UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
**IECS International Doctoral School**

# Deep Emotion Analysis of Personal Narratives

Aniruddha Tammewar

Advisor

Prof. Giuseppe Riccardi

Università degli Studi di Trento

July 2022

# Abstract

The automatic analysis of emotions is a well-established area in the natural language processing ( NLP ) research field. It has shown valuable and relevant applications in a wide array of domains such as health and well-being, empathetic conversational agents, author profiling, consumer analysis, and security. Most emotion analysis research till now has focused on sources such as news documents and product reviews. In these cases, the NLP task is the classification into predefined closed-set emotion categories (e.g. happy, sad), or alternatively labels (positive, negative). A deep and fine-grained emotion analysis would require *explanations* of the trigger events that may have led to a user state. This type of analysis is still in its infancy. In this work, we introduce the concept of Emotion Carriers (EC) as the speech or text segments that may include persons, objects, events, or actions that manifest and explain the emotions felt by the narrator during the recollection. In order to investigate this emotion concept, we analyze Personal Narratives (PN) - recollection of events, facts, or thoughts from one's own experience, - which are rich in emotional information, are less explored in the emotion analysis research. PNs are widely used in psychotherapy and thus also in mental well-being applications. The use of PNs in psychotherapy is rooted in the association between mood and recollection of episodic memories.

We find that ECs capture implicit emotion information through entities and events whereas the valence prediction relies on explicit emotion words such as *happy, cried, and angry.* The cues for identifying the ECs and their valence are different and complementary. We propose fine-grained emotion analysis using valence and ECs. We collect and annotate spoken and written PNs, propose text-based and speech-based annotation schemes for valence and EC from PNs, conduct annotation experiments, and train systems for the automatic identification of ECs and their valence.

**Keywords:** Emotions, Natural Language, Narrative Understanding, Sentiment Analysis

# Acknowledgments

This thesis is an outcome of cumulative efforts of many people including my supervisor, family, friends, peers and professors among others. This thesis would be incomplete without expressing my sincere gratitude to them.

First of all, I would like to thank my supervisor, Prof. Giuseppe Riccardi for all the brainstorming sessions involving exchange of ideas, agreements, and disagreements; for providing me directions towards high quality research at the same time giving me enough freedom to explore, experiment, and fail; and most importantly for believing in me throughout the journey. He was always there for me, not only as a supervisor but also as a friend by making sure I was enjoying my journey in a foreign country. Specially during the lockdown, he kept a close watch on everybody's physical and mental well-being, and always tried to cheer everyone up.

I would like to thank Alessandra for her support, opinions, and suggestions during the initial phase of this journey. I would also like to thank Gabriel and Mahed for all the collaborative work that we have done. This thesis is shaped by insights from our conversations and discussions at SIS lab, thus I extend my thanks to all the labmates including Morena and Carmelo among others. Moreover, I want to express my gratitude to Andrea Stenico, for always fighting to be helpful.

A significant part of my work was only possible because of the collaborators from other universities. I would like to thank Dr Eva-Maria Messner and the students from the Ulm University for the help in data creation and annotation. I would like to thank Prof. Korbinian Riedhammer for hosting me at Technische Hochschule Nürnberg (THN). I extend my thanks to colleagues at THN. In particular, I must mention Sebastian and Farnziska for our deep discussions and tight collaboration during my stay, as well as afterwards.

I was fortunate to make some really good friends in Trento, who helped

me to have a cheerful social life outside research. They were there for providing moral support essential for conducting meaningful research. I must mention Sudipan, Parth, Yiling, Subhankar, Burcu, Nandu, Nasrullah, Mounika, Abhishek and many others, for all the wonderful evenings, memorable trips, pleasant dinner parties, and light-hearted aperitivos!

Finally, I'm grateful to my family, Mummy and Baba for always believing in me and supporting me in my decisions, especially my wife Navu for being my pillar of strength, my niece Spruha for all the gibberish on the calls, and Sam and Vahini for always cheering me up.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 The Context

Emotion analysis from different modalities such as text, speech, video, gestures, and physiological signals has received significant attention from the research community. It has found applications in a wide array of domains such as health and well-being [3, 9, 25], empathetic conversational agents [54, 130], consumer analysis [2, 61], user profiling [148, 89] and security, among others. Most works on emotion analysis have focused on domains such as product reviews[139], social media posts[118], conversations[117], and news[100]. In these scenarios, the most common emotion analysis performed are the utterance and document level emotion recognition in terms of predefined emotion categories (such as happy, angry, or sad) or on a set of numeric scales from emotional valence, arousal, and dominance [129]. However, the semantic information associated with expressed emotions, such as the events that triggered the emotion or the target toward which the emotion is directed, is important to provide a fine-grained understanding of the context that might be needed in real-world applications. This type of deeper understanding of the emotions is still in its infancy. For the explanation of the trigger events, in this work, we introduce the concept of Emotion Carriers (EC) as the speech or text fragments, which may include persons, objects, events or actions that manifest and explain the emotions felt by the narrator during the recollection. We investigate this concept in the context of an important and ubiquitous but often neglected genre of Personal Narratives (PN).

Personal Narratives (PN) are the recollection of events, facts, emotions, or thoughts felt or experienced by the narrator. Recollection of PNs have shown to be beneficial in psychotherapy and general mental well-being in various ways. In psychotherapy sessions, clients share PNs with therapists to provide a rough idea of their orientation toward life and the events and pressures surrounding the problem they are facing. Therapists then analyze the shared PNs, try to understand/identify the emotional state of the client and ask follow-up questions to know details about a particular part from the PN to extract more information from the client that could help better understanding of the situation. The analysis and the actions performed by the therapist depend on the intervention therapy approach. For example, the most widely used Cognitive Behavioral Therapy (CBT) intervention is based on the intuition that more than the events themselves, the emotions are generated depending on how the events are cognitively processed and evaluated. The irrational and dysfunctional beliefs influence this process [101]. In CBT, therapists collect PNs from the clients by asking them questions to identify the event that has caused the patient a certain emotion. Through the answers, in the form of PNs, they try to identify the dysfunctional thoughts and guide the patient on how to change them or find more rational and/or functional thoughts [131]. In Table 1.1, we provide an example of a personal narrative about an emotional episode, shared by a client to a therapist.

As the interventions such as CBT have standard protocols, there has been a tremendous growth in mental well-being applications (making use of such protocols) [31, 5]. Their functionality range from helping in promot-

I had a **difficult** and **weary** week, many commitments and deadlines in the office and I was unable to finish everything as I should have. When **Christmas** approaches, it seems that everything must be over, even if on December 27 we go back to work and there is still time to finish things! I was very **tired** and a little **depressed** in the evening and when I got **home** the **children** would ask me to **play**. I would have liked to be **alone** for a while, but I could not **disappoint** them. Actually I saw that being with them, listening to what happened at **school** and participating in their great **excitement** for **Christmas**, helped me a lot and then I was more **relaxed** and **happier** with my **family**.

Table 1.1: Current Emotion Analyses: A snippet of a spoken Personal Narrative (PN) about an emotional event, shared by a client with a psychotherapist. The original Italian PN is translated to English and has been anonymized and post-processed for better readability. The text is color-coded to represent the perceived emotion valence polarity of the narrator while narrating an event (gray - neutral, green - positive, red - negative), whereas the emotion words and emotion-laden words (that are commonly associated with a positive or negative emotion) are marked in bold. Negative emotion laden words may include *difficult, weary, tired, depressed, alone, disappoint*, whereas the positive emotion laden words may include: *Christmas, home, children, play, school, excitement, relaxed, happier, family.*

ing mental well-being [66]; self-monitoring [136, 7]; providing recommendations for emotion regulation [33]; providing CBT through interactions [45, 84, 66]. In these applications, PNs are collected as a part of the intervention approach (for example, as a part of questionnaire in CBT), or via different tools recommended for general well-being such as journaling (narrate important events from daily life that affected the user's mood). The applications are usually conversational along with some GUI based tools. These PNs contain rich information regarding the user such as the type of events that affect the user, the relationships with the characters involved, and the emotions associated with the events. Thus an in-depth analysis of these PNs may provide psychotherapists or/and the users with important insights about the state of the user.

## 1.2 The Problem

As explained in the Section 1.1, even though rich PNs are collected through well-being applications, mostly these are used only as a part of the protocol. The current emotion analyses systems focus on providing emotion state of the narrator, sentiment analysis, and emotion-laden words from the PNs. As can be observed in the example of PN from Table 1.1, the perceived emotional state (in terms of emotion valence) of the narrator changes as the narration progresses. The emotional valence predicted for the entire PN or for each segment helps keeping track of the emotional state of the user. Another emotion analysis often used for better understanding of the context is of identifying emotion words and emotion-laden words such as *difficult, weary, Christmas, tired, depressed,* among others, from our example. These words are a part of lexicon that are commonly associated with positive or negative emotions.

These analyses provide an overview of the trends in the emotional state of the user and help identify cognitive distortions (dysfunctional thoughts) and provide with a relevant CBT-based/therapeutic tools (such as relaxed breathing, journaling), with the aim of cognitive reconstruction (making thoughts rational). This is a better way of providing self-help, through an application.

A deeper analysis of PNs beyond the specific requirements of the protocol involved in the intervention approach is still missing in these applications, which may prove to be helpful for better assessment of the user's condition. In our example from Table 1.1, with current systems, we can identify the negative emotion valence of the narrator in the first sentence. But there's much more than the negative emotional state. Some basic but very important questions are still unanswered:

- **How can we track the emotional state of the user throughout the narrative and take into account the context ?**

- **What is the reason behind the narrator's emotional state at a specific point ( sentence or clause ) of the narrative ?**

- **Which events, characters, and objects from the narrative affect the user emotionally?**

- **Which event, object, character may have contributed to the state of the user at a specific point ( sentence or clause ) of the narrative ?**

Automatically addressing ( some of ) these questions and provide explanations may help therapists in deciding which events and characters to focus more on, to elicit more relevant information, during a therapy session.

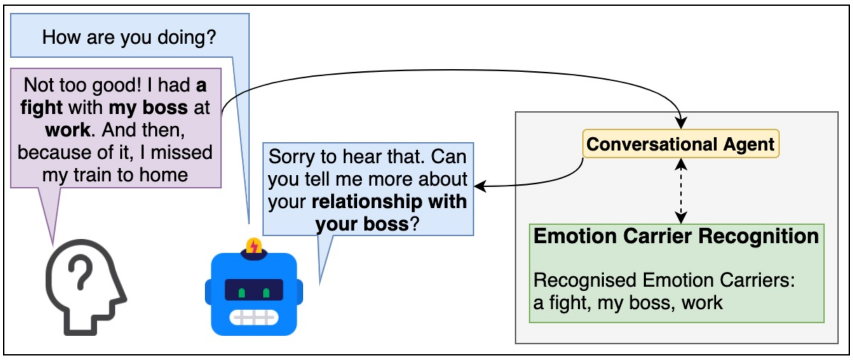

How are you doing?
Not too good! I had a **fight** with **my boss** at **work**. And then, because of it, I missed my train to home
Sorry to hear that. Can you tell me more about your **relationship with your boss**?
Conversational Agent
Emotion Carrier Recognition
Recognised Emotion Carriers: a fight, my boss, work

Figure 1.1: A possible application of automatic Emotion Detection (ED) and Emotion Carrier Detection (ECD): A Conversational Agent (CA), with ED and ECD components, identify the user's emotional state as **"angry"** and three ECs from the user utterance. One of the ECs **"my boss"** is then used to generate a response targeted towards eliciting more information from the user, beneficial for a fine-grain description of the user state of being **"angry"**.

## 1.3 The Solution

In the scenario of PNs, for fine-grained emotional understanding, we propose identification of Emotion Carriers (EC) as the speech or text fragments that manifest or explain the emotional state of the narrator, as an important additional analysis. In our example of PN from Table 1.1, text spans such as *many commitments, deadlines, office, Christmas, children, being with them, my family* could be good candidates for ECs as they manifest the corresponding emotional states of the narrator. The conversational agents of the well-being applications can make use of ECs to ask questions to elicit more information regarding the corresponding EC, for deeper understanding of how the EC is affecting the emotional state of the narrator. A demonstration can be seen in Fig 1.1, where a conversational agent first elicits an emotional narrative from the user and later detects the ECs and use one of them to further elicit more information for deeper understanding of the situation.

With the goal of building an automatic EC identification system and using it for analyzing PNs, in this thesis we perform all the steps starting from defining the concept till building an automated system. First we study deeply the genre of Personal Narratives (PN), it's association with emotions and thus how they are used in psychotherapy. We also study different modalities such as speech and text that may involve in PNs. We then propose the concept of Emotion Carriers in the context of PNs as the text or speech span that manifest the emotion of the narrator during the recollection of the narratives. We back Emotion Carriers from the theory of *Emotion Concepts* from psychology. We compare Emotion Carriers with several different existing emotion analyses schemes, which analyze different aspects of emotions and show how Emotion Carriers and Emotional State when combined provide fine-grained emotion analysis.

Once the concept of ECs and PNs are established, we explore the datasets of PNs that are relevant in our scenario of well-being. There are not many publicly available datasets of PNs. We explore and report details of two publicly available datasets of PNs namely USoM dataset (German) [132, 125] and SEND dataset (English) [102]. Two other publicly available English datasets that are relevant to us but not deeply explored in this study include "Counseling and Psychotherapy Transcripts" published by Alexander Street Press and Motivational Interviewing (MI) session videos [113]. We also study and report 4 other datasets which as per our knowledge, are not available to public yet. These include USoM Elderly dataset (German), PHA-CBT dataset (Italian), MEMOA dataset (Dutch), and User Diaries (Italian). While exploring the datasets there are many factors we study, some of which include: The language used in the personal narratives; The modalities collected during the recollection (any combination of text, speech, video, physiological signals); The demographics of the participants (age, mother tongue, geographical region, profession, or any other relevant information for study); The self and expert annotation provided with the data among others. We further go deeper and perform annotation experiments for human annotation of valence (as emotional state) and Emotion Carriers and build automated systems making use of the three datasets USoM, USoM Elderly, and CBT-PHA.

In one of our early experiments [141], we worked on identification of the narrator's self-assessed valence after recollecting PNs (collected in the USoM dataset). We observed that in different machine learning models (Support Vector Machine, Attention-based neural sequence tagger) trained to predict the valence from PNs, concepts beyond sentiment words (sad, happy) were found to be useful. These concepts included terms such as characters (e.g. grandfather, a friend), locations (e.g. swimming pool) and events (e.g. high school exam). As the task of the models was to predict the emotional state (in terms of valence) of the narrator, we find that these concepts played the role of explaining and carrying the emotional state of the person, which we defined as emotion carriers. Inspired by such evidence, we investigate the possibility of annotating emotion carriers and valence in PNs. We come up with different annotation protocols based on multiple factors such as the modalities considered during the annotation (we experiment with text only and speech + text); freedom provided for the EC-span selection (free span selection of continuous tokens or selection from a preselected list of candidate EC-spans); the amount of context provided during annotation (entire PN, segments: breaking PN into segments and annotate each segment with valence and EC). We begin the annotation experiments with the USoM dataset with narrative level annotation of textual transcripts, and investigate the outcome, the time consumed for the task, and the complexity of the task. We observe high subjectivity and complexity of the task, based on which we also analyze the pain points and come up with other annotation schemes to resolve these issues. When we shift from narrative level annotation to segment level annotation, instead of providing a single valence label to the entire PN, we provide a label for each segment, making the task easier and possible to capture the change in emotions of the narrator as the recollection proceeds, which we call as unfolding of the emotions. Similarly, providing speech during annotation helped the annotators in the task by using the cues from speech. To evaluate the quality of the annotation, we calculate the Inter-Annotator Agreement using different metrics. In the EC detection as a free span selection task, much different from a classification task where there are predefined set of classes, we cannot use the standard Cohen's kappa metric [28] in our scenario as it requires the knowledge of true negatives, which are ill-defined for the span selection task. Instead, we use pairwise Positive Agreement [46]. We come up with different strategies for soft/hard matching of two spans, based on different criteria. We also perform qualitative analysis of the annotation and find interesting observations w.r.t. annotation of emotion words, the valence patterns followed in negative PNs as compared to the positive PNs, the role of neutral valence, among others.

After evaluating the annotation, we try to make use of the annotated data to build automated systems for predicting the valence and ECs from PNs. As the annotation were performed with different strategies, we build different automated systems accordingly. First we explore the text based narrative level prediction of ECs from the transcriptions of the PNs from USoM dataset. We then also try to consider cues from corresponding speech fragments, in addition to the cues from the text with a multimodal system, exploring different fusion strategies to combine features at different stages in the architecture 1) early fusion 2) late fusion 3) decision level fusion. Using the annotation of the CBT-PHA dataset, we also build a system for segment-level detection of valence and ECs. We also jointly train a multi-task model for identification of ECs and valence, and analyze how joint training affects the performances of the individual tasks. Lastly, we compare the human annotation of ECs with the tokens which contribute more in the valence prediction system, with the aim of verifying if the valence and ECs capture the same information or different. For calculating the contribution of tokens toward the model's decision, we make use of Integrated Gradients [140] - a technique commonly used in explainability studies.

## 1.4 Novel Contributions

The major contributions of this thesis include:

- We introduce a new emotion analysis task of "Emotion Carriers" and propose that Emotion Carriers and Valence (Emotional State) combined, provide a fine-grained emotional analysis useful for many applications

- We identify datasets of Personal Narratives relevant for our work on emotion analysis in two languages German and Italian. We propose different annotation schemes to annotate Emotion Carriers and Valence from PNs from the datasets. We annotate the datasets using proposed annotation schemes.

- We perform different quantitative and qualitative analysis of the annotation to assess the quality of annotation. We come up with different Inter-annotator Agreement Analysis metrics to cater to calculate the agreement between annotators. The different metrics measure the agreement based on different criteria such as soft/strict matching of two EC spans, agreement based on valence trajectories.

- We build automated systems for the detection of valence and Emotion Carriers for two languages German and Italian and for different modalities (text only, text + speech).

## 1.5 Structure of the Thesis

We begin our thesis with a survey of the previous relevant works on emotion analysis and automatic narrative understanding and comparatively position our work in the field, in Chapter 2. In Chapter 3, we first go deep into the domain of Personal Narratives, compare it with other domains and explore it's use in psychotherapy while in the second part of the chapter, we introduce the concept of Emotion Carriers, the motivation from psychology, compare and differentiate with other emotion analysis tasks. Next, in Chapter 4, we explore important datasets of PNs relevant for the domain of emotion analysis. We select and work on three of these datasets for further experimentation. Before building the automated systems, we first analyze how humans find the task of EC and valence detection. In chapter 5, we propose different annotation protocols and perform the human annotation experiments on the three datasets. We also propose different evaluation metrics for calculating inter annotator agreement suitable for Emotion Carriers annotation. We analyze the annotation output qualitatively and quantitatively. Once we have the annotated data, we run different experiments to build automated systems, playing with multiple modalities, different segmentation levels, and multi task modeling in Chapter 6. In Chapter 7, we compare human annotation of (tokens from) ECs with the tokens that help valence prediction model in decision making. We find a mismatch between these two sets, proving that Emotion Carriers and Valence explain different things, both of which are important for a complete emotional understanding. Finally we conclude the thesis in Chapter 8 and provide future directions where our work could be utilized or extended.

# Chapter 2

# Related work

In this chapter, we study relevant and established works on different emotion analyses and narrative understanding tasks and position our work in these domains.

## 2.1 Emotion Detection From Speech And Text

The most widely studied emotion analysis task on speech and text data is of Emotion Detection, also called as Emotion Recognition. Emotion recognition has emerged as an important research area which may reveal some valuable input to a variety of purposes. Emotion recognition is the process of identifying emotional state of human by merely depending on personal skills and interpretation. The emotion state representation can be categorized in two classes *Categorical* and *Dimensional* [129]. Categorical representation involves selecting an emotion from a set of predefined discrete emotion categories [19] such as commonly used Ekman's six basic emotions (Anger, disgust, fear, joy, sadness, surprise) [41]. On the other hand, Dimensional representations define a few dimensions with some parameters and specify emotions according to those dimensions. The commonly used dimensions include valence, arousal, and dominance [20, 128]. Automated emotion recognition from speech and text can prove to be useful in different domains and applications such as making Human Computer Interaction (HCI) more natural and empathetic [29, 73], e-commerce websites may utilize product reviews for market research and call centers may use speech emotion recognition for quality control [30]. Emotion recognition is also useful in mental well-being applications to monitor user's emotional state, which in turn can be used to identify stress and depression through various sources such social media posts [59] and personal narratives collected by well-being applications as a part of an intervention or journaling [52].

Emotion state detection does indeed provide important information regarding user's emotional state, which is important for psychotherapy and mental well-being applications, but it does not provide the reasoning behind the emotional state of the user. In our work, in addition to emotional state detection of users in terms of valence score, we also extract the explanation of the emotional state in the form of text or speech fragments from the PNs, which we call as Emotion Carriers (EC). ECs may prove to be important for extracting relevant information from the users to better understand the emotional state.

| |
|---|
| **Personal narrative:** I am generally a person who needs a lot of sleep, but today I was not able to sleep more than 6 hours and I am extremely tired. My eyes hurt and two hours later I have programming [lesson] so I have to be alert. I've already drunk a cup of coffee and although I rarely drink coffee, it had no effect on me. I am not at home so I have limited possibilities as for food. I don't want to do anything too unhealthy such as drinking 10 cups of coffee, tho I may consider drinking another one. |

Table 2.1: Example Personal Narrative from [48], shared with the purpose of seeking advice from the narratee. The task here is to find out *'Which advice-seeking question is more likely to have been asked by the narrator:' 'Q1: Is it even possible to be addicted to coffee?' OR 'Q2: How can I energize myself?'*

## 2.2 Automatic Narrative Understanding

| | **Task** | **Desired Output** |
|---|---|---|
| 1 | Question generation | What do I need to do in 2 hours? |
| | Reading comprehension | |
| | Summarization | I must go for a lesson after getting little sleep. |
| 2 | Ending generation | Lastly, I tried an energizing drink. |
| | Narrative chains, story cloze | |
| 3 | Event2Mind | to learn to code, to be educated |
| | Desire fulfillment | |
| | **Emotion Carriers Detection** | **programming lesson, unhealthy, drink coffee** |
| 4 | Advice-seeking question | How can I energize myself? |
| | **Emotional State (in terms of valence score)** | **3 (10 point Likert Scale)** |

Table 2.2: Categorization of various Automatic Narrative Understanding tasks based on the explicitness of the information to be extracted, as proposed by Fu et al [48]. The entries in the *Task* column are the ANU tasks, while the *Desired Output* is the example output expected from the corresponding task. In the original tasks we add two tasks of our interest, the Emotional state prediction, and Emotion Carriers Detection. The Input for the tasks is the personal narrative from Table 2.1. The authors assumed the second sentence (*"My eyes hurt and two hours later I have programming lesson so I have to be alert."*) to be the answer span for question generation, and the input for Event2Mind (which operates at sentence level).

The field of Automatic Narrative Understanding (ANU) deals with understanding narratives from different perspectives as per the application needs. There have been many works on different ANU tasks but very few have worked on the emotion analysis of PNs.

There have been many previous works on understanding narratives from different perspectives based on the applications. Fu et al [48] broadly categorize these tasks into four categories, based on the explicitness of the information to be extracted. The categories are ordered according to the explicitness, i.e. the first category try to capture the most *'intradiegetic'* apsects while the last category capture the *'extradiegetic'* aspects. We study these categories and try to position our tasks in this categorization.

1. **What happened in the story?** These tasks try to grasp the content explicitly provided in the text. The central idea is to be able to understand the events that occur in the story, identify the characters involved, the time and place (if mentioned) and any other characteristics of the event present in the narrative. This requires general semantic understanding. Some of the prevalent tasks include summarization of the narrative [96, 4], generating questions that are answerable from the text (question generation [38]), or answer a question about the narrative(reading comprehension, [23])

2. **What might happen next?** Often, while trying to understand a story we not only grasp the content provided so far but also try to predict the next event that might happen. Some tasks aim at predicting the likely action trajectory, trying to predict the future. Related tasks include the narrative cloze task [21], the story cloze test [91, 22], and its generative versions [56]. These tasks may require deeper understanding of the narrative and some common sense reasoning along with the semantic understanding.

3. **What can we infer about the characters and entities?** Other important aspects humans try to make inference about are the mental states such as attitudes and desires of the actors involved in the story. This requires logical inference or common sense based reasoning. [122, 123] try to predict the intents and reactions of the actors involved. Whereas, [121] try to find out whether the goals of the actors were fulfilled.

4. **What is the intention of the narrator in sharing their story?** Fu et al.[48] proposed the task of identifying the intention of the narrator behind sharing the story. They try to find out *why* the story is

shared by the narrator by analyzing *how* the story is constructed. In particular, they work on the narratives which are shared with the purpose of seeking advice, as can be seen in the PN example from Table 2.1. The task differs from the previous tasks as the predicted information (the advice-seeking question) is not explicitly mentioned in the narrative. In other terms, it tries to capture extradiegetic aspects of the narrative.

Now we describe in brief, the tasks of Emotional State Detection (in terms of Valence) and Emotion Carriers Detection and try to position them in the above categories.

1. **What is the narrator's Emotional State?** As described earlier, the tracking of user's emotional state may provide insights into the mental health of the user, we propose a task of predicting user emotional state after recollection of a narrative (narrative level). In this thesis, we also explore the valence after narrating each segment of a narrative (segment level), but for the purpose of positioning the task in the ANU categories, we consider only the narrative level analysis. We capture the Emotional State through a numeric valence score. This information may or may not be captured explicitly in the content. Even if the information is present in the form of cues such as "I'm excited" or "I was feeling nervous", some reasoning is required to get a numeric score. Also, the valence depends not only on the content of the current narrative but also the context like previous recollections, the speech context and the external context like the setup in which the PNs are collected. The closest category of tasks would be the $4^{th}$ category as the task tries to capture some extradiegetic aspects. But for the valence prediction task, additional context information is also helpful. [132] performed a similar task on personal narratives.

2. **Which spans of the narrative explain the current emotional state?** In this task, we try to identify the Emotion Carriers, which include entities like characters and events (e.g. “father”, “exam”, etc.) that manifest the valence of the narrator. This information has to be derived from the content of the narrative. Thus the information can be categorized to be intradiegetic. But the extraction requires logical and common sense reasoning. We can categorize this task into the $3^{rd}$ category, provided that we have already predicted the emotional valence.

Consider the eaxample of a Personal Narrative from Fu et al.[48], in the Table 2.1. For this example, Table 2.2 shows the differences between different ANU tasks discussed above.

# Chapter 3

# Personal Narratives and Emotion Carriers

In this chapter we study about Personal Narratives (PN), their properties, and how they are different from other genres. We investigate how PNs are used in psychotherapy and general mental well-being. In the second part, we introduce the concept of Emotion Carriers (EC) in the context of PNs, explain the motivation from the theory of emotion concepts from psychology, and compare with other established emotion analysis tasks.

## 3.1 Personal Narratives

### 3.1.1 Narratives

A narrative is a telling of a true or fictitious event or connected sequence of events or experiences, recounted by a narrator to a narratee [152]. Some examples of fictitious narratives include fairy-tale, story, epic, whereas the true narratives include episode, vignette, travelogue, biography, personal diary, etc. A narrative could be as short as an account of a single event (e.g. a short note from a personal diary or a brief news item) or could be as long as a novel.

### 3.1.2 Personal Narratives

In this thesis, we focus on Personal Narratives (PN), which include recollection of events, facts, emotions, or thoughts felt or experienced by the narrator him/herself. People share PNs in the form of stories to themselves and to others to place daily experiences in context and make meaning of them [86]. Personal narratives contain rich information about the user, useful for various different applications. The field of Automatic Narrative Understanding (ANU) aims at understanding narratives and extracting the information as per the target application needs. Different applications need to analyze narratives from different perspectives. For example, as studied in the section 2.2, [48] investigate a type of personal narrative which people share to seek advice from others. In this scenario, the ANU is tasked to find the narrator's intention behind sharing the narrative, in the form of an advice-seeking question. Table 2.1 shows an example of a PN shared by a narrator for the purpose of seeking advice. Travelogues is another context in which PNs are common, where the bloggers write their positive and negative experiences about a trip they went to. There has been growing number of community forums and applications such as Vent[1], where users post PNs to express their emotions and experiences with others. In these forums, users can read other users' PNs, connect with them, respond to them by sharing their PNs or what they did in a similar situations or wish them in case of a positive experience. As PNs convey emotional information, they are also widely explored in psychotherapy and thus also in well-being applications. In this thesis, we explore emotion analysis of PNs and study how it is used in psychotherapy, with the aim of building an automated emotion analysis system that extracts emotional information useful for psychotherapy and general mental well-being.

[1] https://www.vent.co/

| **Italian (original)** | **English Translation** |
|---|---|
| Mi ha telefonato Maria ieri pomeriggio e mi ha chiesto di vederci oggi a pranzo. Non può immaginare quanto ero felice di questo ma nello stesso tempo molto agitato perché da quando la nostra relazione è finita non abbiamo più trascorso insieme l'intervallo di pranzo, come facevamo quando eravamo insieme. Avevo paura che volesse recriminare e se fosse stato così non avrei saputo difendermi. Invece è stato un pranzo piacevole, lei non sembra più essere arrabbiata con me e questo mi ha molto rassicurato. Però continuo a sentirmi in colpa per il modo in cui è finita la nostra relazione. Ieri mentre eravamo al ristorante pensavo che per il mio stupido tradimento ho perso una ragazza molto intelligente, carina e che mi piace ancora. Ho sbagliato e mi vergogno. | Maria called me yesterday afternoon, she asked me to meet today for lunch. You cannot imagine how happy I was about this but, at the same time, very nervous because since our relationship ended we have not spent the lunch break together, as we did when we were together. I was afraid that she wanted to complain and if that were the case, I would not have been able to defend myself. Instead, it was a pleasant lunch, she no longer seems to be angry with me and this reassured me a lot. But I still feel guilty about the way our relationship ended. Yesterday while we were at the restaurant, I thought that for my stupid betrayal I have lost a highly intelligent, nice girl who I still like. I was wrong and I am ashamed. |

Table 3.1: An emotional experience in the form of a Personal Narrative (PN) shared by a client to an independent psychotherapist, during a session. The PN is post-processed for anonymization and better readability.

### 3.1.3 Complexity

While most NLP tasks focus on domains such as news, microblogs, and reviews, the domain of personal narratives has received a very little attention from the research community. Personal narratives are more complex information sources compared to other domains as they are typically longer, have complex discourse structure, and contain multiple sub-events. Moreover, each sub-event has attributes (such as characters, entities involved in the sub-event), the narrator's reactions and emotions expressed in the narrative. Dealing with this complexity is a challenge that we try to tackle in our tasks.

### 3.1.4 PNs and Other Genres

We compare other genres which are relevant to PNs that we consider in this thesis, yet different. These genres may include reviews of products, movies, hotels, or restaurants; blogs and vlogs about a travel experience; or social media posts. News is another genre which got attention in the emotion analysis studies.

**Social Media Posts** is the most widely used genre in the emotion analysis studies because of the easy availability of the data and the personal nature of the content. Users of a social media platform share information such as events from their personal lives, latest news, their views and opinions about something, and memes. Usually limits are put on the visibility of the post (target audience) based on the content, by the user. The post may involve different modalities such as text, images, and videos. The posts which are most relevant to the PNs considered in our work, are the ones about the personal events. In fact, these posts contain rich information useful for understanding and keeping track of the user's emotional state. However these posts tend to be short in length (some platforms like twitter put limit on the number of words), concise (typically involving a single event), and less complex as opposed to the PNs we study in this thesis, which are long containing multiple events and sub-events and consist of complex discourse structure, as explained in Section 3.1.3.

**Reviews**, for example of products, are usually the *opinions* of the customers about the product after experiencing or using the product. The customers post their views usually on e-commerce platforms, or also on social media in case of extra ordinary cases (very un/satisfactory experience), with the purpose of helping other buyers in making the buying decision by providing them information regarding the positive and negative experiences with the product. The reviews could be written with different structures such as overall rating on a scale of 5 and a list of positive and negative points, a list of adjectives as keywords, or an elaborate experience focusing on different aspects of the products. As the reviews are usually based on the personal experiences of the customers, the structure sometimes consist of a story which may include preface like the reason for buying the product ('I bought this toy as a Christmas gift for my granddaughter'), the emotions and reactions ('She was surprised when she opened the box, as she always wanted that doll. She loved it!'), comment on the overall quality and a specific aspect ('The product is very nice and creative, but is made of very cheap material. I don't think that the plastic is BPA free'), an event/incidence ('The hand of the doll came out within a few days! She was so sad, as she got attached to it. But I'm happy that she is no more playing with the toxic plastic!'), the final verdict and advice to other buyers ('Overall a nice concept but a very poor quality for a huge price tag! Very disappointed! I highly discourage others from buying it.' Although, this type of story-like reviews can be considered as personal narratives, as they might contain emotions, series of events, and characters, for the purpose of this thesis, we do not further explore them. These reviews are focused on the product and also the emotions are expressed to convey the impact level (positive/negative) of the product on the daily life. There is a scope for applying our work on emotion analysis (explained later in section 3.2 of this chapter) on these PNs, which could help the manufacturers in market research, to understand the critical issues in the product so that they can work on fixing the issues while the analysis of positive stories could be used for better marketing. We are interested in PNs that provide useful information for the well-being of the user.

### 3.1.5 PNs in psychotherapy

Rich information provided through PNs can help better understand the emotional state of the narrator, thus PNs are frequently used in psychotherapy [8]. Often, in psychotherapy sessions, clients are invited by therapists to tell their stories/PNs [64]. Through PNs, clients provide therapists with a rough idea of their orientation toward life and the events and pressures surrounding the problem at hand [64]. The recollection and novel interpretation of PNs is a key feature of psycho-therapeutic approaches [149]. Many studies found the mere act of recollection of personal experiences in an emotional way by speaking or writing them down to be therapeutic, which brings about improvements in mental and physical health [109, 108]. The use of narratives in psychotherapy is rooted in the association between mood and recollection of episodic memories [138]. Earlier work showed an interrelation between personal storytelling and self reported affect (mood) as well as mental health and word use in PNs [125, 124]. In Table 3.1, we show a real example of a personal narrative collected by an independent psychotherapist, during a psychotherapy session. We can see that the narrator is sharing rich information which contains recollecting an event which affected the narrator, varying emotions can be observed in the recollection, information about a character from the event and the status of his relationship with the character is also shared. Note that the narrative has been post-processed for anonymization and better readability.

### 3.1.6 Well-being Applications

Cognitive Behavioral Therapy (CBT) is a widely used methodological intervention approach in psychotherapy. It is based on the intuition that it is not the events that directly generate certain emotions in the clients but how these events are cognitively processed and evaluated and how irrational or dysfunctional beliefs influence this process [101]. A commonly used technique in CBT, involves identifying the event that has caused the patient a certain emotion by eliciting PNs from patients using questions from a set protocol. Once dysfunctional thoughts are identified, the patient is guided on how to change them or find more rational and/or functional thoughts [131]. Interventions like CBT have a set protocol, thus making it possible to partially or completely provide or support the therapy using automated applications. There's a growing interest in mental well-being applications, many of which are conversational. These applications elicit PNs from the users, by asking questions, similar to the CBT protocol. The cognitive distortions are identified and a CBT tool or technique is then recommended to the user. Some applications also ask users to maintain a digital diary. Diary is a popular life-logging tool for storing memories and aiding recollection [87, 133]. Diaries have shown to improve adherence by increasing the consciousness of the patients about their condition. They have proven to be very effective in gaining deep insights into a patient's well-being and can be used by a doctor or a therapist to learn about the patient's behavior and routines [53].

## 3.2 Emotion Carriers

In this thesis, we work on analysis of PNs mainly focusing on the emotional aspects. We perform emotion analysis such as emotion state recognition and their linguistic manifestations such as Emotion Carriers. In this section we introduce the Emotion Carriers in the context of PNs.

### 3.2.1 Definition

We define Emotion Carriers (EC) in the context of spoken or written PNs as the speech or text spans that explain and carry the emotions felt by the narrator during the recollection. The spans may include mentions of persons, objects, places, or events that might have affected the emotional state of the narrator, but not the emotion words themselves. In a small fragment of PN, *"I experienced a bit of distress in **the office**, because **talking with colleagues** makes me anxious"*, the ECs 'the office' and 'talking with colleagues' manifest the narrator's emotional state of being 'anxious'. ECs capture fine-grained emotion information, which could be useful in applications like empathetic conversational agents (Fig 1.1 shows a demo application).

### 3.2.2 Motivation - psychology:

According to [98], *"Concepts" are mental representations of categories of entities (natural and artifactual), situations, experience, and action.* Concepts facilitates encoding memory retrieval process, and thus is important and widely studied in psychology and cognitive science. Every person has their own definitions in terms of concepts associated with the surrounding objects and events, helping them parse the inputs like visuals of seeing someone as pleasant or unpleasant based on their concepts of that person and the activity being performed. Concepts are useful for understanding emotions of others, to know how the emotions have come about, and what can be done to alter or celebrate them. Emotional concepts define different emotions for each person differently [98]. There have been several efforts on modeling emotion concepts. One key aspect, the model should be able to account for, is the notion of personal constructs, which allows different persons to have different definitions of the same emotional concept. The Emotion Carriers are conceptualized to allow for personalization in contrast to other models such as a list of basic emotion categories [39], that are universally defined, without any personal aspect to it. Emotion carriers have innate personalization as they are defined for PNs which contain personal sentiments and emotions regarding their own experiences, events, and related participants.

### 3.2.3 Different Emotion Analysis Tasks

Emotion analysis is widely studied in Speech, NLP, and psychology. There are different emotion analysis concepts and tasks, proposed by researchers. In this section, we study well established topics related to emotion analysis, categorize them, and compare with Emotion Carriers.

#### Emotion Detection

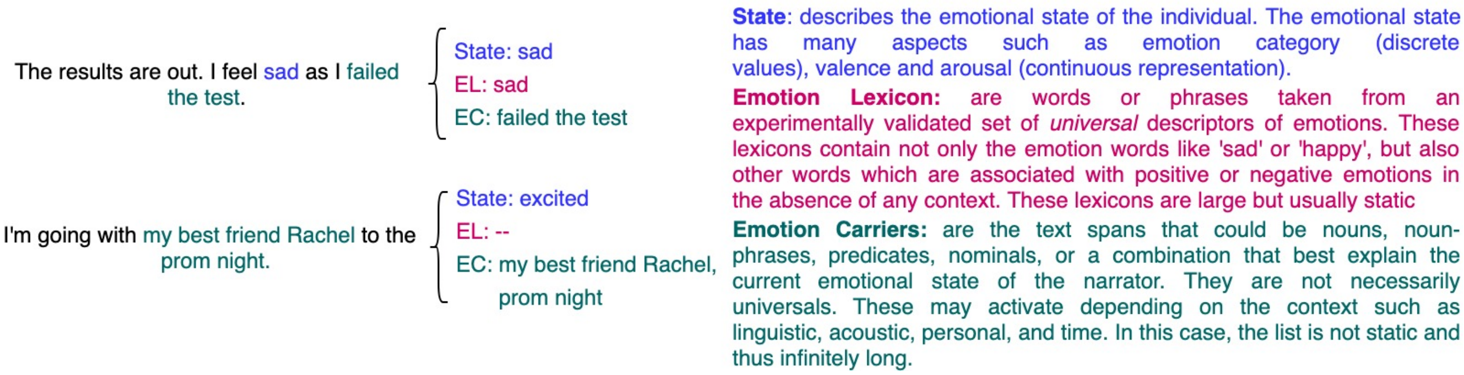
Most works on emotion analysis have focused on the task of identifying the emotional states (commonly known as emotion detection) of the narrator or the author, the characters or the participants involved, and the listener or the reader, using combinations [116] of the text [156], speech [134, 6], visual features such as facial expressions and gestures [70], or bio-signals being analyzed.

Emotional state can be represented in different ways (known as emotion models) based on many parameters such as emotion type and emotion intensity [98, 129]. Emotion models provide structures to define different human emotions using some scores, ranks, or dimensions. [129] have extensively studied literature and listed many different emotion models being used for research. Majority of the emotion models can be broadly classified into two types - *Categorical* and *Dimensional* [19]. The *Categorical* emotion models consists of a predefined list of discrete emotion categories. Researchers define categories as per the purpose of the study or the application. In the literature we find many different propositions on the number of emotion categories ranging from five (Anger, anxiety, disgust, happiness, sadness) proposed by [99] or more commonly used Ekman's six basic emotions (Anger, disgust, fear, joy, sadness, surprise) [41] up to as large as 705 hierarchically structured, fine-grained categories from the recent Vent[2] social media platform, focused on emotional well-being of it's users [85]. The *Dimensional* approach define a set of numeric measurable parameters as the dimensions, and the emotions are defined using these dimensions. Most dimensional emotion models use a subset (1, 2, or 3) of dimensions from - *'valence'* (indicates the extent of positivity or negativity of an emotion), *'arousal'* (indicates the excitement level or physical agitation of an emotion) and *'dominance'* (indicates the level of control over an emotion) [120, 20, 128].

Emotion detection is an important and useful task in diverse domains. It can be used to understand customer satisfaction through reviews, or empathetic response generation by a conversational agent. It can also be used to understand user's emotions behind social media posts. In our scenario of Personal Narratives, a mental well-being application may keep track of the user's emotional state and may also generate a summarized report for a psychotherapist. As explained in the Figure 3.1, the emotion state tells us the emotions felt by the narrator, whereas the ECs explain this emotional state through text or speech spans from the PNs. We find

[2] `https://www.vent.co/`

the emotion state as an important emotion analysis task, it complements the emotion carriers, thus combining the two, provides us with a fine-grained emotion analysis. In the Figure 1.1 (from Chapter 1), we show an use-case of a conversational agent that makes use of emotion state of the user (angry), to generate the first part of the response “sorry to hear that”, whereas using an EC (the boss), it generates the next part of the response “Can you tell me more about your relationship with your boss”, to elicit more information relevant for better understanding of the emotional state.


The results are out. I feel sad as I failed the test.
State: sad
EL: sad
EC: failed the test
I'm going with my best friend Rachel to the prom night.
State: excited
EL: --
EC: my best friend Rachel, prom night

**State**: describes the emotional state of the individual. The emotional state has many aspects such as emotion category (discrete values), valence and arousal (continuous representation).

**Emotion Lexicon:** are words or phrases taken from an experimentally validated set of *universal* descriptors of emotions. These lexicons contain not only the emotion words like 'sad' or 'happy', but also other words which are associated with positive or negative emotions in the absence of any context. These lexicons are large but usually static

**Emotion Carriers:** are the text spans that could be nouns, noun-phrases, predicates, nominals, or a combination that best explain the current emotional state of the narrator. They are not necessarily universals. These may activate depending on the context such as linguistic, acoustic, personal, and time. In this case, the list is not static and thus infinitely long.

Figure 3.1: Comparison of different emotion analysis tasks

### Emotion Lexicon

There have been several efforts to associate emotions and sentiments with words, to build large lexicons. In these lexicons, the associated emotions are generic, in the sense that these are universally accepted word-emotion associations in most common situations. The word *‘love’* would usually be associated with positive emotions whereas *‘hate’* would be associated with a negative emotion.

One line of work in this domain is on **Emotion, Emotion-Related, and Emotion-Laden words**. There have been different propositions by researchers on which group of words to annotate (such as Nouns and Adjectives [150]) and what approach (propositional or componential [103]) to use to assign emotions to the words. [106] performed an in-depth study on various propositions from the literature and come up with an approach that allows them to differentiate the two word types(emotion and emotion-laden words), based on their functions. They identify Emotion Words as words that directly refer to a particular affective/emotional states ("happy", "angry") or processes ("to worry", "to rage"), and function to either describe("she is sad") or express them ("I feel sad"). In this definition, they exclude the Emotion-Related words ("tear", "tantrum", "to scream") that describe behaviors related to particular emotions without explicitly mentioning them. Whereas the Emotion-Laden words do not refer to the emotions directly but instead express("jerk", "loser") or elicit emotions from the interlocutors ("death", "cancer"). These are commonly further sub-categorized into: (a)taboo and swearwords or expletives ("piss", "shit"), (b)insults ("idiot", "creep"), (c)(childhood) reprimands ("behave", "stop") (d)endearments ("darling", "honey"), (e)aversive words ("spider", "death"), and (f)interjections("yuk", "ouch"). The boundaries of these subcategories are not rigid as the words may reflect different emotions depending on the context [106]. Some important lexicons include English Emotion words in [69], Italian emotion words and ratings in [104] and for French words and ratings can be found in [97]. Furthermore many relevant databases for English could be found at - `https://www.reilly-coglab.com/data`

Another well-known and widely used work in this domain is a computer software **Linguistic Inquiry and Word Count (LIWC)**[3] [107], supporting many (more than 15) languages including the ones we explore in this thesis English, German and Italian. There are dictionaries for each language with words categorized in three levels of word-categories. Categories

[3] `https://www.liwc.app/`

of the latest (2022) version can be found in Table 2 of [16]. The top level in the hierarchy includes categories such as Standard Linguistic Dimensions, Psychological Processes, and Personal Concerns. These categories are further divided into sub-categories. The software takes a document as an input, checks each word against the dictionary and returns a percentage for each category, representing how much the document relates to the category. For example, the word *smile* falls into five categories: happiness, positive emotion, cognitive processing, social orientation and psychological distancing. Researchers usually focus only on a subset of relevant categories to the domain, selected manually. For example the works focusing on mental and physical health find these four categories most influential: i) self-referencing words; ii) social words; iii) positive emotion words; and, iv) negative emotion words [125, 109, 62, 119, 127]. LIWC is widely used in psychology research, for tasks such as depression [55] and autism [78] detection. It was first introduced to be able to automatically analyze (from a linguistic perspective, to find trends and traits from) a large number of PNs collected to find correlation between the recollection of emotional PNs and it's health effects, as explained in Section 3.1.

Some other lines of works relevant to the domain of Emotion Lexicons include the works on **Affective Events** and **Sentiment Lexicons** like senti-wordnet. Recently, there has been a growing interest in the identification and analysis of Affective Events (AE)- activities or states that positively or negatively affect people who experience them, from written texts such as narratives and blogs [34, 35] (eg. 'I broke my arm' is a negative experience whereas 'I broke a record' is a positive one.). AffectEventKB is a knowledge base of events extracted from personal stories which were identified from web blogs. Each event is represented using a frame-like tuple which consists 4 fields: Agent, Predicate, Theme, PrepPhrase(PP), and associated with universally accepted affective properties such as emotion polarity [36]. Whereas SentiWordnet is a lexical resource specifically devised for supporting sentiment classification and opinion mining applications [11]. In SentiWordnet, synsets from wordents are enriched by assigning sentiment polarity.

As can be seen in the Figure 3.1, all different emotion lexicons we studied, are static, in the sense that the emotion analysis of words or tuples is performed without considering the surrounding context, which may completely change the meaning of the word or the tuple. For example, the word 'mad' would be considered as positive in the sentence 'He is mad for her' whereas it would be considered negative in a situation like 'The king is mad, he makes rash decisions'. Although we get probabilities for different emotion categories, the probabilities are based on general distribution, and not for our particular context. Whereas with Emotion Carriers, we capture entities and events presented in text spans that explain the current emotional state of the narrator. These spans may not even have any affective properties in the emotion lexicons (neutral), but in our particular context, they carry emotions felt by the narrator.

#### Emotion Cause Extraction

Fine-grained analysis of emotions is still in it's early stages. One such task is of Emotion Cause Extraction (ECE), which primarily aims at identifying an emotion span that represents an emotion, and a single or multiple cause spans from the text that might have triggered or caused the emotion [155, 71], as in this example from [57]: "*< cause > Talking about his honours, < /cause > Mr. Zhu is so < emotion > proud < /emotion >.*" The emotion and cause spans were commonly restricted to clauses [37, 44, 82] until recently a few works explored text spans of any possible length [83, 51]. The ECE task has mainly focused textual genres such as news [147, 58] and microblogs [49] but never focused on spoken or textual PNs. As explained in the Section 3.1.3, PNs have complex structure, consisting of multiple sub-events along with the associated attributes. In such a complex sequence of sub-events, it is difficult to associate the emotion clause with the corresponding event. It was one of the main shortcomings in the work by [57], they call it the problem of cascading events. Moreover, the emotion identified in the ECE task is usually of the characters involved in the text (when the actual event took place), whereas in our scenario the emotions are of the narrator while recollecting the event. Also, except for a very few recent works that identify emotion categories instead of emotion spans [147], ECE, in general requires the emotion to be explicitly expressed in a text span, but in PNs, the narrator's emotions are often implicit and may not have a direct indication in the text.

# Chapter 4

# Personal Narratives Datasets

In Chapter 3 we defined Personal Narratives (PN), their qualitative properties, looked at some examples, and studied how emotion analysis of PNs is a valuable task. In this chapter, we will discuss different datasets consisting of PNs that we use in our work as well as some other datasets that can be explored for the emotion analysis. We will study the datasets from different perspectives such as the target application, the selection of participants, the methods used for the elicitation of PNs, and the annotations provided with the dataset. The datasets and their characteristics are summarized in Table 4.1. The first three datasets USoM, USoM-Elderly, and COADAPT will be part of the analysis and computational models we will investigate in this thesis.

## 4.1 Explored Datasets

In this section, we describe the datasets that we use in our experiments and studies.

| **Dataset** | **Lang.** | **Participants** | **Elicitation** | **Inputs/ modalities** | **Annotation** |
|---|---|---|---|---|---|
| USoM | German | university students (18-36 yrs) | positive and negative situation questions | Audio | self-assessed valence and arousal |
| USoM-Elderly | Getrman | Elderly (60-95 yrs.) | positive and negative situation questions | Audio, Video, physiological signals | self: valence and arousal,<br>experts: continuous live valence and arousal |
| PHA-CBT | Italian | Psychotherapy clients (33-61 yrs.) | ABC questionnaire from CBT protocol | Text | self: emotion category |
| MEMOA | Dutch | Elderly (65-85 yrs) | AMR, LSB, IAPS | Audio, Video, physiological signals | expert: fragmentation<br>expert: Valence of Emotional Memory signals |
| SEND | English | university students | positive and negative events | Audio, Video, physiological signals | self: continuous offline valence,<br>AMT: continuous offline valence |
| User Diaries | Italian | Hypertensive and normotensive | user initiative | Text | self: stress level,<br>experts: stress level |

Table 4.1: Different datasets of Personal Narratives collected by different research groups, relevant for the emotion analysis studies. The first three datasets are further explored in this thesis for experiments related to annotation and building automated tools. The columns from left to right represent, the name of the dataset, the language used in the dataset, the important characteristics of the narrators, the elicitation method used for eliciting the narratives from the narrators, the input modalities (provided with the dataset), and the annotation (self: self-assessed annotation; experts: annotation by experts; AMT: annotation by turkers from amazon mechanical turk) provided with the dataset.

### 4.1.1 The Ulm State of Mind corpus (USoM)

Ulm State-of-Mind in Speech (USoMs) is a database of spoken PNs in German, along with the self-assessed valence and arousal scores, collected by the department of Clinical Psychology and Psychotherapy, University of Ulm [125]. A part of the dataset was used and released in the Self-Assessed Affect Sub-challenge, a part of the Interspeech 2018 Computational Paralinguistics Challenge (ComParE) [132]. The task was to predict the narrator's valence score provided a short speech fragment (8 seconds) of the narrative.

The data consists of 100 speakers (students) (85 f, 15 m, age 18-36 years, mean 22.3 years, std. dev. 3.6 years). The students recollected two negative and two positive PNs, each with a duration of about 5 minutes. As summarized in Fig 4.1, before and after recording each narrative, the participants self-assessed valence (spanning from negative to positive) and

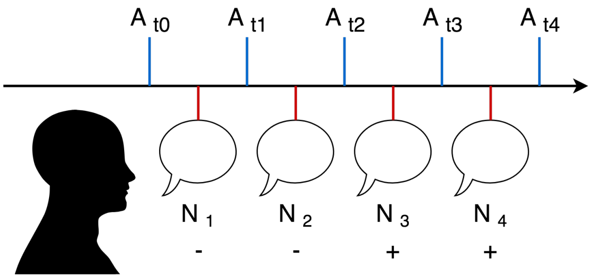


Figure 4.1: Data collection process in the Ulm State of Mind (USoM) corpus: participants were asked to self-report their affect ($A_{t0}$), then recount a negative narrative ($N_1$,$-$), report their affect ($A_{t1}$), recount another negative narrative ($N_2$,$-$), report their affect ($A_{t2}$), recount a positive narrative ($N_3$,+), report their affect ($A_{t3}$), recount a positive narrative ($N_4$,+) and report their affect one final time ($A_{t4}$).

arousal (spanning from sleepy to excited) using the affect grid [128] on a 10-point Likert scale. The narratives were transcribed manually. The number of tokens in PNs vary from 292 to 1536 (mean: 820; std: 208).

Following prompts were used to elicit the narratives 1) Negative narrative: *"Please remember a time in your life when you were facing a seemingly unsolvable problem and report as detailed as possible over the next five minutes"*. 2) Positive narrative: *"Please report of a time in your life were you found a solution, where you felt powerful, happy, and content. Describe that story in-depth over the next five minutes"*.

### 4.1.2 USoM Elderly Dataset

USoM Elderly Dataset consists of German spoken PNs. The "Ulm State of Mind Elderly" cross-sectional study (Dec 2018 through Apr 2019) was conducted by the department of Clinical Psychology and Psychotherapy, University of Ulm. Analogous to USoM [125], they collected German spoken PNs with the purpose of building emotion detection systems, however this time by elderly persons.

Each participant was asked to recollect four experiences from his or her

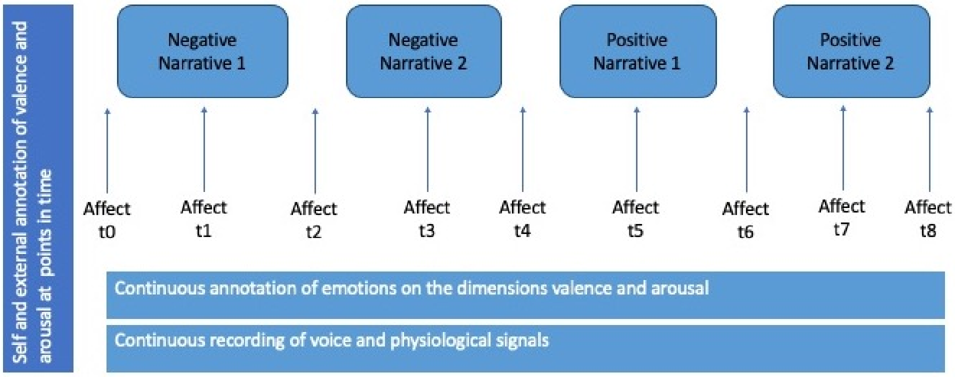

Self and external annotation of valence and arousal at points in time
Negative Narrative 1
Negative Narrative 2
Positive Narrative 1
Positive Narrative 2
Affect t0
Affect t1
Affect t2
Affect t3
Affect t4
Affect t5
Affect t6
Affect t7
Affect t8
Continuous annotation of emotions on the dimensions valence and arousal
Continuous recording of voice and physiological signals

Figure 4.2: USoM elderly data collection process: The self assessed affect values were collected before, after and in the middle of each narrative (eight times), whereas the continuous annotation was performed throughout, using joysticks.

life and was instructed to talk about each of these situations for three minutes, which was captured on audio and video. In the first two experiences, the participants were instructed to talk about a problematic situation (negative narratives, problem situation) and in the other two stories, the participants were asked to remember a situation that included a solution of a problem (positive narratives, solution situation). The physiological activities including parameters such as skin conductance, heart rate, respiratory rate, and blood pressure of the participants were also measured using biosensors.

As described in Figure 4.2, along with the beginning and the end of each narrative, the participants were also interrupted in the middle of the narrative and were asked questions to collect self-assessed valence and arousal based on Russell's core affect [128]. Additionally, an external assessment was conducted. Two independent and trained raters (psychologists with Bachelor's degree) evaluated the participants' perceived valence and perceived arousal during the narration by indicating a position on the Affect Grid via joystick, which was continuously tracked. The values of the valence and the arousal were in the range of [-1000, 1000] and were captured once every 0.5 seconds. We refer to this annotation as "continuous annotation". Moreover, SF-12 and PHQ-8 self-report questionnaires were collected from the participants.

The data includes 88 German-speaking participants (352 PNs), of whom 32 are men (36.4 %) and 56 women (63.6 %), with the age ranging from 60 to 95 years. The majority of the participants lived in small towns or villages. The PNs collected are highly influenced by the regional dialects used by the participants. This poses a major problem for processing data using standard Speech and NLP tools, which are usually trained on non-accented or standard German language. This dataset is not yet publicly available.

Some studies have shown that older adults differ from younger adults in their emotional experiences, regulation and expressions. Older adults experience more complex emotions, express an increased positive affect or have a greater emotional control [75]. With manual analysis we also find the PNs from USoM-Elderly dataset to be more complex and varied as compared to USoM-Young.

### 4.1.3 CBT-PHA dataset

This dataset was collected from the participants aged 33-61 years with mild to moderate levels of stress, anxiety, or depression and were receiving Cognitive Behavioral Therapy (CBT) based psychotherapy through a Personal Healthcare Agent (PHA) mobile application [31, 92]. As a part of the CBT protocol, with ABC (Antecedent, Belief, Consequences) technique, psychotherapists try to identify the event that has caused the patient a certain emotion using a questionnaire to define A) what, when and where the event happened, B) the patient's thoughts and beliefs about the event and C) the emotion the patient has experienced regarding the event. Psychotherapists then identify irrational thoughts of the patient and guide the

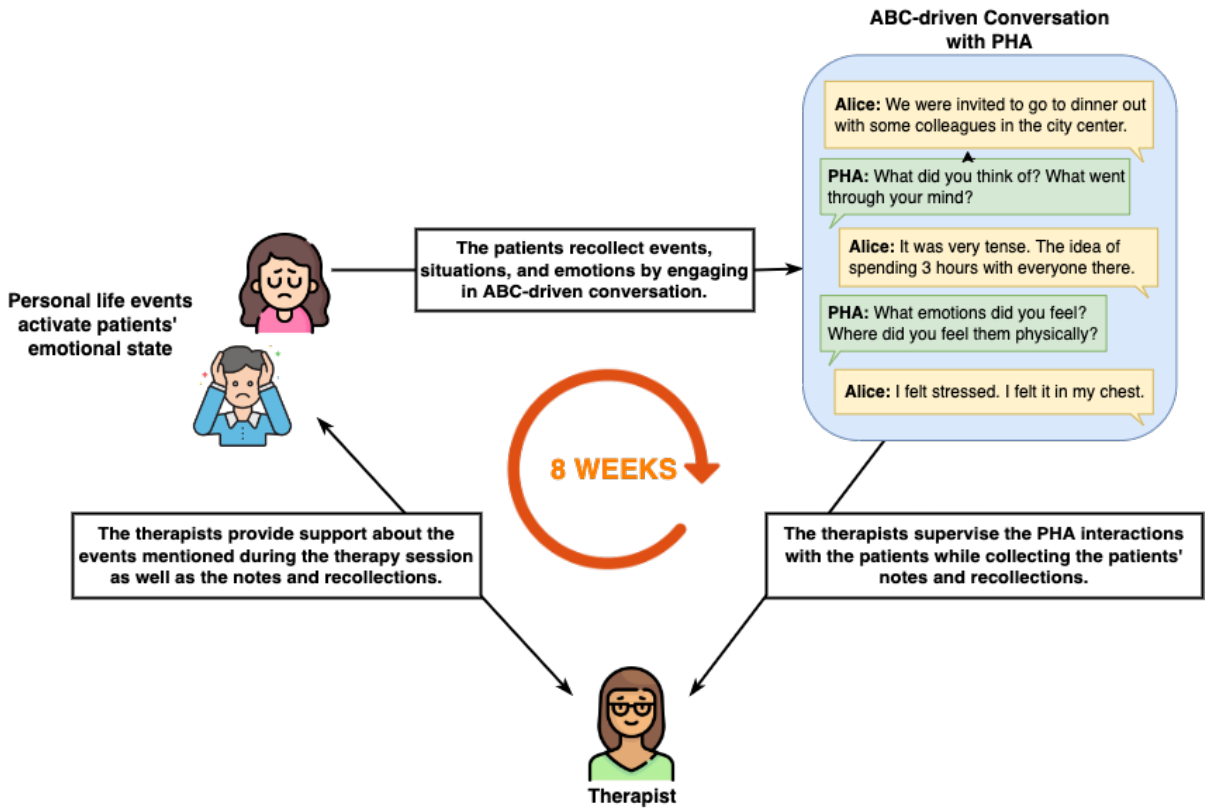

ABC-driven Conversation with PHA
Alice: We were invited to go to dinner out with some colleagues in the city center.
PHA: What did you think of? What went through your mind?
Alice: It was very tense. The idea of spending 3 hours with everyone there.
PHA: What emotions did you feel? Where did you feel them physically?
Alice: I felt stressed. I felt it in my chest.
The patients recollect events, situations, and emotions by engaging in ABC-driven conversation.
Personal life events activate patients' emotional state
8 WEEKS
The therapists provide support about the events mentioned during the therapy session as well as the notes and recollections.
The therapists supervise the PHA interactions with the patients while collecting the patients' notes and recollections.
Therapist

Figure 4.3: CBT-PHA dataset collection process: The participants interact with the mobile personal health care agent (PHA) to share personal recollections of their life events. A one-on-one psychotherapy session is provided once a week for eight weeks. The therapists elaborate on the patients' personal narratives during the therapy session. ABC: antecedents, beliefs, and consequences; PHA: mobile personal health care agent.

patient on how to change them or find more rational and/or functional thoughts [131].

20 Italian speaking patients were provided with the mobile application and a weekly psychotherapy session with a therapist, for a period of 8 weeks for each patient. The data collection was spanned over 3 months in total. The users were asked to write about the daily events that activated their emotional state. For writing the note in the ABC format, the application asked the relevant questions to the users. The users were also asked to select the emotions they felt, from a predefined set, including the six basic

emotions used in psychological experiments (Happiness, Anger, Sadness, Fear, Disgust and Surprise) [40], and two other complex emotional states (Embarrassment and Shame). As the A, B, and C notes are about the same events, they are later combined together to consider them as complete PNs. While providing support to the participants, the psychotherapists also make use of the ABC notes collected from the participant using the PHA application. At the end of the experiment, 224 ABC notes were collected from 20 users giving 92 complete PNs (if any of A, B or C note is missing, it is not considered a complete PN). 18 out of 92 events were provided with an emotion label. The data collection process is summarized in the Figure 4.3, borrowed from [31].

Later, the experiment was further extended for three months with another set of users. Combining the two experiments of 3 months, 481 personal narratives written by 45 Italian speaker users were collected, with the average length of 51 tokens per narrative and overall dictionary size of 5875 tokens. The dataset was collected as a part of the COADAPT (a European Union's Horizon 2020 project). This dataset is not yet publicly available.

This is a unique dataset since it is collected within a longitudinal study (the data for each user was collected for over 8 weeks). With longitudinal emotion analysis we might find interesting trends in the emotional states of the patient. Another possibility is to analyze how the patient's relation with a character or any entity from the PNs, changes over time.

## 4.2 Other Datasets

While we investigate three datasets of PNs for our experiments, there are a few other datasets that include collections of PNs in different domains, collected for different applications or studies related to emotion analysis.

We summarize four such datasets that we think are relevant for fine-grained emotion analysis.

1. **MEMOA dataset:** The Multi-Modal Emotional Memories of Older Adults database [94] consists of positive and negative memories of older adults elicited through two emotion relieving tasks: 1) autobiographical memory recall in the first session and 2) life story books to discuss these memories in depth in the second session. The data is in Dutch. The first sessions were audio taped whereas audio, video, and physiological data were recorded for the second sessions. The authors present an Valence of Emotional Memory (VEM) annotation scheme for emotional memories and the challenges involved in the annotation. Different emotion reliving tasks were used to elicit emotional memories. The session-1 consisted of autobiographical memory recall - a word association task [153], in which two practice words (*grass* and *bread*) and then two emotional words (*sad* and *happy* were presented to which three specific emotional memories had to be recalled by the participant. A photograph or document for each emotional memory were also collected, to be used in session-2. For the session-2, the Life Story Book (LSB)[151, 154, 42] and International Affective Picture System (IAPS)[81] were used for the purpose of elicitation. In LSB, a personalized digital story book containing the pictures and verbal prompts from the first session was created and presented to the participant to elicit positive and negative emotions and their expressions about the memories in greater detail. With IAPS, Six standardized pictures (three sad and three happy) were presented to elicit emotions. The database includes 23 participants (65 to 85 years old) producing 11 hours of audio/video in first sessions and 27 hours in the second sessions. In second sessions, self-reported valence and arousal was collected and physiological signals were captured. This data was later

fragmented and annotated with the VEM annotation scheme and built automatic models for identification of valence using acoustic and lexical features [95].

2. **SEND dataset:** The Stanford Emotional Narratives Dataset [102] is a set of rich, multimodal videos of self-paced, unscripted emotional narratives, annotated for emotional valence over time. The language used in the narratives is English. SENDv1 is available at `https://github.com/StanfordSocialNeuroscienceLab/SEND`. Participants were asked to recollect 3 positive and 3 negatives events from their lives, in a self-paced manner: the participants were alone in the room, and were allowed to talk for as long as they wanted about each event. During the recollection, audio, video and physiological signals were collected, although the physiological signals are not published with the data. Later the participants were presented with their recorded sessions and were asked to annotate the valence they felt while recollecting the event, for the entire duration of the clip, on a continuous scale from -1 (very negative) to +1 (very positive), by sliding a slide-bar. The same clips were also annotated by on avg 20.5 turkers from Amazon Mechanical Turk. The turkers were also presented the same video clips and were asked to annotate the valence that they think the participant might have felt during the recollection. After data filtering, the dataset includes 193 clips from 49 participants (mean age: 23.7 std:7.9). Manual transcriptions of the clips using a professional service is also provided.

3. **User Diaries:** With the aim of building automatic detection of stress levels from the user diaries, [52] collected digital diaries from 10 hypertensive and 10 normotensive adults for 10 days each. Participants were instructed to maintain a periodic electronic diary to capture daily

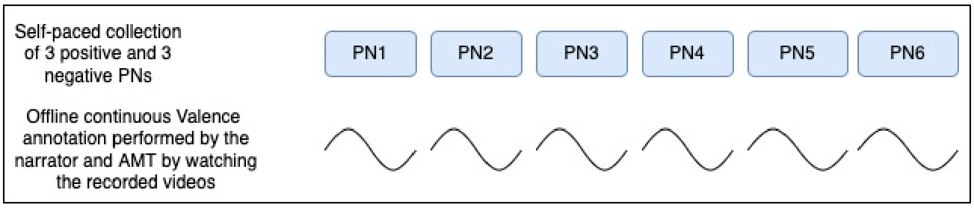

Self-paced collection of 3 positive and 3 negative PNs
PN1
PN2
PN3
PN4
PN5
PN6
Offline continuous Valence annotation performed by the narrator and AMT by watching the recorded videos

Figure 4.4: SEND dataset collection process: First, 3 positive and 3 negative Personal Narratives were recollected by the narrators. The speech, video and physiological signals were recorded using a microphone, camera and bio-sensors. Later, the continuous annotation of the narratives was performed to identify the narrator's perceived valence while recollecting the narratives by the narrators and AMT-turkers, based on the recorded videos.

events, interactions, mood and reflections either as free text, or taking vocal notes which were later transcribed by the system. The participants were also instructed to record their perceived stress levels at regular intervals – twice a day. The data consists of 245 self-annotations of reported stress from the participants. Of these, 154 sessions were annotated as low stress and 91 sessions were annotated as high stress. Physiological signals of the users were also collected during the 10 days of experiments, with an Empatica E3 wearable wristband. The collection also involves 465 text-based diary entries which can be considered as Personal Narratives, with 263 diary entries taken during the sessions annotated as low stress and 202 diary entries taken during sessions marked as high stress. The diary entries were also manually annotated by three psychologists, for the stress levels.

4. Some other datasets include "Counseling and Psychotherapy Transcripts" published by Alexander Street Press. It is a dataset of 4000 therapy session transcriptions on various topics, used as a resource for therapists in-training [1]. The dataset doesn't come with any annotations. [113] collected a dataset of 277 Motivational Interviewing (MI)

[1] https://alexanderstreet.com/

session videos and obtained the transcriptions for each session either directly from the data source, or by recruiting AMT workers.

# Chapter 5

# Annotation Experiments

In this chapter we perform human annotation of different personal narratives datasets discussed in Chapter 4 with emotional states of the narrators felt during recollection of the narrative and the emotion carriers that manifest the emotional states. Multiple annotation protocols are proposed based on the input modalities and the performance of the off-the-shelf NLP tools on the data to cater to differences in the datasets. Written PNs tend to be well structured and contain grammatically well formed sentences as compared to the spoken PNs, thus affecting the performance of the NLP tools. In this chapter we, propose three annotation protocols. First protocol is designed for the annotation of emotion carriers from the textual transcriptions of spoken PNs, and applied to the USoM dataset. We find the main shortcoming of first protocol to be the high complexity and less subjectivity of the annotation task. In the next two protocols we improvise on the first protocol and try to resolve the issues involved. In these protocols the PNs were first divided into smaller meaningful segments, and later annotated with emotional valence and emotion carriers. The second annotation scheme was designed for spoken PNs such as USoM-elderly corpus, to consider speech context while performing the annotation, as an enhancement over text-based annotation of the USoM corpus. Whereas, the third annotation scheme was designed for the textual PNs and the experiments were performed on the PHA-CBT (Italian) corpus. We study each protocol in details in the next sections. The annotations are later used to build automated tools for identification of valence and emotion carriers from PNs.

## 5.1 Protocol 1: Narrative Level EC-Span Annotation of Textual Transcriptions of spoken PNs

### 5.1.1 Motivation

As explained in Section 4.1.1, the USoM corpus includes the narrators' self-assessed ratings of valence and arousal before and after recollecting each PN. In one of our study, we worked on the automatic prediction of the self assessed valence of the narrator after the recollection each PN using the textual transcriptions, while also taking into consideration, the other narratives recounted by the same narrator (as additional context) [141]. While performing the analysis, we observed that in different machine learning models (Support Vector Machine, Attention-based neural sequence tagger) trained to predict valence from transcriptions of PNs, concepts beyond sentiment words (sad, happy) were found to be useful (with the analysis of attention weights in the attention based architecture whereas top tf-idf ngrams, in the case of SVM). These concepts included terms such as characters (e.g. grandfather, a friend), locations (e.g. swimming pool) and events (e.g. high school exam). As the task of the models was to predict the emotional state of the narrator, we observed that these concepts played the role of explaining and carrying the emotional state of the person. We identified these concepts as Emotion Carriers. We performed deep analysis and studied them from psychology perspective and introduced the concept of Emotion Carriers in the Chapter 3.

### 5.1.2 Context

Inspired by such evidence, in this first experiment/protocol, we investigate the possibility of annotating emotion carriers in PNs from the USoM corpus. (This work has been published in the LREC-2020 conference [142].) In this experiment, even though the speech is available with the corpus, as the audio may not be available in other corpora, for better generalizability we perform annotation on the textual transcriptions. In this annotation scheme, even though it is more relevant to us, we do not provide the annotators with a pre-selected list of noun or verb phrases to select from. We give them the freedom to select text segments they feel are most important for our task. We believe that the pre-selection of spans could build bias in annotators towards specific fragments while there could be other text-fragments which are more important emotion carriers. This being an exploratory experiment, we would like annotators to annotate without any bias. Also, being spoken narratives, the automated tools to extract the noun and verb phrases may produce errors, thus affecting the annotation quality. Each narrative is annotated by four annotators. All the annotators are native German speakers and hold a Bachelor's degree in Psychology. They have been specifically trained to perform the task.

### 5.1.3 Guidelines and Protocol

One entire narrative at a time is presented to the annotator, along with the emotion polarity for which it was elicited (Positive or Negative). The annotation task involves the selection of the emotion carrying text spans as perceived by the annotator. We provide annotators with the guidelines to follow while performing the task.

We ask them to select sequences of adjacent words (one or more) in the text that best explain why the narrative is positive or negative for the narrator. We are particularly interested in words that play an important role in the story, such as:

- People (e.g.'mother', 'uncle John', 'my best friend'); Locations ('university', 'our old school'); Objects (e.g. 'guitar', 'my first computer'); Events ('exam', 'swimming class', 'prom night')
- A clause that can include a verb and nouns (e.g. 'Mary broke my heart', 'I lost my guitar', 'I failed the admission exam')

They have to select a minimum of three such text spans for each narrative. We also provide them with the best practices to be followed:

- We ask them to annotate the contentful words ('university', 'mother') preferably over pronouns ('she', 'her', 'it')
- If the same term is present multiple times, they are asked to annotate the first instance of the same concept and to avoid repetition.
- To make sure if something needs to be added or removed from the list of selected fragments, the annotators are asked to make sure:
  - If a person who has not read the narrative can understand why the event was positive or negative just by looking at the list of spans they have selected. If not, they have to check if something is missing.
  - They are asked to ensure that there are no repetitions in the list and that there are no spans, which are not central to the narrative.
- As the annotators already know if the narrative is positive or negative, we ask them to annotate the feelings (emotion words) only if

they are more informative (e.g. ‘feeling of freedom’) than simple positive/negative (e.g. ‘I was happy’).

**Tool:** We provide the annotators with a web-based tool to perform the annotations. The tool is mainly divided into two parts. In one part, we show them a personal narrative and the corresponding sentiment. The annotator can hover over the tokens and select text spans by clicking and dragging over the consecutive tokens. On the right-hand side, they can see the already selected spans and their rankings. They can change the ranking by simple drag and drop.

### 5.1.4 Analysis

#### Output

Table 5.1 shows an example of annotations for a part of a narrative. We observe that sometimes, annotators annotate text-segments representing a similar concept but are at different positions in the text. In the example, the terms *Praktikumsplatz* and *Praktikum* represent the same concept of *internship* but two of the annotators followed the guidelines to select the first occurrence while the other annotator selected the second occurrence of the same concept.

#### Statistics:

In this study, we analyze 239 narratives from 66 participants (the development and test sets from the ComParE challenge) that have been annotated by four annotators each. Note that for 66 participants the total number of narratives should be 264, but in the ComParE challenge, 25 files were removed because of issues like noise.

We observe that the number of annotations (text-spans) annotated by the annotators per narrative vary from 3 to 14 with an average of 4.6,

| **German (original)** | **English (translated)** |
|---|---|
| Okay. Ähm also eine Situation, in der ich mich kompetent gefühlt hab, war, als ich meinen ähm Praktikumsplatz bekommen hab und ähm ich in dem Praktikum dann auch ähm Sachen selbstständig machen durfte und auch die Rückmeldung bekommen hab von den Personen dort, dass das, was ich da so mach, dass das gut ist und dass ähm sie mit mir sehr, mit mit mir sehr zufrieden sind ähm. Und die Gefühle dabei waren natürlich irgendwie Glück, weil man ist davor unsicher, ob man das, was man da macht, ob das so gut ist und ob man das so schafft. Ähm und das hat eben sehr gut funktioniert. Also ich hab mich sehr zufrieden gefühlt, mit mir ähm im Reinen, mit mir glücklich, auch irgendwie so ein bisschen Bestätigung darin bekommen, dass das, was ich mach, gut ist oder das, was ich auch jetzt als Studium gemacht hab, irgendwie passt. Ähm ähm so bisschen so positive Aufregung, also man fühlt sich sehr sehr wach, erregt irgendwie , aber in einer positiven Art und Weise. ... | OK. Um, so a situation in which I felt competent, was when I got my um internship position and er in the internship then I was also allowed to do things independently and also got the feedback from the people there, that what I am doing there, that it is good, and that they are very pleased with me, with me, um. And of course the feelings were kind of lucky, because you are not sure if you, what you are doing, if that is so good and if you can do it that way. Um, and that worked very well. So I felt very satisfied, with me uh, I'm happy, with me, somehow getting a bit of confirmation that what I'm doing is good or what I'm doing now as a study have, somehow fits. Uhm umh so a bit so positive excitement, so you feel very very awake, excited somehow, but in a positive way. ... |

Table 5.1: **PN Annotated Example:** A part of a narrative showing text spans annotated by the four annotators. The intensity of the red color in the background represents the number of annotators who annotated the text-span (varying from lightest for 1 annotator to the darkest for four annotators). It can be seen that some spans are annotated by single annotator while some by multiple. In the text-span *"positive Aufregung"*, two annotators selected the entire span while another one selected only the second word *"Aufregung"*. The annotations contain sentiment words as well as content words.

also that all annotators follow the same pattern from this aspect. We also calculated the number of tokens present in the annotations. The numbers show that three of the annotators (*ann1, ann2, ann3*), on average select a span of 1.5 tokens, while the fourth annotator (*ann4*) selects three tokens (avg.) per annotation. Note that, for all the analysis, we use the spaCy toolkit[1] for tokenization. We observe that many annotations contain punctuation marks, which are considered as separate tokens by spaCy. Thus, we perform the same calculations while ignoring the punctuation tokens. We find that the average number of tokens drops down to 1.1 for the first three annotators, while it drops down to 2.3 for the *ann4*.

We also analyzed the distributions of POS tags, and as expected, found that the most common categories include noun (35%), adjective (30%), verb (15%), and adverb (7%).

#### Inter Annotator Agreement

Commonly used metrics for evaluating the agreement between annotators include variations of $\kappa$ coefficient such as Cohen's [28] for two annotators, Fleiss' [46] for multiple annotators. Unfortunately, calculations for $\kappa$ such as observed and chance agreements involve the knowledge of true negatives, which is not well defined for a text span selection task (eg. in this study, it could mean the number of possible text spans that are not annotated). This makes $\kappa$ impractical as a measure of agreement for text spans annotation.

An alternative agreement measure that does not require the knowledge of true negatives for its calculations is Positive (Specific) Agreement [47]($P_{pos}$ Eq. 5.1). It has previously been shown to be useful in the evaluation of crowdsourced annotations tasks, similar to our's [137, 26].

The Equation 5.1 defines the positive agreement in terms of true positives (TP), false positives (FP) and false negatives (FN). We can see that

[1] https://spacy.io/

| | ann1 | ann2 | ann3 | ann4 |
|---|---|---|---|---|
| ann1 | 1 | 0.344 | 0.417 | 0.125 |
| ann2 | | 1 | 0.389 | 0.106 |
| ann3 | | | 1 | 0.137 |
| ann4 | | | | 1 |

(a) Exact match, position agnostic, token level (mean $F_1$: 0.252)

| | ann1 | ann2 | ann3 | ann4 |
|---|---|---|---|---|
| ann1 | 1 | 0.338 | 0.42 | 0.277 |
| ann2 | | 1 | 0.381 | 0.196 |
| ann3 | | | 1 | 0.308 |
| ann4 | | | | 1 |

(b) Partial match with position, token level (mean $F_1$: 0.320)

| | ann1 | ann2 | ann3 | ann4 |
|---|---|---|---|---|
| ann1 | 1 | 0.397 | 0.483 | 0.402 |
| ann2 | | 1 | 0.439 | 0.264 |
| ann3 | | | 1 | 0.404 |
| ann4 | | | | 1 |

(c) Partial match, position agnostic, token level (mean $F_1$: 0.399)

| | ann1 | ann2 | ann3 | ann4 |
|---|---|---|---|---|
| ann1 | 1 | 0.400 | 0.490 | 0.410 |
| ann2 | | 1 | 0.440 | 0.267 |
| ann3 | | | 1 | 0.413 |
| ann4 | | | | 1 |

(d) Partial match; position agnostic; lemma level (mean $F_1$: 0.403)

Table 5.2: Pairwise Inter-Annotator Agreement scores (F1 measure) with respect to the different matching strategies. We vary the matching criteria for annotations based on three aspects 1) checking if the annotations are exactly same (exact match) or calculating the overlap between them (partial match) 2) positions of the annotations in the text is considered whiles matching (with position) or not (position agnostic) and 3) to calculate the overlap, the tokens in the annotations are matched (token level) or the lemmas of the tokens are matched (lemma level).

the knowledge of *true negatives (TN)* is not required for the calculation of the positive agreement. Also, notice that the equation is similar to the widely used $F_1$-measure [65]. In our experiments, we calculate the positive agreement for each pair of annotators.

$$P_{pos} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5.1}$$

Another problem we face in the task of text spans selection is the annotation of overlapping text fragments. Given the freedom on the lengths and positions of the text spans, two annotators might annotate different but overlapping text spans. The overlapping part could be an important part, thus the annotations should not be discarded completely. For instance, in Table 5.1, *'positive Aufregung'* and *'Aufregung'*, both the spans contain the fragment *'Aufregung'*, which is important to be considered. Thus, we report results on exact matches as well as partial matches, following the work by [68]. For the partial match, they calculate "soft" $F_1$-measure by calculating the coverage of the hypothesis spans. The coverage of a span($s$) with respect to another span ($s'$) is calculated as defined in Equation 5.2, with the help of the number of tokens common in the two spans. The operator|.| counts the number of tokens.

$$c(s, s') = \frac{|s \cap s'|}{|s|} \tag{5.2}$$

Next, a span set coverage $C$ is defined for a set of spans $S$ with respect to another set of spans $S'$ using the Equation 5.3.

$$C(S, S') = \sum_{s_i \in S} \sum_{s_j \in S'} c(s_i, s'_j) \tag{5.3}$$

In order to calculate the soft $F_1$-measure, first soft precision and soft recall are calculated according to Eq 5.4 and Eq 5.5 respectively. Here $S_H$ and $S_R$ are hypothesis and reference spans respectively, and |.| operator counts the number of spans.

$$precision(S_R, S_H) = \frac{C(S_R, S_H)}{|S_H|} \tag{5.4}$$

$$recall(S_R, S_H) = \frac{C(S_H, S_R)}{|S_R|} \tag{5.5}$$

Finally the soft $F_1$-measure is calculated using the standard formula 5.6:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.6}$$

As the personal narratives are long, often some terms are repetitive. In our task, the position of an annotation is not quite important compared to the content. We further try to loosen the criteria for matching by not considering the position of the text fragments. For instance, let us say a narrative contains mentions of *'trip'* at multiple places, like *'we went for a trip to India'* and *'the trip was great'*. If two annotators intend to annotate the word *'trip'*, they have multiple positions to choose from. While from the perspective of discourse, it would be interesting to analyze which position seems more appropriate, for our purpose of extraction of emotion carriers it is less important. Following the same intuition, we also try to match tokens having the same lemma.

Table 5.2 shows the evaluation results based on the various strategies of matching described above. The $F_1$-measure is calculated for all pairs of annotators. For each strategy, we also report the mean of pairwise scores. In the four tables from Table (a) to Table (d) we loosen the matching criteria, thus increasing the scores. We show the results starting from the most strict criteria of exact matching in the table (a), then in the table (b), we show results for partial matching, but the positions of the annotations are taken into consideration. The improvements are most significant in the case of *ann4*, as we saw earlier in Section 5.1.4 that *ann4* usually annotates longer fragments than others. This shows that the *ann4* annotates longer spans, but still contains the important part that other annotators annotate. Later in table (C), we remove the constraint of position, which results in improved scores, showing that even if the annotations by different annotators are different they often contain similar terms/carriers. This also shows that the annotators often ignore the instruction from the guidelines of selecting the first occurrence of the same term (Section 5.1.3). In the table (d), we further try to match more things by considering lemmas instead of tokens, which results in an increment.

### Qualitative Analysis

In this section, firstly we discuss the quality of emotion carriers annotations compared to other annotation efforts which used inter-annotator metrics similar to ours. Secondly, we explore whether emotion carriers consists only of sentiment words.

**How reliable are the annotations?** It is difficult to judge the reliability and the quality of the annotations just by looking at the inter-annotator agreement scores, which are not self-explanatory. If we compare our task with other previous tasks that used a similar metric, we can better understand the complexity of the task and judge the reliability of the annotations. [26] worked on the task of semantic annotations of the utterances from conversations. For example, one of the sub-task annotators had to perform was selecting a text span describing a hardware concept. For a particular concept like *printer*, the annotators could select the spans *'with the printer'*, *'the printer'* or just *'printer'*, all of which are correct. The problem they faced for the selection of span is similar to ours, but the complexity and subjectivity are low, as they work on shorter texts and the annotator has more understanding of the concept to be selected. They use the same metric as ours to evaluate the inter-annotator agreement for the span selection. They achieve $F_1$ scores of 0.39 and 0.46 (for two different subsets of data) for the exact match, whereas 0.63 and 0.7 for the partial match (mean of pairwise agreements between three annotators). Whereas our scores for

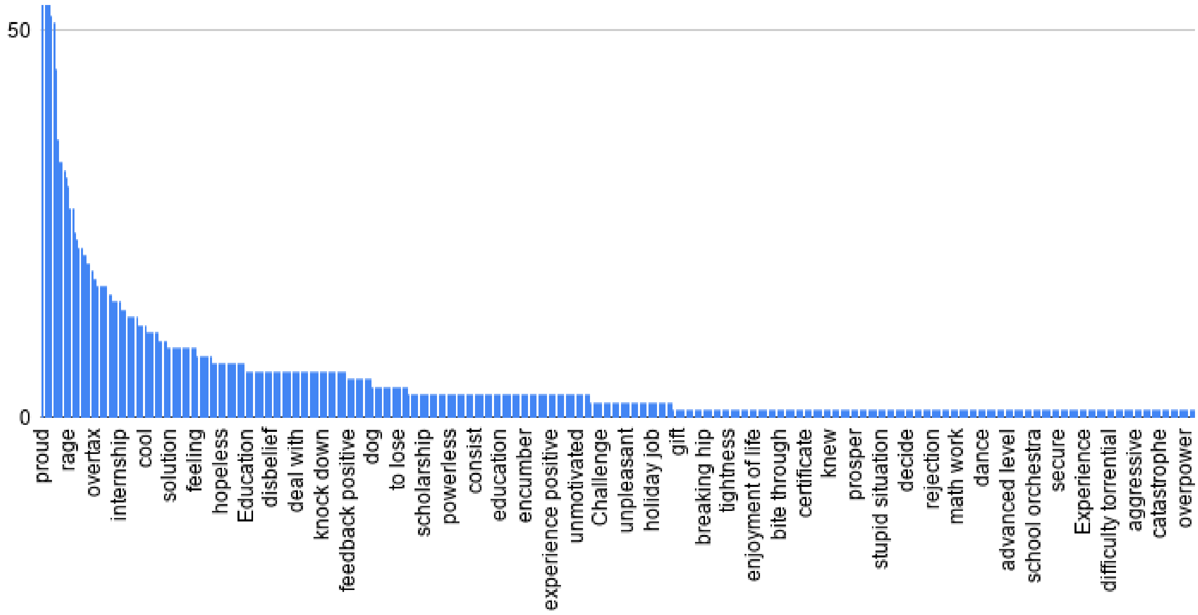


Figure 5.1: Sorted counts of emotion carriers (English translations of annotated German text spans) shared and agreed upon by the annotators. Notice the long tail of singletons. Due to space limitations, we perform binning of the carriers and show only a representative element from each bin.

exact and partial matches are 0.25 and 0.4, which we believe are reasonable given the subjectivity of the task and more number of annotators.

**Are emotion carriers just sentiment words?** As seen in the example from Table 5.1, the annotations include sentiment words as well as content words. In order to further study what is the actual distribution of sentiment words (*angry, joy*) versus content words in emotion carriers, we analyze the annotation of sentiment words across the annotators. For this, we calculate the sentiment polarity of each annotation using the textblob-de library [2], which makes use of the polarity scores of the words from senti-wordnet for German (with simple heuristics), similar to the English senti-wordnet [43]. We find that the trends of using sentiment carrying phrases vary across the

[2] https://textblob-de.readthedocs.io/en/latest

annotators. The fraction of annotations carrying sentiment varies from 24% to 56% (*ann1: 39%; ann2: 24%; ann3: 36%; ann4: 56%*) for the four annotators. For further analysis, we could categorize the annotations into categories inspired by the ones used in the *Psychological Processes* categories of the LIWC.

In addition to the distribution across the annotators, we further analyze if there is a trend in the counts of occurrences of sentiment and content words. In Figure 5.1, we show an interesting observation from the annotations. We plot the histogram of overlaps from partial-matches that we get while evaluating the inter-annotator agreements for all annotator-pairs, with respect to their counts of occurrences. We only show the representative annotations and not all the partial-matches. We can see that the overlaps contain both, emotion words such as *proud, hopeless, disbelief* as well as content words like *scholarship, education, dance.* We notice that there is a long tail of overlaps having only a single occurrence. This shows that a few terms are annotated frequently and agreed upon by annotators, while many terms are unique to specific narratives. As expected, we observe more content words than sentiment words in the tail, while it is surprising to see content words like *internship, solution* appearing in the most frequent words.

## 5.2 Protocol 2: Segment Level Valence and EC-Span Annotation of Spoken PNs

First, we identify issues in the first protocol from Section 5.1, then design a new protocol to resolve the issues, and perform annotation of the USoM-Elderly dataset 4.1.2. Although the annotated ECs captured rich emotional information, based on the low Inter-Annotator Agreement (IAA) and the complex structure of PNs, the annotation task was found to be complex and subjective [143], which may lead to sub-optimal performance of automated tools for the detection of ECs. Later, utilizing this annotation, [144] we worked on the automatic identification of ECs from the textual transcripts, whereas in another work [13], we also tried to improve the system by extracting and using acoustic features for the task (both systems are explained in the next Chapter). However, the different multimodal fusion strategies didn't manage to achieve a significant improvement over the text-based system. This does not follow the trend identified by most of the previous works that have shown acoustic features to help text features in emotion analysis [24].

Although, the reasons for the poor performance of various fusion approaches for combining speech and text are explained in the paper [13], we miss an important aspect regarding the annotation strategy that plays a crucial role in multimodal systems. The annotation of ECs was performed on the text transcriptions of the PNs [143] without having annotators access to the speech of the narrator, thus missing out on the cues from the speech that might help them in the annotation. This does not align with the automated detection task where we try to use the acoustic features for the tokens to identify the ECs that were annotated based only on the text features.

To fill this gap between the human annotation of EC and the EC prediction from spoken PNs, in this protocol, we come up with an improvised speech-based annotation protocol. Here we utilize the information from speech for identifying the ECs as well as valence. Another problem with the previous text-based protocol is that the identified ECs are supposed to explain the sentiment polarity for which the narrative was elicited, which makes an implicit assumption that the valence of the narrator is same as the sentiment polarity throughout the narrative. But the narrator may go through varying emotions during the recollection. In our protocol, we first split the PNs into segments and perform the annotation on these segments, to capture unfolding of the narrator's emotions, as can be seen in the example from Table 5.3 . First, the valence of each segment is annotated and then the ECs that explain the valence are annotated. In this way, the annotation captures the changing emotions of the narrator, within the narrative and the ECs that manifest those emotions.

*"Maria called me yesterday afternoon, she asked me to meet today for lunch. You cannot imagine how happy I was about this, but, at the same time, very nervous, because since our relationship ended we have not spent the lunch break together, as we did when we were together. I was afraid that she wanted to complain and if that were the case, I would not have been able to defend myself. Instead, it was a pleasant lunch, she no longer seems to be angry with me and this reassured me a lot. But I still feel guilty about the way our relationship ended. Yesterday, while we were at the restaurant, I thought that for my stupid betrayal I have lost a highly intelligent, nice girl who I still like. I was wrong and I am ashamed."*

Table 5.3: An example snippet of a spoken personal narrative, anonymized and postprocessed for better readability. The text is color-coded to represent the perceived valence of the narrator while narrating an event (gray - neutral, green - positive, red - negative). It is interesting to see how the emotions expressed by the narrator change while recollecting a series of sub-events.

We propose that providing speech during annotation, breaking the task into two steps: valence and EC annotation, and performing the annotation at the segment level, would help reduce the complexity, subjectivity, and thus the cognitive load of the annotators while also improving the quality of the annotation and capture the unfolding of the emotions.

### 5.2.1 Annotation Steps

In this section, we describe the steps involved and the protocol to annotate valence and ECs from the spoken PNs from USoM-Elderly dataset.

### Transcription

The PNs from the USoM Elderly dataset are transcribed by a professional transcription service. The transcriptions are verbatim and capture fine details such as punctuation, incomplete words, stuttering or repetition of words, pauses, filler words, and dialect. Additionally, they are speaker separated, i.e. a change of speakers is marked. In a subsequent step, accurate time alignment of text to audio is generated by computing forced alignments (FA) using a speaker-adaptive HMM-GMM (Hidden Markov Model, Gaussian Mixture Model) automatic speech recognition system (ASR) based on the one described by [88]. To ensure next to perfect alignments, missing entries in the pronunciation-lexicon, such as incomplete words, dialect, and slang words are generated using a grapheme-to-phoneme tool [15]. For better working of automated NLP tools in the downstream tasks, the transcriptions are preprocessed to remove the incomplete words, filler words, and other metadata such as pauses, speakers, and stuttering.

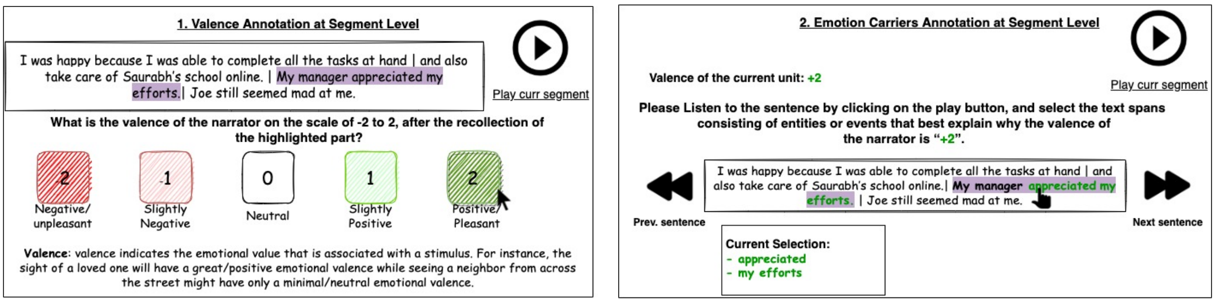
### Segmentation

The transcripts are then segmented into smaller meaningful parts. Ideally, we would like the parts to be functional units, from the speech acts theory [17, 146]. Due to the lack of data annotated with functional units or the presence of any automated tool for such segmentation, we try other levels of segmentation. First, we try sentence segmentation using SpaCy-3[3] (transformers based NLP pipeline for German). We find the resulting segmentation to frequently split at unnatural times, which could be because of the spoken nature of the data. Thus, we instead segment the text using heuristic rules, making use of the cues from punctuation {*.!?*} and

[3] `https://spacy.io/usage/v3`

some typical sequence of tokens such as “und dann” (“and then”) and “aber” (“but”) that indicate natural splitting in spoken language. With the manual analysis, we find that this strategy worked the best for us, and leave prosody-based approaches for future work.

### Annotation



1. Valence Annotation at Segment Level

I was happy because I was able to complete all the tasks at hand | and also take care of Saurabh's school online. | My manager appreciated my efforts.| Joe still seemed mad at me.

What is the valence of the narrator on the scale of -2 to 2, after the recollection of the highlighted part?

**Valence**: valence indicates the emotional value that is associated with a stimulus. For instance, the sight of a loved one will have a great/positive emotional valence while seeing a neighbor from across the street might have only a minimal/neutral emotional valence.

(a) Valence Annotation

2. Emotion Carriers Annotation at Segment Level

Valence of the current unit: +2

Please Listen to the sentence by clicking on the play button, and select the text spans consisting of entities or events that best explain why the valence of the narrator is “+2”.

I was happy because I was able to complete all the tasks at hand | and also take care of Saurabh's school online.| My manager appreciated my efforts. | Joe still seemed mad at me.

Current Selection:
- appreciated
- my efforts

(b) Emotion Carrier annotation

Figure 5.2: Sketches representing the structure of the tools used for the segment level annotation. A segment from the transcript of a PN is highlighted and shown to the user for the annotation, along with local context for better understanding and for disambiguation in cases such as irony. The segments are separated using pipe(|). The audio corresponding to the current segment is played by default, while also providing a provision to listen to any segment by clicking on it. First, the valence annotation task is performed as shown in the Fig a. Then in the EC annotation task, as shown in Fig b, the valence selected by the annotator for the current segment is displayed and asked to select ECs in the form of text-spans, that explain the selected valence.

The annotation of each segment is performed in two phases, valence annotation and EC annotation, using tools as represented in the Figure 5.2.

The **valence annotation** aims to capture the emotional polarity of the narrators in terms of valence while they recollect the events and the intensity of the polarity as expressed on a 5-point bipolar scale from -2 (“unpleasant”) to +2 (“pleasant”) with 0 representing “neutral”, by listening to the audio corresponding to the segment and using the transcript as a support.

The **Emotion Carrier (EC) annotation** task tries to identify ECs - words, chunks, or phrases that contribute to explaining the emotional state of the narrator. For each segment annotated with a non-neutral valence by the annotator, the annotator is asked to identify the text spans (sequences of adjacent words) from the transcript of the segment as ECs that explain why the corresponding segment has a positive or negative valence for the narrator. The focus is particularly on the words that play an important role in the story including entities represented by noun chunks/phrases such as persons, places, objects, and activities and events or actions represented by verb chunks/phrases, rather than emotion laden words such as *happy, sad, and enjoyed.* While annotating the ECs, the annotators are recommended to take into consideration the local context of 2-3 segments, in case disambiguation is needed.

#### Execution

We define an execution strategy to ensure a consistent and high-quality annotation. Four annotators were selected from a pool of graduate students, based on their interests and previous experience with data annotation. The overall annotation task is divided into three phases: training, overlap, and partial-overlap.

The **training phase** started with a training session administered by a psychotherapist, which included explaining the task, the tool, and the annotation guidelines. After each training batch, a consensus meeting is held between all the annotators and the psychotherapist to discuss the differences among the annotators, try to agree on a specific opinion, and modify the guidelines if necessary. We continue the small training batches until we achieve a satisfactory inter-annotator agreement as measured using the evaluation metric explained in Section 5.2.4. We achieved stable agreement after three training batches; the corresponding data is excluded from the analysis that we perform on the collected data.

In the **overlap phase**, all the annotators are given the same data to annotate to ensure that the inter-annotator agreement remains high. After two batches, we concluded that each annotator can now perform annotations separately, without compromising the quality of the annotation.

In the **partial-overlap phase**, we provide different sets of narratives to the annotators, while keeping an overlap of 15% in all the sets. In the end, we get 20% of overlap, i.e. 20% of the data is annotated by all the annotators while 80% of the data is annotated by a single annotator. To ensure the quality of annotation in this last phase, we divide the data into batches and monitor the inter-annotator agreement on the overlapping data.

### 5.2.2 Output

In Table 5.4, we present a PN annotated with valence and ECs using this protocol. The segments are color coded top represent the valence (red: negative **(-2)**; orange: slightly negative **(-1)**; gray: neutral **(0)**; light green: slightly positive**(+1)**; green: positive **(+2)**), whereas the emotion carriers are wrapped in the parentheses. More examples can be found in Appendix A in Table A.2 and Table A.3.

### 5.2.3 Comparison

We compare the speech-based annotation protocol to the text-based protocol from Section 5.1. For ease, we refer to them as *SpP* and *TxtP* respectively. The main difference between the two is the modality used while performing the annotation. Even though audio recordings and the manual transcripts are available, the TxtP performs the annotation on the textual

| | |
|---|---|
| Ich war hier bei den **(Basketballern)**, da waren wir eine **(Clique)** von sieben, acht Basketballern, die auch zusammen Basketball gespielt haben und sich dann irgendwann auch mehr oder weniger um das Management gekümmert haben. ... Und das ging dann so weiter, dass wir dann tatsächlich **(deutscher Meister wurden)** mit den mit den Mädels. Zweimal sogar, dreimal **(deutscher Pokalsieger)**, **(Europapokal gespielt)** haben. Und dann kam halt die Situation, wo es **(finanziell)** ein bisschen eine **(Schräglage)** gab. Und da haben sich dann leider **(zwei Grüppchen gebildet)**. Bei diesen sieben, acht Menschen, die halt früher immer sehr freundschaftlich, eher sogar **(wie Brüder zusammengearbeitet)** haben, **(kam)** es dann tatsächlich **(zum Auseinanderdriften)**. ... Natürlich, wenn man sich gesehen hat, hat man mal hallo gesagt. Aber **(früher)** hatte man sich ja jeden Tag gesehen oder hat, wie das unter Freunden ist, **(viele Sachen zusammen gemacht)**, **(viel zusammen erlebt)**. Und es ist **(total auseinandergegangen)**, **(total auseinander)**. Also zu zwei, drei von **(diesen Menschen)** habe ich leider heutzutage überhaupt **(keine Beziehung mehr)**. ... Und was das Deprimierende ist, dass man vorher mit denen **(alles zusammen gemacht)** hat. Das waren **(Best Friends)**, wie man so schön sagt. ... Und **(vorbei)** ist es. | I was here with the **(basketball players)**, we were a **(clique)** of seven, eight basketball players who also played basketball together and then at some point also more or less took care of the management. ... And that continued in such a way that we then actually **(became German champions)** with the girls. Twice even, three times **(German Cup winner)**, **(played in the European Cup)**. And then the situation arose where there was a bit of a **(financially skew)**. And then, unfortunately, **(two groups formed)**. These seven or eight people, who used to **(work together)** very amicably, more **(like brothers)**, actually **(drifted apart)**. ... Of course, when we saw each other, we said hello. But **(in the past)**, you saw each other every day or, as is the case between friends, **(did a lot of things together)**, **(experienced a lot together)**. And it **(totally fell apart)**, **(totally fell apart)**. So, unfortunately, I **(no longer have any relationship)** at all with two or three of **(these people)**. ... And what's depressing is that you **(did everything together)** with them before. They were **(best friends)**, as they say. ... And it's **(over)**. |

Table 5.4: A negative PN, begins with a positive valence and later shifts to negative valence, and ends in a negative valence.

transcripts, which cannot take into account the valuable information from the speech that could prove to be essential for better understanding of the content and thus ECs. This also causes a mismatch in the EC annotation performed considering only text and the automatic EC detection system which considers both, text and acoustic features [13]. Whereas, in the SpP, the annotators listen to the actual speech of the narrators and perform EC and valence annotation on the corresponding transcripts. This helps annotators better identify ECs using cues from the speech as well as the content.

Another major difference being the context considered in the annotation. In TxtP, the annotation is performed at the narrative level, meaning, the annotator has to read the entire narrative, understand the context from the complex structure and find the ECs from the text that explain the narrator's overall emotion. In the SpP, the annotation is performed at the segment level. The annotator goes through one segment at a time, listens to the segment, annotates the narrator's valence and the corresponding ECs, while also using transcript and local context for reference. The context at the segment level is easy to understand as the complex structure of the PNs is broken into simple segments. The annotation at segment level captures the unfolding of emotions of the narrator, throughout the recollection, which is not possible with the narrative level analysis. In the narrative level analysis, the ECs explain the original sentiment for which the narrative was elicited (positive narrative or negative narrative), whereas in the segment level analysis, the first task is the valence annotation whose polarity could be different from the polarity of the sentiment for which the story was elicited, and the second task explains this valence.

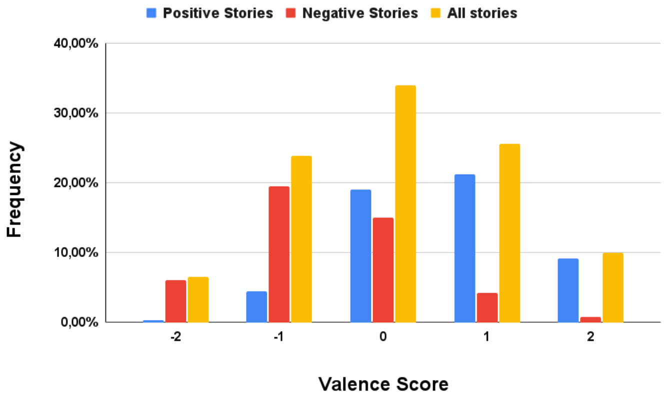


Figure 5.3: Valence score distribution. The blue and red bars represent the distribution of valence scores respectively computed on positive and negative PNs. The yellow bar represents the distribution of the whole dataset.

### 5.2.4 Analysis

#### Statistics

260 PNs from 65 narrators were manually transcribed and used in the annotation experiment. 48 narratives were used for the training phase explained in the Section 5.2.1, which are discarded from the analysis. In total, we analyze 212 PNs collected from 53 narrators. The PNs consist of on avg 370 tokens, 30 segments, and last for about 165 seconds. Whereas each segment contain ∼12 tokens. The entire data contains ∼7000 segments. 20% of the data (42 PNs; ∼1200 segments) was annotated by all annotators as explained in Section 5.2.1, which is further analyzed for calculating the inter-annotator agreement statistics.

#### Label Distribution

Figure 5.3 depicts the valence label distribution of the dataset. On the overlapping examples, i.e. examples annotated by more than one person, we compute the arithmetic mean and round to the nearest integer. Looking at *all stories* series, we observe that the overall distribution appears

Gaussian. The distribution over positive, negative and neutral labels is close to uniform (33% neutral, 33% negative and 34% positive). In Figure 5.3, we report the label distribution computed on *positive stories* and *negative stories*. We observe that the predominant classes are positive and neutral for *positive stories*, and negative and neutral for *negative stories*. This shows that our experiment is accurate and that some stories are not fully positive or negative, but there are parts with opposite polarities.

We will see in subsection 5.2.4(Valence Annotation), how the neutral class is a major source of disagreement among annotators. We provide special attention to the distribution of neutral label. The analysis of the label distribution of narratives with overlapping annotators shows that for 77% of segments at least one annotator selects the neutral class. Moreover, inspecting the cases where annotators disagree, we find that for 70% of the examples include at least one neutral label. This brings important evidence to the fact that neutral class is a difficult to agree on (as will be discussed in the next section).

For the non-neutral segments, we perform EC annotation. We find different trends in the EC annotation by different annotators. The annotators are assigned IDs *ann1,ann2,ann3, and ann4*. The number of ECs annotated per segment vary from as low as 0.2 by *ann1*, 0.67 and 0.7 by *ann2* and *ann3*, while as high as 1.25 by *ann4*, resulting in 4.9, 12.4, 11.1 and 30.5 ECs per PN. Whereas, the number of tokens per EC are 4.95, 2, 1.5, and 2.5 resulting in 24.3, 24.7, 16.7, and 76 EC-tokens per PN. We observe that, *ann1* tends to annotate less ECs but long in length whereas on the other extreme, *ann4* tends to annotate high number of ECs which are also longer in length as compared to *ann2 and ann3*. The inter-annotator agreement metrics used in the next subsection 5.2.4(Emotion Carriers Annotation) caters to these variations by providing flexibility in matching two ECs.

| Labels | seg | cont | seg + cont |
|---|---|---|---|
| -2,-1,0,+1,+2 | 0.29 | 0.16 | 0.26 |
| -2,-1,+1,+2 | 0.41 | 0.78 | 0.38 |
| neg, neu, pos | 0.48 | 0.29 | 0.4 |
| neg, pos | 0.99 | 0.9 | 0.95 |

Table 5.5: Inter-Annotator agreement using Fleiss' $\kappa$. *seg* and *cont* refer to the segment based and continuous annotation. Positive (pos), neutral (neu) and negative (neg) classes are obtained by grouping positive (+1,+2), neutral (0) and negative (-2,-1) valence values. Removing neutral examples, we observe close to perfect agreement.

### Inter Annotator Agreement - Valence Annotation

To assess the quality of the segment-level valence annotation, we compute different inter-annotator agreement (IAA) statistics. We compute IAA in the training, overlap phases of the annotation, and also during the partial-overlap phase on the overlapping part of the batches. We also compare these results with the IAA of the continuous annotation from the USoM Elderly data set (Section 4.1.2), and also with the IAA by combining both the segment-based and continuous annotations. The IAA statistics are shown in Table 5.5. Note that the segment-based annotation involves 4 annotators and the continuous annotation involves 2 annotators, whereas the "segment + continuous" involves 6 annotators.

Since the segment-based annotation is performed by four annotators, we use Fleiss' $\kappa$ to compute the inter-annotator agreement [46]. For the five labels (-2, -1, 0, +1, +2), we observe $\kappa = 0.29$, indicating a fair agreement according to the interpretation table reported in [80].

To inspect the sources of the disagreement, we compute the agreement among only the positive (labels +1 or +2) and negative (labels -1 or -2) examples by removing all the examples in which at least one annotator picked the class neutral (class 0)(remaining data ~34%). The $\kappa$ score increased from 0.29 to 0.41, suggesting that the neutral class is one major source of disagreement and that the remaining disagreement is in identifying the degree of either positiveness or negativeness. However, these degrees are subjective, thus, we remove them from the computation of agreement by grouping the negative (-2, -1), neutral (0), and positive (+1, +2) values in the corresponding negative, neutral and positive classes. With this configuration, the agreement further increases, suggesting that the disagreement on the polarity degrees is responsible for 0.19 points of Fleiss' $\kappa$. Moreover, with this configuration, we again compute the impact of the neutral class on the overall agreement. The results show that the annotators almost perfectly agree ($\kappa = 0.99$) in identifying positive and negative examples but struggle to agree on neutral.

We take a closer look at those examples in which at least one annotator selected the neutral class and observed two main factors: The first factor is the actual ambiguity, that is, examples in which there are several possible different interpretations. The second factor is the presence of both positive and negative aspects within one segment. In this case, there is subjectivity in recognizing the dominant aspect or if positive and negative aspects cancel each other out yielding neutral. For the second case, a better segmentation approach may help.

We then compare segment-based annotation with continuous annotation. To compute and compare the agreement, we chunk the continuous annotation according to the timing information of the segments used in the segment-based annotation. Then, for each segment, we compute the mean of the corresponding scores from the continuous annotation and round it to the nearest integer to obtain the five classes (-2,- 1, 0, +1, and +2). The results are shown in Table 5.5. The inter-annotator agreement of the continuous annotation is lower than the segment-based annotation when neutral is included. Indeed, we observe a greater impact of the neutral class when this is removed from the computation of the agreement.

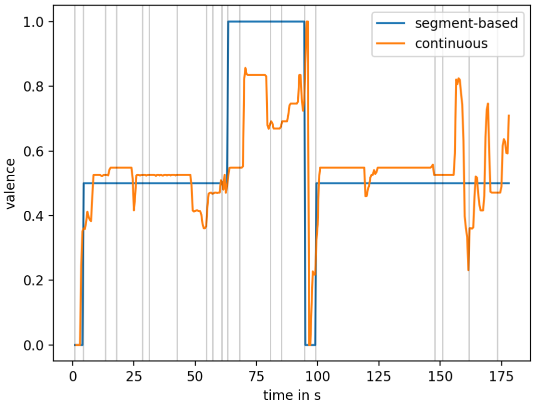


Figure 5.4: *Segment-based* and *continuous* valence annotation of a positive PN; the vertical lines mark the segments.

### Analysis of Valence Trajectories

Above statistics measure the agreement considering each segment as an isolated data point. Since each positive and negative story consists of several consecutive segments, we can also compare the *valence trajectories* for each of the stories, to also consider their agreement w.r.t. time. This is particularly interesting for stories where we observe contradicting segment-level valence, e.g. positive segments in an overall negative story.

We define a valence trajectory as a series of measurements, indexed either by time (*continuous*) or by segment index (*segment-based*), where one trajectory defines a story; stories are roughly the same duration but often of various lengths in the number of segments. The trajectories are suitable for calculating annotator agreement via curve equality measures, as well as for analyses on the time course of valence.

For each *continuous* annotation, we obtain valence values $c(t)$ in $[-1000; +1000]$ sampled at a rate of 0.5s, resulting in about 300 sample points per PN. For each *segment-based* annotation, we obtain valence values $s(i)$ in $[-2; +2]$ for each segment $i$, resulting in about 15 segments of variable length per PN.

Thus, $c$ and $s$ have hugely different lengths for the same PN, which makes them hard to compare. Figure 5.4 shows the continuous and segment-based valence trajectories for a PN; the vertical dividers mark the segment boundaries.

We propose two approaches for comparing $c(t)$ and $s(i)$ using the start and end times of the segments:

1. *Continuous*: If $c(t)$ is the reference series, we sample from $s$ by mapping $t$ to the corresponding segment.

2. *Segmental*: If $s(i)$ is the reference series, for each segment $i$, we extract average of the corresponding values from $c$.

In both cases, we map the continuous *values* to the discrete *class labels*, and normalize both trajectories, which will be described below.

As a measure of agreement between two curves, we compare Root Mean Square Error (RMSE) and Dynamic Time Warping (DTW)[14]. We find temporal shifts between the continuous and segment-based sequences (cf. Figure 5.4), which we attribute to the response time of the continuous annotators during live annotation. For this reason, we use DTW in addition to RMSE because it better maps the similarity of two sequences as it is inherently less prone to shifts; we apply per-trajectory mean and variance normalization prior to computation.

Table 5.6 shows the results indicating the mean and standard deviation of RMSE and DTW for the entire data set. For the inter-annotator agreement between continuous and segment-based, we achieve a mean RMSE of 0.48±0.14 for *segmental* reference and 0.49±0.15 for *continuous* reference. This means that the average mean error is about half a rating point, which in our case corresponds to one valence class (-2, -1, 0, +1, +2). This confirms the results from Section 5.2.4; the agreement in the stimuli (pos, neg) is high with variations in its subclasses. We find that DTW is almost the

| **Measure** | **Segmental** | **Continuous** | **Seg.-based only** |
|---|---|---|---|
| RMSE | 0.48±0.14 | 0.49±0.15 | 0.40±0.11 |
| DTW | $7.22 \times 10^{-2}$<br>$\pm 4.11 \times 10^{-2}$ | $1.82 \times 10^{-3}$<br>$\pm 1.17 \times 10^{-3}$ | $7.23 \times 10^{-2}$<br>$\pm 4.93 \times 10^{-2}$ |

Table 5.6: Annotator agreement for valence trajectories by means of RMSE and DTW; the last column shows the agreement among the segment-based annotators only.

same for *segmental* reference and segment-based annotators only, showing that the agreement among continuous and segment-based annotators is comparable to the one among only segment-based annotators. For *continuous* reference, DTW is significantly smaller than for *segmental* reference, although it was slightly higher in RMSE. We attribute this to the fact that DTW benefits from long matching sequences in *continuous* reference (cf. Figure 5.4 first and last third of the signal). Furthermore, the higher RMSE caused by the shifts is compensated for in DTW.

The results above include time-shifting the continuous signal to mitigate the delays due to response time. We obtained 1.5s as the optimal value with a minimal improvement of the agreement by 0.1 RMSE, while DTW inherently remains the same. For normalization to a range of $[-1; 1]$, we use a separate normalization to $[0; 1]$ and $[-1; 0]$ for the positive and negative value ranges, respectively. In this way, we correct for the possibility that an annotator may deviate more in one of the two ranges than in the other during continuous annotation. We achieve a 0.6 higher RMSE with this normalization method than with standard min-max normalization. To further improve normalization, we tried to find the annotators' "felt" neutral position of the joystick during continuous annotation. Thereby, we found a miscalibration of the joystick and a threshold range for continuous neutral valence in $[-30, 30]$. We solved this by shifting the zero point and mapping the threshold range to zero, increasing the agreement by 0.2 for RMSE.

We use valence trajectories to also gain insights into different common trends and traits of the narrators and narratives. The analyses is presented in Appendix B

**Inter Annotator Agreement - Emotion Carriers**

For calculating Inter-Annotator Agreement (IAA) for EC annotation, we use the metrics used in the first protocol, positive agreement, as explained in section 5.1.4, with different parameters for better matching of two EC spans. The IAA is calculated separately for each pair of annotators and the overall result is calculated by taking mean of the pairwise IAA. We experiment with different parameters for the matching of ECs and show the IAA results in the Table 5.7. The first row $a$, has the strictest criteria for matching two ECs, which matches the two ECs only if they contain exactly same tokens and at the same position in the narrative. Whereas the criteria in the last row (e) is most lenient for matching two ECs, which performs partial/soft matching, allows matching of the same ECs present at different positions in the narrative, and also uses matching of lemmas

| | **Parameters** | **USoM-Elderly**<br>**f1 - mean** | **USoM**<br>**f1 - mean** |
|---|---|---|---|
| a | Exact, T, token | 16.69 | NA |
| b | Exact, F, token | 18.46 | 25.2 |
| c | Partial, T, token | 36.49 | 32.0 |
| d | Partial, F, token | 51.34 | 39.9 |
| e | Partial, F, lemma | 59.10 | 40.3 |

Table 5.7: Inter annotator agreement (IAA) for the annotation of USoM-Elderly corpus, based on the positive agreement metric, with different parameter configurations. For comparison, the corresponding IAA for the annotation of USoM corpus from section 5.1.4 is provided in the last column. The agreement scores is calculated for each pair of annotators, the mean of the scores are presented in the table. For each configuration, [Parameters:(Matching strategy:Exact, Partial); (Position: considered(T), agnostic(F)); (lexical level: token, lemma)]

instead of tokens.

In the last column of the table, we compare the IAA results of the USoM-Elderly EC annotation to that of the USoM-Young annotation from Section 5.1. As we loosen the matching criteria, we see increment in the IAA of both the corpus. Although the trend is similar, from row *c* onward, the IAA of the USoM-Elderly is significantly higher than the IAA of USoM. The higher IAA indicates higher quality of the annotation. It also supports our claim that with the speech based, segment level annotation, the complexity and the subjectivity of the task reduces. Using the textBlob-de library[4], we also find ECs that carry sentiment polarity are significantly less (18%) in this annotation, compared to the USoM (39%), which further show quality improvement. Moreover, we find the majority of the EC tokens are Noun, Verb, Adverb, Determiner and Adjectives, which shows that the ECs belong to entities, represented by noun chunks and events, represented by verb chunks, which aligns with our expectations.

## 5.3 Protocol 3: Segment Level Valence and EC-Span Annotation of Written PNs

The first two annotation protocols were designed for spoken PNs by either considering their textual transcripts or speech context. Spoken nature of the data makes it difficult to parse with off the shelf NLP tools, thus making the task of segmentation and pre-selection of EC candidates difficult. In this annotation experiment, we design a protocol to annotate written PNs, which are structurally well-formed. Specifically, we annotate the PNs from the CBT-PHA dataset (concatenation of responses to the ABC questionnaire) explained in section 4.1.3, with segment level valence and emotion carriers.

[4] `https://textblob-de.readthedocs.io/en/latest`

### 5.3.1 Annotation Steps

In this section, we describe the steps involved in the experiment including preprocessing and segmentation of narratives, selection of candidates, and annotation of each segment.

#### Segmentation

First we segment the PNs into Functional Units using our in-house functional unit segmenter built for Italian as opposed to the hierarchical segmentation performed on the USoM-Elderly dataset (refer to section 5.2.1). The tool performs well on the written PNs, as it is able to consider cues such as punctuation and capitalization. Being a minimal communicative unit in human communication, the functional unit reduces the possibility of presenting both positive and negative emotions in the same segment, in turn reducing the problem of confusion on neutral label as experienced in the USoM-Elderly dataset (explained in section 5.2.4).

#### Candidates Selection

With good performance of dependency parser and POS (part of speech) tagger on the written PNs, it is possible identify noun-chunks and verb-chunks to be considered as a pre-selected list of candidates, from which ECs have to be selected. We use SpaCy toolkit with Italian model for identifying noun-chunks using the built-in function. Whereas, for identifying verb-chunks, we wrote rules based on dependency relations and POS tags. We include adverbs to be the part of verb-chunks as well. We exclude chunks consisting only of auxiliary verbs without another head verb. Once we get the list of noun-chunks and verb-chunks as candidates, we further filter out the chunks whose chunk-heads belong to a predefined set of emotion words. This helps, tackles the problem of selecting emotion words as emotion carriers as faced in the previous protocols. Providing a candidate list to select ECs from, gives additional control over what annotators select as emotion carriers, compared to the free span selection in the previous two protocols.

#### Annotation

Similar to the previous segment based annotation protocol, in this protocol as well the task is performed at the segment level and is divided into two sub-tasks, Valence annotation and EC-annotation. One segment at a time is presented to the annotator, and is asked to select the valence as felt by the narrator while recollection on a 5 point bipolar scale -2, -1, 0, +1, +2. The annotators were asked to adopt the point of view of the narrator. In the second part, for the non-neutral valence (positive or negative), the annotator is also presented with the pre-selected candidates highlighted in the text, and are asked to select the spans that help explain their selection of valence as ECs. While performing the task, the annotator has access to the entire narrative (as context), including the corresponding ABC questions (as explained in Section 4.1.3). Providing context helps better understanding whenever a segment is ambiguous without context. The annotators are advised to look at the context only in the case of confusion. Similar to the previous protocol from section 5.2, this protocol also captures unfolding of the emotions.

#### Execution

We recruited 3 Italian native speaker annotators from a pool of graduate students based on their research interests and previous experience with data annotation. Similar to the previous experiment, as explained in the section 5.2.1, we divide the execution in three phases Training, Overlapping and Partial-overlapping. The training phase begins with training of the annotators through a session to explain the task, the guidelines, and the tool. Later they practice on small batches to improve the Inter Annotator Agreement (explained in section 5.3.3), and the differences are discussed in the consensus meeting. Guidelines are modified if required. Later, overlap and partial-overlap phases are conducted, to finally get a 20 % overlap, meaning 20% of the data is annotated by all annotators, while 80% data is annotated by only a single annotator.

### 5.3.2 Output

In Table 5.8, we present an Italian PN annotated with valence and ECs using this protocol. The segments are color coded to represent the valence (red: negative **(-2)**; orange: slightly negative **(-1)**; gray: neutral **(0)**; light green: slightly positive**(+1)**; green: positive **(+2)**), whereas the emotion carriers are wrapped in the parentheses.

### 5.3.3 Analysis

Using the 481 Personal Narratives, 4273 Functional Units (FU) were annotated. The majority of the FUs, 60%, were annotated as neutral, while 13% and 27% of them were labeled as positive and negative respectively. The Inter-Annotator Agreement (IAA), computed with the Fleiss' $\kappa$ coefficient [46], on the valence annotation is 0.67 (Substantial) on the 5-point scale results, and 0.73 (Substantial) on the 3-point scale (obtained by regrouping the values into three groups of *positive* {*1,2*}, *negative* {*-2,-1*} and *neutral* {*0*}). Furthermore, the IAA on the examples that were labeled with a non-neutral polarity by all annotators is 0.98 (Almost Perfect).

In the EC selection task, out of 4452 EC-candidate spans in the FUs that were labeled with a non-neutral sentiment polarity, 1991 spans (45%) were selected as EC by the annotators, resulting in 2551 EC tokens (tokens

<table>
<tr>
<td>Questi pochi giorni di ferie (<b>stanno passando</b>) nella maniera più leggera possibile. | Stiamo passando questi giorni in (<b>famiglia</b>) giocando alla play station oppure (<b>guardando</b>) (<b>serie</b>) televisive come se non ci fosse un domani.| Facciamo (<b>pranzo</b>) e (<b>cena</b>) con (<b>i nonni</b>) che ci evitano, come famiglia, di avere anche la “preoccupazione” di cucinare. | Ho pensato che domani ricomincerò (<b>la solita routine</b>) ,| pensando che dopo tre giorni ripartirò nuovamente | Devo ottimizzare il tempo ed il “mood” per (<b>ricaricare</b>) per bene le batterie. | Sto (<b>cercando</b>) di essere il più spensierato possibile,|anche se oggi già penso che (<b>domani si ricomincia</b>) a (<b>lavorare</b>). | Non sono ansioso di lavorare, |ma qualche altro giorno di (<b>svago</b>) (<b>avrebbe giovato</b>) e non poco. |</td>
<td>These few days of vacation (<b>are passing</b>) in the lighest possible way. | We are passing this days with our (<b>family</b>), either playing play station or (<b>watching</b>) TV (<b>series</b>) as if there’s no tomorrow.| We have (<b>lunch</b>) and (<b>dinner</b>) with (<b>grandparents</b>) who spare us, as a family, to have “concerns” about cooking. | I thought that tomorrow I’m going to begin again (<b>the same routine</b>) , | thinking that after three days I’m going to start again. | I have to optimize the time and the “mood” to properly (<b>recharge</b>) the batteries. | I’m (<b>trying</b>) to be as cheerful as possible, | even if today I already think that (<b>tomorrow one starts again</b>) to (<b>work</b>). | I’m not anxious about work, | but some additional days of (<b>leisure</b>) (<b>would have been beneficial</b>) a lot. |</td>
</tr>
</table>

Table 5.8: An example of Personal Narrative from the CBT-PHA dataset (Original content in Italian in the left column whereas Translated content in English in the right column), annotated with Valence and Emotion Carriers. The Narrative is formed by concatenating responses to the ABC questionnaire by the client (more details on data collection: section 4.1.3). The segments are color coded to represent the intensity of the valence, whereas the Emotion Carriers are wrapped in parentheses.

in the EC-span) and the EC dictionary size of 962. The IAA on the EC annotation is 0.4 (Fair), computed by considering each EC-candidate as an example to annotate where the labels are *yes* if it is an EC, and *no* otherwise. It is worth noting that providing a list of pre-selected EC-candidates, makes it possible to use Fleiss' $\kappa$ as an agreement measure as opposed to the more complex Positive Agreement metric used in other protocols as explained in sections 5.1.4 and 5.2.4.

| **Polarity** | **Freq.** | **EC** | **non-EC** |
|---|---|---|---|
| *Positive* | 13% | 566 (28%) | 736 (30%) |
| *Negative* | 27% | 1425 (72%) | 1725 (70%) |
| *Neutral* | 60% | - | - |

Table 5.9: The distribution of valence polarity and Emotion Carriers (EC) in the annotated dataset of Personal Narrative at functional unit level.

The statistics regarding the labeled ECs and the sentiment distribution are presented in Table 5.9.

# Chapter 6

# Automatic Emotion Analysis Systems

In this chapter, we exploit the annotated data to train automated systems for the detection of the Emotion Carriers and Valence from the narratives. Similar to the annotation protocols, we experiment with different combinations of input modalities (text, speech, text+speech), levels of contexts (segment, narrative), and EC-span selection criteria (free, list of candidates).

Starting with the USoM dataset (annotated with ECs), in Section 6.1, we train a model to identify EC-spans from the textual transcripts of the spoken PNs. As the data was spoken in nature, we move on to Section 6.2 where we improve the model by utilizing speech features, along with the textual features. After exploring the spoken PNs from the USoM corpus, later in Section 6.3, we work on written PNs from the CBT-PHA dataset and train segment level models for identifying perceived valence and ECs from a preselected list of EC candidate spans. In the Section 6.4, we explore joint learning for multi task modeling of valence prediction and EC detection from the CBT-PHA dataset, to study if the two tasks can trained jointly and provide automatically an EC-based explanation of the valence prediction.

## 6.1 Narrative Level EC-Span Detection from Textual Transcripts

Similar to annotation protocols, we begin with a text based system for better generalizability as speech input may not be available in all datasets. In this section, we build a system to automatically detect Emotion Carriers from Personal Narratives from USoM dataset (section 4.1.1) annotated with the Protocol-1 (explained in section 5.1. The modality used for the detection (as well as during the annotation) is text (from the transcriptions); narrative level context was considered while performing the annotation, whereas in the detection, we try both sentence level as well as the entire narrative level; the annotators and thus the system try to predict free text spans as ECs. (This work has been published in the INTERSPEECH 2021 conference [144])

### 6.1.1 Task Formulation

The annotated data can be represented with the *IO* encoding, as shown in the example from Table 6.1. We consider the document as a sequence of tokens, where each token is associated with the label *I* if it is a part of an EC, and the label *O* if it is not. A continuous sequence of tokens with label *I* represents an EC. In the third row *Annotation*, we show the manual annotation by four annotators. It can be observed how the annotators perceive ECs differently, showing the high subjectivity of the task.

To recall, in the preprocessing step, we first perform tokenization using the spaCy toolkit [63]. Next, we remove punctuation tokens from the data. Based on initial experiments, we found that removing punctuation helps improve the performance of the models. The number of annotations (ECs) identified by the annotators per narrative varies from 3 to 14 with an average of 4.6. On average, the number of tokens per EC consists of 1.1

<table>
<tr><td><b>PN fragment:</b></td><td>Und</td><td>ähm</td><td>die</td><td>Gefühle</td><td>dabei</td><td>waren</td><td>dass</td><td>man</td><td>sich</td></tr>
<tr><td><b>Gloss:</b></td><td>And</td><td>um</td><td>the</td><td>feelings</td><td>there</td><td>were</td><td>that</td><td>you</td><td>yourself</td></tr>
<tr><td><b>Annotation:</b></td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td></tr>
<tr><td><b>PN fragment:</b></td><td>einfach</td><td>freut</td><td>und</td><td>glücklich</td><td>ist</td><td>dass</td><td>man</td><td>eine</td><td>Familie</td></tr>
<tr><td><b>Gloss:</b></td><td>easy</td><td>pleased</td><td>and</td><td>happy</td><td>is</td><td>that</td><td>you</td><td>a</td><td>family</td></tr>
<tr><td><b>Annotation:</b></td><td>O|O|O|O</td><td>I|O|O|O</td><td>I|O|O|O</td><td>I|O|O|I</td><td>O|O|O|I</td><td>O|O|O|O</td><td>O|O|O|O</td><td>O|O|O|O</td><td>I|I|I|O</td></tr>
<tr><td><b>Translation:</b></td><td colspan="9"><i>And uh, the feelings were that you are uh just pleased and happy that you have a family ...</i></td></tr>
</table>

Table 6.1: A small text fragment from a PN annotated with emotion carriers. The first row reports the original German words from the PN, the second row shows the corresponding English translation, while the third row shows the annotations. The annotation is performed by 4 annotators, thus for each token, there are 4 IO labels. For the token "Familie" the annotation is I|I|I|O, which means that the first three annotators classified it as *I* while the forth as *O*. The intensity of the red color in the background for the PN fragment also highlights the number of annotators who annotated the token (from lightest for 1 annotator to the darkest for all 4 annotators).

tokens for three annotators, while the fourth annotator identified longer segments consisting of 2.3 tokens (avg.). On average, a narrative consists of 704 tokens, while a sentence consists of 22 tokens. The sentence splitting is performed using the punctuation provided in the original transcriptions. We find that only 7.3% of the tokens are assigned the label *I* by at least one annotator. This shows that the classes *I* and *O* are highly imbalanced, which could result in inefficient training of the models. With further analysis, we notice that only 34% of the total sentences contain at least one EC, while the remaining 66% sentences do not contain any carrier marked by any annotator.

We pose the detection of emotion carriers from a given PN as a sequence labeling problem. The final goal is the binary classification of each token into classes *I* or *O*. As seen in Table 6.1, the task of selecting EC text spans is subjective. Each annotator has a different opinion toward the spans to be selected as ECs, so it is challenging to identify an annotation as valid or invalid. For each token, we have annotations from four annotators with *IO*

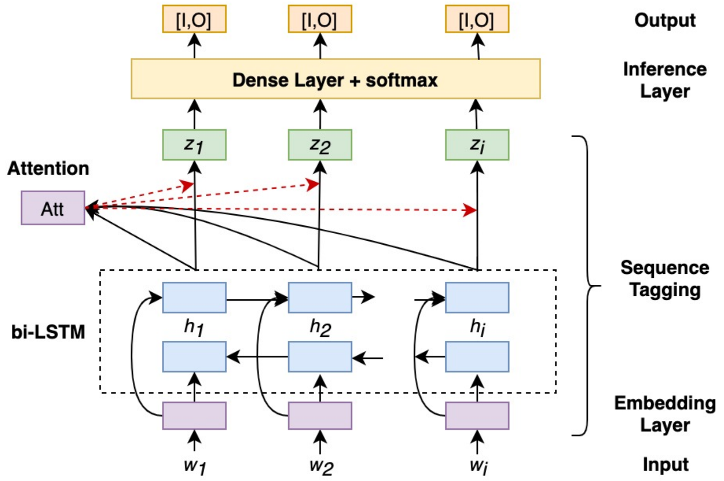


Figure 6.1: bi-LSTM based DNN architecture for ECR. In the output, [I,O] represent the probabilities for the classes *I* and *O*.

labels. Some annotators may agree on the annotation, but it is infrequent that all four annotators annotate the token as *I*. In the example from Table 6.1, the token "glücklich" is annotated *I* by two annotators, three annotators agree that the token "Familie" is an EC, while a few tokens are marked as EC by only one annotator. In this scenario, it is difficult to provide a hard *I* or *O* label. To tackle this problem, we model the problem of Emotion Carrier Detection (ECD) as providing scores to the label *I*, representing the likelihood that that token is a part of an EC. Label distribution learning (LDL) [50] can effectively capture the label ambiguity and inter-subjectivity within the annotators. We use LDL with a sequence labeling network and the KL-Divergence loss function. The advantage of LDL is that it allows to modeling the relative importance of each label. For evaluation, we use different strategies to select the final *IO* labels.

### 6.1.2 Model

We use sequence labeling architecture relying on biLSTM with attention, similar to [135]. As shown in Figure 6.1, the input text is first passed through the embedding layer to obtain the word embedding representation for each token. We use 100-dimensional pre-trained GloVe [110] embeddings. To encode the sequence information, we then use two stacked bidirectional LSTM layers with a hidden size of 512. We also use attention mechanism [157] along with the bi-LSTM, where attention weights $a_i$ represent the relative contribution of a specific token to the text representation. We compute $a_i$ at each output time $i$ as follows:

$$a_i = softmax(v^T tanh(W_h h_i + b_h)) \tag{6.1}$$

$$z_i = a_i.h_i \tag{6.2}$$

where $h_i$ is encoder hidden state and $v$ and $W_h$ are learnable parameters of the network. The output $z_i$ is the element-wise dot product of $a_i$ and $h_i$.

Finally, the output is passed through the inference layer consisting of two fully connected layers with 50 units each, and a softmax layer to assign probabilities to the labels for each word. We also use layer normalization and two dropout layers with a rate of 0.5 in the sequence and inference layers [10]

During training, we use the Kullback-Leibler Divergence (KL-DIV) as the loss function [77] and the Adam optimizer [72] with the learning rate of 0.001.

### 6.1.3 Experiments

As described in Section 6.1.1, there is class imbalance in the data. The class $I$ tokens are very infrequent compared to the class $O$. This may result in a bias toward class $O$ in the classifier. Another problem we have to deal with is the length of the narratives. The narratives are very long with an average length of 704 tokens. Standard machine learning and bi-LSTM based architectures are not efficient in dealing with very long contexts.

To address these challenges, we experiment with different levels of segmentation of the narratives and apply strategies to select proper train and test sets. We train and test the sequence-labeling models at narrative and sentence levels. In the **narrative** level, we consider the entire narrative as one sequence, while in the **sentence** level, we consider one sentence as a sequence. In this way, we analyze how the length of a sequence affects the performance of the model. Also, note that at the sentence level, the model does not have access to other parts of the narrative. We study how limited access to context affects performance.

The sentence-level sequences are further considered in two ways : 1) **SentAll:** all the sentences are considered 2) **SentCarr:** only sentences containing at least one EC are considered. *SentCarr* reduces the class imbalance as we remove all sentences that do not contain any token tagged as class *I*. In a real-world scenario, we would have to extract carriers from the entire narrative or all the sentences, as we do not know beforehand which sentences contain the carriers. Thus, we use *SentCarr* only for training, but in the test set *SentAll* is used.

We also experiment with another sequence labeling model based on **Conditional Random Fields (CRF)** [79], a widely used machine learning algorithm for sequence-labeling problems in NLP, such as Part of Speech tagging. We use sklearn-crfsuite library [1] for the CRF implementation. For the CRF model, we use the context window of ±3 with features such as the token, its suffixes, POS tag, prefix of POS tag, sentiment polarity.

### 6.1.4 Evaluation

In this section, we propose different evaluation strategies for the ECD task. Note that even though our model is trained to predict the probability dis-

[1] `https://tinyurl.com/sklearn-crf`

tribution of the classes, our final goal is to assign one of the two classes from *I* and *O*. For all evaluations, for the ground truth, we consider that a token is annotated (i.e. *I* in the *IO* tags) if at least one of the annotators has annotated it. Similarly, for the output, we consider the output as *I* if the probability assigned crosses the minimum threshold of 0.25, which is equivalent to one of four annotators tagging the token as *I*. In all evaluations, we do five-fold cross-validation with the leave one group (of narrators) out (LOGO) strategy. For each training session, we split the data into train, dev, and test sets without any overlap of narrators in the three sets.

| **Data** | | | **Results (F1)(std)** | |
|---|---|---|---|---|
| **Segmentation** | **#train** | **#test** | **class-I** | **micro** |
| SentCarr | 1737 | 367 | 53.2(4.7) | 93.7(0.8) |
| **train: SentCarr**<br>**test: SentAll** | 1737 | 1533 | **34.9(3.4)** | **96.6(0.5)** |
| SentAll | 6378 | 1533 | 31.2(3.9) | 96.6(0.5) |
| Narrative | 191 | 48 | 34.6(5.0) | 96.7(0.4) |
| CRF(SentCarr; SentAll) | 1674 | 1582 | 34.2(4.1) | 96.2(0.3) |

Table 6.2: Results of bi-LSTM based models with different data segmentation. Notice how the the number of data-points vary as we change the segmentation.

### Token Level

The token level evaluation measures the performance of predicting *I* or *O* class for each token in a sequence. We use this metric to evaluate our models with different data segmentation strategies. We are concerned more about the prediction of the class *I*, as we are interested in applications of ECD such as Conversational Agents, where it is important to find one or more important carriers to start a conversation with the narrator. Thus,

| sr | Parameters | Prec(std) | Recall(std) | F1(std) | IAA (F1) |
|---|---|---|---|---|---|
| a | Exact, F, token (w/ stopwords) | 32.6(3.3) | 52.1(3.6) | 40.0(2.9) | 25.2 |
| b | Exact, F, token | 42.3(4.6) | 67.2(4.2) | 51.7(4.0) | NA |
| c | Partial, T, token | 37.4(4.0) | 51.3(0.4) | 43.1(2.7) | 32.0 |
| d | Partial, F, token | 59.4(5.5) | 83.6(4.7) | 69.2(3.5) | 39.9 |
| e | Partial, F, lemma | 61.8(6.4) | 86.5(4.2) | 71.8(4.3) | 40.3 |

Table 6.3: Evaluation based on the agreement metrics (positive agreement) with different parameter configurations. For each configuration, the corresponding inter-annotator agreement (IAA) score is in the last column (in terms of F1 score).[**Parameters**:(Matching strategy:Exact, Partial); (Position: considered(T), agnostic(F)); (lexical level: token, lemma)]

we show the F1 score of class *I* and weighted average (micro) of F1 score of *I* and *O*.

As discussed earlier, considering a real-world scenario, we need the model to perform well on the *SentAll* or *Narrative* data. In Table 6.2, we find that the model trained on *SentCarr* performs best on the *SentAll*. For further evaluation we use this model, thus recognition would be done on the sentences and not the entire narrative at once. Note that the performance of the *Narrative* strategy is only slightly worse, suggesting that the task is not affected much by the length of the context available. The CRF model is the worst performing one.

Using the *SentCarr* model, we extract the continuous sequences of tokens that are tagged as *I*. These text spans are considered as the ECs recognized by the model. In the metrics in the next section, we evaluate the model by comparing this set of carriers with the set of manually annotated reference carriers.

### Agreement Metrics (Narrative Level)

We evaluate the performance of the models using the metrics that were used to evaluate the inter-annotator agreement between the four annotators (pair-wise) in section 5.1.4 and 5.2.4, based on the *positive (specific) agreement* [47]. This evaluation is important as it compares the performance of the system with the inter-annotator agreement, which can loosely be considered as human performance.

We also explore the different criteria to decide whether two spans match or not, as used in the original metrics. Different parameters in the metrics while matching two spans include Exact vs Partial matching, position in the narrative, and tokens vs lemma. We remove stopwords from the annotation as we are interested in the content words.

**Results:** Table 6.3 summarizes the evaluation using the Agreement metrics. As expected, with the loosening of the matching criteria, the results improve. A similar trend is observed in the inter-annotator agreement. When we move from *a* to *b*, we are removing the stopwords from the predicted and reference carriers. This improves the results significantly. The reason behind this is the fact that the reference annotations, which were also used for training the model, as mentioned earlier, contain all the tokens that are tagged by at least one annotator. As noticed in the corpus analysis, in the annotations, one of the annotators usually annotates longer spans than others. We also observed that many annotations also contain punctuation and stopwords. To understand this issue, let us consider an example of concept annotation. For a concept like a printer, the annotators could select spans ‘with the printer’, ‘the printer’ or just ‘printer’. With our strategy for creating reference annotations, we end up selecting the longest span ”with the printer” which contains stopwords like *with, the.* However, this might not be the case in the model’s output (as the training data also contain concepts marked by only one annotator). To reduce this effect, one way is to remove the stopwords (strategy b) and another is to use the partial match (strategy c). While both strategies improved the scores, the improvement with strategy b is more significant than with strategy c. We notice a significantly large jump in the model's performance from *c* to *d*, compared to the inter-annotator agreement. Our intuition is that this could be because the model is trained at the sentence level, thus the position in the narrative is not taken into consideration, resulting in recognition of multiple occurrences of the same carrier. Additionally, the performance further improves when we match lemmas instead of tokens (from *d* to *e*).

#### Recognized at least *k* carriers

In the context of a target application of the ECD such as human-machine dialogue, an EC could be used as a trigger for a machine to start a conversation with a human. To begin a conversation, we would need at least one EC to talk about. In this evaluation metric, for each narrative, we measure if at least *k* carriers from the reference are recognized by the model. A carrier is considered recognized if it is an exact match. When matching, we remove stop-words. We perform two evaluations, considering and not considering the position of the carrier in the narrative. The results are described in Table 6.4. For our goal of starting a conversation about a particular carrier, the results seem overwhelmingly good. However, an important question remains, how many of the recognized carriers are useful for a conversation?

**Sentiment vs Content Carriers:** The annotations include sentiment words as well as content words. To study if the model is biased towards the recognition of ECs with sentiment words (angry, joy) versus content words

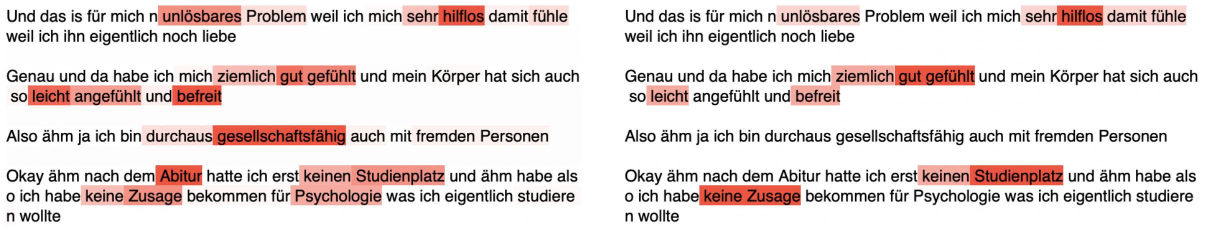
| **Type of carriers** | **Posn** | **at least k recognized** | | |
|---|---|---|---|---|
| | | **k=1** | **k=2** | **k=3** |
| all carriers | Yes | 99.1(1.1) | 95.4(1.6) | 86.1(3.6) |
| | No | 99.1(1.1) | 97.0(1.7) | 88.6(5.1) |
| content carriers | Yes | 95.0(2.1) | 75.2(6.2) | 48.9(7.2) |
| | No | 95.4(1.6) | 80.3(4.5) | 56.4(8.0) |
| sentiment carriers | Yes | 78.6(4.4) | 52.4(7.9) | 29.2(6.1) |
| | No | 80.4(4.2) | 54.9(7.3) | 33.4(6.8) |

Table 6.4: Evaluation based on the fraction of narratives (%) in which at least k carriers are recognized correctly by the model. The values are represented in the format *mean(std)* across five folds. The evaluation is performed separately for all the carriers, only content carriers and only sentiment carriers. In the *posn* column, *yes* represents position considered while *No* means position agnostic

(internship, parents) in ECs, we further divide the annotations (reference and predicted) into sentiment and content carriers and perform the similar evaluation on them separately. For this analysis, we calculate the sentiment polarity of each annotation using the textblob-de library (similar to section 5.1.4) If the score is 0 the carrier is considered a content carrier, otherwise a sentiment carrier. We find that more than 60% of the Emotion Carriers are classified as content carriers.

We find that on average more than half of the annotations are classified as content carriers. The manual analysis of the annotations shows that the classification using textblob-de is not perfect. While it can recognize the content carriers properly, we see some examples of sentiment-carriers are also being classified into the content-carriers. Some examples of correctly recognized content carriers include *Bachelorarbeit (bachelor thesis), Magenprobleme (stomach problems), Durchhaltevermögen (stamina)* while an example of sentiment-carriers that are classified as content-carriers include *unzufrieden (unsatisfied)*

Next, we do the evaluation based on the *at least k recognized* metric for each group independently. In Table 6.4, we compare the results for the content and sentiment carriers. We observe a decline in the performance compared to the evaluation of all carriers. Nevertheless, we find that in 95.4 % of the narratives (position agnostic), we can predict at least 1 emotion carrier, which is a requirement for starting a conversation.



Und das is für mich n unlösbares Problem weil ich mich sehr hilflos damit fühle weil ich ihn eigentlich noch liebe

Genau und da habe ich mich ziemlich gut gefühlt und mein Körper hat sich auch so leicht angefühlt und befreit

Also ähm ja ich bin durchaus gesellschaftsfähig auch mit fremden Personen

Okay ähm nach dem Abitur hatte ich erst keinen Studienplatz und ähm habe als o ich habe keine Zusage bekommen für Psychologie was ich eigentlich studiere n wollte

(a) Model's Output

Und das is für mich n unlösbares Problem weil ich mich sehr hilflos damit fühle weil ich ihn eigentlich noch liebe

Genau und da habe ich mich ziemlich gut gefühlt und mein Körper hat sich auch so leicht angefühlt und befreit

Also ähm ja ich bin durchaus gesellschaftsfähig auch mit fremden Personen

Okay ähm nach dem Abitur hatte ich erst keinen Studienplatz und ähm habe als o ich habe keine Zusage bekommen für Psychologie was ich eigentlich studiere n wollte

(b) Ground Truth

Figure 6.2: Heatmap of sentences from narratives annotated with emotion carriers; highlighting tokens with model's output and ground truth probabilities. Notice the wider range of scores in the Model's Output as compared to only four possible scores in the Ground Truth.

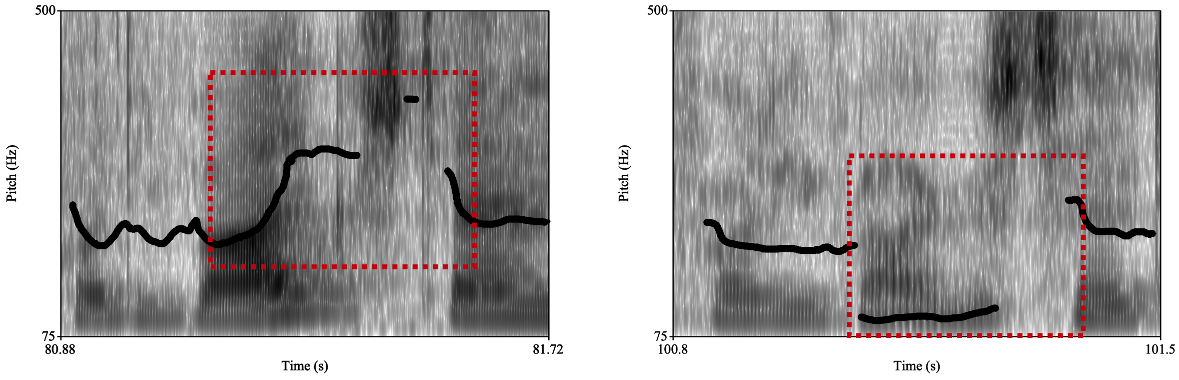
### Qualitative Analysis

Figure 6.2 shows example sentences from a test set with a heatmap showing the model's predicted score and ground truth probabilities for each token. In most cases, the probability distribution in the model's output seems to follow similar trends to that of the ground truth probabilities. We also observe frequent cases of false positives, where the model assigns a high probability to class $I$ even when the ground truth label is $O$, as can be seen in the third example. This behavior could be a result of training the model at the sentence level with the *SentCarr* strategy, where all the sentences in the training set contain at least one EC, biasing the model towards that distribution.

## 6.2 Detecting Emotion Carriers by Combining Acoustic and Lexical Representations

In the first experiment 6.1, we built an automated system using the PNs from USoM corpus 4.1.1, annotated using protocol-1 5.1. The annotation and the model, both used cues from textual transcriptions for the identification of ECs, completely ignoring the speech. In this section, we extend this experiment to also consider the cues from acoustic features of the tokens from corresponding speech signals, and study the improvements. We try different fusion techniques to combine the acoustic and lexical representations at different levels. [This work has been published in ASRU-2021, in collaboration with and lead by *Sebastian P. Bayerl*[13]]

### 6.2.1 Motivation



(a) emotion carrier

(b) non emotion carrier

Figure 6.3: Spectrograms with *f0*-contour of the phrase: "**vor die Wahl gestellt**,"(Translation: "made me choose"). (a) was marked as an EC, showing signs of emotional speech and the voice cracking in the center part whereas (b) was taken from the same recording session, but was not marked as EC. While this is an anecdotal example, statistical analysis revealed significant differences in f0, energy, and shimmer on nouns marked as EC when comparing them with all other nouns in the dataset.

People express and communicate emotions consciously as well as subconsciously. This is done by modifying the manner of speaking, the content of a conversation or written text, facial expressions, gestures, or even the way of walking. The combination of these signals, especially speech and text, has successfully been used to determine the emotional state of a person making a statement or telling a narrative [90, 115]. In the previous experiment, we worked on the automatic detection of emotion carriers from transcriptions of spoken PNs. However, relying only on lexical features leaves out the possibility of the same lexical content conveying different things based on acoustic context.

Ivanov *et al.* showed that there is a relationship between meaning-bearing parts of utterances and their acoustic properties [67]. Following up on that research, we have found evidence supporting distinct prosodic profiles for emotion *vs* non-emotion carriers: Figure 6.3 compares the spectrograms of two occurrences of the phrase ‘‘vor die Wahl gestellt”; “made me choose”; while (a) was annotated as an emotion carrier, (b) was not. The strong rise in fundamental frequency (f0), as well as the strong fluctuations at the beginning of Figure 6.3a, indicates emotional speech [105]. In contrast, the same phrase that was not marked as an EC has a very flat f0 contour (*cf.* Figure 6.3b). While the figure provides only anecdotal and motivational evidence, in this paper we provide ample evidence of the complementarity of acoustic and lexical information.

### 6.2.2 Method

We follow the approach of finding representations for EC from different feature spaces. As EC are a word-, and phrase-level concept, we try to find appropriate representations from the linguistic and acoustic input feature space. The representations are then used in uni-modal experiments as well as multi-modal fusion experiments for EC recognition.

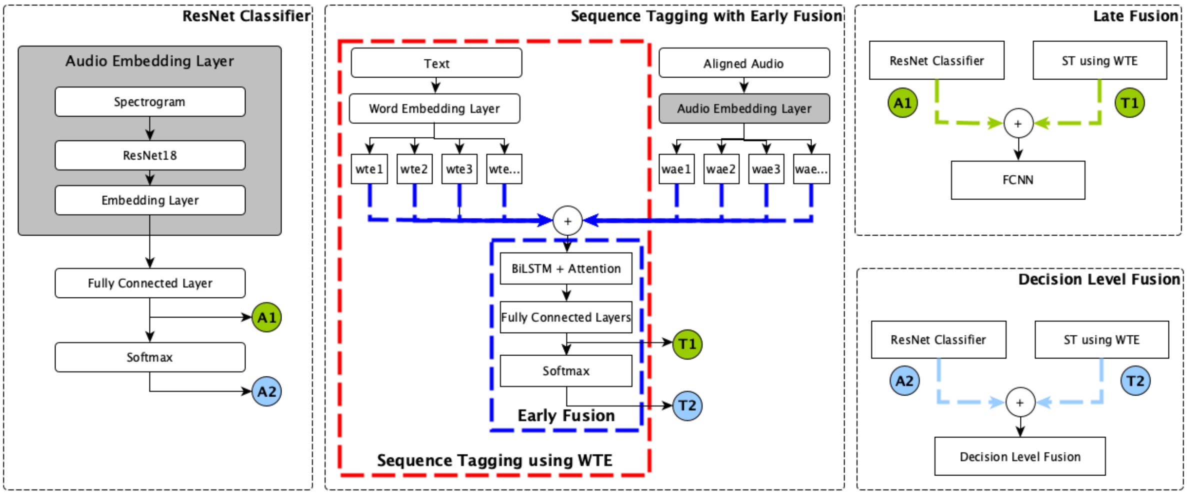


Figure 6.4: Overview of the neural network architecture used in the experiments. The left part shows the ResNet classifier containing the audio embedding layer which is used to extract word-based acoustic embeddings (WAE) for each word. The central part contains the sequence tagging (ST) architecture that can operate using either word-based textual embeddings (WTE), WAE, or a combination of WTE and WAE in an early fusion (EF) approach. The right part of the figure is depicting the late fusion (LF) and decision level fusion (DLF) systems. Inputs for the LF are taken after the fully connected layer in the ResNet and the ST, using only WTE as inputs, as logits (A1, T1) and for DLF after the Softmax layer (A2, T2) returning normalized probabilities.

### Word-Based Textual Embeddings (WTE)

For word-based textual embeddings, we use 100-dimensional pre-trained GloVe word embeddings trained on the German Twitter corpus [111], from the preveios experiment. A total of 656 (10.2%) words were not present in the pre-trained embeddings. The word-embeddings where fine-tuned on the actual task inside the cross-validation loops.

### Word-Based Acoustic Embeddings (WAE)

The previously performed feature analysis revealed differences between EC and non EC with respect to handcrafted acoustic features. This motivated us to use embeddings based on convolutional neural networks (CNN) and handcrafted acoustic features as described in [112]. Our early stage experiments failed to produce good results on an utterance level with the German EmoDB dataset [18] on the speech emotion recognition (SER) task.

The failure to produce good results with CNNs and handcrafted acoustic features led us to explore other network architectures and input features. We decided to use a ResNet architecture which was successfully applied to a number of speech applications such as speaker recognition, and SER and has been shown to produce good embeddings [27, 145]. ResNets are fully convolutional neural networks (FCN) and can handle inputs of different sizes (or lengths, respectively) due to a global pooling layer at the end of the convolutional part of the neural network. The network we use is very similar to the one described in [60] and consists of 18 convolutional blocks (ResNet18). Its architecture was adapted by removing the initial max-pooling layer to keep more features prior to the residual blocks as the expected inputs are already relatively small. The dimensionality of the embedding layer was reduced from 1000 to 512 and the final classification layer was altered to match the number of classes (2).

As acoustic input features for the ResNet, we extract 40-dimensional Mel-frequency cepstrum coefficients (MFCC) with a window length of 0.025 s, a frameshift of 0.01 s, along with 1st and 2nd order moments, stacking them to a tensor with three dimensions (frequency x time x moments) and apply z-score normalization. Those acoustic input features are then used to train the acoustic only classifier and to extract neural acoustic word embeddings from the trained acoustic encoder. The word-based acoustic contexts are extracted using the aforementioned FAs.

The network was pre-trained on short utterances from the German EmoDB corpus to differentiate between neutral and emotional speech [18]. This is done to learn filters that are already primed to extract features from speech that are important to classify emotional speech. Both pretraining and training were done using stochastic gradient descent with a cross entropy loss function. To overcome the class imbalance problem, an oversampling strategy was applied as it had proven to be the best performing technique in our experiments.

To obtain word-based acoustic embeddings (WAE), we froze to resulting model to act as an acoustic word encoder. We feed the word-based acoustic context to the model and extract WAE from the embedding layer of the model.

#### Sequence Tagging

We model the task of detecting emotion carriers as a binary sequence labeling problem using both modalities, with targets encoded as $I$ if the token is part of an EC and $O$ otherwise, similar to the first experiment 6.1.1. For this task, we adopt a bidirectional Long Short-Term Memory (LSTM) neural network with an attention-based sequence tagging (**ST**) architecture from section 6.1

#### Fusion

There are two main challenges in combining multiple modalities: How to combine features of different dimensionality and valuation (different vector space), and at which stage to combine the streams. In general, three different kinds of approaches can be differentiated: early, late, and decision-level fusion.

**Early and late fusion** In early fusion (EF), features for each modality are extracted separately, i.e. each modality represents a view of the same concept. The resulting feature vectors are then combined, e.g. by concatenation or stacking, and then treated as a single input channel. In late fusion

(LF), each modality has its own model and is often trained independently. The outputs of those classifiers are used as input to another classifier that combines them for an overall best prediction. EF as used in this paper can be found at the center of Figure 6.4 and the LF approach can be found at the top right

**Decision Level Fusion** In our experiments the sequence tagger using WTE is trained as a regression problem with the Kullback–Leibler (KL) divergence, predicting the probability of a token being an emotion carrier. For this, the best decision threshold was experimentally found to be $p_{db} = 0.15$ for lexical features only. This motivated us to explore a rather heuristic late fusion approach: a *rule-based cascaded classifier based on posterior probabilities.* Applying a similar technique to the normalized probabilities in the output of the ResNet classifier, we can find a decision threshold and then merge the decisions, defining decision states around these thresholds. This way it is possible to leverage long-range lexical information as well as local acoustic information. We define the lexical-based ST model to be the primary model and the ResNet classifier to be the disambiguator, leveraging local acoustic information. In our merging approach, we define an $\epsilon$ parameter that indicates how certain a classifier is with the decision if a token is an EC. The decision boundary (DB) is defined by setting a probability value $p_{db}$.

We only consider the probability for the EC to determine certainty. If the normalized probability of a token being an EC $p_{ec}$ is within the epsilon interval ($p_{db} \pm \epsilon$) the classifier is considered to be uncertain regarding a positive decision of a token being an EC. $p_{ec} > p_{db} + \epsilon$ is considered to be certain. Those certainty indicators are computed for both models separately. Merging is then done by checking certainty indicators: If the lexical model is certain, the token is considered to be an EC. If the lexical model is uncertain and the acoustic model is certain, the token is also considered as an EC. In all other cases, the token is not considered as an EC. We call this heuristic *decision level fusion* (DLF).

### 6.2.3 Experiments

Results for single modality, fusion experiments, and baselines, are reported in Tab. 6.5. We report metrics for class *I*. The equal priors baseline constitutes random guessing with no knowledge about the actual class distribution with $p_I = p_O = 0.5$, resembling a fair coin toss whereas the class priors baselines resembles a heavily biased coin with $p_I = 0.066$ and $p_O = 0.934$ (The annotated data contains only 6.6% I labels as compared to O).

#### Training Details

All experiments were performed using five-fold cross-validation with consistent folds across all experiments. The folds were split by speaker to ensure no speaker in the test set was present in the training set and hyperparameters were tuned on separate development folds, as is common when working with acoustic data and small datasets. Tab. 6.5 contains results for single modality classification using a ResNet classifier as well as ST using either WTE or WAE as inputs. We report one result for an EF experiment concatenating WTE and WAE to a single word vector as well as a logit-based LF experiment combining the ResNet classifier and the ST using WTE as inputs only. Lastly, we show our overall best results, obtained with DLF and oracle results. Oracle results are obtained by a fictitious fusion of classifiers, which is considered to be right, if at least one of the contributing classifiers (ResNet18 and ST WTE), predicted the correct label. Details of the proposed neural network architectures can be found in Fig. 6.4.

| **Model** | **Features** | **Prec-I** | **Recall-I** | **F1-I** |
|---|---|---|---|---|
| Baseline | equal priors | 6.6 | 50.0 | 0.12 |
| Baseline | class priors | 6.6 | 6.6 | 0.07 |
| ResNet18 | MFCCs | 19.4 (5.3) | 64.6 (14.6) | 0.29 (0.05) |
| ST | WAE | 7.6 (2.3) | 40.3 (9.0) | 0.13 (0.03) |
| ST | WTE | 37.9 (6.9) | 46.7 (6.4) | 0.41 (0.03) |
| ST EF | WTE, WAE | 35.3 (6.4) | 44.3 (5.9) | 0.39 (0.04) |
| FCNN LF | logits | 25.6 (3.6) | 52.5 (9.2) | 0.34 (0.01) |
| **DLF** | **post. prob.** | **42.3 (5.3)** | **51.2 (6.4)** | **0.46 (0.05)** |
| Oracle | - | 70.6 (6.1) | 67.4 (4.6) | 0.69 (0.03) |

Table 6.5: Precision, Recall and F1 scores for class $I$ of EC detection. We report results for different models trained using combinations of modalities (acoustics and lexical) with early (EF), late (LF) and decision level late fusion (DLF) using posterior probabilities. For LF, only the best performing experiment using a fully connected neural network (FCNN) is shown. Baseline results are included for equal priors with $p_I = p_O = 0.5$ representing a fair coin toss and class priors with $p_I = 0.066$ and $p_O = 0.934$. The results are in the format: *mean(std)*; computed over the five folds.

#### Results

Direct word-level EC detection using only MFCC features (ResNet18) improved results compared to both random baseline classifiers using class priors for both actual class priors in the dataset as well as equal class priors. It can therefore be assumed that useful representations can be extracted from the embedding layer of the ResNet classifier. The analysis of the word-based acoustic embeddings produced by the ResNet system also looked promising. Fig. 6.5 contains a t-distributed Stochastic Neighbor Embedding (t-SNE) plot of embeddings marked as EC vs. embeddings not marked as EC. The plot shows that there is potential to differentiate EC from non-EC tokens in this low-dimensional projection.

While we achieved good results with the ST using WTE only, results for the WAE failed to perform better than the ResNet classifier that solely relied on local acoustic information. It barely improved results compared to the random classifiers' expected baseline precision. We, therefore, decided to not use the ST with WAE in late fusion experiments and rather use the ResNet classifier in LF.

The EF experiment combining WAE and WTE performed worse than WTE alone as described in this paper and only slightly improved previous WTE only results (Sec 6.1.4). The LF experiment using logit outputs from the ResNet classifier and the ST using WTE improved the ST using WTE only in terms of recall, but lowered the precision which lead to overall worse results w.r.t. F1-I. Experiments with Logistic Regression to model the probability of a word being an EC using the logit outputs of the ST using WTE and the ResNet classifier did not improve over the LF experiment with the FCNN.

Unfortunately, the experiments using the standard EF and LF approaches couldn't improve over the already strong textual system (ST WTE). However as shown in Fig. 6.3 and the feature analysis, there definitely is evidence that acoustic information can help with the detection of EC. Our experiments with the word-level ResNet classifier could not completely convince but still beat all statistical baselines as a stand-alone system. Lastly oracle results presented in Tab. 6.5 show that the combination of the ResNet classifier and the ST using WTE still has a lot of room for improvements.

This led us to explore the rather heuristic decision level fusion (DLF) approach described in  6.2.2 and yielded the best overall results. The decision boundary was tuned for the ResNet classifier only since the ST using WTE was already trained using KL divergence with a tuned decision boundary at $p_{db} = 0.15$. The DB for the ResNet classifier was determined using 5-fold cross-validation. Results are reported in Tab. 6.5 (DLF). The decision boundary for certainty of the ResNet classifier was found to be $p_{DB} = 0.75$ with $\epsilon = 0.05$.

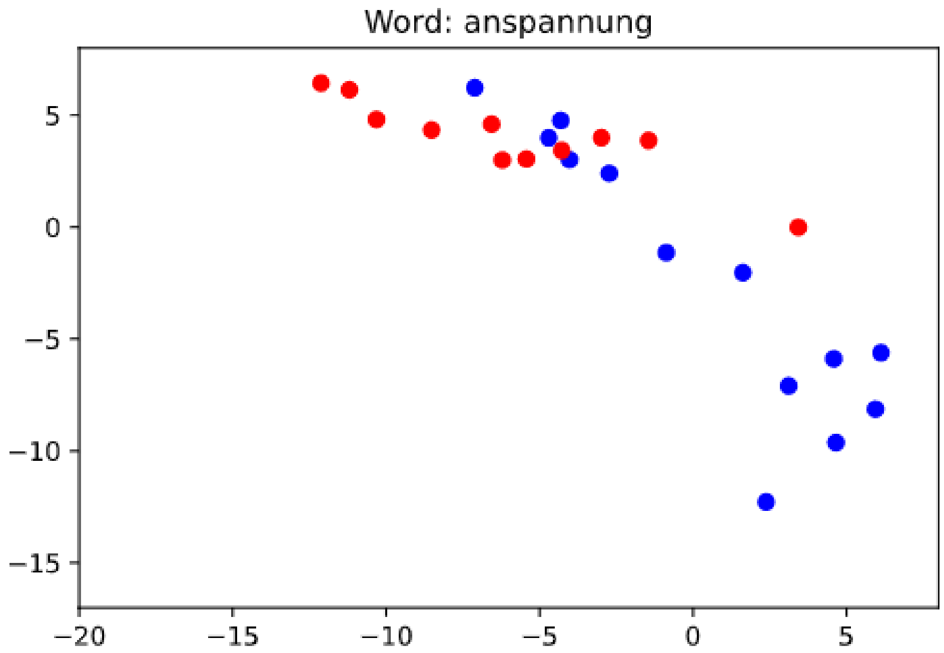


Figure 6.5: t-SNE plot for the word-based acoustic embeddings of the German word "Anspannung" (English: tension). Red dots represent tokens marked as EC while blue are non EC.

### 6.2.4 Discussion

With a strong lexical baseline and the promising results from previous experiment (6.1.3), we were convinced that ordinary fusion strategies would help to improve our results. The high recall on the acoustic ResNet18 system was encouraging. However, the results for EF and LF experiments suggest that simply adding the acoustic representations, extracted from the ResNet, adds a lot of entropy that the system in its current architecture can't handle, yielding worse accuracy than the textual system.

The analysis of the extracted representations and our knowledge about the existence of acoustic cues led us to explore heuristic ways to combine the modalities. The final DLF experiments show that the accuracy of the lexical model with its knowledge about context and content of a narrative could be improved by relying on local acoustic information in case of uncertainty.

Combining acoustic and lexical modalities yields higher accuracy than the uni-modal approaches to this difficult task if done the right way. We could show that local acoustic information alone is not reliable to detect EC but helps to improve results when combined with a text-based system that captures long-range semantic relations.

Another issue we find with experiments is the difference between the EC identification strategies used for the human annotation and the automated system. The annotation was performed only using the cues from text, whereas in the automated system we tried to predict the same ECs using cues from speech and text. To resolve this difference and to reduce the complexity of the annotation, we come up with a modified human annotation strategy as explained in Chapter 5, in Sect 5.2, which provides access to both speech and text while performing the human annotation.

## 6.3 Valence and Emotion Carrier Detection from Written Personal Narratives

In this section, we build automated systems for the detection of valence and ECs for each functional unit of the PNs from CBT-PHA dataset (of written narratives) (4.1.3) annotated using Protocol-3 (5.3). This dataset and thus experiment is different from previous experiments, as the annotators/model is provided with an automatically preselected list of EC-candidates for each functional unit. Also, the level of context considered in the annotation as well as in the automated system is of function unit, without access to the other parts of the PNs.

[This work was done in collaboration with and lead by S. Mahed Mousavi [93] based on the outcome of our annotation experiment explained in Sect-5.3.]

### Valence Prediction

[The valence prediction part of the experiment was performed by *Roccabruna et al.* [126] ]

We summarize the experiments we performed to predict valence of the narrator from the written PNs collected in the CBT-PHA dataset (4.1.3, annotated by humans in sect-5.3. Similar to the annotation protocol, we trained a valence prediction model to predict the polarity at the level of functional units. The model is based on the AlBERTo architecture [114] with a three-heads output layer, instead of the original two-heads fully connected layers, to predict the valence of each FU over the 3-label output space of *negative, positive* and *neutral.* {+1,+2} scores were merged into 'positive' class, {-1,-2} scores were merged into the 'negative' class, whereas the 0 score is represented as the 'neutral' class We split the training set of the SENTIPOLC16 dataset [12][2] into training and validation sets of 90% and 10%, in a stratified manner. We then used the training set to fine-tune the model in the first step, and the validation set in the next step for hyper-parameter optimization and selecting the best model using the Optuna framework [1]. Using the obtained hyper-parameters[3], the model was then further fine-tuned on the CBT-PHA dataset of with annotated functional units extracted from PNs. The results of these experiments are presented in Table 6.6.

### Emotion Carrier Detection

We trained a baseline model to assess the EC annotation on the CBT-PHA dataset for the task of EC detection. The approaches used in previous experiments 6.1 6.2 do not fit with our case, since the annotators were asked to select the EC from a predefined set of candidates, rather than

[2] SENTIPOLC16 is a dataset of tweets in the Italian language

[3] `learning_rate=6.599e-05, weight_decay=0.0215, warmup_steps=0.899, num_epochs=11`

| **Model** | **F1** | **Prec.** | **Rec.** |
|---|---|---|---|
| *AlBERTo_SP16* | 0.64 | 0.63 | 0.70 |
| *AlBERTo_opt_SP16* | 0.63 | 0.62 | 0.71 |
| *AlBERTo_opt_SP16+PN* | **0.76** | **0.76** | **0.76** |

Table 6.6: Macro F1, Precision, and Recall of the sentiment prediction models optimized in different settings. *AlBERTo_SP16* is the vanilla AlBERTo model fine-tuned on SENTIPOLC16; *AlBERTo_opt_SP16* is the model optimized utilizing validation split; and *AlBERTo_opt_SP16+PN* is the *AlBERTo_opt_SP16* further fine-tuned on the training set of CBT-PHA dataset. All evaluation results are obtained using the test split of the PNs dataset.

selecting any token-span in the text. Thus, in our case the model is tasked to classify each EC-candidate span as EC or non-EC.

The first part of the architecture computes the tokens embedding of each functional unit. Afterwards, we extract the encoded representation of the EC-candidate tokens and perform max-pooling, which takes the maximum value for every dimension of the vector encoding, producing the vector representation of the EC-candidate. The vector representation is then given as input to the classification layer (dense layer + softmax) yielding the probability distribution over the EC and non-EC classes. To compute the embeddings, we experimented with bi-LSTM with attention and AlBERTo, a pre-trained BERT-based model for the Italian language [114]. In the experiments with the AlBERTo model, we experimented concatenating the representation of the $[CLS]$ token with the EC-candidate representation, to better consider the context during the classification.

The results of these experiments, summarized in Table 6.7, indicate that the outperforming baseline combination is obtained by using the AlBERTo model for the input representation with the concatenation of the $[CLS]$ token.

| **Model** | **F1** | **Prec.** | **Rec.** |
|---|---|---|---|
| *bi-LSTM + attn.* | 0.66 | 0.70 | 0.66 |
| *AlBERTo Emb.* | 0.69 | 0.69 | 0.69 |
| *AlBERTo Emb.+*$[CLS]$ | **0.70** | **0.70** | **0.70** |

Table 6.7: Results of EC Detection experiments on the test set. All scores are measured with the "macro" average strategy. The AlBERTo-based architecture with the concatenation of $[CLS]$ token achieves the best performance.

## 6.4 Joint Learning of Valence and Emotion Carriers Detection from Written PNs

In the previous experiment (6.3), we explored AlBERTo (a BERT-based model) for identification of both Valence and Emotion Carriers separately. In this section we explore possibility of jointly training both the tasks. Our hypothesis is that if the valence detection and emotion carriers detection tasks are related, they can inform each other in a multi-task model setting. Moreover, via joint training we would get a further confirmation of EC-based explanation of the valence prediction. The EC detection involves classification of each EC-candidate from a list provided with each utterance (as Yes - candidate is an EC or No - candidate is not an EC), whereas the valence prediction task is a classification task to identify one of the three valence classes (positive, negative, and neutral) for the utterance. It is difficult to classify all EC candidates and the valence, at once, using a BERT based architecture, as explored in the Section 6.3. We explore GePpeTto [32] (a GPT based model for Italian) generative model to test our hypothesis. We first train GePpeTto based models for individual tasks of EC detection and Valence prediction separately. Later, we train a joint model to study the performance gain.

### 6.4.1 Emotion Carriers Detection

For the EC detection task, we first generate strings concatenating the inputs and outputs in a specific format. Once we have the data represented in the form of strings, we fine-tune the GePpeTto language model on these strings. The input and output makes use of different separator tokens for different purpose and are represented in the following format:

```
Input:<|begin|> Functional Unit <|endoffu|> candidate-1 <|endofcand|>
candidate-2 <|endofcand|>...candidate-n <|endofcand|> <|endofinp|>
Output: label-1 label-2 ... label-n <|endoftext|>
```

For better understanding, an actual example would look like:

```
 Input: <|begin|> e ho provato nausea <|endofsent|> ho provato
<|endofcand|> nausea <|endofcand|> <|endofinp|>
Output: Y N <|endoftext|>
```

The functional unit is one segment from the narrative whereas the candidate-i are the corresponding EC-candidates. The label-i represents the true class of the candidate-i i.e. Y for Yes or N for No. $< |begin| >, < |endoffu| >$ $, < |endofcand| >, < |endofinp| >$ are the separators defined to separate different parts of the input and output. In the example, the segment or functional unit from a narrative is *'e ho provato nausea'*, while there are two EC-candidates *'ho provato'* and *'nausea'*. The corresponding labels for the EC-candidates are Y (tes) and N (No). The input and output string are concatenated before using them for fine-tuning the Language Model (LM). Once the language model is fine-tuned on this data, we use the input strings as prompts and task the LM to predict generate output strings. We then compare the generated output with the ground truth EC-candidate labels and evaluate the performance. We obtain Precision, Recall and F1 scores of 0.58, 0.57, and 0.58 respectively, calculated using macro average strategy, as presented in Table 6.8.

### 6.4.2 Valence Prediction

Similar to the EC detection model, we first represent the input and output related to the valence prediction task in a string format and fine-tune the GePpeTto language model on the training set. The Input and the Output are represented in the format:

```
Input:<|begin|> Functional Unit <|endoffu|>
Output: Valence <|endoftext|>
```

An actual example would look like:

```
Input: <|begin|> e ho provato nausea <|endoffu|>
Output: neg <|endoftext|>
```

In the input there is only the functional using and the output consists of the corresponding valence class pos (positive), neg (negative), or neu (neutral). The input and output string are concatenated before using them for fine-tuning the Language Model (LM). Once the language model is fine-tuned on this data, we use the input strings as prompts and task the LM to predict generate output strings. We then compare the generated output with the ground truth valence classes and evaluate the performance. We obtain Precision, Recall and F1 scores of 0.60, 0.59, and 0.57 respectively, calculated using macro average strategy.

### 6.4.3 Joint Model for EC Detection and Valence Prediction

To assess the possibility of the two tasks helping each other in a multi-task setting, we explore combining input and output representations from both tasks, fine-tune the GePpeTto LM and evaluate the two tasks based on the

| | **Task** | **Prec** | **Recall** | **F1 score** |
|---|---|---|---|---|
| **independent** | EC Detection | 0.58 | 0.57 | 0.58 |
| **independent** | Valence Prediction | 0.60 | 0.59 | 0.57 |
| **Joint-learning** | EC Detection | 0.56 | 0.55 | 0.56 |
| | Valence Prediction | 0.75 | 0.65 | 0.68 |

Table 6.8: Evaluation of the generative-based EC Detection and Valence Prediction tasks, independently and in a joint-learning setting. It can be observed that Valence Prediction is benefiting greatly from the additional context of EC candidates, in the joint setting. The EC Detection on the other hand is not able to gain benefit from the valence context.

output generated. One strong cue from valence prediction useful for EC detection could be in the case of neutral valence. In the joint setting, the model may learn that in the case of neutral valence, the output label for all EC-candidates should be N (No), as the annotation of EC candidates was performed only on segments with non-neutral valence. The Inputs and the Outputs are combined from the two tasks and represented in the format:

```
Input:<|begin|> Functional Unit <|endoffu|> candidate-1 <|endofcand|>
candidate-2 <|endofcand|>...candidate-n <|endofcand|> <|endofinp|>
Output:Valence <|endofval|> label-1 label-2 ... label-n <|endoftext|>
```

A real example would look like:

```
Input: <|begin|> e ho provato nausea <|endofsent|> ho provato
<|endofcand|> nausea <|endofcand|> <|endofinp|>
Output: neg <|endofval|> Y N <|endoftext|>
```

Note that the input is same as that of the EC detection task, whereas in the output, the valence information and the EC output labels are concatenated using the $< |endofval| >$ separator token. Similar to the individual tasks, the input and output is concatenated and the LM is fine-tuned on the concatenated strings.

The input is used as a prompt to LM and generated output is compared with the expected output. We evaluate the outputs corresponding to the two tasks, separately. As shown in Table 6.8, for the valence prediction task, we obtain Precision, Recall and F1 scores of 0.75, 0.65, and 0.68 respectively, calculated using macro average strategy. We observe an increment of 0.11 in the F1 score as compared to the single task setting. Whereas for the EC detection task, we obtain Precision, Recall and F1 scores of 0.56, 0.55, and 0.56 respectively, calculated using macro average strategy. The F1 score is dropped by 0.02 when compared with the independent model. Overall, we find that the joint modeling significantly improves the performance of valence prediction while slightly damages the performance of the EC detection. The context of EC-candidates which represent entities and events from the segment is present in the joint modeling but absent in the standalone task of valence prediction. This additional focus on the context might have helped in the valence prediction as an EC-based explanation. Whereas the additional context of valence doesn't seem to help the EC detection.

# Chapter 7

# Correlation Between Emotion Carriers and Valence - an Explainability Study

In the previous chapter 6, we built different automatic systems for detecting Emotion Carriers and Valence, using outcomes of different annotation experiments performed in Chapter 5. We also performed quantitative and qualitative evaluation and analysis of these automated systems. In this chapter, with a deep analysis of DNN based model from section 6.3, we try to find the correlation between Emotion Carriers and Valence, specifically to check whether these concepts capture the same information or if the information captured is complementary. We perform the analysis on the CBT-COADAPT dataset (4.1.3), using the human-annotated ECs from the annotation experiment (5.3) and the automatic valence prediction using the model 6.3.

[This work was published in WASSA workshop in ACL-2022, in collaboration with and lead by S. Mahed Mousavi and Gabriel Roccabruna [93].]

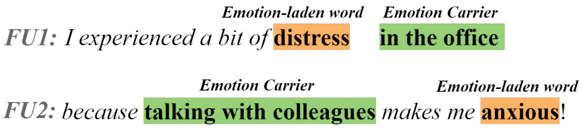

Emotion-laden word
Emotion Carrier
FU1: I experienced a bit of distress in the office
Emotion Carrier
Emotion-laden word
FU2: because talking with colleagues makes me anxious!

Figure 7.1: Example of a sentence consisting of two Functional Units (FU1, FU2), the basic units of annotation. Emotion-laden words in each Functional Unit manifest a sentiment explicitly while Emotion Carriers describe the events, persons or objects conveying emotions.

## 7.1 Overview

We defined the Emotion Carriers (3.2.1) PNs as the speech or text spans that explain and carry the emotions felt by the narrator during the recollection. In this chapter, we analyze and compare the tokens that help the automated system trained to identify valence of functional units (FU) from PNs (sect 6.3) with the token spans selected by human annotators as ECs (that explain the valence) (5.3). We identify the tokens that contribute to the model's prediction according to their attributions given by Integrated Gradients [140], an Explainable-AI technique. Our comparative analysis shows the human annotator identifies the tokens that explain an event or its participants as the carrier of emotions and valence, which clearly convey the activation of the emotional state in the narrator, even though they are not *explicitly* manifesting an emotion. Meanwhile, the DNN model bases its decision mostly on a limited set of tokens which belong to the category of emotion-laden words (see Figure 7.1 for an example).

## 7.2 Valence Prediction Decision Explainability

We investigate the explainability of the automatic valence prediction by comparing the tokens influencing the prediction with those selected by the human judges as ECs. In order to detect the tokens crucial to the model's prediction, we use the attribution assigned to each token by the Integrated Gradients [140] technique. Integrated Gradients (IntGrad) is an attribution method for Explainable AI which builds on top of the classic backward gradient analysis. Given our valence prediction model $f(FU)$, where FU is the functional unit $FU = \{w_1, w_2, .., w_n\}$ and $w_i \in R^d$ are the token embeddings , the backward gradient is given by:

$$BackwardGrad_j(w_i) = \frac{\partial f}{\partial w_{ij}} \tag{7.1}$$

measuring how much perturbing the input token $w_i$ by an infinitesimal amount along dimension $j$ affects the output of function $f$. The IntGrad method extends this by computing the integral of the derivative along the path connecting a baseline token $w'$, which is a neutral element, to the input point $w$:

$$IntGrad_j(w_i) = (w_{ij} - w') \int_{\alpha=0}^{1} \frac{\partial f(w' + \alpha(w_{ij} - w'))}{\partial w_{ij}} \, d\alpha \tag{7.2}$$

where $\alpha \in [0, 1]$ draws a linear path, from the baseline token to the input token, along which the gradients are integrated. In our studies, we used a zero vector for the baseline token $w'$, and the open-source library Captum [74] for efficient IntGrad computation. In cases that a token is split into several sub-tokens by the tokenizer of our model [76], we average the Integrated Gradients attributions of the subtokens, to get the attribution of the whole token.

### 7.2.1 Token Analysis based on IntGrad Attributions

Using the test set samples for which the model predicts the valence polarity correctly, we employ two approaches regarding the explainability analysis. In the first approach, we extract the tokens influential or crucial to the prediction process of the model based on their Integrated Gradients (IntGrad) attributions, and study whether or not they belong to the spans annotated as EC by the human annotator.

In order to identify tokens crucial to the model's prediction we experimented with two different thresholds for the IntGrad attribution:

- **Greater than 0 (G0)**: This baseline is based on the fact that each token with a positive IntGrad attribution value has a positive influence on the prediction. Nevertheless, tokens with small IntGrad attributions have a marginal contribution and thus they are noisy for our analysis;

- **Lower Bound (LB)**: This threshold is obtained uniquely for each FU and is measured by consecutively masking each token in the FU, with a zero-vector embedding, in a descending order of IntGrad attributions until a change in the polarity prediction is observed. The IntGrad attribution of the last masked out token is then selected as the LB threshold.

The results of this analysis using the two mentioned threshold policies are presented in Table 7.1 and Figure 7.2. The analysis indicates that although 67.9% of the EC tokens (tokens in ECs selected by human annotators) have a positive contribution to the model's prediction, more than 60% of the tokens with an attribution above the thresholds do not overlap with the EC tokens. Nevertheless, the majority of EC tokens with an attribution higher than the thresholds are EC-heads, regardless of the threshold policy. Furthermore, the distributions of the Content Words (CW), i.e. nouns, verbs and adjectives, confirm our previous assumption that **G0** threshold is noisy since 54% of tokens above this threshold are non-CWs, while this number is smaller than 20% for the tokens with an IntGrad attribution higher than the **LB**. The CWs in LB and G0 groups

| **Threshold (Thr.)** | **G0** | **LB** |
|---|---|---|
| *#Tokens with*<br>*IntGrad A.>Thr.* | 482<br>(46% CW) | 109<br>(81% CW) |
| *#Tokens w. IntGrad A.>Thr.*<br>*in EC-span* | 141<br>29.3% | 43<br>39.5% |
| *#Tokens w. IntGrad A.>Thr.*<br>*that are EC-heads* | 82<br>18.1% | 32<br>29.3% |

Table 7.1: The analysis of tokens influencing the model's prediction based on two different policies for the IntGrad attribution (IntGrad A.), namely Greater than 0 (G0) and Lower Bound (LB). Regardless of the threshold policy, the tokens inside the EC-span that contribute to the model's prediction are less than 40%.

are distributed as 52% nouns, 27% verbs, 21% adjectives, and 47% nouns, 40% verbs and 13% adjectives, respectively.

In the next step, we further analyzed the valence polarity distribution of CWs by using the OpeNER[1] lexicon-based sentiment model. The results, presented in Table 7.2, show that the percentage of non-neutral CWs in the ECs is less than 5%, while more than 40% of the influential tokens, i.e. tokens with attributions over the **LB** threshold, represent a positive or negative polarity. This remarks the importance of emotion-laden words, such as *anxiety*, *fear* and *worry*, for the model in predicting the valence, and suggests that the model mostly focuses on the tokens that explicitly convey emotions, and the ECs (as the implicit manifestations of emotions) are less significant in its decision process.

### 7.2.2 Contribution of ECs to the Model's Decision

For the second approach, we evaluate the influence of the ECs selected by the human annotators in the decision process of the model. For this

[1] `https://www.opener-project.eu/`, This publicly available lexicon was semi-automatically created starting from 1,000 manually controlled keywords

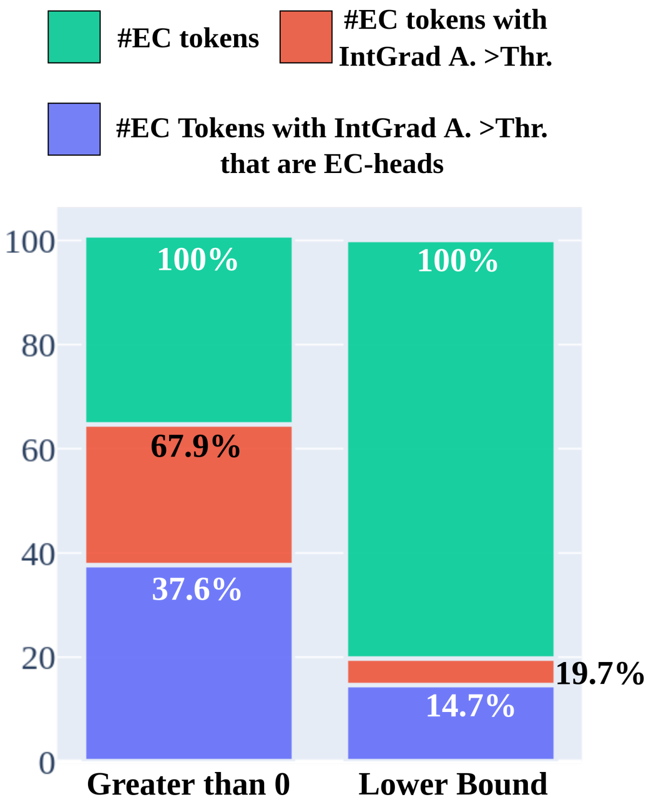


Figure 7.2: The percentage of the tokens in EC-spans with an Integrated Gradient attribution (IntGrad A.) higher than the threshold (Thr.). The majority of EC tokens with an attribution higher than the Lower Bound are EC-heads.

purpose, we mask out the EC-span in the Functional Unit with the highest IntGrad attribution, and measure the drop in the confidence score for the initially predicted valence polarity. The confidence score represents the probability assigned by the model to a given class, which in our case the classes can be either *positive* or *negative*. In the next step, we extend this analysis to the token level and measure the drop in the confidence score caused by masking out the EC-head with the highest IntGrad attribution, as well as all EC-heads present in the corresponding FU.

The results, shown in Table 7.3, present the strong contribution of emotion-laden words that explicitly manifest the sentiment on the model's decision. Furthermore, the confidence drop caused by masking the EC-span

| Token set | Positive | Negative | Neutral |
|---|---|---|---|
| *CW in G0* | 3.9% | 13.6% | 82.5% |
| *CW in LB* | 10.3% | 29.9% | 59.8% |
| *CW EC tokens* | 0.7% | 4.0% | 95.3% |

Table 7.2: The polarity distribution of the Content Words (CW) with IntGrad attribution higher than the different thresholds. The results indicate that the majority of CWs in EC tokens are neutral and they do not represent any emotions explicitly. The polarity was retrieved using the OpeNER sentiment lexicon for the Italian language.

| Masked Content in FU | Conf. Score Drop |
|---|---|
| *EC-Span w. highest IntGrad A.* | 0.15 |
| *EC-Head w. highest IntGrad A.* | 0.09 |
| *EC-Heads in FU* | 0.14 |
| *Token w. highest IntGrad A.* | 0.55 |
| *Emotion-laden Words* | 0.36 |

Table 7.3: The drop in the confidence score of the predicted polarity caused by masking out selected contents in Functional Units. The results show that the Emotion-laden words have a stronger influence than the tokens selected as ECs by the human annotator.

is higher than masking only the head of the corresponding EC, suggesting that all the tokens in the EC-span contribute to the prediction confidence. However, the highest drop is achieved by masking the most influential token (the token with the highest IntGrad attribution) and emotion-laden words, respectively. These results once again support the findings of the previous analysis, suggesting the importance of tokens that explicitly manifest an emotion in the decision process of the model.

## 7.3 Findings

In this experiment we studied whether the valence prediction decision of DNN models can be explained by Emotion Carriers, spans of text that convey and carry emotions. We have focused our study on Personal Narratives which encompass real-life events and experiences that activate the emotional state of the narrator. The valence prediction model is based on AlBERTo architecture [114], whereas the EC annotation is performed by human annotators. We have investigated whether the decision of the model is based on the Emotion Carriers that the human annotator selected to explain the valence of the text. We find that the model focuses on explicit emotion words such as emotion laden words to base it's decision, however the Emotion Carriers contain words describing actions and events that have activated the emotional state of the narrator. The findings suggest that the valence and EC capture different and complementary emotion information from the PNs and thus the two concepts when combined provide a deeper and fine-grained emotion analysis of the PNs and the narrators.

# Chapter 8

# Conclusion and Future Work

We find conventional emotion analysis systems mainly focus on emotion state detection, which only captures the surface emotions from a predefined set of categories (sad, happy) or numeric values on dimensions (valence, arousal), but the semantics of the emotions such as what triggered the emotions, the actions and the entities that manifest the emotions are not identified. We introduced a new concept of Emotion Carriers (EC) that captures semantics of emotions by providing explanations for the emotional state. We proposed that the Emotion State and Emotion Carriers combined provide deep and fine-grained emotion analysis. We explore this analysis in an important but infancy domain of Personal Narratives (PN). We explored different datasets of PNs collected for the purpose of emotion analysis. We enriched the datasets with the human annotation of ECs and emotion state (in terms of valence) with different annotation protocols. We first found the annotation task to be complex and subjective, but the complexity was reduced with the steps such as breaking the task into smaller parts, providing speech of the narrator while performing the annotation. We also built speech-based, text-based, and multimodal automated valence and EC detection systems making use of the annotated data. The high performance of the automated systems prove their usability in the downstream application such mental well-being agents.

We see two direct applications of our Emotion analysis for the future works.

First, in the context of conversational mental well-being applications. As already presented in the thesis, a potential application of fine-grained emotion understanding is to detect the user's emotional state and emotion carriers that manifest the emotional state and generate an automated response using this information to elicit more relevant information from the user, relevant for better understanding of the emotional state. This information can further be summarized and provided to the therapist or caregivers.

Second application is applicable for longitudinal experiments, similar to the CBT-PHA dataset. As we have observed, PNs often include mentions of other people. We plan to leverage this and emotion information to identify the relation dynamics of the narrator and the person, over the time. The characteristics that we study may include how the valence of the narrator changes when s/he mentions the person's name, by identifying the valence at the functional unit level. If the trend is towards negative valence, and also the mention of the person is an EC, it could be a strong signal that the person is affecting the user negatively. This may help the therapist to be able to provide a proper advice, accordingly. (Ex. to try and avoid that person). Also the associated ECs from those functional units may help understand how that person is affecting the user. (Ex. by being unreasonably strict at work.) Whereas a positive trend could mean that the person is affecting the user positively. And, if appropriate, the therapist may suggest the user to try and interact more with that person. A relationship status, at any given point in time, would represent the state of a relation using the emotional valence associated with the mentions of the person. Another extension could be to use the valence trajectory of the relation state to predict the future states.

# Bibliography

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] Yaara Benger Alaluf and Eva Illouz. Emotions in consumer studies. *The Oxford Handbook of Consumption*, page 239, 2019.

[3] Hassan Alhuzali and Sophia Ananiadou. Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 339–347, 2019.

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

[5] Felwah Alqahtani and Rita Orji. Insights from user reviews to improve mental health apps. *Health informatics journal*, 26(3):2042–2066, 2020.

[6] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from

speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[7] Joaquin A Anguera, Joshua T Jordan, Diego Castaneda, Adam Gazzaley, and Patricia A Areán. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. *BMJ innovations*, 2(1), 2016.

[8] Lynne E Angus and John McLeod. *The handbook of narrative and psychotherapy: Practice, theory and research.* Sage, 2004.

[9] Mario Ezra Aragón, Adrián Pastor López Monroy, Luis Carlos González-Gurrola, and Manuel Montes. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 1481–1486, 2019.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[11] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[12] Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*, 2016.

[13] Sebastian P. Bayerl, Aniruddha Tammewar, Korbinian Riedhammer, and Giuseppe Riccardi. Detecting emotion carriers by combining acoustic and lexical representations. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 31–38. IEEE, 2021.

[14] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[15] Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451, 2008.

[16] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. 2022.

[17] Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer, 2017.

[18] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[19] Rafael A Calvo and Sunghwan Mac Kim. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543, 2013.

[20] Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the workshop on natural lan-*

*guage processing in the 5th information systems research working days (JISIC)*, pages 37–43, 2014.

[21] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, 2008.

[22] Snigdha Chaturvedi, Dan Goldwasser, and Hal Daume III. Ask, and shall you receive? understanding desire fulfillment in natural language text. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[23] Danqi Chen. *Neural Reading Comprehension and Beyond*. PhD thesis, Stanford University, 2018.

[24] Ming Chen and Xudong Zhao. A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. In *Proc. Interspeech 2020*, pages 374–378, 2020.

[25] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660, 2018.

[26] Shammur Absar Chowdhury, Arindam Ghosh, Evgeny A Stepanov, Ali Orkan Bayer, Giuseppe Riccardi, and Ioannis Klasinas. Cross-language transfer of semantic annotation via targeted crowdsourcing. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[27] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech 2018*, pages 1086–1090, 2018.

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

[29] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[30] Neiberg Daniel, Elenius Kjell, and Laskowski Kornel. Emotion recognition in spontaneous speech using gmms. In *Proceedings of the 9th ISCA International Conference on Spoken Language Processing*, 2006.

[31] Morena Danieli, Tommaso Ciulli, Seyed Mahed Mousavi, and Giuseppe Riccardi. A conversational artificial intelligence agent for a mental health care app: Evaluation study of its participatory design. *JMIR Form Res*, 5(12):e30053, Dec 2021.

[32] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. Geppetto carves italian into a language model, 2020.

[33] Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182, 2020.

[34] Haibo Ding and Ellen Riloff. Acquiring knowledge of affective events from blogs using label propagation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[35] Haibo Ding and Ellen Riloff. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018*

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, 2018.

[36] Haibo Ding and Ellen Riloff. Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[37] Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6343–6350, 2019.

[38] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, 2017.

[39] P EKMAN. Expression and the nature of emotion. *Approaches to emotion*, pages 319–343, 1984.

[40] Paul Ekman. Are there basic emotions? 1992.

[41] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[42] Teuntje R Elfrink, Sytse U Zuidema, Miriam Kunz, and Gerben J Westerhof. Life story books for people with dementia: a systematic review. *International psychogeriatrics*, 30(12):1797–1811, 2018.

[43] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[44] Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. A knowledge regularized hierarchical approach for emotion cause analysis. 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)*, pages 5614–5624, 2019.

[45] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

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

[47] Joseph L Fleiss. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, pages 651–659, 1975.

[48] Liye Fu, Jonathan P Chang, and Cristian Danescu-Niculescu-Mizil. Asking the right question: Inferring advice-seeking intentions from personal narratives. In *NAACL-HLT (1)*, 2019.

[49] Kai Gao, Hua Xu, and Jiushuo Wang. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications: An International Journal*, 42(9):4517–4528, 2015.

[50] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[51] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer, 2015.

[52] Arindam Ghosh, Evgeny A Stepanov, Morena Danieli, and Giuseppe Riccardi. Are you stressed? detecting high stress from user diaries. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000265–000270. IEEE, 2017.

[53] Eva Gjengedal, Sissel Lisa Storli, Anny Norlemann Holme, and Ragne Sannes Eskerud. An act of caring–patient diaries in norwegian intensive care units. *Nursing in Critical Care*, 15(4):176–184, 2010.

[54] Raman Goel, Sachin Vashisht, Armaan Dhanda, and Seba Susan. An empathetic conversational agent with attentional mechanism. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4, 2021.

[55] Yuan Gong and Christian Poellabauer. Topic modeling based multimodal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 69–76, 2017.

[56] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480, 2019.

[57] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific, 2016.

[58] Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. World Scientific, 2018.

[59] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages. 2014.

[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[61] Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, and Anat Rafaeli. Predicting customer satisfaction in customer support conversations in social media using affective features. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 115–119, 2016.

[62] Thomas Holtgraves. Text messaging, personality, and the social context. *Journal of research in personality*, 45(1):92–99, 2011.

[63] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[64] George S Howard. Culture tales: A narrative approach to thinking, cross-cultural psychology, and psychotherapy. *American psychologist*, 46(3):187, 1991.

[65] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 2005.

[66] Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106, 2018.

[67] Alexei V Ivanov, Giuseppe Riccardi, S Ghosh, S Tonelli, and E A Stepanov. Acoustic Correlates of Meaning Structure in Conversational Speech. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, page 4, 2010.

[68] Richard Johansson and Alessandro Moschitti. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 67–76, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[69] Philip Nicholas Johnson-Laird and Keith Oatley. The language of emotions: An analysis of a semantic field. *Cognition and emotion*, 3(2):81–123, 1989.

[70] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[71] Arunima Khunteta and Pardeep Singh. Emotion cause extraction - a review of various methods and corpora. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 314–319, 2021.

[72] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[73] Jesper Kjeldskov and Connor Graham. A review of mobile hci research methods. In *International Conference on Mobile Human-Computer Interaction*, pages 317–335. Springer, 2003.

[74] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

[75] Stefanie Kremer and Louise den Uijl. Studying emotions in the elderly. In *Emotion measurement*, pages 537–571. Elsevier, 2016.

[76] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

[77] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[78] Manoj Kumar, Rahul Gupta, Daniel Bone, Nikolaos Malandrakis, Somer Bishop, and Shrikanth S. Narayanan. Objective Language Feature Analysis in Children with Neurodevelopmental Disorders During Autism Assessment. In *Proc. Interspeech 2016*, pages 2721–2725, 2016.

[79] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

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

[81] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.

[82] Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218, 2019.

[83] Xiangju Li, Wei Gao, Shi Feng, Daling Wang, and Shafiq Joty. Span-level emotion cause analysis by bert-based graph attention network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3221–3226, 2021.

[84] Yuting Lin, Carina Tudor-Sfetea, Sarim Siddiqui, Yusuf Sherwani, Maroof Ahmed, Andreas B Eisingerich, et al. Effective behavioral changes through a digital mhealth app: Exploring the impact of hedonic well-being, psychological empowerment and inspiration. *JMIR mHealth and uHealth*, 6(6):e10024, 2018.

[85] Nikolaos Lykousas, Constantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. Sharing emotions at scale: The vent dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 611–619, 2019.

[86] Paul Henry Lysaker, John Timothy Lysaker, and Judith Thompson Lysaker. Schizophrenia and the collapse of the dialogical self: Recovery, narrative and psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, 38(3):252, 2001.

[87] Jana Machajdik, Allan Hanbury, Angelika Garz, and Robert Sablatnig. Affective computing for wearable diary and lifelogging systems: An overview. In *Machine Vision-Research for High Quality Processes and Products-35th Workshop of the Austrian Association for Pattern Recognition. Austrian Computer Society*, pages 2447–2456, 2011.

[88] Benjamin Milde and Arne Köhn. Open source automatic speech recognition for german. In *Proceedings of ITG 2018*, 2018.

[89] Saif Mohammad and Svetlana Kiritchenko. Using nuances of emotion to identify personality. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 27–30, 2013.

[90] Claude Montacié and Marie-José Caraty. Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, pages 541–545, 2018.

[91] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

[92] Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. Would you like to tell me more? generating

a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, 2021.

[93] Seyed Mahed Mousavi, Gabriel Roccabruna, Aniruddha Tammewar, Steve Azzolin, and Giuseppe Riccardi. Can emotion carriers explain automatic sentiment prediction? a study on personal narratives. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 62–70, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[94] Deniece S Nazareth, Michel-Pierre Jansen, Khiet P Truong, Gerben J Westerhof, and Dirk Heylen. Memoa: Introducing the multi-modal emotional memories of older adults database. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 697–703. IEEE, 2019.

[95] Deniece S. Nazareth, Ellen Tournier, Sarah Leimkötter, Esther Janse, Dirk Heylen, Gerben J. Westerhof, and Khiet P. Truong. An Acoustic and Lexical Analysis of Emotional Valence in Spontaneous Speech: Autobiographical Memory Recall in Older Adults. In *Proc. Interspeech 2019*, pages 3287–3291, 2019.

[96] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

[97] Paula Niedenthal, Catherine Auxiette, Armelle Nugier, Nathalie Dalle, Patrick Bonin, and Michel Fayol. A prototype analysis of the french category "émotion". *Cognition and Emotion*, 18(3):289–312, 2004.

[98] PAULA M NIEDENTHAL. Emotion concepts. *EMOTIONS*, page 587.

[99] Keith Oatley and Philip N Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50, 1987.

[100] Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, 2020.

[101] Horea-Radu Oltean, Philip Hyland, Frédérique Vallières, and Daniel Ovidiu David. An empirical assessment of rebt models of psychopathology and psychological health in the prediction of anxiety and depression symptoms. *Behavioural and cognitive psychotherapy*, 45(6):600–615, 2017.

[102] Desmond C Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594, 2019.

[103] Charles E Osgood. On the whys and wherefores of e, p, and a. *Journal of personality and social psychology*, 12(3):194, 1969.

[104] VandaLucia Zammuner University of Padua IV. Concepts of emotion:" emotionness", and dimensional ratings of italian emotion words. *Cognition & emotion*, 12(2):243–272, 1998.

[105] A. Paeschke, Miriam Kienast, and W. Sendlmeier. F0-contours in emotional speech. *Psychology*, 1999.

[106] Aneta Pavlenko. Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and cognition*, 11(2):147–164, 2008.

[107] James W Pennebaker. Linguistic inquiry and word count: Liwc 2001.

[108] James W Pennebaker. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162–166, 1997.

[109] James W Pennebaker and Janel D Seagal. Forming a story: The health benefits of narrative. *Journal of clinical psychology*, 55(10):1243–1254, 1999.

[110] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[111] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[112] Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustin Gravano. Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6484–6488, Barcelona, Spain, May 2020. IEEE.

[113] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, 2016.

[114] Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. AlBERTo: Italian BERT Language Un-

derstanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR, 2019.

[115] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[116] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[117] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[118] Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California, June 2016. Association for Computational Linguistics.

[119] Sarah D Pressman and Sheldon Cohen. Positive emotion word use and longevity in famous deceased psychologists. *Health Psychology*, 31(3):297, 2012.

[120] Sreeja PS and G Mahalakshmi. Emotion models: a review. *International Journal of Control Theory and Applications*, 10:651–657, 2017.

[121] Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. Modelling protagonist goals and desires in first-

person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369, 2017.

[122] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, 2018.

[123] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, 2018.

[124] Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister. How did you like 2017? detection of language markers of depression and narcissism in personal narratives. In *Proc. Interspeech 2018*, pages 3388–3392, 2018.

[125] Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. State of mind: Classification through self-reported affect and word use in speech. In *Proc. Interspeech 2018*, pages 267–271, 2018.

[126] Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. Multi-source multi-domain sentiment analysis with bert-based models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 581–589, Marseille, France, June 2022. European Language Resources Association.

[127] Aubrey J Rodriguez, Shannon E Holleran, and Matthias R Mehl. Reading between the lines: The lay assessment of subclinical depression from written self-descriptions. *Journal of personality*, 78(2):575–598, 2010.

[128] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[129] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26, 2018.

[130] Samiha Samrose, Kavya Anbarasu, Ajjen Joshi, and Taniya Mishra. Mitigating boredom using an empathetic conversational agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

[131] Diego Sarracino, Giancarlo Dimaggio, Rawezh Ibrahim, Raffaele Popolo, Sandra Sassaroli, and Giovanni M Ruggiero. When rebt goes difficult: applying abc-def to personality disorders. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 35(3):278–295, 2017.

[132] Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B. Pokorny, Eva-Maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, and Stefanos Zafeiriou. The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Proc. Interspeech 2018*, pages 122–126, 2018.

[133] Abigail J Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. Do life-logging technologies support memory for the past?: an experimental study using sensecam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90. ACM, 2007.

[134] Nancy Semwal, Abhijeet Kumar, and Sakthivel Narayanan. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–6. IEEE, 2017.

[135] Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, 2019.

[136] Jorge A Solís-Galván, Sodel Vázquez-Reyes, Margarita Martínez-Fierro, Perla Velasco-Elizondo, Idalia Garza-Veloz, and Claudia Caldera-Villalobos. Towards development of a mobile application to evaluate mental health: systematic literature review. In *International Conference on Software Process Improvement*, pages 232–257. Springer, 2020.

[137] Evgeny A Stepanov, Shammur Absar Chowdhury, Ali Orkan Bayer, Arindam Ghosh, Ioannis Klasinas, Marcos Calvo, Emilio Sanchis, and Giuseppe Riccardi. Cross-language transfer of semantic annotation via targeted crowdsourcing: task design and evaluation. *Language Resources and Evaluation*, 52(1):341–364, 2018.

[138] Jennifer A Sumner, James W Griffith, and Susan Mineka. Overgeneral autobiographical memory as a predictor of the course of depression: A meta-analysis. *Behaviour research and therapy*, 48(7):614–625, 2010.

[139] Xiao Sun, Chongyuan Sun, Changqin Quan, Fuji Ren, Fang Tian, and Kunxia Wang. Fine-grained emotion analysis based on mixed model for product review. *International Journal of Networked and Distributed Computing*, 5(1):1–11.

[140] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[141] Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Modeling User Context for Valence Prediction from Narratives. In *Proc. Interspeech 2019*, pages 3252–3256, 2019.

[142] Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Annotation of emotion carriers in personal narratives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1517–1525, Marseille, France, May 2020. European Language Resources Association.

[143] Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Annotation of emotion carriers in personal narratives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1517–1525, 2020.

[144] Aniruddha Tammewar, Alessandra Cervone, and Giuseppe Riccardi. Emotion Carrier Recognition from Personal Narratives. In *Proc. Interspeech 2021*, pages 2501–2505, 2021.

[145] Dengke Tang, Junlin Zeng, and Ming Li. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In *Proc. Interspeech 2018*, pages 162–166, 2018.

[146] Jenny A Thomas. *Meaning in interaction: An introduction to pragmatics.* Routledge, 2014.

[147] Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3975–3989, 2021.

[148] Svitlana Volkova and Yoram Bachrach. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578, 2016.

[149] Lynette P Vromans and Robert D Schweitzer. Narrative therapy for adults with major depressive disorder: Improved symptom and interpersonal outcomes. *Psychotherapy research*, 21(1):4–15, 2011.

[150] Anthony FC Wallace and Margaret T Carson. Sharing and diversity in emotion terminology. *Ethos*, pages 1–29, 1973.

[151] Gerben J Westerhof and Ernst T Bohlmeijer. Celebrating fifty years of research and applications in reminiscence and life review: State of the art and new directions. *Journal of Aging studies*, 29:107–114, 2014.

[152] Wikipedia contributors. Narrative — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Narrative&oldid=923338483`, 2019. [Online; accessed 16-April-2022].

[153] J Mark Williams and Keith Broadbent. Autobiographical memory in suicide attempters. *Journal of abnormal psychology*, 95(2):144, 1986.

[154] Teresa Wills and Mary Rose Day. Valuing the person's story: use of life story books in a continuing care setting. *Clinical Interventions in Aging*, 3(3):547, 2008.

[155] Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, 2019.

[156] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[157] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, 2017.

# Appendix A

# USOM-Elderly Examples Annotated with Valence and Emotion Carriers

In this appendix, we present some Personal Narratives (PN) from the USoM-Elderly dataset, annotated with Valence and Emotion Carriers (EC). Examples are presented in the Table format, where the first (left) column is the transcription of the PN in the original language, German, whereas the the second (right) column shows the translation of the PN into English.

The **Valence** on the bipolar scale, from -2 to +2 is represented by color-coding the text:

- red - negative **(-2)**
- orange - slightly negative **(-1)**
- gray - neutral **(0)**
- light green - slightly positive **(+1)**
- green - positive **(+2)**

The **emotion carriers** are wrapped in the parentheses.

**Note:** The examples are provided to give a sense of the USoM-elderly data and the annotation protocol for EC and valence. As the narratives are long, we show only the essential part required to understand the narrative structure and show the excluded part using "...".

| | |
|---|---|
| Ich war hier bei den **(Basketballern)**, da waren wir eine **(Clique)** von sieben, acht Basketballern, die auch zusammen Basketball gespielt haben und sich dann irgendwann auch mehr oder weniger um das Management gekümmert haben. ... Und das ging dann so weiter, dass wir dann tatsächlich **(deutscher Meister wurden)** mit den mit den Mädels. Zweimal sogar, dreimal **(deutscher Pokalsieger)**, **(Europapokal gespielt)** haben. Und dann kam halt die Situation, wo es **(finanziell)** ein bisschen eine **(Schräglage)** gab. Und da haben sich dann leider **(zwei Grüppchen gebildet)**. Bei diesen sieben, acht Menschen, die halt früher immer sehr freundschaftlich, eher sogar **(wie Brüder zusammengearbeitet)** haben, **(kam)** es dann tatsächlich **(zum Auseinanderdriften)**. ... Natürlich, wenn man sich gesehen hat, hat man mal hallo gesagt. Aber **(früher)** hatte man sich ja jeden Tag gesehen oder hat, wie das unter Freunden ist, **(viele Sachen zusammen gemacht)**, **(viel zusammen erlebt)**. Und es ist **(total auseinandergegangen)**, **(total auseinander)**. Also zu zwei, drei von **(diesen Menschen)** habe ich leider heutzutage überhaupt **(keine Beziehung mehr)**. ... Und was das Deprimierende ist, dass man vorher mit denen **(alles zusammen gemacht)** hat. Das waren **(Best Friends)**, wie man so schön sagt. ... Und **(vorbei)** ist es. | I was here with the **(basketball players)**, we were a **(clique)** of seven, eight basketball players who also played basketball together and then at some point also more or less took care of the management. ... And that continued in such a way that we then actually **(became German champions)** with the girls. Twice even, three times **(German Cup winner)**, **(played in the European Cup)**. And then the situation arose where there was a bit of a **(financially skew)**. And then, unfortunately, **(two groups formed)**. These seven or eight people, who used to **(work together)** very amicably, more **(like brothers)**, actually **(drifted apart)**. ... Of course, when we saw each other, we said hello. But **(in the past)**, you saw each other every day or, as is the case between friends, **(did a lot of things together)**, **(experienced a lot together)**. And it **(totally fell apart)**, **(totally fell apart)**. So, unfortunately, I **(no longer have any relationship)** at all with two or three of **(these people)**. ... And what's depressing is that you **(did everything together)** with them before. They were **(best friends)**, as they say. ... And it's **(over)**. |

Table A.1: A negative PN, begins with a positive valence and later shifts to negative valence, and ends in a negative valence.

| | |
|---|---|
| Ja, ist eigentlich der frühe Tod meiner Mutter. Der hat mich sehr getroffen. ... die ist ganz **(elend gestorben)** ... Und das hat mich natürlich wahnsinnig mitgenommen. Und **(als sie tot war)**, hab ich insgeheim, ich will net sagen, dass ich froh war aber so eine gewisse **(Erleichterung)**. Also, das ist eigentlich **(ein Gefühl)**, das ich mir **(selbst nicht gestattet)** hab, ja? Eigentlich war sie **(erlöst)**, wenn man so will. ... Schlimm war **(die Zeit, bis sie tot war)**. ... Weil man wusste, da ist **(nix zu retten)**. Das **(geht diesen Weg)**. Und sie ist sehr **(früh verstorben)**, ich war damals gerade mit dem Studium fertig. Und das war ein völliges **(Gefühl der Hilflosigkeit)**. Aber als sie tot war, hab ich mich irgendwo **(erleichtert gefühlt)**. Und das hab ich **(mir)** eigentlich **(nicht gestattet)**, **(dieses Erleichtertsein)**, ja? Hab mich eigentlich **(geschämt)**. Gut, also Gott sei Dank habe ich immer ein glückliches Leben geführt. Ich habe nicht so viele negative Erinnerungen. Also was dann **(so tief im Gedächtnis geblieben)** ist. Es gab die eine oder andere negative Erfahrung am Arbeitsplatz , aber das hat mich nicht mitgenommen. Da habe ich immer gewusst, wie ich es abstellen kann. Da hatte ich immer das Gefühl, das kann ich ändern. | Yes, it's actually the early death of my mother. That hit me very hard. ... she **(died)** quite **(miserably)** ... And that, of course, took a lot out of me. And **(when she was dead)**, I secretly, I don't want to say that I was happy, but I **(felt)** a certain **(relief)**. So that's actually **(a feeling)** that I **(didn't allow myself)**, yes? She was actually **(redeemed)**, if you will. ... **(The time until she was dead)** was terrible. ... Because you knew there was **(nothing you could do)**. That's **(going that way)**. And she **(died very early)**, I had just finished my studies at that time. And that was a complete **(feeling of helplessness)**. But when she was dead, I **(felt relieved)** somewhere. And I actually **(didn't allow that to myself)**, **(that feeling of relief)**, yes? I was actually **(ashamed)**. Well, thank God I've always led a happy life. I don't have so many negative memories. So what then has **(remained so deeply in the memory)**. There was the one or other negative experience at work , but that didn't take me away. I always knew how to turn it off. I always had the feeling that I could change that. |

Table A.2: A negative PN, begins with a neutral to negative valence and later shifts to negative valence, and ends on a neutral note.

| | |
|---|---|
| Ja, also da fällt mir grade so aktuell was ein. Ich bin grüne Dame neuerdings im Krankenhaus, und hatte eine **(Begegnung mit einem älteren Herrn)**. Es war halt die ersten Male, als ich da war, für mich noch ein bisschen neu alles. Und **(der Mann)** war **(nicht so gut gelaunt)** und hat sich **(beschwert)** übers Essen und übers Personal und so weiter. ... **(Die Aufgabe)**, die ich da drin sehe, ist den Leuten einfach nur einen kurzen Moment ein bisschen **(eine Entspannung zu geben)** oder eine Abwechslung. Und dann hat er mir erzählt, das ist alles schlecht da und er fühlt sich überhaupt nicht wohl, und ... ich hab ihn dann gefragt, was er denn Zuhause so für Hobbys hat. Dann fing er an von seinem Hund zu erzählen. Und auf einmal hab ich gesagt “Ach ja, haben Sie einen Hund?” - “Nein, das ist nicht meiner, das ist der vom Schwiegersohn.” Aber ab dem Moment **(hat dieser Mann)** auf einmal **(angefangen zu lächeln)**. Das war so **(außergewöhnlich)** ... wie kann man in so kurzer Zeit so mies drauf sein, und jetzt wenn ich ihn über seinen Hund befrage ... da ist er **(auf einmal so entspannt)** gewesen. Und dann **(fing er an zu erzählen)** und da hat er überhaupt **(nichts Negatives mehr erzählt)**, sondern eher wirklich **(nur noch schöne Sachen)**, was er alles so macht und wie viel er schon gearbeitet hat und ja. | Yeah, so something just came to my mind right now. I am volunteer recently in the hospital, and had an **(encounter with an elderly gentleman)**. The first times I was there, everything was still a bit new for me. And **(the man)** was **(not in such a good mood)** and **(complained)** about the food and about the staff and so on. ... **(The job)** that I see in there is just to **(give people)** a brief moment of a little bit of **(a relaxation)** or a change of pace. And then he told me it’s all bad there and he doesn’t feel good at all ... I then asked him what his hobbies are at home. Then he started to tell me about his dog. And suddenly I said, “Oh yeah, do you have a dog?” - “No, it’s not mine, it’s the son-in-law’s.” But from that moment on, **(this man suddenly started smiling)**. This was so **(extraordinary)** ... how can you be in such a bad mood in such a short time, and now when I ask him about his dog ... he was **(so relaxed all of a sudden)**. And then **(he started to talk)** and he **(didn’t say anything negative at all)**, but rather **(just really nice things)**, what he does and how much he has already worked and yes. |

Table A.3: A positive PN, begins with a neutral to slightly positive or negative valence and later shifts to positive valence, and ends on a positive note.

# Appendix B

# Valence Trajectory Insights

Examination of the top matching curves resulting from the trajectory analysis allowed us to make the following observations about positive PNs (pPNs) and negative PNs (nPNs). We observe an increase or decrease in valence over time for pPNs and nPNs, respectively, showing that the narrator gets more deeply involved in the emotion as the narrative progresses (cf. Figure 5.4 and B.1).

There are prominent dips in the valence curves where the valence abruptly jumps from the positive or negative to neutral . An example of this can be seen in Figure 5.4 at 95s, where the valence falls from the maximum positive state (1.0) to the neutral state (0.0). This can be attributed to the interviewer's interruptions for self-assessment as explained in Figure 4.2, suggesting that the narrator often falls out of the emotion at this point and then has to re-enter the emotional state.

We observe two valence patterns in nPNs with examples in Figure B.1:

- *Pattern 1*: the narrator starts neutral or positive, then dives into the negative emotion and dwells in it at the end.
- *Pattern 2*: the narrator starts neutral or positive, then dives into the negative emotion and increased valence at the end.

These nPNs patterns can have different reasons. We suspect that people

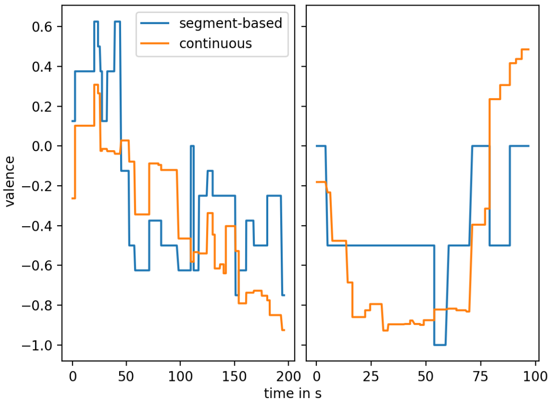


Figure B.1: Valence trajectories examples for nPN *Pattern 1* (left) and nPN *Pattern 2* (right). The blue curves represent the segment-based annotation, while the orange ones represent the continuous annotation, whereby the reference for both is w.r.t. segments. The original and translated contents of the trajectories can be found in Table A.1 and A.2

find it harder to talk about negative events than positive ones, or that they are more uncomfortable holding the negative emotion. The high fluctuations in valence trajectories for nPNs (cf. Figure B.1, *Pattern 1* and Figure B.2) support this hypothesis. Upon closer examination of the nPNs content, we find that the positive or neutral beginning is related to the fact that the narrator is not yet immersed in the emotion and that negative events often develop from positive events. An example of this can be found in Table A.1, where the narrator begins talking about the successful basketball management with his friends, which then develops into a negative story of how those friends fell apart. The positive or neutral ending, on the other hand, is often characterized by lessons learned, dealing with emotions, repression, discomfort, overplaying emotions, and self-care. An example of this can be found in Table A.2, where the narrator talks about the death of his mother and concludes that apart from that, however, he generally had a happy life.

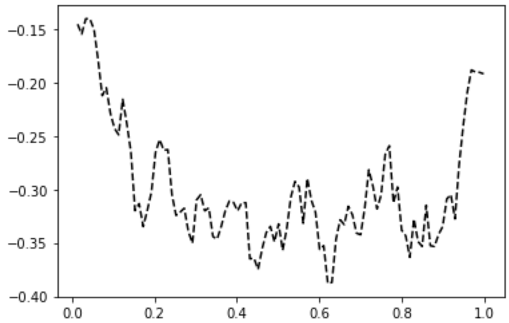


Figure B.2: Mean valence over time for all nPNs. Time is normalized in [0, 1] and mean segment-based valence is obtained at 100 sample points. We find strong fluctuations and a peak at the beginning and end of nPNs.

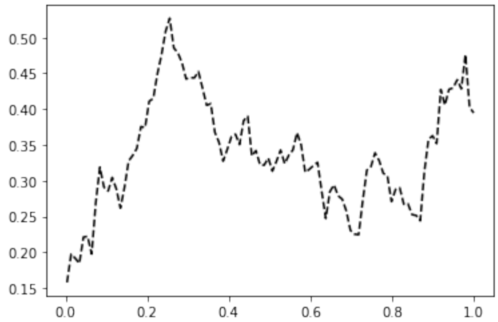


Figure B.3: Mean valence over time for pPNs. Time is normalized in [0, 1] and mean segment-based valence is obtained at 100 sample points. We find a peak in the first and second half of pPNs.

In pPNs, we find the valence patterns that:

- *Pattern 1*: the narrator starts neutral or positive, then in the first half of the narrative the positive emotion rises to a peak, after which the emotion levels off and stabilizes around some positive level.

- *Pattern 2*: the narrator starts neutral or positive, then in the first half of the narrative the positive emotion rises to a peak, after which the emotion levels off, but rises to another peak in the second half of the narrative.

Upon closer examination of the pPNs content, the peaks are related to the highlights of the narrative. If we observe multiple peaks (*Pattern 2*), this is usually related to the self-assessment interruptions of the narrative and the fact that the narrator resumes the emotion afterwards. Otherwise, if there is only one peak in the first half (*Pattern 1*), the narrator does not resume the emotion with the same intensity after self-assessment. Figure 5.4 shows an example of a positive story that starts in neutral and reaches its climax in the first half of the narrative stabilizing at a lower positive level afterwards. In Figure B.2 and Figure B.3, we show the statistical existence of the nPNs and pPNs patterns by plotting the mean valence over time for all nPNs and pPNs, respectively.