219] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[220] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proceedings of the British Machine Vision Conference (BMVC)*. Swansea, UK: BMVA Press, September 2015, pp. 41.1–41.12.

[221] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: IEEE, 2016, pp. 770–778.

[222] F. K. Pil and M. Holweg, "Linking product variety to order-fulfillment strategies," *Interfaces*, vol. 34, no. 5, pp. 394–403, 2004.

[223] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[224] L. Stappen, A. Baird, M. Lienhart, A. Bätz, and B. Schuller, "An estimation of online video user engagement from features of continuous emotions," *Frontiers in Computer Science*, 2022.

[225] P. Geurts, "Pattern extraction for time series classification," in *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*. Freiburg, Germany: Springer, 2001, pp. 115–127.

[226] B. Schuller, M. Lang, and G. Rigoll, "Automatic emotion recognition by the speech signal," in *Proccedings of SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics (WMSCI)*. Orlando, USA: ISAS, 2002, pp. 381–386.

[227] D. P. Doane and L. E. Seward, "Measuring skewness: a forgotten statistic?" *Journal of Statistics Education*, vol. 19, no. 2, pp. 1–18, July 2011.

[228] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, May 1992.

[229] P. H. Westfall, "Kurtosis as peakedness," *The American Statistician*, vol. 68, no. 3, pp. 191–195, July 2014.

[230] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, September 2018.

[231] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Annals of Biomedical Engineering*, vol. 41, no. 2, pp. 349–365, February 2013.

[232] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. 2039–2049, June 2000.

[233] G. Palshikar, "Simple algorithms for peak detection in time-series," in *Proceedings of the 1st International Conference on Advanced Data Analysis, Business Analytics and Intelligence*. Ahmedabad, India: IIMA, 2009, pp. 2–12.

[234] S. Amiriparian, A. Awad, M. Gerczuk, L. Stappen, A. Baird, S. Ottl, and B. Schuller, "Audio-based recognition of bipolar disorder utilising capsule networks," in *2019 International Joint Conference on Neural Networks (IJCNN).* Budapest, Hungary: IEEE, 2019, pp. 1–7.

[235] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH).* Graz, Austria: ISCA, 2019, pp. 221–225.

[236] Z. Guan, L. Chen, W. Zhao, Y. Zheng, S. Tan, and D. Cai, "Weakly-supervised deep learning for customer review sentiment classification," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI).* New York, USA: AAAI, 2016, pp. 3719–3725.

[237] L. Stappen, N. Cummins, E.-M. Meßner, H. Baumeister, J. Dineley, and B. Schuller, "Context modelling using hierarchical attention networks for sentiment and self-assessed emotion detection in spoken narratives," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6680–6684.

[238] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, March 1989.

[239] V. Pandit and B. Schuller, "On many-to-many mapping between concordance correlation coefficient and mean square error," *arXiv preprint arXiv:1902.05180*, 2019.

[240] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[241] N. Kriegeskorte and T. Golan, "Neural network models and deep learning," *Current Biology*, vol. 29, no. 7, pp. 231–236, April 2019.

[242] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[243] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010, pp. 807–814.

[244] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, June 2014.

[245] L. Prechelt, “Early stopping-but when?” in *Neural Networks: Tricks of the trade*, L. Prechelt, Ed. Springer, 1998, pp. 55–69.

[246] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[247] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, March 1994.

[248] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, June 2017.

[249] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proccedings of the 3rd International Conference on Learning Representation (ICLR)*. San Diego, USA: ICLR, 2015.

[250] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, “Human attention in visual question answering: Do humans and deep networks look at the same regions?” *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, October 2017.

[251] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International Conference on Machine Learning (ICML)*. Lille, France: PMLR, 2015, pp. 2048–2057.

[252] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, pp. 68–80.

[253] L. Stappen, G. Rizos, and B. Schuller, “X-aware: Context-aware human-environment attention fusion for driver gaze prediction in the wild,” in *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*. Utrecht, the Netherlands: ACM, 2020, p. 858–867.

[254] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5585–5599, 2018.

[255] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS).* Montreal, Canada: ACM, 2014, p. 3104–3112.

[256] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: ACL, October 2014, pp. 1724–1734.

[257] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Las Vegas, USA: IEEE, 2016, pp. 21–29.

[258] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[259] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* London, UK: ACM, 2018, pp. 1049–1058.

[260] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.

[261] S. Hamieh, V. Heiries, H. Al Osman, and C. Godin, "Multi-modal fusion for continuous emotion recognition by using auto-encoders," in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge and Workshop (MuSe), co-located with the 29th ACM International Conference on Multimedia (ACMMM).* Virtual Event, China: ACM, 2021, p. 21–27.

[262] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe), co-located with the 28th ACM International Conference on Multimedia (ACMMM)*, Virtual, 2020, pp. 27–34.

[263] H.-J. Yang, G.-S. Lee, J.-H. Kim, and S.-H. Kim, “Multimodal fusion with attention mechanism for trustworthiness prediction in car advertisements,” in *Proceedings of the 9th International Conference on Smart Media and Applications (SMA)*. Jeju, Korea: ACM, 2020.

[264] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification,” *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[265] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth, “Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition,” in *2008 IEEE International Conference on Multimedia and Expo (ICME)*. Hannover, Germany: IEEE, 2008, pp. 1333–1336.

[266] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, “Achieving human parity in conversational speech recognition,” *arXiv preprint arXiv:1610.05256*, 2016.

[267] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi.” in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 498–502.

[268] A. Baird and B. Schuller, “Considerations for a more ethical approach to data in ai: on data representation and infrastructure,” *Frontiers in Big Data*, vol. 3, pp. 25–36, 2020.

[269] L. C. De Silva, T. Miyasato, and R. Nakatsu, “Facial emotion recognition using multimodal information,” in *Proceedings of 1997 International Conference on Information, Communications and Signal Processing (ICICS)*, vol. 1. Singapore, Singapore: IEEE, 1997, pp. 397–401.

[270] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[271] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “Elan: a professional framework for multimodality research,” in *5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006, pp. 1556–1559.

[272] J. M. Girard and A. G. Wright, “Darma: Software for dual axis rating and media annotation,” *Behavior Research Methods*, vol. 50, no. 3, pp. 902–909, june 2018.

[273] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing annotations for visual object detection," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*. Toronto, Canada: AAAI, 2012, pp. 40–46.

[274] K. Konyushkova, J. Uijlings, C. H. Lampert, and V. Ferrari, "Learning intelligent dialogs for bounding box annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, USA: IEEE, 2018, pp. 9175–9184.

[275] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. Cancun, Mexico: IEEE, 2005, pp. 381–385.

[276] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2286–2294, 2009.

[277] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 279–294, 2015.

[278] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. Geneva, Switzerland: AAAC, 2013, pp. 85–90.

[279] S. Mariooryad and C. D. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *Introduction to Information Retrieval*, vol. 100, pp. 2–4, 2008.

[280] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proceedings of the INTERSPEECH 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology (L2WS)/Speech and Language Technology in Education (SLaTE)*. Tokyo, Japan: ISCA, 2010.

[281] A. Baird, L. Stappen, L. Christ, L. Schumann, E.-M. Meßner, and B. W. Schuller, "A physiologically-adapted gold standard for arousal during a stress induced scenario," in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with*

the 29th ACM International Conference on Multimedia (ACMMM). Changu, China: ACM, 2021.

[282] E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia: ISCA, 2005, pp. 813–816.

[283] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[284] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 3, pp. 306–307, 1979.

[285] M. J. Zaki and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. New York, NY: Cambridge University Press, 2014, pp. 187–191.

[286] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[287] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, 2000.

[288] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100–108, 1979.

[289] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.

[290] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 0, pp. 2825–2830, November 2011.

[291] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[292] P. Tzirakis, S. Zafeiriou, and B. Schuller, "Real-world automatic continuous affect recognition from audiovisual signals," in *Multimodal Behavior Analysis in the Wild*. Elsevier, 2019, pp. 387–406.

[293] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Efficient modeling of long temporal contexts for continuous emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Cambridge, UK: AAAC, 2019, pp. 185–191.

[294] R. Li, J. Zhao, J. Hu, S. Guo, and Q. Jin, "Multi-modal fusion for video sentiment analysis," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop (MuSe), co-located with the 28th ACM International Conference on Multimedia (ACMMM)*. Seattle, USA: ACM, 2020, pp. 19–25.

[295] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[296] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. Sydney, Australia: IEEE, 2010, pp. 911–916.

[297] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the Conference on Artificial Intelligence (AAAI)*. Virtual: ACM, 2020, pp. 8992–8999.

[298] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Project Report, Stanford*, vol. 1, no. 12, pp. 1–6, January 2009.

[299] "Twitter us airline sentiment," 2015, https://www.kaggle.com/crowdflower/twitter-airline-sentiment.

[300] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "Semeval-2013 task 2: Sentiment analysis in twitter," in *Proceedings of the Seventh International Workshop on Semantic Evaluation, Second Joint Conference on Lexical and Computational Semantics (SEM)*. Atlanta, USA: ACL, 2013, pp. 312–320.

[301] J. S. Lim, M.-J. Choe, J. Zhang, and G.-Y. Noh, "The role of wishful identification, emotional engagement, and parasocial relationships in repeated viewing of

live-streaming games: A social cognitive theory perspective," *Computers in Human Behavior*, vol. 108, pp. 106 327–106 329, July 2020.

[302] L. Stappen, J. Thies, G. Hagerer, B. W. Schuller, and G. Groh, "Graphtmt: Unsupervised graph-based topic modeling from video transcripts," *arXiv preprint arXiv:2105.01466*, 2021.

[303] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems*. Hong Kong, China: ACM, 2013, pp. 165–172.

[304] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Los Angelos, California: ACL, 2010, pp. 804–812.

[305] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model for aspect extraction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: ACL, 2017, pp. 388–397.

[306] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content," in *12th International Workshop on the Web and Databases*. Rhode Island, USA: ACM, 2009, pp. 1–6.

[307] A. Schofield and D. Mimno, "Comparing apples to apple: The effects of stemmers on topic models," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 4, pp. 287–300, 2016.

[308] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[309] S. Sia, A. Dalmia, and S. Mielke, "Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too!" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Virtual: ACL, 2020, pp. 1728–1736.

[310] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[311] R.-G. Radu, I.-M. Rădulescu, C.-O. Truică, E.-S. Apostol, and M. Mocanu, "Clustering documents using the document to vector model for dimensionality reduction," in *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*. Virtual: IEEE, 2020, pp. 1–6.

[312] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 3, pp. 211–225, 2015.

[313] D. W. Matula, "k-components, clusters, and slicings in graphs," *SIAM Journal on Applied Mathematics*, vol. 2, no. 3, pp. 459–480, May 1972.

[314] D. R. White and M. Newman, "Fast approximation algorithms for finding node-independent paths in networks," *SSRN Electronic Journal*, pp. 1–9, June 2001.

[315] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*. Shanghai, China: ACM, 2015, pp. 399–408.

[316] S. Blair, Y. Bi, and M. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Applied Intelligence*, vol. 50, no. 1, pp. 138–156, Jul. 2020.

[317] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada: Curran Associates Inc., 2009.

[318] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden: ACL, 2014, pp. 530–539.

[319] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*. Melbourne, USA: IEEE, 2003.

[320] F. Ringeval, E. Marchi, M. Mehu, K. Scherer, and B. Schuller, "Face reading from speech – predicting facial action units from audio cues," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. Dresden, Germany: ISCA, 2015, pp. 1977–1981.

[321] T. Bänziger and K. R. Scherer, “Introducing the geneva multimodal emotion portrayal (gemep) corpus,” in *Blueprint for affective computing: A sourcebook*, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford University Press Oxford, UK, 2010, pp. 271–294.

[322] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille, France: JMLR, 2015, pp. 448–456.

[323] A. Landowska, “Uncertainty in emotion recognition,” *Journal of Information, Communication and Ethics in Society*, vol. 17, no. 3, pp. 273–291, September 2019.

[324] D. Heaven, “Why faces don’t always tell the truth about feelings,” *Nature*, vol. 578, no. 7796, pp. 502–504, February 2020.

[325] B. R. Duffy, “Fundamental issues in affective intelligent social machiness,” *The Open Artificial Intelligence Journal*, vol. 2, no. 1, pp. 21–34, June 2008.

[326] C. Reynolds and R. Picard, “Affective sensors, privacy, and ethical contracts,” in *Extended abstracts of the 2004 Conference on Human Factors and Computing Systems*, Vienna, Austria, 2004, pp. 1103–1106.

[327] J. Diesner and C.-L. Chin, “Gratis, libre, or something else? regulations and misassumptions related to working with publicly available text data,” in *Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHI-CA2)*. Portoroz, Slovénie: ELRA, 2016, pp. 13–17.

[328] J. P. Daries, J. Reich, J. Waldo, E. M. Young, J. Whittinghill, A. D. Ho, D. T. Seaton, and I. Chuang, “Privacy, anonymity, and big data in the social sciences,” *Communications of the ACM*, vol. 57, no. 9, pp. 56–63, September 2014.

[329] J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, *Privacy, big data, and the public good: Frameworks for engagement*, J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Eds. Cambridge University Press, 2014.

[330] I. Hutchby, M. O’Reilly, and N. Parker, “Ethics in praxis: Negotiating the presence and functions of a video camera in family therapy,” *Discourse Studies*, vol. 14, no. 6, pp. 675–690, December 2012.

[331] M. Koutsombogera and C. Vogel, “Ethical responsibilities of researchers and participants in the development of multimodal interaction corpora,” in *2017 8th IEEE*

International Conference on Cognitive Infocommunications (CogInfoCom). Debrecen, Hungary: IEEE, September 2017, pp. 277–282.

[332] F. Eyben, F. Weninger, L. Paletta, and B. W. Schuller, "The acoustics of eye contact: detecting visual attention from conversational audio cues," in *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction, International Conference on Multimodal Interaction (ICMI)*. Sydney, Australia: ACM, 2013, pp. 7–12.

[333] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines," *American Psychologist*, vol. 70, no. 6, pp. 543–556, September 2015.

[334] M. G. Anderson and P. F. Brown, "The economics behind copyright fair use: A principled and predictable body of law," *Loyola University of Chicago Law Journal*, vol. 24, p. 143, 1993.

[335] P. Svoboda, M. Hradiš, L. Maršík, and P. Zemcík, "Cnn for license plate motion deblurring," in *2016 IEEE International Conference on Image Processing (ICIP)*. Phoenix, USA: IEEE, September 2016, pp. 3832–3836.

[336] V. Stodden, F. Leisch, and R. D. Peng, *Implementing reproducible research*, V. Stodden, F. Leisch, and R. D. Peng, Eds. CRC Press, 2014.

[337] M. A. Ullah, M. M. Islam, N. B. Azman, and Z. M. Zaki, "An overview of multimodal sentiment analysis research: Opportunities and difficulties," in *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. Dhaka, Bangladesh: IEEE, 2017, pp. 1–6.

[338] European Organization For Nuclear Research and OpenAIRE, "Zenodo," 2013. [Online]. Available: https://www.zenodo.org/

[339] Z. M. Ibrahim, H. Wu, A. Hamoud, L. Stappen, R. J. Dobson, and A. Agarossi, "On classifying sepsis heterogeneity in the icu: insight using machine learning," *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 437–443, 2020.

[340] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions," *Journal of Cybersecurity*, vol. 4, no. 1, pp. 1–20, March 2018.

[341] S. Barocas, M. Hardt, and A. Narayanan, “Fairness in machine learning,” in *Neural Information Processing Symposium, Tutorials Track (NeurIPS)*. Los Angeles, USA: NeurIPS, 2017, p. 2017.

[342] N. T. Lee, “Detecting racial bias in algorithms and machine learning,” *Journal of Information, Communication and Ethics in Society*, vol. 16, no. 3, pp. 252–260, August 2018.

[343] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, “Improving fairness in machine learning systems: What do industry practitioners need?” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow, UK: ACM, 2019, pp. 1–16.

[344] L. Stappen, F. Brunn, and B. Schuller, “Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel,” *arXiv preprint arXiv:2004.13850*, 2020.

[345] T. Zhu, D. Phipps, A. Pridgen, J. R. Crandall, and D. S. Wallach, “The velocity of censorship: High-fidelity detection of microblog post deletions,” in *USENIX Security Symposium*. Washington D.C., USA: ACM, 2013, pp. 227–240.

[346] S. Rukavina, S. Gruss, H. Hoffmann, J.-W. Tan, S. Walter, and H. C. Traue, “Affective computing and the impact of gender and age,” *PloS one*, vol. 11, no. 3, p. e0150584, 2016.

[347] S.-G. Jung, J. An, H. Kwak, J. Salminen, and B. J. Jansen, “Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race,” in *Twelfth International AAAI Conference on Web and Social Media (AAAI)*. Palo Alto, USA: AAAI, 2018.

[348] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an _obviously_ perfect paper),” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 4619–4629.

[349] B. Yao and L. Fei-Fei, “Modeling mutual context of object and human pose in human-object interaction activities,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. San Francisco, USA: IEEE, 2010, pp. 17–24.

[350] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[351] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[352] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao *et al.*, "Interpretability of deep learning models: A survey of results," in *2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SCI)*. Dehradun, India: IEEE, 2017, pp. 1–6.

[353] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[354] D. Doran, S. Schulz, and T. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," in *CEUR Workshop Proceedings*, vol. 2071, 2018.

[355] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[356] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, and W. H. Chan, "Segment-level joint topic-sentiment model for online review analysis," *IEEE Intelligent Systems*, vol. 34, no. 1, pp. 43–50, 2019.

[357] S. Ruder, P. Ghaffari, and J. G. Breslin, "A hierarchical model of reviews for aspect-based sentiment analysis," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMLP)*. Austin, Texas: ACL, 2016, pp. 999–1005.

[358] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, pp. 68–80.

[359] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.