Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7004–7013
Marseille, 20-25 June 2022

Annotation of Valence Unfolding in Spoken Personal Narratives

A. Tammewar[2], F. Braun[1], G. Roccabruna[2], S. P. Bayerl[1], K. Riedhammer[1], G. Riccardi[2]
[1] Technische Hochschule Nürnberg Georg Simon Ohm, Keßlerplatz 12, 90489 Nürnberg, GERMANY
[2] Signals and Interactive Systems Lab, DISI, University of Trento, ITALY
{franziska.braun, sebastian.bayerl, korbinian.riedhammer}@th-nuernberg.de
{aniruddha.tammewar, gabriel.roccabruna, giuseppe.riccardi}@unitn.it

Abstract

Personal Narrative (PN) is the recollection of individuals' life experiences, events, and thoughts along with the associated emotions in the form of a story. Compared to other genres such as social media texts or microblogs, where people write about experienced events or products, the spoken PNs are complex to analyze and understand. They are usually long and unstructured, involving multiple and related events, characters as well as thoughts and emotions associated with events, objects, and persons. In spoken PNs, emotions are conveyed by changing the speech signal characteristics as well as the lexical content of the narrative. In this work, we annotate a corpus of spoken personal narratives, with the emotion valence using discrete values. The PNs are segmented into speech segments, and the annotators annotate them in the discourse context, with values on a 5 point bipolar scale ranging from -2 to +2 (0 for *neutral*). In this way, we capture the unfolding of the PNs events and changes in the emotional state of the narrator. We perform an in-depth analysis of the inter-annotator agreement, the relation between the label distribution w.r.t. the stimulus (positive/negative) used for the elicitation of the narrative, and compare the segment-level annotations to a baseline continuous annotation. We find that the neutral score plays an important role in the agreement. We observe that it is easy to differentiate the *positive* from the *negative* valence while the confusion with the *neutral* label is high.

Keywords: Personal Narratives, Emotion Annotation, Segment Level Annotation

1. Introduction

The Personal Narrative (PN) is the recollection of individuals' life experiences, events, and thoughts along with the associated emotions in the form of a story. PNs could be shared with others in different ways such as by meeting and telling them the story directly or by calling them, by posting them on social media, or by writing a blog. Different genres or domains have different structures of PNs and may involve the use of different modalities such as speech, writings, facial expressions, or gestures for conveying emotions. Social media posts tend to be concise and more specific about the events and emotions. Spoken personal narratives, on the other hand, have a more complex structure. They are long and contain descriptions of multiple sub-events, characters involved and the emotions felt (Tammewar et al., 2020). Rich information provided through PNs can help better understand the emotional state of the narrator, thus PNs are frequently used in psychotherapy and mental well-being applications. In psychotherapy sessions, therapists often ask clients to narrate events, in the form of PNs, that affected their mental state (Howard, 1991). Digital personal diaries (aka *journaling*) are nowadays a common tool to recollect and store personal narratives in digital form (Ghosh et al., 2017; Jeong and Breazeal, 2016; Eisenstadt et al., 2021). These narratives can be used to analyze and track the user's emotional state.

Emotion recognition is a well-grounded field of research in the natural language and speech processing communities. There are two commonly used ways to represent the emotion states: categorical and dimensional. Categorical representation classifies emotions into an established set of emotion categories such as *anger, disgust, fear, joy, sadness, surprise*, whereas dimensional representation uses numeric valence and arousal scores measuring the narrator's degree of negativity or positivity and the degree of excitement (low/mid/high), respectively (Sailunaz et al., 2018).

"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 1: 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.

Most work on emotion annotation and detection focuses on domains such as social media posts (Preoţiuc-Pietro et al., 2016; He et al., 2017; Dai et al., 2015), news (Bhowmick et al., 2009), reviews (Chuttur and Pokhun, 2021) and conversations (Poria et al., 2018; Chatterjee et al., 2019a; Chatterjee et al., 2019b), which are more structured as compared to spoken personal narratives.

In (Schuller et al., 2017), the task of valence prediction was performed on random 8 seconds fragments from spoken PNs, to predict the self-assessed valence of the narrator at the end of the recollection. Whereas, Tammewar et al. (2019) used the transcripts of the same corpus to identify the narrators' valence using the whole narratives as input. The emotional state of the narrator changes as the story unfolds as can be seen in the example from Table 1. A story that was elicited with a negative stimulus starts with a neutral to positive valence, later changes the state multiple times, and ends on a negative note. Thus, providing a single valence score for the entire narrative cannot capture the unfolding of the story. Also, it can be seen that a random fragment of the narrative cannot represent the narrator's emotional state at the end of the narrative.

In this work, we segment spoken PNs from the USoM Elderly Dataset (details in section 4.2). Annotators assign a valence score to each segment by listening to the segment and taking into consideration the surrounding context. This way we are better able to capture the unfolding of the story and thus the relevant changes in the emotional state of the narrator.

The segment-level emotion analysis could prove to be useful for mental well-being applications in tracking thought process and the emotions of the user. This data can then be analyzed to identify cognitive distortions such as "filtering", where it is important to identify if the user is focusing more on negatives and ignoring positives.

Moreover, we analyze the annotation and find that it is most difficult to distinguish the neutral valence from positive and negative.

Our novel contributions are:

- We introduce a scheme for annotating emotion valence in the discourse context and apply it to a corpus of spoken PNs.
- We perform inter-annotator agreement using different metrics and compare the discrete segment-based annotation to a baseline continuous annotation.

2. Related Work

There has been some work in the field of spoken PNs from the perspective of emotion analysis. The USoM (Ulm State-of-Mind) corpus (Rathner et al., 2018) consists of spoken PNs of undergraduate students of psychology collected in a laboratory setup. Four PNs per participant were elicited using negative and positive stimuli. Self-assessment was conducted before and after each PN to capture the state-of-mind of the participants in terms of valence and arousal. In (Schuller et al., 2017), this data was split into segments of 8 seconds to identify the corresponding narrator's self-assessed state of mind after narrating the PN. As can be seen in the example from Table 1, not all fragments of the PN represent the final state of the narrator but may change over time as the narrator recollects the narrative. Furthermore, the time-based segmentation ignores sentences and thus possible semantic boundaries. In a follow-up study, Tammewar et al. (2019) worked on the transcriptions of the same data, to identify the valence of the narrator using the entire narrative.

Another line of work where emotion annotation on segments is performed, rather than the entire text or story, is in the domain of conversations. A number of data sets have been released where each turn of the conversations is marked with coarse or fine-grained emotion classes (Busso et al., 2008; Poria et al., 2018; Rashkin et al., 2019). Welivita et al. (2021) presented an extensive study on conversational data sets with annotated emotions. The data sets have mainly been exploited for building empathetic response generation by conversational agents, such as (Roller et al., 2021). The conversational data is naturally segmented into utterances and consists of multiple speakers. Thus, they are quite different from PNs, although even conversations may involve the unfolding of stories.

3. USoM Elderly Dataset

In this section, we briefly describe the USoM Elderly Dataset of 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 ("young"), they collected German spoken PNs with the purpose of building emotion detection systems, however this time by elderly persons.

Every participant was asked to recollect four experiences from his or her 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 bio-sensors.

As described in Figure 1, 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 (Russell, 2003). Additionally, an external assessment was con-

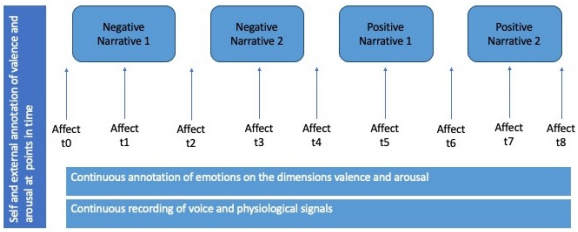

Figure 1: USoM elderly data collection process: The self assessed affect values were collected before, after and in the middle of each narrative (eight times), whereas the continuous annotation was performed throughout, using joysticks.

ducted. 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" and the annotators as "continuous annotators".
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.

4. Segment Level Valence Annotation

In this section, we explain the steps involved and the protocol to annotate valence at the segment level of the PNs from the USoM Elderly Dataset.

4.1. Transcription

The PNs from the USoM Elderly dataset were 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 was 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 Milde and Köhn (2018). To ensure next to perfect alignments, missing entries in the pronunciation-lexicon, such as incomplete words, dialect, and slang words, were generated using a grapheme-to-phoneme tool(Bisani and Ney, 2008). For better working of automated NLP tools in the downstream tasks, the transcriptions were preprocessed to remove the incomplete words, filler words, and other metadata such as pauses, speakers, and stuttering.

4.2. Segmentation

The transcripts were then segmented into smaller meaningful parts. Ideally, we would like the parts to be functional units, from the speech acts theory (Bunt et al., 2017; Thomas, 2014). Due to the lack of data annotated with functional units or the presence of any automated tool for such segmentation, we tried other levels of segmentation. First, we tried using SpaCy-3[1] sentence segmentation (transformers based NLP pipeline for German). We found the resulting segmentation to frequently split at unnatural times, which could be because of the spoken nature of the data. Thus, we instead segmented the text using heuristic rules, making use of the cues from punctuation $\{.!?\}$ and 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 found that this strategy worked the best for us, and leave prosody-based approaches for future work.

4.3. Annotation

The annotation task aimed to capture the emotional polarity of the narrators while they recollected the events and the intensity of such polarity as expressed on a range from unpleasant to pleasant, by listening to the audio and using the transcript as support. A 5-point bipolar scale from -2 ("unpleasant") to 2 ("pleasant") with 0 representing "neutral" was used to label the valence of each segment.
The annotators were asked to adopt the point of view of the narrator (putting themselves in their shoes) as some events might be considered irrelevant by the annotator which are meaningful for the narrator. The annotators may consider the neighboring context before and after the current segment. This helped in assessing the contribution of the annotation segment to the event described in the narrative. We refer to this annotation as "segment-based annotation" and to the annotators as "segment-based annotators".

4.4. Execution

We defined a scheme 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 was 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 was held between all the annotators and the psy-

[1] https://spacy.io/usage/v3

chotherapist to discuss the differences amongst the annotators, try to agree on a specific opinion, and modify the guidelines if necessary. We continued the small training batches until we achieved a satisfactory inter-annotator agreement as measured using the evaluation metric explained in Section 5.2. 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 were 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 provided 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 was annotated by all the annotators while 80% of the data was annotated by a single annotator. To ensure the quality of annotation in this last phase, we divided the data into batches and monitored the inter-annotator agreement on the overlapping data.

We plan to make the corpus available to the public for research purposes. A few examples of PNs annotated with valence can be found in the Appendix in Table 4, 5, and 6.

5. Analysis

We now analyze the newly collected segment-based annotations and compare them to two continuous valence annotations that are collected in the USoM Elderly data set.

5.1. 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 4.4, 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 $\sim$12 tokens. The entire data contains $\sim$7000 segments. 20% of the data (42 PNs; $\sim$1200 segments), was annotated by all annotators as explained in Section 4.4, which is further analyzed for calculating the inter-annotator agreement statistics.

5.2. Inter-Annotator Agreement

To assess the quality of the segment-based annotation, we computed different inter-annotator agreement (IAA) statistics. This was computed in the training and overlap phases of the annotation, and also during the partial-overlap phase on the overlapping part of the batches. We also compared these results with the IAA of the continuous annotation from the USoM Elderly data set (Section 3), and also with the IAA by combining both the segment-based and continuous annotations. The IAA statistics are shown in Table 2. Note that the segment-based annotation involves 4 annotators and the continuous annotation involves 2 annotators, whereas the "segment + continuous" involves 6 annotators.

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 2: Inter-Annotator agreement using Fleiss' κ. *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.

Since the segment-based annotation was performed by four annotators, we used Fleiss' κ to compute the inter-annotator agreement (Fleiss, 1971). 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 (Landis and Koch, 1977).

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

We took 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. We then compared segment-based annotation with con-

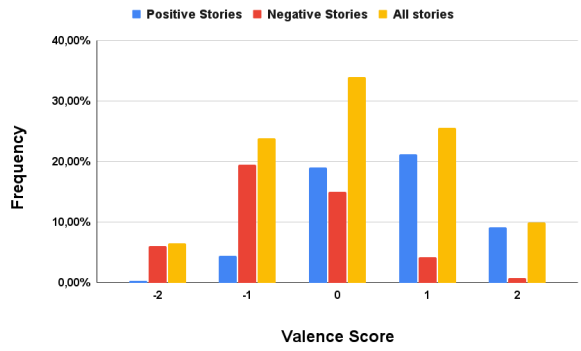

Figure 2: Valence score distribution. The blue and red bars represent the distribution of valence scores respectively computed on positive and negative stories. The yellow bar represents the distribution of the whole dataset.

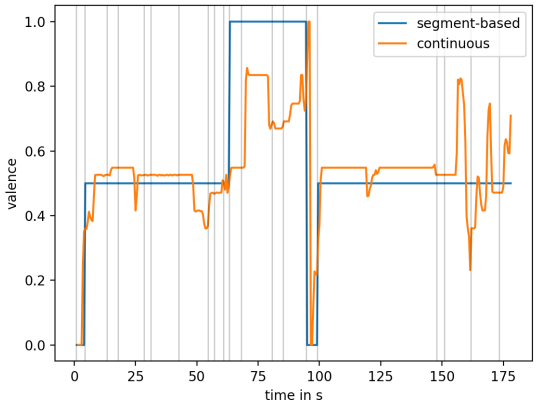

Figure 3: *Segment-based* and *continuous* valence annotation of a positive PN; the vertical lines mark the segments.

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

5.3. Label Distribution

Figure 2 depicts the label distribution of the dataset. On the overlapping examples, i.e. examples annotated by more than one person, we computed the arithmetic mean and rounded 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 2, 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.

As we discussed in subsection 5.2, the neutral class is a source of disagreement. The analysis of the label distribution of narratives with overlapping annotators shows that for 77% of segments at least one annotator chose the neutral class. Moreover, inspecting the cases where annotators disagree, we found that for 70% of the examples include at least one neutral label. This brings additional evidence to the fact that neutral is a hard to agree on.

5.4. Analysis Based on Valence Trajectories

The 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 in terms of time. This is particularly interesting for such stories where we observe contradicting per-segment 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 regarding the 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 3 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.

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

Table 3: Annotator agreement for valence trajectories by means of RMSE and DTW; the last column shows the agreement among the segment-based annotators only.

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)(Berndt and Clifford, 1994). We find temporal shifts between the continuous and segment-based sequences (cf. Figure 3), 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 3 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; the agreement in the stimuli (pos, neg) is high with variations in its subclasses. We find that DTW is almost the 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 3 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.

6. Conclusion

Use of Personal Narratives (PN) is growing in well-being applications, but the research on PNs is still very limited. We proposed a novel segment level emotion annotation scheme for spoken personal narratives that includes a training and verification (overlap) phase, followed by a phase that includes partial overlap to monitor quality. We statistically show the significance of the confusion in the annotation of neutral valence. We prove our annotation quality using different inter-annotator agreement metrics, using discrete class based agreement as well as using valence trajectories.

We believe this work can be extended to identify emotion carriers as defined by (Tammewar et al., 2020) from segments, instead of the entire narrative. This way we will be able to show emotion carriers for the emotions at the different parts of the recollection of the narratives, which in turn could be used by mental well-being applications to have meaningful conversations with the user.

7. Acknowledgements

The research leading to these results has received funding from the European Union – H2020 Programme under grant agreement 826266: COADAPT.

We would also like to thank Eva-Maria Messner (Clinical Psychology and Psychotherapy, University of Ulm), for providing us the USoM Elderly dataset for the purpose of this research.

8. Bibliographical References

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

Bhowmick, P. K., Basu, A., and Mitra, P. (2009). Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Comput. Inf. Sci.*, 2(4):64–74.

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017). Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and

Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., and Agrawal, P. (2019a). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019b). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Chuttur, Y. and Pokhun, L. (2021). An evaluation of deep learning networks to extract emotions from yelp reviews. In *Progress in Advanced Computing and Intelligent Engineering*, pages 55–67. Springer.

Dai, W., Han, D., Dai, Y., and Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management*, 52(7):777–788.

Eisenstadt, M., Liverpool, S., Infanti, E., Ciuvat, R. M., Carlsson, C., et al. (2021). Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: Systematic review and meta-analysis. *JMIR Mental Health*, 8(11):e31170.

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

Ghosh, A., Stepanov, E. A., Danieli, M., and Riccardi, G. (2017). Are you stressed? detecting high stress from user diaries. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000265–000270. IEEE.

He, Y., Yu, L.-C., Lai, K. R., and Liu, W. (2017). Yzu-nlp at emoint-2017: Determining emotion intensity using a bi-directional lstm-cnn model. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 238–242.

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

Jeong, S. and Breazeal, C. L. (2016). Improving smartphone users' affect and wellbeing with personalized positive psychology interventions. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 131–137.

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

Milde, B. and Köhn, A. (2018). Open source automatic speech recognition for german. In *Proceedings of ITG 2018*.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). 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.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Rathner, E.-M., Terhorst, Y., Cummins, N., Schuller, B., and Baumeister, H. (2018). State of mind: Classification through self-reported affect and word use in speech.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., et al. (2021). Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110 1:145–72.

Sailunaz, K., Dhaliwal, M., Rokne, J., and Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al. (2017). The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring.

Tammewar, A., Cervone, A., Messner, E.-M., and Riccardi, G. (2019). Modeling User Context for Valence Prediction from Narratives. In *Proc. Interspeech 2019*, pages 3252–3256.

Tammewar, A., Cervone, A., Messner, E.-M., and Riccardi, G. (2020). Annotation of emotion carriers in personal narratives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1517–1525.

Thomas, J. A. (2014). *Meaning in interaction: An introduction to pragmatics*. Routledge.

Welivita, A., Xie, Y., and Pu, P. (2021). A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Appendix: Annotation Examples

In this Section, we present some PNs from the USoM-Elderly dataset, annotated with Valence.
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 text is segmented into segments using pipes (|).
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)

Note-1: The examples are provided to give a sense of the USoM-elderly data and the annotation protocol for 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 4: 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 5: 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 6: 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.