EmoProgress: Cumulated Emotion Progression Analysis in Dreams and Customer Service Dialogues

Eileen Wemmer[1,2], **Sofie Labat**[2,3], **Roman Klinger**[2,4]
[1]Institut für Arbeitswissenschaft und Technologiemanagement, University of Stuttgart, Germany
[2]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[3]LT3, Language and Translation Technology Team, Ghent University, Belgium
[4]Fundamentals of Natural Language Processing, University of Bamberg, Germany
eileen.wemmer@iat.uni-stuttgart.de
sofie.labat@ugent.de
roman.klinger@uni-bamberg.de

Abstract

Emotion analysis often involves the categorization of isolated textual units, but these are parts of longer discourses, like dialogues or stories. This leads to two different established emotion classification setups: (1) Classification of a longer text into one or multiple emotion categories. (2) Classification of the parts of a longer text (sentences or utterances), either (2a) with or (2b) without consideration of the context. None of these settings, does, however, enable to answer the question which emotion is presumably experienced at a specific moment in time. For instance, a customer's request of "My computer broke." would be annotated with anger. This emotion persists in a potential follow-up reply "It is out of warranty." which would also correspond to the global emotion label. An alternative reply "We will send you a new one." might, in contrast, lead to relief. Modeling these label relations requires classification of textual parts under consideration of the past, but without access to the future. Consequently, we propose a novel annotation setup for emotion categorization corpora, in which the annotations reflect the emotion *up to the annotated sentence*. We ensure this by uncovering the textual parts step-by-step to the annotator, asking for a label in each step. This perspective is important to understand the final, global emotion, while having access to the individual sentence's emotion contributions to this final emotion. In modeling experiments, we use these data to check if the context is indeed required to automatically predict such cumulative emotion progressions.

Keywords: emotions, appraisals, progression analysis, dialogue, customer interactions, dreams

1. Introduction

The most popular subtask in emotion analysis is to assign emotions from a predefined model to textual units. Often, this classification task is formulated as supervised machine learning, for which manually annotated corpora are leveraged. Such resources differ in various ways (Bostan et al., 2020), one variable being the granularity of annotations. Some corpora use sentences as the unit of interest (Alm et al., 2005), others rely on utterances in conversations (Hsu et al., 2018) or whole tweets (Schuff et al., 2017). While fine-grained, sequential annotations reflect changes in the emotional content on a sentence-by-sentence level, they do not necessarily describe the cumulative emotional progression as presumably experienced by an interlocutor or reader. This is, because each label only describes the sentence it pertains to (Alm et al., 2005; Aman and Szpakowicz, 2007). Traditionally, the sentences in the text "*Maria felt she was being followed. She stopped and tilted her head.*" may be labeled as [*fear, neutral*], since the second sentence does by itself not express any emotion. Yet when considering the first sentence, the protagonist would likely still be fearful in the second sentence and the annotations reflecting the overall emotions in the text would be [*fear, fear*].

Similarly, context offered to annotators usually includes the entire text both before and after the current sub-unit (Zahiri and Choi, 2017; Poria et al., 2019a; Labat et al., 2023), or annotations are not gathered on a sufficiently fine-grained level (Chatterjee et al., 2019). In this sense, existing resources do consider the emotion on a global level *or* on a fine-grained level in context. We argue that this reflects only to a limited degree the actual emotional state of a participant who develops an emotion throughout a narrative, event, or interaction.

To accurately describe the progression of emotions in a text, annotations need to satisfy three central criteria in regards to (1) granularity, (2) context, and (3) scope: (1) They must be sequential to capture changes in the underlying emotional content, (2) they must both take into account and (3) reflect not only the emotion expressed in the unit they are assigned to, but also the emotional content of the text leading up to the current part.

The need for the exclusion of the text following the current annotation unit becomes evident when we add an additional sentence to the example above: "*Everyone who knew Maria recognized this as a sign that she was to burst out laughing and sure enough she did, as she turned to hug*

her sister." Given this context, the annotator might have assessed the situation to be joyful rather than scary as it was the case in the second sentence ("She stopped and tilted her head."), which would change the second label from *fear* to *joy*.

We aim at studying how each sentence in a discourse contributes to the global emotion. Thus, we propose a new corpus of customer-agent interactions and dream reports following an annotation procedure in which we uncover text step-by-step[1]. For each part, annotators are asked to judge the emotional content of the current text up to and including the most recently revealed part, both in terms of emotion category and event appraisal. This incremental approach ensures that future context is disregarded, while providing all necessary previous context for the annotation of the current state of the overall emotional progression. Our experiments aim at understanding the impact of contextual information for the prediction of progressively aggregated emotion labels. We find that previous context is not required for categorical emotion classification, though its impact varies depending on the emotion class.

2. Related Work

We briefly review emotion theories (Section 2.1) and discuss related corpora (Section 2.2).

2.1. Emotion Models in Psychology

Emotion models that are relevant for emotion analysis can be separated into two groups: categorical and dimensional models. Categorical models describe emotions in terms of a fixed set of separate categories. Prominent examples that have found application in computational analyses are the basic emotion models by Ekman (1992) and Plutchik (2001). Ekman identified *anger*, *surprise*, *disgust*, *joy*, *fear*, and *sadness* as basic emotions. Dimensional emotion models describe emotions in terms of values along a pre-defined set of axes, thus placing them in a continuous space. One commonly employed representative of this group is the Circumplex Model of Affect (Posner et al., 2005), which distinguishes emotions in terms of valence and arousal. Another class of dimensional emotion models are appraisal theories (Ellsworth and Scherer, 2003), which distinguish emotions by the subjective, cognitive evaluations (i.e., the appraisals) that arise when an individual is faced with an event (Moors, 2017; Scherer, 2009). Different appraisal theories exist (Ellsworth and Scherer, 2003; Scherer, 2009; Roseman and Smith, 2001; Roseman, 1984). Notably, Smith and Ellsworth (1985); Ellsworth and Smith (1988) show the correlation of certain appraisals with categorical emotions and find that categorical emotions can be distinguished through appraisals.

For dreams, domain-specific sets of emotion categories have been developed to make dreams quantifiable for statistical analyses (Domhoff, 1996; Schredl, 2010). One commonly used set of classes stems from the Hall/Van De Castle System of Content Analysis (Domhoff, 1996). Overall, the emotions encountered in dreams have been found to reflect those experienced in waking life (Gilchrist et al., 2007), yet, biased towards negative emotions (Nielsen et al., 1991; Hartmann et al., 2001).

Customer service dialogues are often analyzed with a focus on detecting whether the customer expresses a positive or negative sentiment (Park et al., 2021). For the broader task of emotion recognition in conversations (Poria et al., 2019b), annotations featuring emotion classes are more common (Poria et al., 2019a; Li et al., 2017; Zahiri and Choi, 2017), though customer service corpora featuring emotion categories are also beginning to emerge (Labat et al., 2023, 2024).

2.2. Corpora for Emotion Analysis

Corpora for emotion analysis usually feature a set of texts, or instances, that are manually labeled for the emotional content they express. They vary in multiple regards, such as the underlying emotion models, whether they are gathered in a sequential manner or on instance level, or what domain the instances were sourced from. This section introduces corpora that are similar to the ones gathered in this work in one or multiple of these aspects. For a comprehensive overview of corpora up to 2018, we refer to Bostan and Klinger (2018).

One early contribution of sentence-level emotion annotations is the blog corpus by Aman and Szpakowicz (2007). The authors note that "there is often a dynamic progression of emotions [...] in the conversation texts and blogs" (Aman and Szpakowicz, 2007, p. 198). However, the labels they gathered only describe the sentence they are assigned to, not the current state of an emotion progression. Another similar example of annotations on sentence level is the Tales corpus (Alm et al., 2005). Similarly, annotations reflect only the sentence-level emotions in isolation.

Sequential annotations are also commonly found in conversation corpora. Our customer service dialogue corpus is a reannotation of the EmoWOZ-CS corpus (Labat et al., 2022a, 2024), translated from Dutch to English. While the original annotations are contextualized in the entire conversation, emotions would only be annotated for a turn if they were either implicitly or explicitly expressed in that turn.

[1]The EmoProgress corpus is available at https://lt3.ugent.be/resources/emoprogress/.

Hence, not only do the annotations take into account future context, they also do not represent the current state of the emotional progression at the part they are tied to.

Another resource with categorical emotion labels and valence–arousal–dominance scores is the EmoTwiCS corpus on Dutch customer service exchanges on Twitter (Labat et al., 2022b, 2023). The annotations in this dataset pertain to passages in the customers'turns that express the labeled emotion or the entire turn if no one part explicitly expressed underlying emotion, hence again reflecting a different context and scope (Labat et al., 2020).

Other examples of emotion corpora for conversations include EmoryNLP (Zahiri and Choi, 2017), EmotionLines (Hsu et al., 2018) and its multimodal successor MELD (Poria et al., 2019a) from TV show conversations. EmotionLines and MELD both feature annotations on utterance level that were contextualized on both past and future utterances. The DailyDialog corpus, which was sourced from websites for English learners, also features utterance level annotations (Li et al., 2017).

Another particularly relevant conversational corpus was gathered for the 2019 SemEval shared task on contextual emotion detection in text that is based on conversations with a conversational agent (Chatterjee et al., 2019). In this corpus, instances consist of three consecutive utterances labeled with categorical emotions. The assigned label describes the emotional content of the last utterance, while the previous two utterances serve to contextualize the expressed emotion.

3. Corpora Creation

We now describe the data acquisition and preprocessing in Section 3.1, and the annotation procedure described in Section 3.2. The corpus EmoProgress is available at https://lt3.ugent.be/resources/emoprogress/.

3.1. Data Gathering and Preprocessing

We build EmoProgress out of dream self-reports and customer service interactions (which we abbreviate as CS). We are interested in the emotion of the dreamer and the customer who we both assume reappraise events as they progress. While dreams usually contain either a continuous event or a series of events, customer service dialogues are underlying a reappraisal of the overall situation as more information becomes available through the course of the interaction.

Our customer service subset is based on a translated version of the EmoWOZ-CS dataset (Labat et al., 2024). This corpus was gathered in a Wizard-of-Oz setting, in which a person acted as a customer service chatbot and aimed at steering the emotion trajectory of the conversation towards a pre-defined sentiment. The participants were provided with a fictional description of an event and were asked to solve the issue that occurred with the help of the chatbot. All instances follow a format of alternating turns by the service agent and the customer. Therefore, the unit of annotation in the customer interaction domain corresponds to bi-turns of an agent's turn followed by a customer's turn. We refer to this unit as a *part*.

For our dream corpus subset, we use data from http://www.dreambank.net/ (Domhoff and Schneider, 2008). Before annotation via crowdsourcing, we manually cleaned the corpus for sensitive topics, profane words[2], words that are not in English, or that differ from the standard file format on DreamBank. We define a *part* (the annotation unit) as one sentence. We further only consider instances that consist of 4–10 parts and maximally 1000 words. This ensures that emotions can evolve, while still limiting the annotation time per instance.

Inspired by Bostan et al. (2020), we leveraged the NRC lexicon to count potentially emotionally charged words (Mohammad and Turney, 2013) in both corpora and sample instances for annotation proportionally to the relational emotion word count. For dialogues, we only consider emotion words in customer turns.

3.2. Annotation Procedure

We gather annotations through SoSciSurvey[3] and recruit and pay annotators through Prolific[4]. Annotators are first given details on the annotation procedure, goals, data security and payment of the task. We then gather their consent and give more detailed instructions on the annotation task.

While dream reports largely resemble event descriptions, for customer service dialogues we were interested in how the events referred to in the conversations may be appraised and how those appraisals may change over the course of their interaction with the service agent. Hence, we instruct annotators to focus on the events addressed in the conversation, such as the reason the customer contacted customer service or any measures the service agent took.

Table 1 shows all collected variables. The emotion label set has been developed starting with the same label set as Labat et al. (2024), extended to be appropriate also for dreams by aggregating the original labels with Ekman's basic emotions.

[2] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt, accessed 22.12.2022

[3] https://www.soscisurvey.de/

[4] https://www.prolific.co

Variable	Question Formulation	Values
Emotion Category	The events (in the dream) made the person dreaming/customer feel...	[Emotions]
	To the customer/dreamer, the events in the dream/in the conversation...	
Pleasantness	...were pleasant.	{1,...,5}
Familiarity	...were familiar.	{1,...,5}
Effort	...required a lot of energy to deal with (within the dream).	{1,...,5}
Own Responsibility	...were caused by their own behaviour (in the dream).	{1,...,5}
Others' Responsibility	...were caused by somebody else's behaviour (in the dream).	{1,...,5}
Chance Control	...were the result of outside influences (within the dream) of which nobody had control.	{1,...,5}
	(In the dream,)The dreamer/customer...	
Event Predictability	...could have predicted the occurrence of events (in the dream).	{1,...,5}
Attention	...paid attention to the events (in the dream).	{1,...,5}
Consequence Anticip.	...felt that they anticipated the consequences of the events (in the dream).	{1,...,5}
Own Control	...had the capacity to affect the events (in the dream).	{1,...,5}
Confidence	How confident are you about your judgements for the dream/conversation you've just read?	{0,...,4}

Table 1: Variables collected in the annotation study. Formulations in parentheses only apply to dreams. As emotion categories, we use *joy*, *admiration*, *gratitude*, *relief*, *desire*, *fear*, *anger/annoyance*, *sadness/disappointment*, *surprise/confusion*, and *neutral*.

Annotators are shown the first part (i.e., sentence/bi-turn) along with the question of how the events made the dreamer/customer feel. They are first asked to pick the most fitting categorical label, before rating each of the ten appraisal dimensions on a five-point scale. Afterwards, the next part is added to the previous text and the questions are asked again. This procedure is visualized in Figure 12 in the Appendix. Annotations of the previous part are pre-selected for the current part to lower cognitive load and to allow annotators to focus on the changes the newly displayed part introduced to the emotional content of the overall text. Upon completion of an instance, annotators are asked to indicate their confidence in the overall annotations they had just submitted on a five-point scale. As an attention check halfway through the survey, annotators would be shown a prompt to select a specific emotion. Submissions with failed attention checks are rejected.

Annotators judged up to 24 sentences/bi-turns in each survey. Due to glitches on the survey platform, four instances were not fully annotated. In these cases, the authors of this paper re-annotated the full instance. To obtain data to estimate the inter-annotator agreement, we ensure that two people annotate the same instances in a subset of the data. We collect 46 conversations (271 parts) and 45 dreams (264 parts) for IAA calculations.

We pay participants £9/h. Altogether, including prestudies and additional annotations to calculate inter-annotator agreement, the creation of the corpus cost £683.33, including £170.83 service fees.

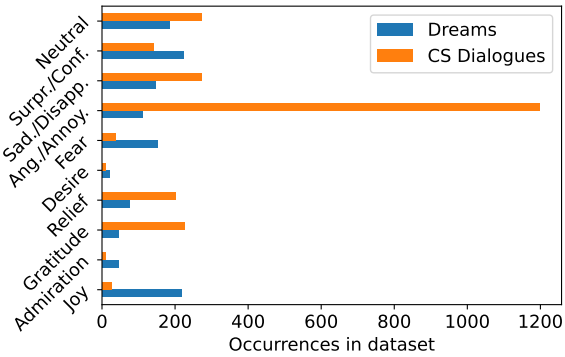


Figure 1: Overall label distribution over all parts and annotations for both corpora

4. Corpora Description and Analysis

We will now describe our collected corpus. Example annotations are shown in Table 2a for dreams and Table 2b for conversations.

4.1. Dataset Description

EmoProgress contains 149 dreams and 339 customer care dialogues, with a total of 890 and 2010 annotated parts, respectively. On average, the dreams contain 5.97 and the dialogues 5.92 parts (1.85 and 1.7 standard deviation, respectively).

Figure 1 shows the label distribution of all annotations. Aside from the classes specifically designed to capture emotions in customer service interactions (*admiration*, *gratitude*, *relief*, *desire*), the label

1. I am flying in a cloud tunnel and I see out the opening at the other end.
2. Some flying creature is there and is coming toward me.
3. I think, "I sure hope it's a good creature"!
4. I will it to be so even as I avert my eyes for fear it is a horrible, hurtful creature.

Part	1	2	3	4
Emotion	J	S/C	F	F
Appraisal				
Pleasantness	5	3	2	1
Familiarity	1	1	1	1
Effort	5	5	5	5
Own Resp.	1	1	1	1
Others' Resp	1	5	5	5
Chance Control	5	5	5	5
Attention	5	5	5	5
Cons. Antic.	1	1	1	1
Own Control	2	2	2	3
Event Predictability	1	1	1	1

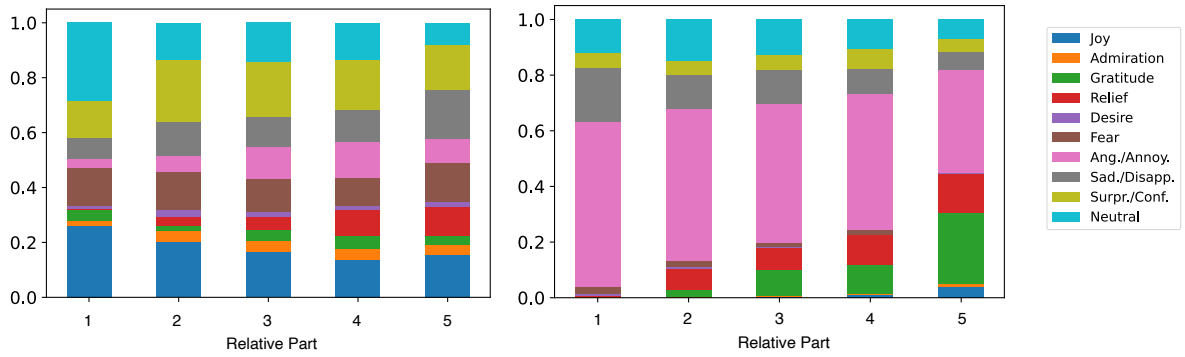
(a) Dream report annotation example.

1. *Bot:* Hello, my name is Chatty9 😊 How can I help you today?
   User: I paid extra for a deluxe breakfast and it was terrible.
2. *Bot:* At BookYourHolidays we always guarantee the best quality.
   User: Are you even listening to what I'm saying?
3. *Bot:* I like listening to what you have to say.
   User: I'm just saying breakfast was terrible.
4. *Bot:* Would you like to leave a review about breakfast at your hotel?
   User: No, thanks.

Part	1	2	3	4
Emotion Class	S/D	A	A	A
Appraisal				
Pleas.	4	3	3	3
Famili.	1	4	4	4
Effort	1	1	1	1
Own Resp.	1	1	1	1
Others' R.	5	5	5	5
Chance Ctrl.	3	3	3	3
Attention	3	2	2	2
Cons. Antic.	3	2	2	2
Own Control	3	2	2	2
Event Pred.	3	2	2	2

(b) Customer service dialogue annotation example.

Table 2: Example instances of EMOPROGRESS.



(a) Dreams

(b) Customer Service Dialogues

Figure 2: Normalized label distributions progression of all instances. The five parts are calculated relative to the length of the instance and do not correspond to the annotation units.

distribution is, by and large, balanced. Classes reflecting one of the basic emotions (Ekman, 1992) and *neutral* were chosen more frequently. Figure 1 shows a prevalence of *anger/annoyance* for the customer service domain.

4.2. Progressions of Categorical Emotions

Figure 2 shows the relative occurrence of emotion labels over the progression of texts. Figure 2b puts the prevalence of *anger/annoyance* annotations observed in Figure 1 into perspective, with a decrease in this emotion over time. We observe similar effects for *sadness/disappointment*, mirrored by an increase in *relief* and *gratitude*.

For dreams, overall trends are more subtle. While the prevalence of *joy* and *neutral* decreases as reports progress, *relief* and *anger/annoyance* become more frequent. The average of unique emotion categories per instance is 3.35 for dreams (std 1.31) and 2.75 for CS dialogues (std 1.13). This shows that the less pronounced changes in the dream dataset are not due to more static annotations for each dream, but rather a result of a more varied set of different progressions.

Hence, we conclude that while progressions can be found in both domains, there are domain-dependent differences. This holds not only for the overall distribution of emotions, but also in how they change over the course of texts.[5]

[5] An analysis of the progression of appraisal annotations is available in the appendix.

	F_1	
Emotion	Dreams	CS
Joy	.48	.29
Admiration	.27	.0
Gratitude	.12	.45
Relief	.23	.5
Desire	.0	—
Fear	.41	.0
Ang./Annoy.	.42	.72
Sad./Disapp.	.56	.16
Surpr./Conf.	.49	.15
Neutral	.32	.38
Micro F_1	.43	.52
Macro F_1	.33	.29

Table 3: Inter-Annotator F_1 scores for binarized emotion category annotations. For *desire* in CS, not enough samples were available to calculate the inter-annotator F_1.

4.3. Inter-Annotator Measures

In this section, we take a closer look at the inter-annotator agreement for both corpora in regards to their emotion label and appraisal annotations. All scores are calculated for six fixed pairs of annotators per domain that labeled the same instances. Given the length (in parts) of the included instances, we opt to collect data for inter-annotator measures using with pairs of annotators. This way, we kept the workload of individual crowdworkers manageable. Each pair was assigned the same subset of either dreams or conversation. This method of collecting data for analysis renders both Cohen's (Cohen, 1960) and Fleiss' (Fleiss, 1971) κ unsuitable for investigating inter-annotator agreement, since they both assume that the same annotator has seen all instances. Furthermore, Krippendorff's Alpha focuses on expected differences across all labels instead of single-label agreement (Krippendorff, 2004). We therefore choose to report F_1 scores over these popular metrics, as they do not take the expected chance into account and therefore are better suited to our collected data. All reported scores are averaged.

Table 3 shows the inter-annotator agreement in terms of F_1. For dreams, we observe a lower score for those emotion classes that were adopted from the customer service domain and not aggregated with basic emotions, namely *admiration*, *gratitude*, *relief*, and *desire*. All other emotion labels reach acceptable scores of >0.4.

For CS dialogues, the bias towards *anger/annoyance* annotations observed in Figure 2b results in a high inter-annotator agreement of F_1=0.72 for that particular class. In total, four of the ten classes reach F_1 scores of >0.3. This results in a higher macro F_1 score for dreams than for dialogues.

	Pearson r	
Appraisal	Dreams	CS
Pleasantness	.62	.56
Familiarity	.41	.13
Effort	.41	.35
Own Resp.	.27	.27
Others' Resp.	.29	.38
Chance Control	.20	.22
Event Predict.	.06	.29
Attention	.08	.12
Cons. Antic.	.41	.08
Own Control	.08	−.09
Avg	.30	.24

Table 4: Average Pearson's r for each appraisal dimension in the evaluation set of both domains (via Fisher z-transformation)

Setup		
Single	Context	Label
p_1	p_1SEP	l_1
p_2	p_1SEPp_2SEP	l_2
p_3	p_1SEPp_2SEPp_3SEP	l_3
p_4	p_1SEPp_2SEPp_3SEPp_4SEP	l_4

Table 5: Tasks for a four-part instance with parts $p_1 \ldots, p_4$ and corresponding labels $l_1 \ldots, l_4$ in the single and context setup. In the single setup, we only show the part p_i. In the context setup, we present the classifier with all parts up to p_i.

Table 4 shows average Pearson's r between annotator pairs of the same instances. The overall average correlation is moderate for both domains. Only few dimensions show an agreement >.5. This shows that annotating fine-grained appraisals in an incremental setup is a challenging task.

5. Importance of Context and Order for Progression Prediction

This section introduces the setup and results of an experiment we conducted to get insights into the importance of the previous context for automatic classification purposes.

5.1. Experimental Setup

Each label describes not only the emotional content of one part, but also takes into account the previous parts. This means that even if no emotion (*neutral*) is expressed in a part, if it becomes clear from the text leading up to it that the subject is experiencing an emotion, the part would still be labeled with that emotion. Thus, we hypothesize that removing prior

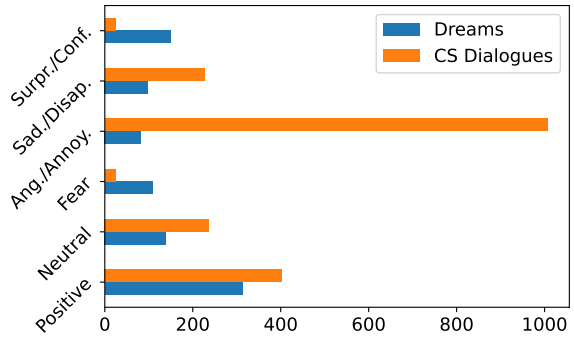


Figure 3: Absolute occurrences of labels by class after sampling and aggregating for the experiment.

Data	Setup	Acc.	Std.	Macro F_1	Std.
CS	single	.61	.04	.29	.03
CS	context	.60	.04	.26	.03
Dreams	single	.44	.06	.34	.10
Dreams	context	.44	.06	.32	.07

Table 6: Accuracy and Macro F_1 across all trained setups. Metrics are averaged over folds and repetitions, standard deviations are calculated between folds of the same iteration and averaged over all repetitions.

context would negatively affect the classification abilities of a trained model.

To test this, we fine-tune RoBERTa with default parameters (Liu et al., 2019) in two setups for each domain. For the *single* setup, the task is to predict the associated emotion class in isolated parts. For the *context* setup, the entire text sequence up to and including the part in question is provided to the classifier during training and evaluation. Table 5 illustrates the differences between setups. If our hypothesis holds, we would expect the latter classification task to yield better results. From instances annotated more than once, we take one annotation randomly.

As discussed in Section 4, the data distribution is imbalanced. Classes not describing basic emotions are particularly rare. To combat the effects of the skewed data distribution in our experiment, we aggregate the classes *joy*, *admiration*, *gratitude*, *relief*, and *desire* to one class, referred to as *positive*. Figure 3 shows the resulting distribution.

We evaluate according to a stratified 10×10-fold cross-validation setup with splits along instances (not parts). Splits are fixed between all setups.

5.2. Experiment Results

Table 6 shows the overall performance of all setups, averaged over these ten iterations and all ten folds per iteration. As expected for a skewed dataset, the accuracy is higher than the corresponding macro F_1. Macro F_1 is comparably low with the highest score of .34 for the emotion prediction in dreams in the single-part setup. Similarly, the context-free setup outperforms its context-dependent counterpart in terms of overall F_1. Despite the smaller dataset, performance in terms of F_1 is overall higher for dreams. Average standard deviations are reasonably low for all cases.

Table 7 shows the results split by category. Particularly challenging in CS data are the classes *sadness/disappointment*, *fear*, and *surprise/confusion*. While the latter two classes have by far the least amount of samples in the dataset, the number of parts with *sadness/disappointment* annotations for customer service dialogues is comparable to the number of *neutral* annotations. Yet, *neutral* is predicted more reliably in both setups. Overall, whenever there is a difference in scores for customer service dialogues, the single-part setup performs better than its context counterpart, though only by a difference of <.01 for all classes. With a difference of .13, there only seems to be an actual difference in prediction quality for the *neutral* class. In this case, it seems to be advantageous for the system to not consider previous context.

Focusing on dreams yields a slightly different picture. Differences in F_1 scores between setups remain small, though for four out of six classes the contextualized prediction performs better. The single-part prediction performs better by .09 for *anger/annoyance* and <.01 for *surprise/confusion*. Notably, these are the only two classes for which F_1 drops below .2 in the contextualized setup. This may indicate that a simplification from considering context to not considering it may help when the former task is too complex.

While *anger/annoyance* represents the minority class for dreams and performance issues may stem from this, *surprise/confusion* is the second most frequent label in the underlying data. *Surprise/confusion* is apparently easier to recognize without context. This makes intuitively sense, because surprise is not an emotion that develops over time but is instead expressed independently of previous events. Another reason for the single setup performing better could therefore be that the parts themselfes hold enough information to support the achieved classification results.

The overall greater stability in the F_1 scores over all classes for dreams also helps to contextualize the lower Macro F_1 for customer service dialogues shown in Table 6, indicating it likely stems from the data imbalance in the label distribution. Overall, the majority class yields the best results over all setups, indicating that the limited amount of data may be the bottleneck.

Finally, Figure 4 shows the average macro F_1

	CS			Dreams		
	Sin.	Con.	Δ	Sin.	Cont.	Δ
Positive	**.63**	.62	.004	.62	**.63**	−.01
Neutral	**.35**	.22	.13	.28	**.32**	−.03
Fear	.0	.0	.0	.33	**.35**	−.02
Ang./Annoy.	**.73**	.73	.0	**.23**	.13	.09
Sad./Disap.	**.0**	.0	.0	.34	**.35**	−.01
Surpr./Conf.	.0	.0	.0	**.22**	.16	.06
Average	.29	.26	.02	**.34**	.32	.02

Table 7: F_1 score by emotion class, averaged over all folds and iterations. The better score between setups is highlighted.

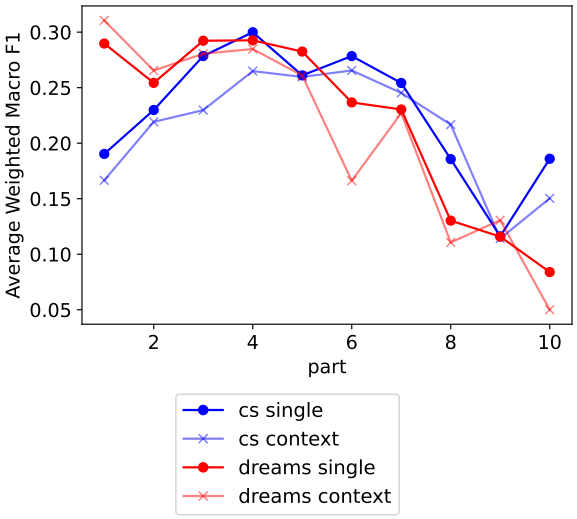


Figure 4: Performance on weighted macro F_1 of the classifiers over the progression of parts. The scores are averages over folds per part and weighted based on how many instances reached the given length. Averages over iterations were done without weighting.

scores over the progression of texts. We first calculated the per-part macro F_1 score for each fold. These scores were then weighted by how many parts had reached the given length. This weight is constant for the first four parts, as all texts in the test set reached the minimum length. After normalizing using the length distribution over all folds, the average is then calculated over all ten iterations.

We observe that, for dreams, the contextualized case starts out with slightly better performance than its uncontextualized counterpart. After three sentences, it falls behind the classifier working on isolated parts and only outperforms it once more after nine parts. Notably, the standard deviation also increases for this classifier at part 9 (not shown in depiction). Similarly, on the customer service domain, considering context yields worse results for all parts but one. Overall, this means that the amount of previous context we consider does not seem to have a positive impact on performance when contrasted with the isolated alternative.

6. Conclusion & Future Work

In this work, we gathered emotion corpora on dreams and customer service dialogues that were labeled through crowdsourcing for the current status of the emotion progression in terms of appraisal scores and categorical emotions. These corpora differ from previous resources in their combination of granularity, context, and scope in that the gathered annotations are sequential and each label reflects what emotion was experienced up to that point. To this end, we developed a novel, incremental annotation task in which the instances are revealed to annotators part-by-part and annotations were gathered for the whole instance up to each part at every step. This ensured that annotators could consider previous context, while not taking into account future parts.

Despite the complexity of this task, annotators reported a high confidence in the annotations they provided and we achieved overall acceptable inter-annotator measures across domains and annotation types. We conclude that the incremental annotation task is a suitable way of gathering cumulative progression annotations.

We furthermore confirmed that emotions do progress over the duration of texts, both in terms of categorical emotions and appraisal scores. Through an analysis of the occuring annotations both over the progression of the underlying text and overall, we found that the progression of emotional content is domain-dependent. In addition, instances are varied in how they progress, making emotion progression prediction a non-trivial task. We then looked into the impact previous context had on the performance of emotion progression classification by finetuning and evaluating a RoBERTa model on parts with removed context and available prior context. Contrasting the results shows little difference between both setups, though the results suggest this may be due to the amount and skewedness of available training data. We found evidence that the importance of context may depend on the emotion class. Repeating the experiment with a different machine learning architecture that is better able to leverage previous context, or gathering more data before repeating the experiment may help gain further insights into its predictive value.

In our experiments, we did not study the role, impact, or value of appraisal annotations. This forms one item for important future work. Several non-basic emotions can be considered programs or scripts that develop out of a sequence of events and associated emotions. Our dataset allows to study these relations, and the evaluation of appraisals is

a means to quantify them. Jointly with emotion predictions, this also has the potential to increase the predictive quality of automatic text analysis models.

Future work may further discern the differences between the gathered sequential labels in the original dataset by Labat et al. (2024), which was contextualized by the whole conversation, yet focused each annotation on only the part it pertained to. This could help to get more insights into how the differences in annotation task and setup influence the final annotations.

Finally, with our work we showed the feasability of the general annotation setup. This enables to gather more data in the same domain and in other domains. That may enable better performing and more robust models, and help to get a clearer picture of the importance of context in progression classification.

Ethics statement

The two datasets we used to gather annotations are publicly available resources. The original EmoWOZ-CS corpus was collected through a Wizard-of-Oz experiment, a well-established technique which inherently results in a low level of deception. Since a low level of deception was involved in the experiment after participants had provided their informed consent, the authors debriefed their participants about the wizard setup at the end of the experiment. Given this new information, participants had the option to withdraw their data from the corpus without any repercussions. The whole experimental setup of this corpus was approved by the Ethics Committee of the Faculty of Arts and Philosophy at Ghent University.

The DreamBank corpus, on the other hand, contains dream reports from a variety of different resources ranging from previous studies to long-time dream journals of individuals. Most importantly to remark in this sense is that the dream reports were willingly shared by their dreamers/authors. Since we noticed that quite a large amount of dreams dealt with negative topics, we manually removed any dream reports that could be upsetting to annotators. Topics that were removed included, e.g., sexual depictions, racism, murders/death, descriptions of butchered animals, etc.

Finally, potential annotators were first informed about the annotation task at hand on Prolific. Interested participants were then referred to the actual questionnaire on SoSciSurvey, in which they received more detailed instructions along with an example annotation. It was clearly emphasized that participation to our study was voluntary and could be withdrawn at any point. Moreover, we ensured that no personal information was collected on the annotators we recruited through Prolific.

Acknowledgements

This research has been supported by the German Research Foundation (DFG), project CEAT, KL 2869/1-2. It also received funding from the Flemish Government under the Research Program Artificial Intelligence - 174Z05623 (2023) and the Research Foundation Flanders (FWO-Vlaanderen) with grant number 1S96322N. We would also like to thank the anonymous reviewers for their constructive comments and insightful feedback.

7. Bibliographical References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer Berlin Heidelberg.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. 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, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48. Association for Computational Linguistics.

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

G. William Domhoff. 1996. The Hall/Van de Castle system of content analysis. In *Finding Meaning in Dreams: A Quantitative Approach*, pages 9–37. Springer US.

G William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on Dreambank.net. *Consciousness and Cognition*, 17(4):1238–1247.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Phoebe C Ellsworth and Klaus R Scherer. 2003. Appraisal processes in emotion. In *Handbook of affective sciences*, pages 572–595. Oxford University Press.

Phoebe C Ellsworth and Craig A Smith. 1988. From appraisal to emotion: Differences among unpleasant feelings. *Motivation and emotion*, 12(3):271–302.

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

Sandra L. Gilchrist, John A. Davidson, and Jane Shakespeare-Finch. 2007. Dream emotions, waking emotions, personality characteristics and well-being–A positive psychology approach. *Dreaming*, 17:172–185.

Ernest Hartmann, Michael Zborowski, and Robert Kunzendorf. 2001. The emotion pictured by a dream: An examination of emotions contextualized in dreams. *Sleep and Hypnosis*, 3(1).

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. SAGE Publications.

Sofie Labat, Naomi Ackaert, Thomas Demeester, and Véronique Hoste. 2022a. Variation in the expression and annotation of emotions: A Wizard of Oz pilot study. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 66–72, Marseille, France. European Language Resources Association.

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2020. Guidelines for annotating fine-grained emotion trajectories in customer service dialogues (version 1.0). Technical report, LT3, Faculty of Arts, Humanities and Law, Ghent University (Ghent, Belgium).

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2023. EmoTwiCS : A corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter. *Language Resources and Evaluation*.

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2024. A customer journey in the Land of Oz: Leveraging the Wizard of Oz technique to model emotions in customer service interactions. Manuscript submitted for publication.

Sofie Labat, Amir Hadifar, Thomas Demeester, and Véronique Hoste. 2022b. An emotional journey: Detecting emotion trajectories in Dutch customer service dialogues. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 106–112, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Agnes Moors. 2017. Appraisal theory of emotion. In *Encyclopedia of Personality and Individual Differences*, pages 1–9. Springer International Publishing.

Tore A Nielsen, Daniel Deslauriers, and George W Baylor. 1991. Emotions in dream and waking event reports. *Dreaming*, 1(4):287.

Sunghong Park, Junhee Cho, Kanghee Park, and Hyunjung Shin. 2021. Customer sentiment analysis with more sensibility. *Engineering Applications of Artificial Intelligence*, 104:104356.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting*

of the Association for Computational Linguistics, pages 527–536. Association for Computational Linguistics.

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

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

Ira J. Roseman. 1984. Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology*, 5:11–36.

Ira J Roseman and Craig A Smith. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. In *Appraisal processes in emotion: Theory, methods, research*, pages 3–19. Oxford University Press.

Klaus R. Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351.

Michael Schredl. 2010. Dream content analysis: Basic principles. *International Journal Of Dream Research*.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299.

A. Appendix

A.1. Overall Appraisal Score Distribution

Figures 5 and 6 show the overall distribution of appraisal annotations for each dimension and domain. For customer service dialogues, we can observe a heavy bias toward low *own responsibilty* in 5, which is mirrored by a bias toward high *others' responsibility*. Overall, the conversation were also more frequently rated as low in *pleasantness* and *familiarity*. Figure 6 shows that dreams were overall rated as low in *event predictability*, *attention*, and *consequence anticipation*. All these biases are plausible within their domains, given that customer service is usually contacted when something on the company's part went wrong, while dreams are often erratic in nature. This lends further credibility to the gathered annotations.

A.2. Progressions in Appraisal Values

Figure 7 shows how the mean appraisal scores progress over the course of the dream and CS dialogue datasets. In both datasets, attention is overall high. For dialogues, we furthermore observe a bias toward high scores for effort and other's responsibility. The latter is mirrored by particularly low mean scores for own responsibility in the same domain. In both domains, we observe a slight increase in attention, effort and other's responsibility as the underlying text progresses. Overall, scores stay relatively constant.

We therefore conclude that for most appraisal dimensions, values depend more on the underlying domain than on how far a text has progressed, though some dimensions may have characteristic developments. The stability in scores raises the question of whether this is based in a lack of progressions within instances, or a variety of changes between them. To investigate this, we looked into the mean absolute changes in annotated values between individual parts per appraisal dimension. While these stay below one for almost all cases in both domains, we report that after the minimum length of four parts, the mean changes sum up to more than one for eight out of ten appraisal categories for dreams and three out of ten for CS dialogues. This means, that after four parts, annotators will on average have changed the score by one for eight appraisal dimensions for dreams and three for CS dialogues. After the mean length which rounds to six parts, this sum exceeds one for all dimensions in dreams and is greater than two for three out of six appraisal dimensions. This implies that for six out of the ten underlying appraisal dimensions, the progressions could on average not have been captured through only two annotations. For CS dialogues, the sum exceeds one for six out of ten appraisal dimensions and two for only one. Overall, this indicates that dreams carry more dynamic appraisal progressions than CS dialogues, though both domains do display changes.

A.3. Mean (Dis-)Agreement Between Annotators for Appraisals

Figure 8d shows the (dis-)agreement of annotators over parts in terms of mean difference between scores. As annotations were made on a five-point scale, the upper bound of this measure is 4. A difference of $\geq$2 can be seen as a general disagreement between annotators, as this would place them either in opposite halfes of the scale, or on an extreme and a neutral position. For the first four parts - the minimum length - the mean differences stay between 0.5 and 1.6, indicating an overall agreement. After that, as the number of samples that reach the required length drops and the average scores get less reliable, the mean differences diverge.

A.4. Instance Length and Overall Emotion Class

Figure 1 shows the distributions of instance lengths by the class that was assigned to the last part of an instance. Since each label represents the emotion up to and including the part they are assigned to, this corresponds to the label describing the overall instance. For dreams, the mean length of instances largely stays at around five for all last classes but *relief*, *anger/annoyance*, *surprise/confusion* and *neutral*, which feature higher mean lengts. For customer service dialogues, more negatively valanced classes describing the whole conversation seem to be correlated with longer texts.

A.5. Example Annotations

Figures 10 and 11 show example annotations from the dataset. Mirroring the according text below, the y-axis represents the parts top-to-bottom. The x-axis represents the annotation decision of each of the three visualized annotators in terms of Likert-score per appraisal dimension and emotion category.[6]

A.6. Incremental Annotation Task

Figure 12 shows the annotation task for the first three sentences of an example dream. Each time a new sentence gets revealed, annotators are first asked for the overall categorical emotion which describes the dreamers experience, before being asked to judge a set of appraisal dimensions on a five-point likert scale.

[6] Emojis designed by OpenMoji. License: CC BY-SA 4.0, https://openmoji.org/

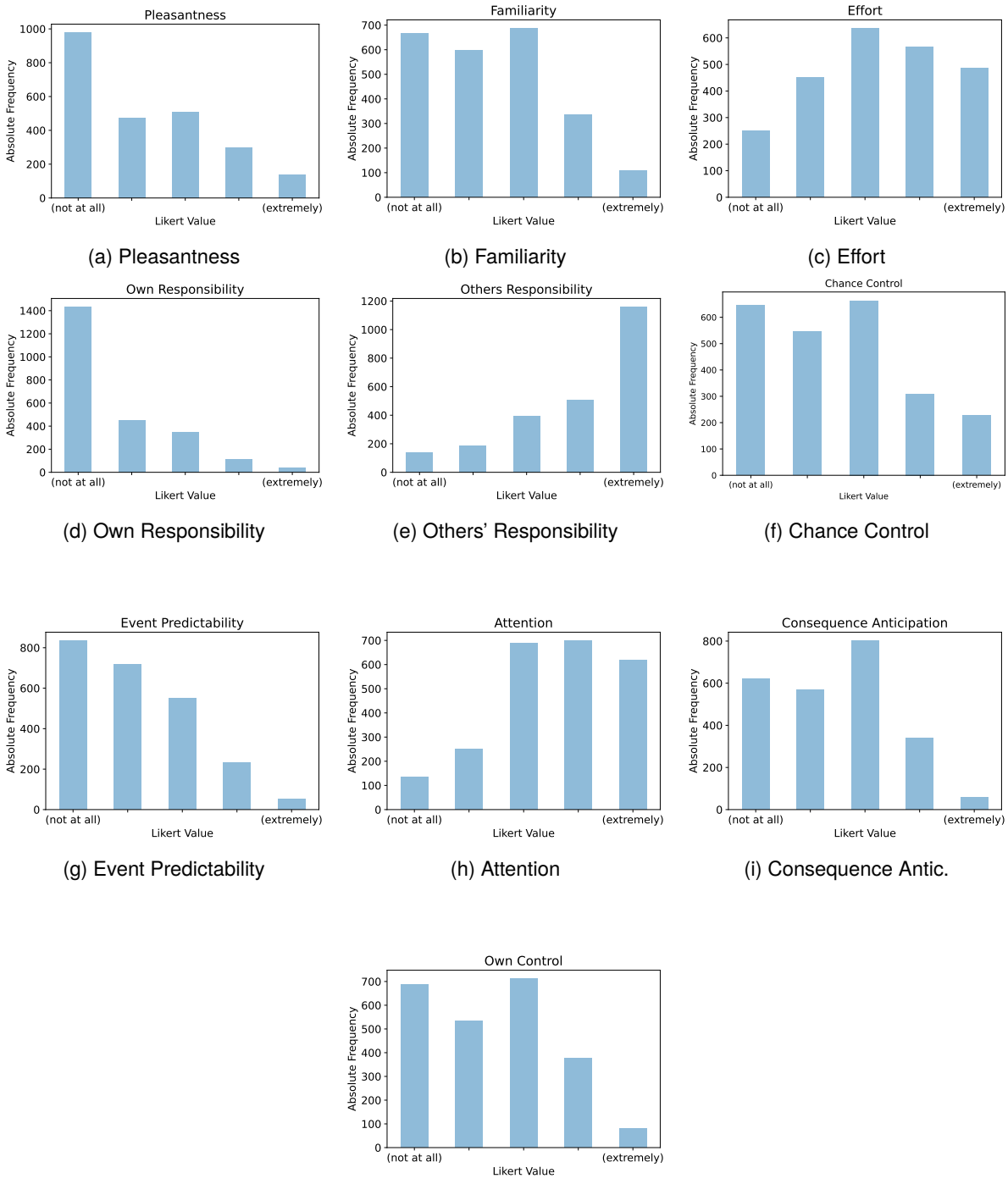
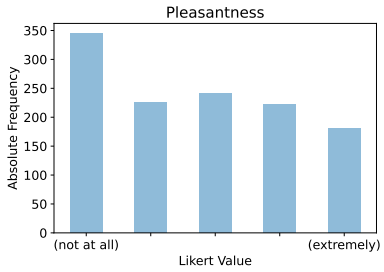
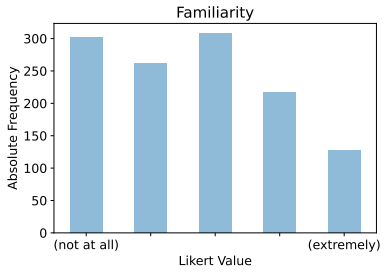


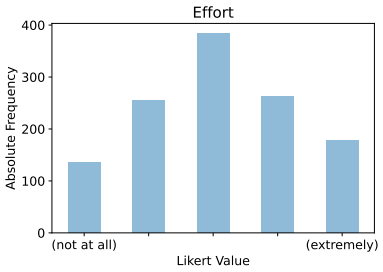
Figure 5: Score distributions for each appraisal dimension over all annotations and parts for customer service dialogues.



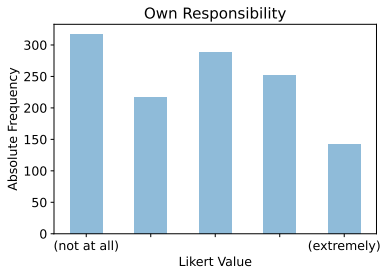
(a) Pleasantness



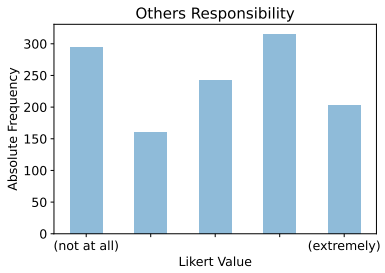
(b) Familiarity



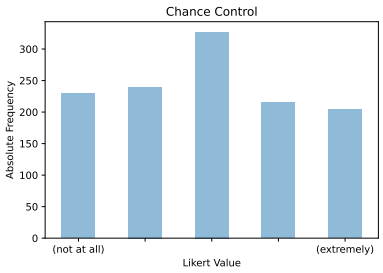
(c) Effort



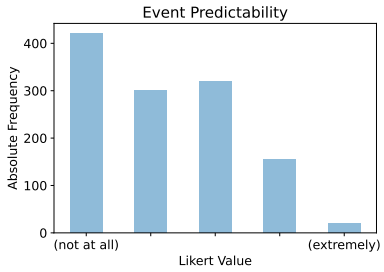
(d) Own Responsibility



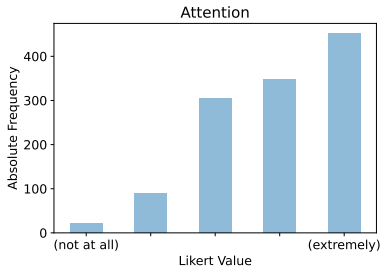
(e) Others' Responsibility



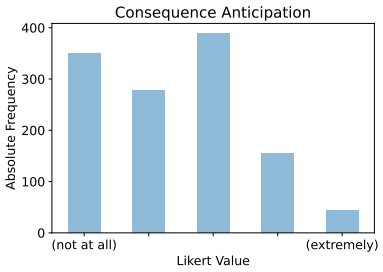
(f) Chance Control



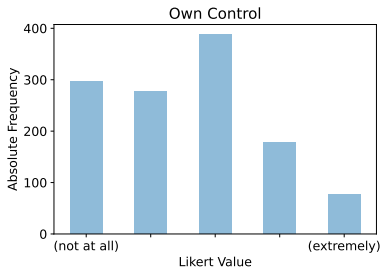
(g) Event Predictability



(h) Attention

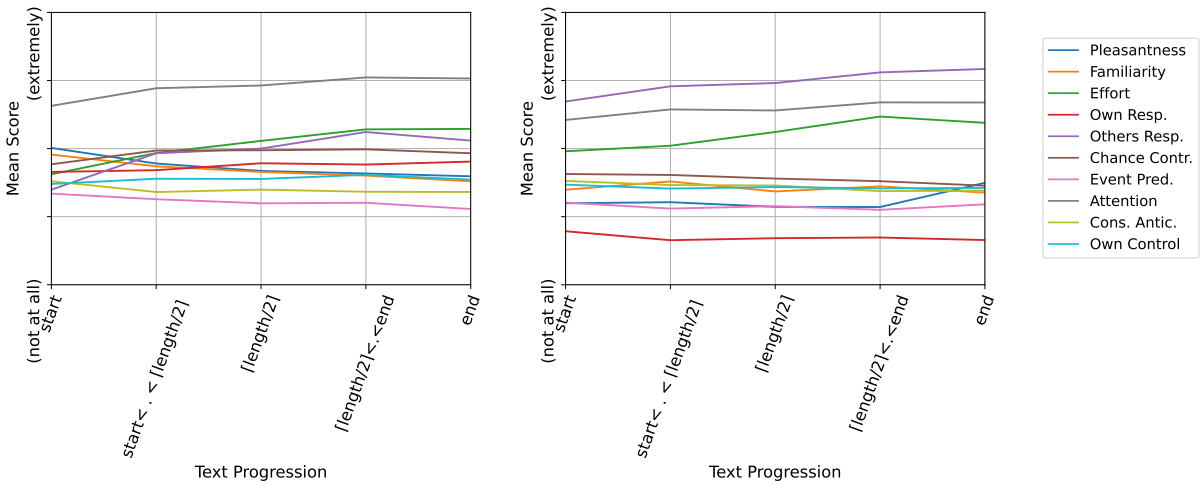


(i) Consequence Antic.



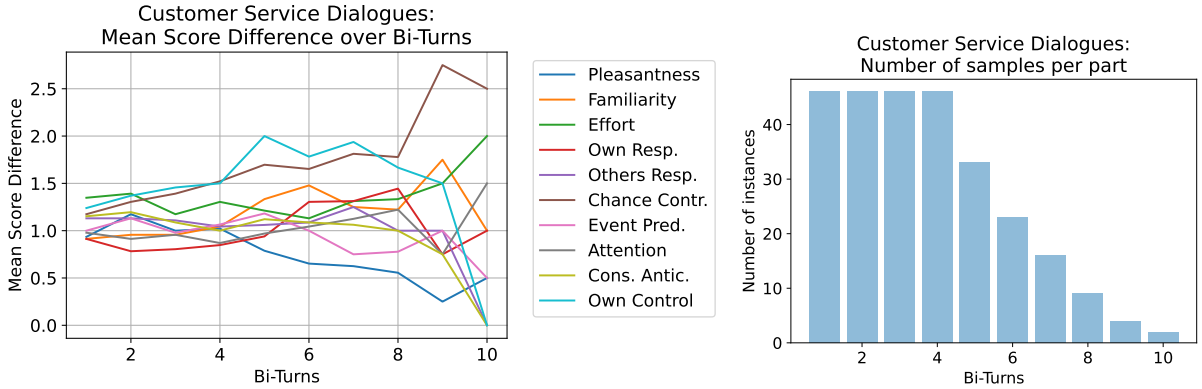
(j) Own Control

Figure 6: Score distributions for each appraisal dimension over all annotations and parts for dreams.



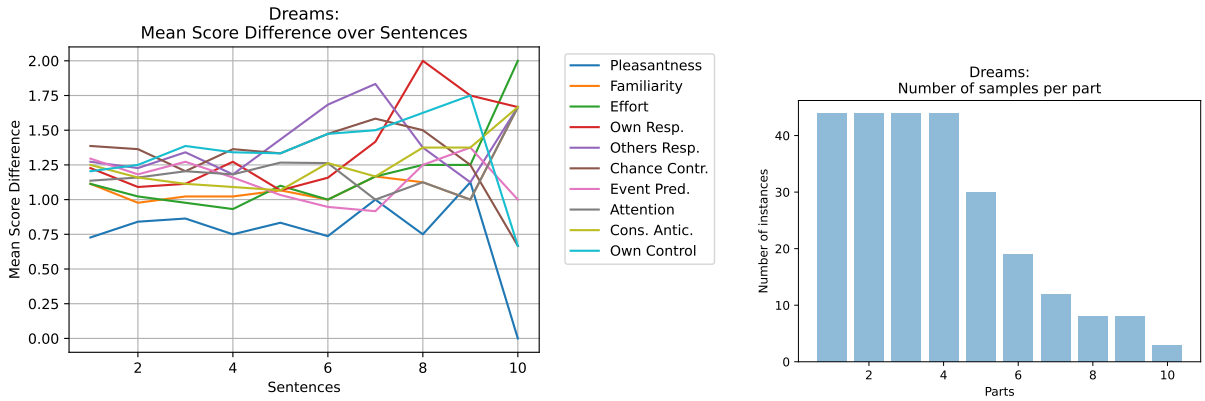
(a) Dreams (b) Customer Service Dialogues

Figure 7: Mean appraisal scores over the progression of all instances.



(a) The mean difference in scores for each appraisal category and part for customer service dialogues.

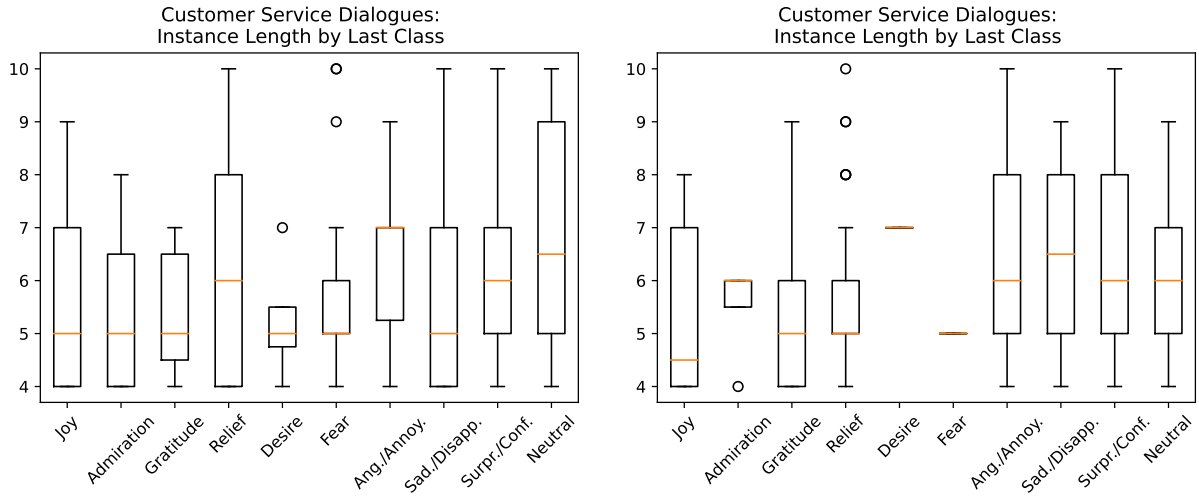
(b) The amount of available evaluation samples per part for customer service dialogues.



(c) The mean difference in scores for each appraisal category and part for dreams.

(d) The amount of available evaluation samples per part for dreams.

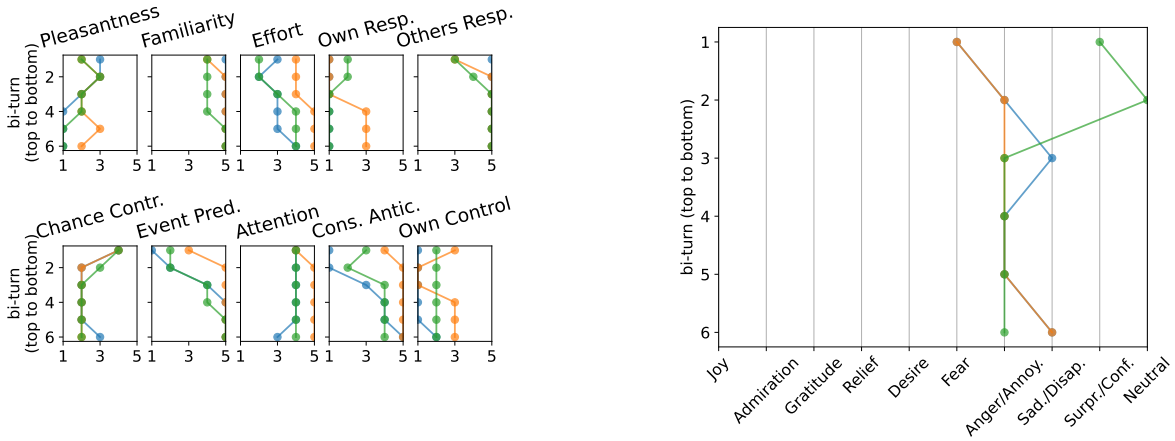
Figure 8: The mean difference of appraisal annotations over the course of texts for both evaluation datasets.



(a) Dreams

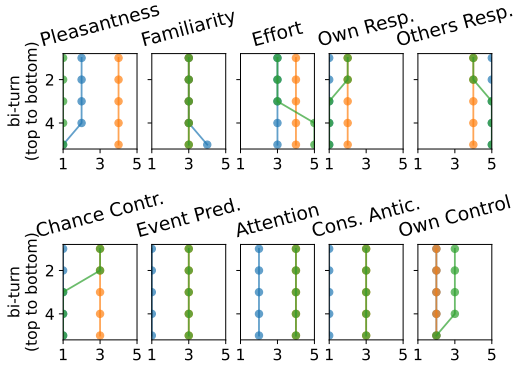
(b) Customer Service Dialogues

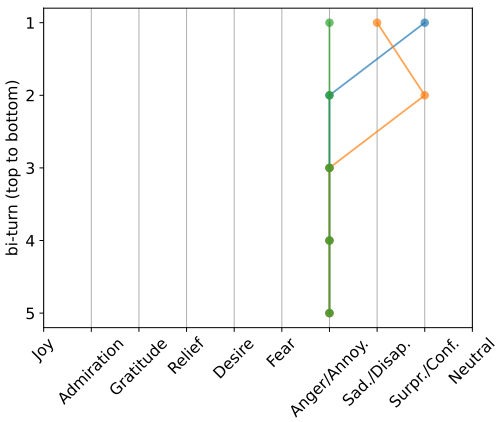
Figure 9: Instance lengths by emotion label for the last part.



1. I see my mother rushing to help Aunt Rosalie.
2. My mother straightens up some clothes and puts them where she wants them, even though Rosalie has different ideas.
3. My mother is pushy and controlling.
4. I turn to Aunt Millie and say, "She is so annoying".
5. Aunt Millie says, "You should have seen her as we grew up"!
6. I see my mother's intense, determined, angry face and I think, "How sad, that was my mother, my 'nurturing' part".

Figure 10: Three example annotations for the dream with identifier *d_b_1966*. Annotations are visualized top-to-bottom as the text progresses. Each color represents one individual annotator.





1. ==ADMIN== Hello, my name is Chatty10 How can I help you today?
   ==PART== Hello, it says on the site that my package has been delivered, but I haven't received anything.
2. ==ADMIN== Oh, that's not supposed to happen.
   ==PART== I found it in the mailbox, but it was all wet.
3. ==ADMIN== Could you please give me your order number?
   ==PART== the rest is also damaged
   order number is 33221100
4. ==ADMIN== Your package was delivered according to our information.
   ==PART== but it's broken
5. ==ADMIN== Are there any other problems?
   ==PART== it's broken

Figure 11: Three example annotations for the customer service dialogue with identifier *c288_9*. Annotations are visualized top-to-bottom as the text progresses. Each color represents one individual annotator.

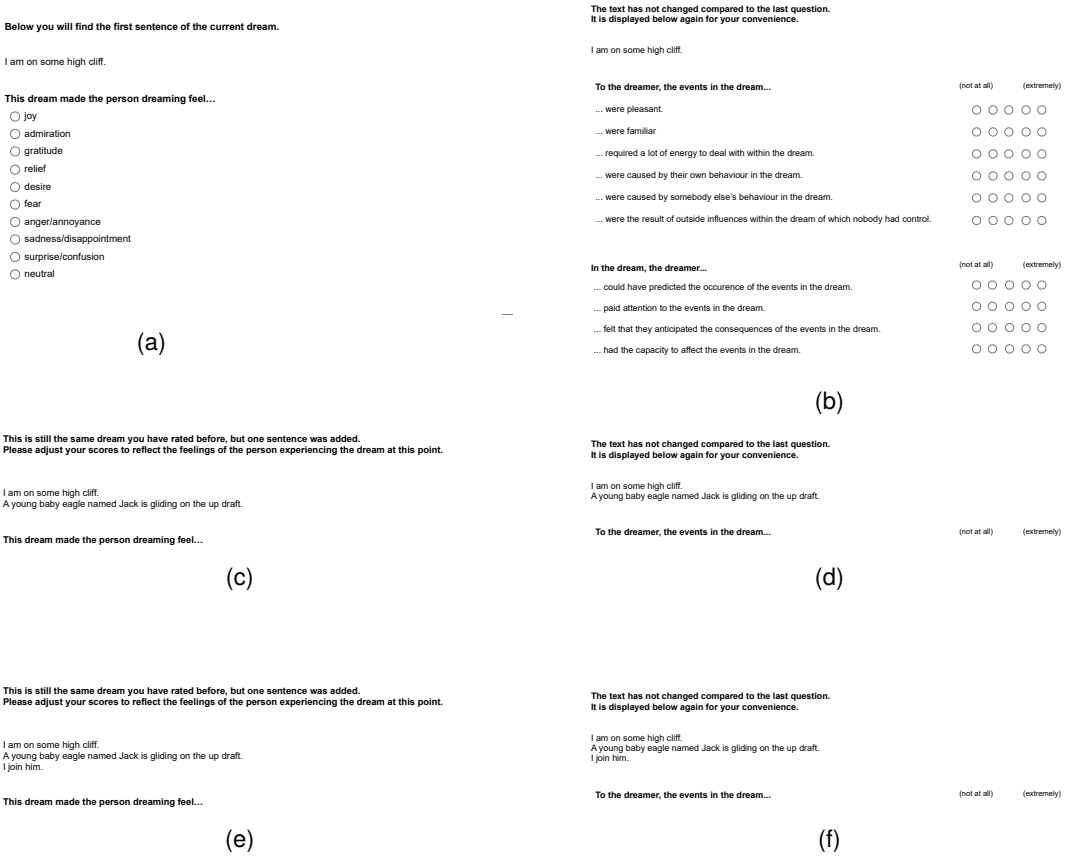


Figure 12: The first three steps in the annotation procedure for an example dream in order. Annotators were only shown the text up to and including the part they were currently rating. Answer options stayed the same as displayed in Figures 12a and 12b and previous answers were preselected for the next part.