Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung

Pfaffenwaldring 5b
70569 Stuttgart

Masterarbeit

Progressions of Emotion Annotations in Conversations and Dreams for Emotion Analysis

Eileen Wemmer

Studiengang:	Informatik
1. Prüfer:	PD Dr. Roman Klinger
2. Prüfer:	Prof. Dr. Sebastion Padó
Betreuer:	Sofie Labat (Ghent University)
begonnen am:	04.11.2022
beendet am:	04.05.2023

Abstract

Emotion Analysis describes the field of study concerned with the extraction of explicit emotional content implicitly contained in text. To this end, supervised machine learning approaches are commonly employed which rely on annotated data for training and evaluation. This data dictates the tasks trained models will be able to solve and as a consequence, multiple corpora containing emotion annotations have been gathered in the past. These corpora commonly vary in multiple ways, including their underlying domain, the scope of each annotation, and whether they feature emotion representations in terms of categories or their position in a vector space spanned by interpretable dimensions. However, to the best of our knowledge, none of the previously gathered corpora allow for a fine-grained analysis and prediction of the progression of emotional content over the course of texts. While sequential annotations representing the emotional content of each part a text is comprised of exist, they usually only pertain to the parts they are attached to and do therefore not accurately reflect the current emotions expressed in the overall text at this point. Similarly, contextualized annotations that take into account the previous content of a text exist, yet they are not gathered in sequences and therefore also don't allow for the analysis of the changes in emotional content. This thesis aims to close that gap by gathering progressional labels, that are both sequential and contextualized by the previous text, through a novel, incremental annotation task. In a crowdsourcing setup, texts are revealed to annotators part-by-part and they are asked to annotate the emotional content in terms of categorical labels and appraisal dimensions up to the current point. This yields a set of sequential labels that represent the development of emotional content up to the part of the text they are assigned to. We gather progression annotations for both dream reports and customer service dialogues and find the novel incremental annotation setup to be suitable for their collection. Analyses of the data show that changing progressions exist in both domains, though they are more varied for dream reports than for customer service dialogues in terms of how often annotations change throughout one instance. We do not uncover any clear tendencies in progressions for either domain, implying a rich variation in changes between instances. Leveraging the gathered data to study the degree to which simple sequential models are able to learn to make use of this progressional information, we show a consistent increase in performance for models trained on intact categorical progressions that is too small to be conclusive. This motivates further research, as experiments with stronger baseline systems could help get clearer insights into the matter.

Zusammenfassung

Das Forschungsgebiet der Emotionsanalyse beschäftigt sich mit der Extraktion expliziter emotionaler Inhalte, welche implizit in Texten zum Ausdruck gebracht werden. Hierfür werden in der Regel überwachte maschinelle Lernverfahren eingesetzt. Ihr Training und ihre Beurteilung basieren auf annotierten Daten. Diese Daten schränken ein, welche Aufgaben trainierte Modelle lösen können. Entsprechend wurden in der Vergangenheit zahlreiche Korpora mit Emotionsannotationen gesammelt. Diese unterscheiden sich in vielerlei Hinsicht, wie der zugrundeliegenden Domäne, der Granularität der einzelnen Annotationen sowie darin, ob sie Emotionen in Form von Kategorien oder deren Position in einem Vektorraum mit interpretierbaren Dimensionen darstellen. Nach unserem Kenntnisstand erlaubt jedoch keiner der bisher gesammelten Korpora eine Analyse und Vorhersage des Verlaufs emotionaler Inhalte in Texten. Zwar existieren sequentielle Annotationen, welche jeweils den emotionalen Inhalt jedes Teils des Textes wiederspiegeln, jedoch beziehen sich diese in der Regel nur auf eben den Teil, welchem sie zugeordnet sind. Entsprechend spiegeln sie nicht die Emotionen wider, welche der Gesamttext bis zu diesem Punkt ausdrückt. Ähnlich verhält es sich mit kontextualisierten Annotationen, welche zwar den vorhergehenden Text berücksichtigen, aber welche bislang nicht in als sequentielle Annotationen gesammelt wurden. Das Ziel dieser Arbeit ist es, diese Lücke zu schließen. Hierzu werden durch ein neuartiges, inkrementelles Vorgehen Annotationen durch Crowdsourcing gesammelt, welche sowohl sequentiell den emotionalen Inhalt des Textes für jeden Teil widerspiegeln, als auch den Kontext des vorangehenden Textes berücksichtigen. Hierbei werden den Annotatoren Texte nach und nach offenbart und sie werden gebeten, den emotionalen Inhalt bis zum aktuellen Punkt zu beschreiben - sowohl durch Kategorien als auch durch Ordinalwerte entlang kognitiver Bewertungsdimensionen (Appraisals). Daraus ergibt sich eine Menge sequentieller Annotationen, welche jeweils den emotionalen Inhalts des gesamten Textes bis zu jenem Teil darstellen, dem sie zugeordnet sind. Annotationen werden hierfür sowohl für Traumberichte als auch für Kundendienstdialoge gesammelt. Die Analyse der Daten zeigt, dass sich in beiden Domänen im Verlaufe der Texte Änderungen in den emotionalen Inhalten ergeben, wobei diese bei Traumberichten häufiger auftreten als bei Kundendienstdialogen. Eindeutigen Tendenzen konnten in den Verläufen beider Domänen nicht gefunden werden. Dies deutet darauf hin, dass die Verläufe zwischen den einzelnen Texten variieren. Anschließend nutzen wir die gesammelten Daten um zu untersuchen, inwieweit einfache sequenzielle Modelle diese Verlaufsinformationen nutzen und vorhersagen können. Unsere Ergebnisse zeigen eine konsistente Leistungssteigerung für Modelle, welche während des Trainings Zugriff auf den korrekten Verlauf von Emotionen hatte gegenüber solchen, die dies nicht hatten, welcher jedoch zu klein für eine eindeutige Interpretation ist. Zukünftige Experimente mit stärkeren Modellen könnten hier für mehr Klarheit sorgen.

Contents

1. Introduction

Emotions accompany us throughout our lives: From childhood on we learn to recognize emotions in others [26] and to express our feelings through language [9]. Therefore, it comes as no surprise that emotions have long been the subject of study in psychology, which has yielded a large number of theories that try to describe what emotions are, what their purpose is, and how they develop. These theories are called emotion theories and have since been leveraged by researchers in the domain of computer science with the goal to develop automated systems that are able to classify or quantify which emotions people express. When applied to written text, this area of research is known as Emotion Analysis. Currently, the most powerful and established systems fall under the category of supervised machine learning, a subset of artificial intelligence that relies on annotated data for training and evaluation.

For Emotion Analysis, this data usually comes in the form of text samples for which humans have explicitly assigned the emotion(s) they contain. These explicitly assigned targets for the following training tasks are called labels or annotations, and they limit what tasks the system will be able to perform later on. For example, if a system is only shown examples labeled "joy" and "sadness" during training, it will not know how to recognize "anger" later on or be able to quantify *how much* sadness a text expresses. In addition, emotion prediction tasks have been found to be highly domain-specific [13] - a system that learned to judge the emotional content of news headlines might not perform well when performing the same task for tweets. All of this motivates the need for corpora, or annotated datasets, for Emotion Analysis in general, and a substantial amount of effort was dedicated to the development of suitable resources [11].

However, to the best of our knowledge, none of the corpora gathered so far feature annotations that reflect the current state of the *emotional progressions* in texts, meaning the way the expressed emotions change and develop over the course of a text. Annotations are usually gathered for the entire textual unit, or instance, that will later be subject to automated evaluation. Even when annotations were gathered on a more fine-grained level, for example for every sentence in fairy tales [4], or by explicitly highlighting what parts of customer responses in the exchange with customer service have evoked which emotions [47, 48], these annotations only reflected the emotional content expressed in that part. However, not only do the emotional semantics of parts of a text depend on context, but they also describe an emotional progression that depends on the prior information specifically. If we consider the sentence "Today, I did not get any phone calls.", the sentence might be labeled neutral when asked for the writer's emotion. However, this changes if we consider this sentence in different contexts. Rating the current emotional experience of the following texts up to each sentence with either "joy", "neutral", or "sadness" might yield displayed labels:

Sentence	**Label**	**Sentence**	**Label**
It's my birthday!	joy	This morning I got up and watered my plants.	neutral
Today, I did not get any phone calls.	sadness	Today, I did not get any phone calls.	neutral
But when I came home, all my friends were there!	joy	It's been a busy week at the call center, so I was glad to catch a break.	joy

This example shows how the emotional content interpretation of the middle sentence "Today, I did not get any phone calls." inherently changes with the annotation task: When considered in isolation, it may be labeled neutral. When adopting the label of the overall instance it may be labeled joy in both cases, as both reports detail an overall happy story for the writer. The same holds true when considering the preceding and following parts as context, as has been done for other contextualized sequential annotations [44]. In this case annotators would know that no phone calls were great news for the person in the second text and that the according emotion annotation would likely be joy rather than neutral. Yet, when considering the label as a representation of the emotional content up to that point - as a snapshot of the overall emotional progression of the presented text up to that point - that information is not yet available for the line about phone calls, hence joy is not assigned in the second case and sadness takes its place in the first. This emotion information would therefore not be captured with either of the other annotation setups. Annotations contextualized by only the preceding text have also been gathered before [18], but not in a sequential manner. While these labels carry the context information of the previous text, there is only one data point to leveraged. Therefore, these annotations are also unsuitable to capture, analyze and predict the emotional progressions found in text.

In this thesis, we aim to fill this gap by gathering progressional annotations, which can be characterized as sequential in nature *and* contextualized by the content leading up to each part. We do this to then analyze the potential of progressional emotion information for Emotion Analysis through a preliminary computational experiment. To this end, we propose a novel, incremental annotation task that reveals texts to annotators part by part. For each part, annotators will be asked to judge the emotional content up to and including the last added part. The incremental approach is employed to ensure later developments in an instance do not influence annotators' judgment of the current emotional content.

With this, we aim to answer three central questions:

Q1 Is the proposed method of gathering annotations through an incremental task suitable to acquire information about the progression of emotional content in a text?

Q2 How do emotion annotations build up over the progression of a text?

Q3 To what extent can this progressional information be leveraged for Emotion Analysis?

To this end, we gather two corpora on different domains, namely customer service dialogues and dream reports. Analyzing the data of two domains allows us to compare the results and to better judge which effects are due to domain-specific, underlying dynamics. Furthermore, we gather two sets of progressional annotations for each instance in the corpora: One is based on categorical emotion models, where annotators are asked to pick an emotion word that best describes the contained emotions. The other is based on a numerical description of emotions. In this case, annotators judge to what extent certain cognitive evaluations (appraisals) about the events in the texts apply from the experiencers' point of view. Both annotation sets are then analyzed for both domains to answer the first two research questions. After that, we train models on both intact and disrupted categorical emotion progressions and compare the results. As the second set of models does not have access to the correct sequential or contextual information, we expect the first set of models to have an advantage in predicting emotion progressions. This would evidence that the added scope of emotion information in progressional annotations can be leveraged by sequential machine learning models.

2. Background and Related Work

To learn to extract the expressed emotional content from text, we need to answer two central questions first: For one thing, we need to define what emotions are. Psychological research has addressed this question repeatedly and from multiple viewpoints. This has led to a set of so-called emotion models, that aim to answer that exact question. We will turn to these emotion models to define what annotations we want to gather and, consequently, what exactly we want to be able to extract from the texts in our target domains. To this end, we introduce a set of relevant emotion models in section 2.1.1.

The second question we need to answer is how can we extract any type of information from text. This falls into the area of research called Natural Language Processing (NLP). More specifically, we are interested in how approaches in NLP have already been employed to analyze texts with regard to their emotional content. This sub-field of NLP is known as Emotion Analysis. In section 2.2, we introduce the relevant research to answer both these questions to define a starting point for our own work and methodology, which we will introduce in the next chapter.

2.1. Emotions and Emotion Models

For this work, emotion models can help us define what exactly we want to ask annotators: One of the central goals of this thesis is to construct two corpora: one for customer service dialogues and one for dreams. These emotion corpora should carry annotations, or labels, that explicitly describe the progressions of the emotional content both implicitly and explicitly expressed through text. To gather these corpora, we want to use crowdsourcing to label our texts. This means that instead of training a small group of experts, we want to ask a multitude of non-experts to judge the emotional content of our texts. Yet, as "emotion" is a fuzzy term, we would likely get a very diverse and hard-to-compare set of labels from an approach that simply asks "What emotion was person x experiencing?". For example, the sentence "I was writing a love letter to my boyfriend!" could yield a whole host of viable answers, if we chose to ask an open-ended question like the one just posed. These could for example include "happy", "love", or "admiration". For our dataset, we want to limit and define what annotations could look like, to enable later analyses and hopefully gather enough data for each type of annotation to enable machine learning (ML) models to learn to recognize them. In addition, when asking the question above, we would likely get an answer that describes the emotion in one or more terms, or categories. This disregards the fluidity of emotions: Rather than labeling certain emotions as just applicable or not applicable, we could also describe them along a set of dimensions that could be more or less pronounced for different texts. An example of this could be to ask "How positive did person x feel?" about our example above. Both the categorical and the dimensional approaches have been explored in various emotion theories, some of which have particularly attracted attention for their applicability in Emotion Analysis [89]. We will briefly introduce them in the following and refer to Troiano et al. [89] for a more comprehensive overview of emotion models in the context of Emotion Analysis.

2.1.1. Emotion Categories: Models of Basic Emotion

One of the most popular emotion models for Emotion Analysis is the model of basic emotions by Ekman [27], who names six basic emotions: Anger, Surprise, Disgust, Enjoyment, Fear, and Sadness. He poses that emotions are the product of evolution, designed to help us deal with "fundamental life-tasks". This evolutionary view on emotions that can be distinguished by expression is in line with Darwin's view of emotions [22]. According to Ekman, these emotion distinctions are based on their expression, especially through facial movements, though he notes that this is probably not the only factor separating them. This means that in this theory, what distinguishes one emotion from another is how a person expresses said emotion. When transferred to NLP, this is an important assumption, that we often implicitly make when trying to infer emotional states from texts [89]: We assume, that a person expresses their emotions, and that we can distinguish - classify - these emotions based on how they were expressed. For example in the domain of customer service dialogues, unless a customer clearly states how they feel, we need to infer this from the way they express themselves, for example through word choice.

Plutchik's model of basic emotion, depicted in 2.1, not only adds two categories to the ones proposed by Ekman, namely Anticipation and Trust, but also introduces the similarity between categories, mixed emotions and intensity [69]. The wheel can be interpreted as a cone with one dimension depicting the intensity of the emotion, making Plutchik's model three-dimensional. On the wheel, this is represented by saturation and position relative to the center. The further out an emotion is placed, the less intense it is. For example, it places Rage as a more intense version of Anger, with Annoyance as the less intense variant. In addition, closeness indicates similarity between emotions. Two emotions placed opposite of one another, such as Joy and Sadness, therefore implies that two emotions are opposites. With this grouping of emotions, Plutchik tackles the problem we introduced at the beginning of this section: It allows to account for a variety of emotion concepts that we can describe with language, such as the picked examples of "love", "happiness" and "admiration" to be related back to basic emotions through their positioning on the wheel and the basic emotions that they are a combination of. For example, it poses love as a low-intensity mix of joy and trust, as shown in fig. 2.1. Like Ekman and Darwin, Plutchik describes evolution as a driver of emotion and describes their expression as one of the main criteria that separates emotions.

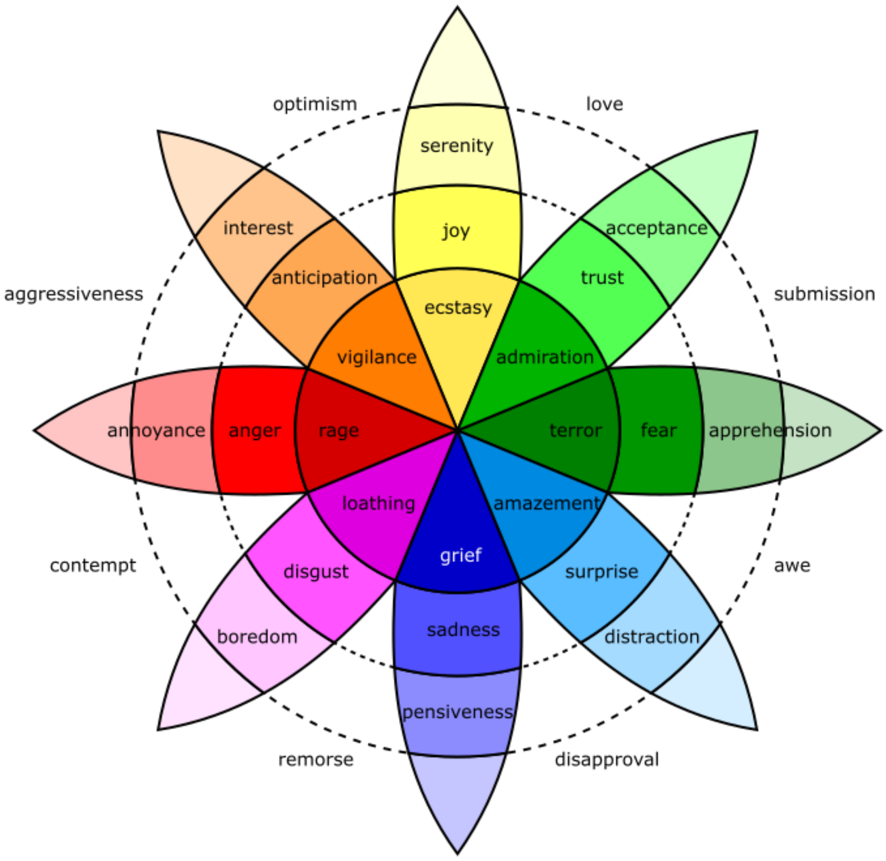


Figure 2.1.: Plutchik's wheel of emotion. Downloaded from [68], adapted from [69].

2.1.2. Emotions Along Axes: Dimensional Emotion Models

Dimensional models describe emotion along a predefined set of intuitively interpretable, bipolar axes. While Plutchik's wheel of emotion incorporates some of the aspects of dimensional models, by introducing similarity and intensity, the Circumplex Model of Affect approaches the modeling of emotion from the opposite direction [72] . It suggests that any emotion can be described along two continuous, bipolar dimensions: **V**alance (high-low pleasantness), and **A**rousal (high-low activation). At times, the third dimension **D**ominance is added. The resulting model model of affect is sometimes referred to as the VAD model, by the first letters of the dimension it considers [79]. The description if emotions with continuous values implies that in this model categorical emotions can be placed in the space spanned by these dimensions, which also means that they are not inherently discrete. The description of emotions through a set of numerical values enables the modeling of computational tasks that extract emotional content from text as a regression rather than a classification problem. When the goal is to extract valance from text instead of emotion, a task falls into the category of Sentiment Analysis, which is often regarded as a related, but separate task to Emotion Analysis [60].

We find another set of dimensional emotion models in the so-called appraisal theories of emotions. Proposing that the two dimensions often used to describe emotions (valance and arousal) are not sufficient to describe their commonalities and differences, appraisal theories introduce more dimensions to the differentiation [86]. Instead of focusing on how emotions are expressed, what function they serve, or how they effect the person experiencing the emotion, appraisal theories describe emotions as and differentiate between them based on the subjective, cognitive evaluations - the appraisals - an individual makes when faced with an event [58, 81]. While different appraisal theories differ in multiple ways, one notable difference between them is the quantity and selection of appraisal dimensions they consider [28, 81, 78, 77]. One example of such an appraisal dimension is *pleasantness*, which corresponds to the valence dimension in the VAD model. Another such dimension is certainty. While coming home to a party for your birthday is probably pleasant for most people, certainty can help us differentiate between different pleasant emotions that might arise. If a person knew it would happen, organized it, and therefore is certain of the reason their friends are in their living room, they will likely be happy about seeing them there. If, on the other hand, their friends kept it a secret and upon first seeing them the person is not quite certain about what is happening, they might instead feel surprised. This example also reveals how appraisals account for differences in emotional reactions between individuals to the same event: We assumed that most people would find a birthday party pleasant. If the person in question appraises birthday parties as unpleasant, they might instead feel fear or anger at the prospect of getting through the event when presented with a group of people in their living room.

While the sectioning in this work might suggest categorical and dimensional models to exist in spite of one another, the different emotion models presented are not necessarily in conflict with each other, as Ekman and Smith and Ellsworth explicitly state [27, 86]. One of the probable differences between his proposed basic emotions that he explicitly mentions, other than expression, is appraisals. Ekman also notes "distinctive universals in antecedent events" [27] as one of the nine key factors that distinguish basic emotions. According to Scherer's Component Process Model, appraisals are the first step in a series of systematic changes in a variety of systems that are intertwined - the emotion components - and that together form the emotion [81]. He relates this to categorical emotions by posing them as an interpretation of those parts of these changes that become conscious. In the same vein, he explains the assignment of emotion words to the actual emotions. Smith and Ellsworth [86, 29] investigate the correlation of certain appraisals with categorical emotions and find that categorical emotions can be distinguished through appraisals. They consider appraisals as a way of describing the similarities and differences between categorical emotions, posing the question "[...] in what ways can emotions be "opposite"?" [86] - a question that could be linked back to Plutchik's wheel of emotion displayed in fig. 2.1, where opposite emotions are positioned on opposite sides of the wheel.

While all mentioned emotion models can and have been employed for Emotion Analysis, categorical and dimensional models allow for different kinds of analyses, as becomes apparent in section 4.1.2 and section 4.1.3. Since this thesis seeks to get an insight into how emotions progress in texts, gathering both numerical and categorical progressions therefore promises broader insights. Although outside the scope of this thesis, gathering appraisals in particular would allow for an analysis of if and to what extent appraisals and categorical emotions progress in conjunction.

2.2. Natural Language Processing and Emotion Analysis

As introduced before, Emotion Analysis deals with a subset of tasks that are common in NLP. Therefore, advances in NLP are usually leveraged to improve performance and approach new problems in the realm of Emotion Analysis. This section gives an overview of the resources and approaches that have been popular in Emotion Analysis. More specifically, we focus on those works that are particularly relevant to this thesis because they pertain to sequential emotion annotations or predictions, feature work with appraisals, or focus on one of the two domains we work with. To this end, we first introduce generally relevant corpora in in section 2.2.1, excluding domain specific work. We then discuss basic approaches of NLP that have been popular for sequence prediction tasks in Emotion Analysis in section 2.2.2. In section 2.2.3, we finally consider works specific to dialogues and dreams.

2.2.1. Corpora for Emotion Analysis and Their Evaluation

This section gives a brief overview of corpora for Emotion Analysis that are particularly relevant to this thesis due to the annotations they contain or because of insights from the data-gathering process that are of interest. Corpora relevant to the dialogues and dreams are introduced in section 2.2.3. From there, we introduce common metrics used to evaluate the reliability of emotion annotations that will later serve us to evaluate the data we gathered to help answer the first research question posed in Chapter 1. We refer to Bostan and Klinger [11] for a comprehensive overview of corpora for Emotion Analysis up to 2018.

Corpora and Approaches to Data Collection

Some previously gathered corpora feature sequential labels that describe the emotional content on a more fine-grained level than for the whole text. One example of this is the corpus built on blog entries by Aman and Szpakowicz, that features annotations on sentence level [6]. Similar to our motivation, the authors chose this granular annotation scheme because, as they state, "there is often a dynamic progression of emotions [...] in the conversation texts and blogs" [6]. However, the categorical labels and intensity values featured in their corpus represent the emotional content of only the sentences they are assigned to and do not, as the annotations we gather, the emotional content of the entire text up to that point. Another example for annotations on sentence level can be found in the corpus gathered by Alm et al. [4]. The categorical labels for in this corpus were gathered from fairy tales with the intention of training a classifier that would allow text-to-speech systems to read texts from the domain with the appropriate intonation. To this end, annotations reflect the emotions found in each sentence. In addition, corpora based on conversations often feature sequential annotations. We introduce some of them in section 2.2.3.

Besides categorical labels, we also gather appraisal scores for both domains. Generally, the annotations of a corpus dictate the tasks trained models can be used for. An exception to this are works that use the idea that named emotion categories have numerical equivalents in dimensional models to either learn mappings between these representations [14, 13], or that leverage existing mappings [64]. Interest in appraisal theory for Emotion Analysis has risen over the last few years. Some corpora with appraisal annotations have been gathered so far, though less than for categorical emotions and VAD annotations. The APPReddit corpus features posts from the social media platform Reddit with four-point appraisal annotations in five different categories [87]. Most other corpora with appraisal annotations feature event descriptions, which ensures the presence of an event to be appraised. One example of this is the x-enVENT corpus gathered by Troiano et al. through expert annotations [88]. The annotations feature appraisal annotations of 22 appraisal dimensions on a five-point scale, along with categorical emotion labels and span labels highlighting the experiencer and appraised event. The crowdsourced crowd-enVENT corpus by Troiano et al. [89] also features annotations for both categorical emotion, and scores for 21 appraisal dimensions on a five-point scale. However, in this setting, the reports were annotated by both the original experiencer and a second party. In their analyses of this data, they found that while second-party annotators agree more often with each other than with the original experiencer, yet their findings still support the claim that inference of appraisal values on the base of text alone is possible. Hofmann et al. [42, 41] gathered binary appraisal scores for seven appraisal dimensions and performed computational experiments that showed not only the viability of appraisal prediction from text but also that appraisal prediction as an intermediate goal could bring improvements for emotion classification tasks. To the best of our knowledge, all annotations for appraisals that are available so far refer to the whole instance.

Inter-Annotator Measures

One common way of judging the quality of annotations is by analyzing inter-annotator agreement (IAA) for categorical data, and inter-annotator correlation for continuous and ordinal annotations [89, 12]. While for the latter case, general correlation measures, such as Pearson's r are often employed, categorical annotations have yielded statistics specific to this use case. One such measure is Cohen's κ [20]. It is commonly employed, as it controls for agreement by chance between two annotators. Since this correction relies on statistics specific to the two annotators whose agreement is under scrutiny, it cannot be employed on a set of twice-annotated data points, that have not been rated by the same two people [32]. While its generalization, Fleiss' κ [31] can take into account more than two annotators, the basic assumption remains that all of these annotators labeled the same instances. Similarly to inter-annotator correlation, the alternative consists of resorting to more general measures of "agreement". Another option that does not suffer from this constraint, and that has still be found to represent IAA well is the F-score [43]. Originally designed to evaluate the performance of classifiers, when judging IAA, we need to assign one annotation as the gold standard and check the other annotation against it. When precision and recall are weighted the same, we obtain the F1-measure, which is calculated as the harmonic mean of both values:

$$F1 = \frac{2 \cdot (precision \cdot recall)}{precision + recall}$$

Precision and recall, in turn, are defined for binary classification problems and respectively measure what percentage of identified targets actually were targets and what percentage of all actual targets were identified. A typical approach for extending these measures to multi-class classification problems is one-hot encoding the labels for each class to obtain a binary one-vs-rest classification problem for each class, followed by averaging. Micro averaging refers to the average over all samples of all classes. In a multi-class setting, it is therefore equivalent to accuracy [37]. For macro averaging, we first compute one average per class, then average over the class scores. This form of averaging weights all classes equally.

Inter-annotator measures for emotion annotations are often low to moderate [90]. One reason for this can be found in the subjective nature of emotion perception [83]. This is evidenced by the findings of Hofmann et al. [41], who showed that IAA for appraisal annotations increases when annotators are presented with the associated categorical emotion along with the texts they were to annotate. Still, event descriptions have been found to contain enough information for appraisal annotation and classification by themselves [89]. Troiano et al. [90] further showed, that emotion intensity, reported confidence in annotations and IAA are correlated. In addition, Buechel and Hahn [15] showed, that the perspective that annotators are asked to take during annotation has an impact on IAA. They compared annotations from the perspective of the reader, text, and writer and found the latter to yield the best results. Mohammad and Turney [56] give an overview of some of the key challenges in crowdsourcing emotion annotations. These include the possibility of cheating, which motivates measures to filter out inattentive annotators, or such that employ bots. They also highlight the importance of easy-to-understand task formulations, and find that the formulation of annotation guidelines has an impact on IAA.

2.2.2. Basics of NLP and Machine Learning for Emotion Analysis

This section introduces relevant machine learning frameworks, and the use of lexicons for Emotion Analysis tasks. It serves as a primer for the utilized methods in the remainder of this thesis.

Lexicons

An early approach to linking emotions to text consisted of building lexicons that explicitly link certain words to certain emotions. Popular representatives of this approach include the Linguistic Inquiry and Word Count (LIWC) [66], a lexicon-based program which is still popular in psychological research to this day, and the NRC lexicon [56]. The idea behind these lexicons is that certain words are related to certain emotion concepts. For example, the word "laugh" could be associated with joy. When encountering the sentence "I laugh loudly.", we would look up every word, encounter the word "laugh", associate it with joy after a lexicon lookup and hence, categorize the whole sentece as expressing joy. While these resources are easy to use and results are clearly interpretable, as scores are derived directly from a word level, they are unable to take into account context that goes beyond word usage and any out-of-vocabulary words can not be leveraged.

LSTMs

Advancements in NLP also brought about shifts in the approaches commonly used for Emotion Analysis tasks. At the WASSA shared tasks on emotion intensity in 2017 [57] and implicit emotion in 2018 [45], most teams opted to tackle the respective tasks with approaches based on either the Long Short-Term Memory (LSTM) [40, 35], or Convolutional Neural Network (CNN) [52] architectures. Both LSTMs and CNNs are able to consider sequences, such as words in a sentence or sentences in a text. However, they greatly vary in *how* context is considered. CNNs were first introduced for image recognition purposes and therefore focus on learning local patterns. To this end, during training they learn filters, that slide over the input. Each filter learns a certain pattern, that is compared to the input at inference time. The information of these filters is then combined and further processed through convolutions, pooling, and further connected layers. For Emotion Analysis, this could, for example, mean that a filter learns to recognize the word "cry" in proximity to certain trigger words that indicate the crying is indicative of joy rather than of sadness, such as "happiness" or "surprise".

LSTMs, on the other hand, process sequences by sharing information between steps. To this end, for every part in a sequence, an LSTM cell receives multiple inputs: Information from the computations of the previous cell and the current sequence information. They process the current input x_t and previous information through internal "gates". The previous information comes in the form of the output h_{t-1} and the cell state C_{t-1} of the last cell. The flow of information in LSTM cells is visualized in fig. 2.2.

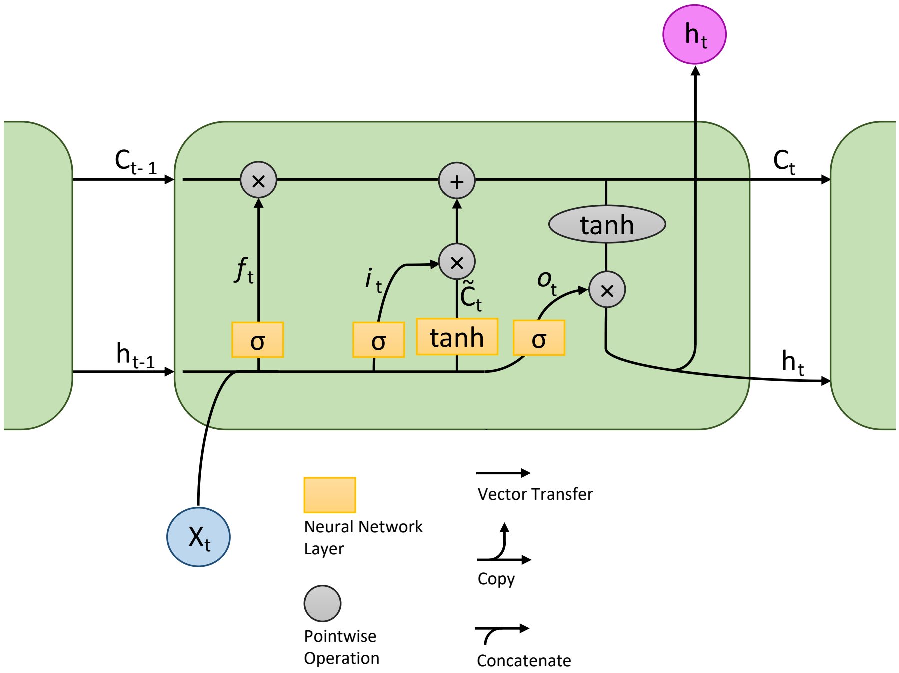


Figure 2.2.: The internal flow of an LSTM cell, adapted from [63].

Each gate features a set of trainable parameters in the form of neural network layers, which consists of a set of trainable weights W and biases b, as well as an activation function. The forget gate modulates the state of the last cell C_{t_1} in that it decides what part of that information should be kept or forgotten based on the last output and the current input:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.1}$$

The input gate learns to decide how the current step should impact the internal state and which values should be updated:

$$i_t = \sigma\left(W_i \cdot [h_{t-i}, x_t] + b_i\right) \tag{2.2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2.3}$$

From there, the cell state is updated:

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \tag{2.4}$$

Finally, the output gate controls to what degree different parts of the updated cell state should be output:

$$o_t = \sigma\left(W_o \cdot [h_{t-i}, x_t] + b_o\right) \tag{2.5}$$

$$h_t = o_t \tanh(C_t) \tag{2.6}$$

Through these mechanisms, LSTMs are able to leverage the information and computations of the previous inputs in a sequence.

Transformers

The latest WASSA shared task in 2022 [8] featured a track for emotion prediction, in which participants were to classify emotion based on Ekman's classes from essays. In this task, only one team opted for CNNs, all other teams relied in pre-trained transformer models [91], such as BERT [23] and popular variations [80, 54]. Transformers leverage context by learning to pay attention to the parts of an input sequence that are important for each part of an input through so-called attention heads. BERT was designed based on transformers as an all-purpose word encoder, that is usually pre-trained on a large datasets with tasks like next sentence predictions and subsequently fine-tuned for each specific task [23]. With transformers and BERT in particular, context can be better leveraged over long distances better than with LSTMs [76]. Since its introduction, BERT and its variations have become widespread in NLP tasks in general [76], and Emotion Analysis in particluar [1] and improved baselines in many areas. Often, systems like BERT are used to compute embeddings, which are vector representations for inputs based on latent representations in a trained network, as they are contextualized: The embeddings computed for a word not only carry information about the word, but also about its context, which enables different embedding representations even for homographs. Homographs are words that are spelled the same, but have different meanings, such as the word "can" ("can do", "a can"). Other than the ability to represent words the model did not encounter during training, this is the main advantage over previously popular encodings like GloVe [67]. To leverage these advantages even for longer units than words, approaches like Sentence-BERT [75], also referred to as sentence transformers, are able to compute high-quality embeddings for entire sentences at a time.

2.2.3. Emotions and Emotion Analysis for Dreams and Customer Service

Supervised machine learning approaches rely on data to be trained and evaluated. For Emotion Analysis in particular, the performance of these models is highly dependent on the domain of the source texts [13]. This partly motivates the choice of gathering corpora on various domains in this thesis: By training a model on two different domains, we hope to get insights that go beyond the domain-specific restrictions and dynamics by comparing results between domains. Overall, this motivates the need for corpora in a variety of domains with a variety of annotations.

The emotions encountered in dreams have been found to resemble those experienced in waking life [36]. Yet, the emotions experienced in dreams are biased toward negative emotions, both in comparison to the emotions experienced while awake [61] as well as overall [39]. As emotions play a central role in dreams, there are sets of emotion categories that have been designed specifically for this domain in the context of dream content analysis, which aims at statistically analyzing dream reports, and to this end, describing said reports in a statistically analyzable manner [25, 82]. One of the most widely used systems is the Hall/Van De Castle System of Content Analysis, which uses a categorical approach to describe the contents of dreams [25]. However, it only considers emotions if they are either explicitly stated in the dream report, or a physiological reaction is being reported. Since explicit mentions of emotional states are rare and due to the bias toward negative emotions, this system consists of five negative emotions, namely Anger, Apprehension, Sadness, and Confusion, and one single positive category, namely Happiness.

Dreams have been automatically evaluated for their emotional content before, using approaches such as LIWC [16, 7, 59], a dictionary-based system for Emotion Analyses, which is widely used in psychology [66]. Another strand of work focused on the representation of texts to aid emotion or sentiment analysis on dreams [5, 73]. Many of the works in Emotion Analysis that focus on dreams are based on the Hall/Van De Castle emotion categories, be it to build their own dictionary based on the rules and categories defined there [34], via explicit analysis of the structure and context of a dream report [33], or using state-of-the-art transformer-based models for automated annotation [10]. Overall, despite their rich emotional content, dreams have not often been the focus of research with the goal of Emotion Analysis. To the best of our knowledge, no previous work has focused on how emotional content develops during dream reports.

Emotion Recognition in Conversations (ERC) is a task that focuses on the extraction of emotional content from conversational data and that has increasingly attracted attention in recent years [71]. In contrast to Emotion Analysis, ERC is not generally limited to text as an input modality, though a subset of works in ERC do focus on Emotion Analysis from conversation. Multiple corpora from different domains have been gathered so far, that can serve as basis for training models for Emotion Analysis in conversations.

Most relevant to this thesis is the customer service dialogue corpus by Labat et al. [49], as it serves as the basis for the customer service dataset gathered here [49]. It was gathered in a Wizard of Oz setting, motivated by the findings in their pre-study with a similar setup [46]. Participants were presented with complaints that were tight to a certain start sentiment and were introduced to solve them with a chatbot. They were aware that complaints were fictional and that the chatbot they interacted with was not actually operated by the company the scenario was tied to. A working student impersonated the chatbot and was introduced to navigate the emotion trajectory, toward a certain end sentiment. The original conversations were gathered in Dutch, this thesis works with translated English versions of these conversations. Besides a set of categorical emotion labels that serve as a basis for the categorical emotion annotations gathered in this thesis, the dataset features VAD annotations for each customer utterance. These annotations are contextualized in the sense that annotators had access to the entire conversation, yet annotations were only associated with a turn if the emotion was expressed in that particular turn. This means, a neutral sentence in an overall joyful text would still be labeled neutral instead of joy, thereby not representing the current emotional content of the whole text up to that point. Therefore, the previously gathered annotations differ from the ones we seek to gather in this thesis.

Labat et al. also gathered the EmoTwiCS dataset for the analysis of emotion trajectories in customer service interactions in Dutch Twitter exchanges [48] that features categorical emotion annotations and VAD scores. This dataset was gathered with expert annotations rather than crowdsourcing. The authors trained students for the annotation tasks and obtained good reliability scores. The emotion labels in this dataset are made for customer turns and are tied to the passages in the text (spans) that express the labeled emotion [47]. Therefore, in contrast to the annotations we gather, the labels in this dataset again only pertain to the part they describe and not to the overall up to that part. For ERC, Wang et al. [92] introduced the modeling of emotion recognition over various turns as a Sequence Tagging problem. Instead of going through a sequence and making predictions one at a time, this means considering the entire sequence at once to predicting the corresponding sequence of expressed emotion. Leveraging the data of their EmoTwiCS dataset [48], Labat et al. followed this approach and found that using Conditional Random Fields [51] improved prediction performance [50].

Other examples include EmoryNLP [93], EmotionLines [44] and its multimodal successor MELD [70] from TV show conversations. EmotionLines and MELD both feature annotations on utterance level that were contextualized on both the preceding and following utterances. As motivated in the introduction, this differs from the annotations we gather in this thesis. The DailyDialog corpus, which was sourced from websites for English learners [53] was manually annotated with Ekman's set of categorical emotions plus an Other class on utterance level. Annotations reflect the emotions expressed in each utterance.

Another particularly relevant conversational corpus was gathered for the 2019 SemEval shared task on Contextual Emotion in text in 2019 [18] that is based on conversations with a conversational agent. In this corpus, instances consist of three consecutive utterances and are labeled with the most fitting class out of four contenders, namely Happy, Sad, Angry, and Others. This assigned label describes the emotional content of only the last utterance, the previous two utterances only serve to contextualize the expressed emotion. While these labels consider context, they only portrait the emotional content at one point of the conversation and are therefore unable to capture progression. In addition, with a length of three utterances, any possible progression could only contain a single change. The submissions to this shared task most commonly relied on LSTMs or bidirectional LSTMs (BiLSTMs) [84], a related architecture able to capture not only the previous but also the following parts of sequence as context. Additionally, the majority of participating teams considered the output of multiple trained systems for their final prediction.

For customer service dialogues and in the overall direction of this thesis, the previously mentioned works by Labat et al. are most closely related to the research questions and topics of this thesis [48, 47, 49, 46].

3. Methodology

This chapter focuses on the methodology utilized to answer the research questions defined in chapter 1. To this end, the first section describes the data gathering process and answers three central questions: How do we choose our domains and why are they eligible? How do we source the instances that will be annotated? What steps are necessary to clean the texts? The second section in this chapter introduces the annotation guidelines and answer another three questions: What labels do we want to gather exactly and how do we communicate this to the study participants? What other details do we need to consider? Lastly, the last section introduces the machine learning setup for the experiment which will help answer the third research question posed in chapter 1.

3.1. Corpus Creation: Data Acquisition and Preprocessing

One of the main goals of this thesis consists of gathering two corpora that form the basis for analyses and the training of machine learning models. These corpora coincide in their general structure: Both are comprised of a variety of different instances, meaning different texts. Each text is split into parts - sentences for dreams and bi-turns for customer service dialogues - and each of those parts is annotated using the same set of questions. Importantly, these annotations reflect the emotional content of the entire text up to and including that part. Viewed in combination, these labels describe the progression of emotional content over the course of each text. This consistency in data structure and content helps the later analysis, as training and evaluation tasks are more easily transferable and comparable. To this end, this section focuses on the first necessary steps in building a corpus: Data acquisition and cleanup.

3.1.1. Domain Choices

In this section, we answer the question of how and why we chose our domains. The core difference between the corpora we gather is the difference in their underlying domains. While for the customer service corpus translated chat exchanges between customers and service agents were considered [49], the other corpus was built on dream reports. The differences that come with textual content from separate domains are manifold and include a difference in linguistic structure, formulations and content. All of these differences will be relevant for the task of extracting emotional content from text, as models might need to adjust to these differences. However, they also influence the annotations.

To illustrate this, let us consider science books as an underlying domain. Most chosen texts from this domain will likely result in static emotion annotations: Close to all instances would likely be continuously annotated as neutral, as the language and content commonly found in science books is chosen specifically to transport knowledge in a factual, non-emotional manner. For example, the text

> "Determining the distribution status of fungi is fraught with difficulties.
> Many areas of the world have not been explored for fungi, and documentation from tropical Africa is especially limited.
> It is premature to state unequivocally that any species is "endemic" until we have more data on the diversity of fungi from understudied areas.
> For this treatise, if a species was described as new from São Tomé or Príncipe and it has not yet been reported from elsewhere, we recognize the taxon as a putative endemic and annotate as such in the Appendix."

from [17, p. 193] does arguably not carry emotion. When asking for emotion categories, the expected annotation sequence for this example would therefore likely end up being *neutral, neutral, neutral, neutral*. Since our research questions are geared toward progression annotations, we aim to include sequences that feature non-static progressions, as instances can otherwise be sufficiently described with a single label. Additionally, part of the annotations we gather are appraisal scores, which are defined as a cognitive evaluation of events. The example above does not contain any events that could be appraised and the example would therefore be inherently unsuitable for these annotations.

Therefore, to get insights into how the emotional content changes over the course of texts, any underlying domain should at least meet the following criteria: Texts in the domain should generally...

- ... describe, refer to, or express emotional content.
- ... contain one or more events that can be appraised.
- ... be limited or limitable in length, so that a suitable amount of instances can be gathered.
- ... be expected to contain changes in the emotion categories and appraisals they contain, as to yield non-static progressions.

Customer service dialogues and dreams arguably both fit these criteria.

For customer service dialogues, customers are likely to enter the conversation with an issue. This means something - an event - happened and they are looking for help resolving it. These events are likely emotional in some way, as they were important enough to the customer to reach out. Still, when selecting instances we take care to further focus on events that are more likely to carry emotion, as described in section 3.1.3. Additionally, depending on how the service agent acts, this domain is promising in terms of changing emotions and appraisals over the course of the interaction [48]. In the corpus we build on, service agents were specifically instructed to evoke a certain change in the customer's sentiment.

As dreams have been found reflect waking emotions [36], it is not surprising that they are also often emotional. In addition, dreams have been described as "hav[ing] some resemblance to plays" [25, chapter 2], suggesting the common presence of a story of sorts and hence, that of events. When considering the progression of emotions in events, dreams pose an interesting case, as they do not always follow the rules of typical every day life. The possibilities of swift changes in the dreams may also reflect as swift changes in the present emotions, hence yielding a very different set of progressions than those found in other domains.

3.1.2. Raw Data Acquisition

After fixing the underlying domains, we needed to acquire of raw data to annotate. Obtaining texts consists of roughly three steps: First, we must source a base set of data to choose from. In a second step, a subset of that data that can be used for annotations is selected by filtering out unsuitable instances. The third step consists of data preparation and cleanup, in which we adjust the data into the final format that will be annotated and check the results of step two. While clearly separable, these tasks do at times have to be performed in iterations. In the following, we take a closer look into how these steps looked in detail for both our domains.

Dreams

For dreams, we decided to source our base data from `http://www.dreambank.net/`. With over 20,000 dreams, DreamBank not only offers a great number of dreams, but also a web API that allows us to select dreams based on several criteria, including the number of words [24]. It is furthermore co-authored by G. W. Domhoff, a well-established researcher in the area of dream content analysis, and has served as the basis for automized dream-content analysis before and is claimed to be the largest dream database [33, 10]. All dreams stored in this online collection are categorized into "dream series", that divide the content by their origin. As DreamBank itself obtained the dreams from different sources, from online blogs over studies to originally handwritten dream reports, the series' scope reflects those. Some dream series reflect a specific author in a specific time or age-range, others more broadly refer to a certain group of people or a certain experiment setting. DreamBank offers some background information for these dream series and, for some, previous annotations of different kinds. None of those annotations, however, match the task at hand and therefore they could not be repurposed for our dataset.

Customer Service Dialogues

The customer service dialogues were based on a translated version of a previous dataset gathered by Labat et al. [49]. This data was gathered in a Wizard Of Oz setting, in which the participants were given scenarios and the task to resolve the issue with customer service. While participants were told the contacted agent was a chatbot, this chatbot was in reality fully controlled by a student worker. Throughout these chats, the service agents behaved in different ways to steer the conversation towards a given end sentiment; While some tried to help the customer, others refused help or went as far as to reply with inappropriate comments to the customer. This special setup of the data makes it particularly interesting for the analysis of emotion progression. In typical customer service interaction we might usually find a service agent that tries to be helpful, yielding more predictable progressions that would likely either stay static or lean toward a more positive affect over time [48]. With this setup, however, we may also encounter the opposite, as people might get more upset over the course of the conversation.

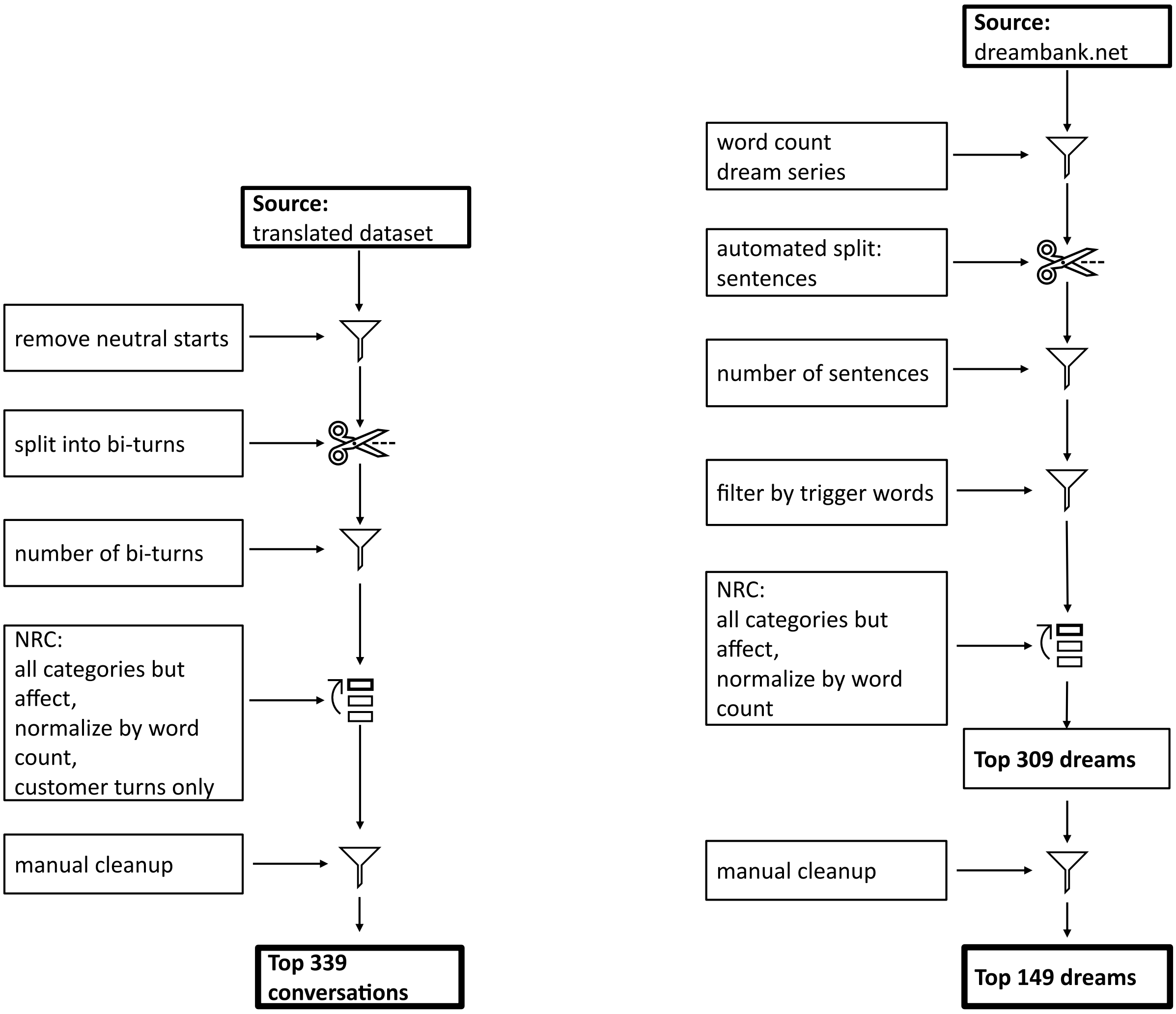
3.1.3. Data Preparation

After gathering the base data, the collected instances had to be processed and cleaned. While the data preparation pipelines look similar overall, the different domains posed varying challenges. Figure 3.1 provides an overview of the entire data preparation process for both dreams and conversations.

Dreams

Before downloading any dreams, we first decided which dream series to consider. For our purposes, some of them needed to be exclude upfront for a variety of reasons, including:

- **Language:** Some of the dream series feature non-English dream reports. For this work, we will limit the reports we consider to those in English.
- **Expected overlap:** Some long running dream series have been assigned to both split-up lists and a whole collection. Considering both would lead to unnecessary overlap.
- **Upsetting content:** Some dream series are clustered based on the group of dreamers or content. While most bare no greater risk of being upsetting than others some do, such as series from veterans or dreams about pregnancy and abortion.
- **Formatting:** Some dream series include additional annotations or notes. While we could have chosen to include and preprocess these, there are enough dream series without additional formatting that did not require this added step.



(a) Customer Service Dialogues (b) Dreams

Figure 3.1.: The preprocessing pipeline for both conversations and dreams.

After filtering the dream series, dreams were gathered from the remaining series. When using the random sample function, DreamBank offers the word count of each dream as a parameter to filter by. Setting this to return all dreams in each series that contain between 0 and 1000 words yields over 24,000 dreams before any filtering - an easily large enough basis for all following filtering steps.

For dream reports, a natural way of defining a "part" is to look at the sentences comprising the reports. We define a length range of four to ten parts, meaning we will only consider texts that contain between four and ten sentences. While the lower bound served to make sure we can observe progressions, the upper bound will help making sure the annotation task will not take too much time per instance. This mattered, as if the annotation time for each instance was too high, it would have impact the amount of total instances we were be able to gather, as annotators get paid depending on the time annotations take and the budget for our project was limited. Less instances, however, mean less train data and less testing data, which in turn has implications for both the upper bounds of the adaptation as well as the statistical significance of our results.

We therefore proceeded to split the gathered texts into sentences, using a regular expression. In the next step, we filtered out all dreams containing less than four or more than ten sentences. As splitting a few thousand texts with a regular expressions is bound to result in some superfluous or missed splits, there is a good chance some of the instances we filtered out here would actually still fall in the four to ten sentence range. As we have access to far more raw instances than required, we accepted this and proceeded to correct wrong splits that have not been filtered out by hand in a later step.

As nightmares frequently contain events that could possibly be triggering or upsetting to annotators, and the bias toward negative emotions found in dreams suggests the possibility of a relatively high amount of nightmares in a collection of dreams, we are needed to make sure no dreams which are too graphic reached the annotators. Additionally, some of the dreams on DreamBank have originally been noted down in journals many years ago. This suggests that in addition to the possibly upsetting dreams, reports might contain potentially upsetting personal views. To combat both, we chose to filter the dreams twice: Once automatically, based on a list of trigger words and once manually later one. The trigger words collected by Luis van Ahn [2] served as a basis for the filtering of the data we had selected based on length. The list contains many terms or variations of terms that are not relevant for the filtering of dreams. Also, many of the words are not upsetting by themselves, but might be based on context. To avoid over-filtering, we cleared the list of the latter category of words by manually deleting them before using it as a basis for filtering. In the next step, we simply checked our instances against the occurrence of any of the leftover trigger words and remove any instances from consideration that contain any of them. As with the filtering by length, we accept that this heuristic might be too simplistic, but still work with it in favor of keeping the pipeline simple. All dreams were later checked manually.

With over 9,000 instance still left, a subset needed to be selected for annotation. Ideally, we wanted to choose dreams that clearly contain emotion and at least some that have a clear emotion progression. This is important, as not only did we aim to gather a new dataset, we also do so through a novel, iterative labeling task. Therefore, we are not only interested in the labels themselves, but also in gauging whether the annotation setup is suitable to gather progression data. As a rough approximation of the emotional content found in each of the dreams, we decided to rank them by the amount of emotionally charged words they contain, using the NRC lexicon [56], akin to the sampling strategy used by Oberländer et al. [62]. This ranking was done in two steps:

1. For each dream, we counted the number of words in the report that express any of the categorical emotions, namely *anticipation, trust, disgust, anger, fear, joy,*

sadness, surprise according to the NRC. We did not regard the other two sets of words, which denote those terms that express positive or negative sentiment. This has two main reasons: For one, while we gathered categorical emotion annotations, we were not going to label affect. Secondly, the expected overlap between words occurring in both the sentiment and category set would mean we might consider the same words more than once, despite the sentiment adding no information.

2. We then normalized the amount of contained words that express emotions by the total word count, which we stored from our original query to DreamBank. This step helped to not favor longer texts over shorter ones.

These normalized scores were then used to rank all dreams in descending order, presenting those that were most likely to contain emotional content at the top. How many instances needed to be selected depended on multiple factors, including the time annotations would take. Details on this can be found in section 3.2.4. The exact number is, however, not crucial for the discussion of the next preprocessing steps.

Once the top instances were selected in this manner, the previously mentioned manual cleanup ensued. In this step, we checked for both structure and content. The structure check mainly consisted of inspecting the automatically split sentences and editing those splits wherever needed. If these adjustments led to an instance no longer falling into the four to ten sentence range, it was discarded. For content, the dreams were checked on their possibility to be upsetting to annotators and rigorously removed from consideration, if they contained or dealt with one of the following themes:

- sexual depictions
- racism
- otherwise derogatory
- murders, death threats, detailed death descriptions, shootings
- descriptions of butchering animals

This list is not exhaustive and merely reflects recurring upsetting themes. Dreams were also removed, if their structure was unsuitable for the annotation task due to other circumstances:

- hard to understand content, either because the report relies on knowledge of context not presented in the dream itself or because of how it was written
- sudden change of scenery or sudden reference to a completely different set of events
- references to other dreams
- references to waking up, facts from the real world outside the dream, or the dreamer commenting on the dream

The last reasons for instance removal was specifically to help the gathering of appraisal scores and served to make sure the scope of what needed be evaluated was clear for annotators. This is discussed in more detail in section 3.2.2. The penultimate reason was a practical decision: Data gathered on instances that essentially detail more than one unrelated storyline contain more than one set of unrelated label-progressions. Even when considering this point as an inherent part of the domain and and accepting it on a conceptual level, this would increase the complexity of our data and therefore, the complexity of the prediction task. For this work, we only gathered a small-scale dataset that was unlikely to contain enough example for our model to adapt to this added complexity. We therefore choose to exclude those instances.

The dreamer of one dream series habitually split her sentences with semicolons. These were split into sentences ending in full stops if and only if they could syntactically stand on their own. To obtain 149 instances, 309 instances had to be manually checked, meaning that about 50% of the data had to be removed. This is not surprising, as many particularly upsetting themes are among those most commonly ones found in dream reports, such as being attacked and pursued, sexual experiences, and killing someone [21]. The manual cleanup concluded the data preparation for dreams and left us with a set of ready-to-annotate instances.

Customer Service Dialogues

The base data for the customer service dataset has been gathered in a controlled manner and been preprocessed before [49], allowing for a more straight-forward preprocessing. This is reflected in the lower number of steps in the data preparation pipeline, which is visualized in fig. 3.1

Among the annotations that are already attached to the conversations is one label that marks the sentiment of the customer at the beginning of the conversation. To increase the chance of obtaining emotional conversations in our final selection of conversation, we removed all instances that are marked with a neutral start up front.

When deciding upon what a "part" is for dreams, the natural choice seemed sentences. In the setup of a chat dialogue between a customer service agent and a customer, there are more options to choose from messages, sentences, and turns to name a few. The focus of the emotion annotations will be placed on the emotional progression of the customer, as will be further discussed in section 3.2.2. Therefore, labeling any part that does not contain a response from the customer will not hold new information about the customer's emotional state, and therefore, for the annotation task. In considering the customer turns for annotations, we follow the guidelines for customer service annotation detailed by Labat et al. [47].

With this in mind, we made use of the preexisting annotations in the dataset, which marked each message as either sent by the service agent or by the customer. As every chat conversation starts with the chatbot introducing itself and asking the same opening question, the conversations can be naturally split into bi-turns. Each bi-turn contains a set of consecutive messages from the chatbot (at least one), followed by a consecutive set of messages from the customer. A bi-turn ends once the service agent sends a message again. If the conversation ended with a turn from the service agent which is not followed by an answer from the customer, the last turn was cut, as it contained no new information on the emotional state of the customer. Since the annotation task was set up to only show the relevant information up to the current bi-turn, a possible sudden end did not influence the annotations. Additionally, we added three special tags: "==ADMIN==" indicates the start of a turn by the customer service agent, "==PART==" indicates the start of a turn by the customer and "==NEWMESSAGE==" indicates that the following was sent in a separate message. These tags serve to maintain this contextual information for later processing.

With the same reasoning as for dreams, we chose to consider instances that are four to ten parts in length, with the difference that one part now represents a bi-turn. After filtering the instances for length, we employed the same procedure to score and rank conversations based on the NRC lexicon that was already described for dreams, with one difference: Keeping in mind that the focus lies on the customer's emotional progression and assuming that this can only be inferred from their responses, we only scored their turns and normalized the NRC scores by the word count in the customer turns.

Finally, the customer service instances posed an additional challenge: The texts contained references to companies, that existed in real life, paired with made-up scenarios and customer service that was not affiliated with those companies and, at times, was instructed to elicit a negative sentiment. For example, the dataset contained instances, in which the customer claimed a well-known online shop had sold them counterfeit watches and instances, where the supposed chat bot of a well-known travel company made fun of a passengers weight. To avoid releasing these to annotators and risk confusion and any of these being released without context, we chose to replace all company names with made-up names. This posed a challenge because the company names themselves carried context information that needed to be retained. Let us for example consider the following phrases from the dataset:
"Hello, I noticed today that the internet is not working (we have companyA). Turns out they are doing major maintenance work in the street."

In this example, companyA serves as a placeholder for the actual company name. Not only is the sentence unusual in structure without the name of the company, but also knowing that the company that is being contacted is a telecommunications company helps understanding the overall situation. Additionally, some company names also serve to communicate the actual product. For example, simply replacing the company name in the phrase *"I'm dissatisfied with my companyB."* makes it very confusing, while *"I'm dissatisfied with my Airbnb."* is easily understandable. Taking this into account, we replaced all company names while making sure to

1. come up with company names that contain what they do.
2. replace formulations that refer to the product rather than the business with whole phrases that manage to maintain the meaning.

For example, the three telecommunications companies in the dataset were renamed *TelcoIT*, *Telco25* and *YourTrustyTelco* respectively, as they sometimes occurred in the same instances. Hotel chains, on the other hand, never occurred together, so they were all renamed *"BestHotels"*. For Airbnb specifically, multiple replacements were defined. Using a simple string replacement for these adaptations worked well, as the usage of the company name as a Stand-In for the product was consistently marked by the possessive pronouns or articles. This way, *"the Airbnb"* became *"the place on RentYourHolidayHome"*. Some special cases, like *"the Airbnb host"* were simply caught by an earlier regular expression that changed them accordingly, for example into *"the host from RentYourHolidayHome"*.

These replacements were part of the manual cleanup stage, the final step in the preprocessing pipeline for the customer service data depicted in fig. 3.1a. A few instances were removed in this step, either due to trouble with the data itself, such as confusing repetitions or sudden ends of the conversations, or due to references or language-based jokes that would likely have been lost due to either the translation or the different cultural background of the annotators.

3.2. Annotation Guidelines and Survey Details

In this section, we turn our attention to the annotation process and the questions of what annotations we chose to gather and how we formulated then according survey questions. We needed to design two separate surveys - one for each domain - that were centered around the incremental gathering of emotion annotations that is at the core of this thesis. In this section, we take a closer look at the development of these survey, with a focus on the questions that were posed. The outcomes of this section are the question sets displayed in fig. A.6, A.10 and A.11.

3.2.1. The Finished Survey

We start this section by reporting its result: To get an overview of the annotation process and to put the developed annotation guidelines into perspective, we first introduce the finished survey and all included platforms that annotators got to interact with.

For annotations, we decided to use two platforms: The survey was implemented and executed on `https://www.soscisurvey.de/`. SoSci is a German page that offers free usage for research purposes and offers all the features we required for the iterative label task. In addition, we used `https://www.prolific.co`, a platform that links researchers to participants. Prolific screens participants, which promises a higher chance of obtaining annotations from participants that have performed well in surveys in the past. In addition, all contributions by participants on Prolific are paid. By sourcing annotators from Prolific, we hoped to obtain more motivated and thorough participants than on other, possibly more common, crowd-sourcing platforms. That way, we aimed to combat the extent of some of the problems commonly associated with crowd-sourced data. Further details on payment and participant screening are reported section 3.2.4.

While the wordings and some displayed details differ between the two domains, the overall flow is fixed and depicted in fig. 3.2. All survey pages can be found in Appendix A.1.

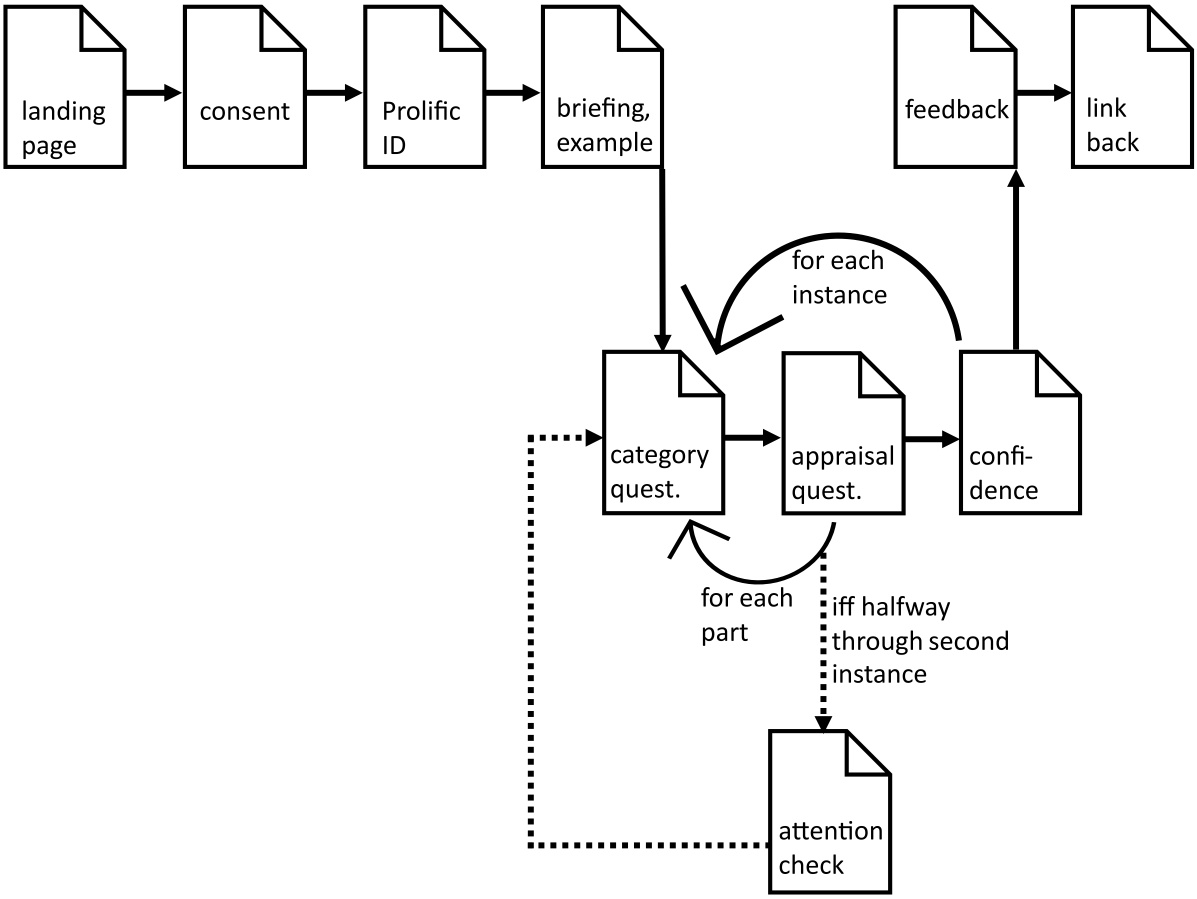


Figure 3.2.: The conceptual flow of the survey for both dreams and customer service dialogues.

While the annotation task was at the center of our survey, we could not just show it to participants upfront. Before that, we needed to give an introduction, get their consent, and explain the task. The welcome page, depicted in fig. A.1 and A.2 for dreams and customer service dialogues respectively, served as a landing page that provided participants with the necessary context for the consent they were asked to give on the next page, as depicted in fig. A.3.

Both landing pages looked similar overall. Yet, the customer service data warranted a few additional warnings due to the nature of its content that we placed in the "General Information" section. The information on Prolific shown to eligible participants mimiced this landing page but contained less detail.

Once the participants gave consent, they were asked to paste their Prolific ID into a text field. A Prolific ID is an identifier given to each member of the platform. This was merely implemented to assure all IDs would be available for payment purposes later on. Prolific IDs were also automatically recorded using the URL.

From there, participants were introduced to the task in more detail. To this end, they were shown the texts illustrated in fig. A.5 or A.4, for either dreams or customer service dialogues respectively. To ensure participants pay attention, they could not simply skip this page, as the "Next" button, that will take them to the actual survey, only appeared after 60 seconds. A red disclaimer at the top of this page makes participants aware of this. The description was kept short and concise to make sure the task was clear and the description will more likely be fully read and understood. The given examples then practically demonstrated the iterative nature of the task and provided example reasoning for each step.

Once participants clicked the "Next" button, the actual annotation task started. The interfaces are depicted in fig. A.6a, A.6b, A.10, and A.11. For each part, participants were first asked about the categorical emotion labels and then about the appraisal values. Until the current instance was finished, a new part was added to the previous text every time the categorical questions were shown anew. For the annotators' convenience, and to lower mental load, their previous answers were pre-selected for each question. This way, annotators could focus on the *difference* between the previous part and the newly added one.

To make sure participants were still paying attention, one attention check was conducted per survey. This also helped to flag possible bots. For both dreams and customer service dialogues, these attention checks were displayed instead of the new part with the categorical question. The categorical question lent itself better for this task than the appraisal questions because (1) only one value needed to be set and (2) the new parts were always displayed first for this question. Therefore, if a participant wanted to answer this question for the current part, they had to read the new part carefully and would therefore definitely notice the attention check. We did include the whole instance up to the current part in the interface for appraisal questions, but there is a good chance even participants who were paying attention would not have re-read the added part they had just read for the previous question, and would therefore fail the attention check. Structurally, the attention checks resembled what the participant would have expected to see, including the pre-selected answer they had given to the categorical question for the last part. To this end, attention checks were never displayed as a first part, but rather halfway through the second one. In place of a new part, a message would appear informing them of the attention check and would asking them to check "admiration". We chose this category because we assumed it would be less frequent than other categories, therefore lowering the chance of an annotator passing "by chance" by having admiration be the label they had selected for the last part. A depiction of attention checks during dreams and customer service dialogues can be found in fig. A.7 and A.8 respectively. After the "Next" button was pressed, the preset option returned to the value that was set for the previous part and the new part was displayed along with the previous text.

After the completion of each instance, annotators were made aware of the current instance's end and asked for their confidence in the annotations they have just provided, as depicted in fig. A.9.

Once the survey is completed, participants are asked for feedback as shown in fig. A.12, before they are finally sent to the last page that provides them with a link back to Prolific for payment purposes. All reported feedback can be found in Appendix A.4.

3.2.2. Categorical Emotion Labels

For the categorical labels, we decided to start with the list of categorical labels that had already been used previously to label the customer service data [49]. These were selected based on the high number of their occurrences in customer service exchanges on Twitter [48], as well an additional positive category (admiration), a particularly common category from the original annotations in the Wizard of Oz data collection [46]. This would allow for a comparison between the previous annotations and the ones gathered through the newly designed, iterative task, although this comparison is out of scope for this thesis. The labels that had been previously used are *Neutral*, *Anger*, *Annoyance*, *Disappointment*, *Fear*, *Confusion*, *Desire*, *Relief*, *Gratitude*, *Joy*, and *Admiration*.

As discussed in the beginning of this chapter, the set of labels was to be the same between the two datasets, hence the selected categories also needed to be suitable to describe the emotional progressions of dreams. Self-annotations quickly showed that some labels, such as admiration and gratitude, would likely not occur frequently, if at all, for dreams. Additionally, some emotions that might occur in dreams, such as *Sadness* or *Surprise* were not reflected in this label set.

We assumed dreams could contain the whole spectrum of emotion, as they reflect waking emotions [61]. Therefore emotion categories such as those described by Ekman or Plutchik that aim at capturing the whole range of basic emotions seemed like a good fit. In the spirit of keeping the emotion labels used in the previous annotation task on the customer service data and keeping the labels fixed between domains, we decided to enrich the previous labels by aggregating basic emotion categories with the already present ones - wherever possible. Disgust could not be easily aggregated and was therefore left out. As all of Ekman's emotion categories appear on the same base level of intensity in Plutchik's wheel of emotions, we focus on those to not overcomplicate the label set. These were then roughly clustered by sentiment in the order positive, negative, and neutral. This yielded the final set of possible labels for both domains, as depicted in fig. A.6.

While for dreams, the associated question reads *"This dream made the person dreaming feel..."*, for customer service dialogues it was posed as *"The events made the costumer feel. . . "* As depicted in fig. A.6b, what *events* the question referred to was not only explained before the survey started, as shown in fig. A.4, but was also clarified again before each instance.

3.2.3. Appraisal Dimensions

As introduced in chapter 2, appraisals describe the subjective, cognitive evaluation of salient events. For this reason, when picking our domains, we already formulated the presence of events as a core prerequisite of any texts we would consider. When developing the survey to obtain appraisal annotations, we needed to sharpen our understanding of what possible events are, which appraisal categories might apply and be of interest for each of the domains and lastly, how to communicate all of this as concisely as possible to our survey participants.

Dream reports are usually written out as event reports, with the dreamer describing how the story of each dream unfolded. There are two main reasons a report might not follow this logic:

1. The report features details that include references to waking up, previous dreams or judgments they made about the dream while awake.
2. There are jumps in the dream and/or multiple dreams are reported in one instance.

When explaining the manual cleanup process, we noted both of these as reasons for exclusion from the base dataset and detailed our reasoning behind excluding those reports that include jumps or multiple dreams. We are now taking a closer look at the first reason for exclusion.

One example for the first category is the following dream report:

> Between 12 and 12:30 I had another dream which was almost like a daydream.
> I dreamed I was receiving a large sum of money and I don't know whether the dream lasted any more than five minutes and the entire action consisted of not of the dispersal of this money, but of my receiving this money.
> It was a very ceremonious affair where great sums of money were being handed over to me with a great deal of legal matters being transferred.
> Anyway, it was almost like a funeral, so ceremonious.
> There were a lot of people there.
> I was very happy about it, very cheerful.

> Although I was very cheerful about the transfer of this money, it really wasn't a happy dream.
> I can't say that I woke in great cheer or anything.
> It seemed to have the same pall that the second dream had.

This report starts and ends with judgments about the actual dream. If we were to ask about the events, this might make it unclear if we refer to the events in the dream from the perspective of the actual life of the dreamer, or within the context of the dream. In this example, we see very clearly that these two interpretations may vary greatly - even for categorical emotions, as the report goes from describing the dream-ego as "happy" and "cheerful", right before judging the whole dream as not happy. If we decide to ask about the impact of the dream in respect to real life, we could ask participants to consider them "as if they were real life events". But often, dreams come with their own set of circumstances: While a horse showing up at my doorstep might elicit a high *response urgency* when embedded in real life, as I don't have the space or resources to care for a horse, in the context of a dream I might live on a farm and may therefore not be bothered by it. This setting would therefore require the annotators to make assumptions about the real life of the dreamers that they have no information about, hence introducing a higher uncertainty. Additionally, this would still leave sentences like the very last one as possibly confusing and superfluous to the task. We therefore choose to look at dreams as self-contained reports, disregarding any real-life information. This elicits the need to deal with instances such as the ones above, either through removal or editing. We chose the former to avoid further increasing the complexity of the manual cleanup task and for easier reproducibility, as this required only one decision (removal vs non-removal) as opposed to editing (which sentences to remove).

For customer service dialogues, the instances are not textual event descriptions. Rather, they refer to events that are being appraised and we are interested in how these appraisals may change through the interaction with the service agent. As the actual events might vary, we instructed annotators to focus "on the events addressed in the conversation", as shown in fig. A.4. To clarify, we added examples for such events as "the reason the customer contacted customer service, or any measures the service agent takes to resolve the issue". To help annotators, we repeated this explanation of events on every page of the actual questionnaire for dialogues, as shown in fig. A.6b and A.11.

When choosing appraisal dimensions, we initially took the base set of twenty-one appraisal dimensions used by Troiano et al. [89] and their formulations into consideration. For time reasons, after running some trials, we found that collecting around ten appraisal dimensions is feasible. Among the twenty-one dimensions mentioned above are some that were mapped from Hofmann et al. [42], who had used the seven categories *Attentional Activity*, *Certainty*, *Anticipated Effort*, *Pleasantness*, *Responsibility*, *Control*, and *Circumstance*. The paper goes on to show that appraisal scores are viable for Emotion Analysis. Furthermore, these dimensions are based on the appraisal dimensions used by Smith and Ellsworth [86], who were able to distinguish fifteen emotions based on their associated appraisals in their work. With this prior work, we have a strong case for choosing these dimension in conjunction as a base set of appraisal dimensions. We followed the mapping and formulations as described by Hofmann et al. [42] and considered the following dimensions:

- Attention
- Own Responsibility
- Consequence Anticipation
- Own Control
- Pleasantness
- Chance Control
- Effort

Having chosen a base dimension set based on the predicted potential for machine learning approaches to leverage them, we then considered the other fourteen dimensions to find which ones we might add to this base set based on their fit to our domains.

For this, we considered each dimension regarding two questions: Is the dimension equally relevant for both domains? Do we expect this dimension to change over the duration of the instances in each domain, yielding non-static progressions? Adapting the question formulations defined by Hofman et al. [42] to match the domains at hand helped finding those appraisal dimensions that would likely apply to dreams and customer service dialogues alike. With this in mind, we discarded candidates like *response urgency* (*"To the customer, the events in this conversation inspired urgent action."*), as they were likely to stay static for the customer service instances. The reasoning here is that customers tend to contact customer service to inspire urgent action on the company's part, not on their own. Similarly, dimensions like *suddenness* could be used for dreams (*"To the dreamer, the events in the dream were sudden or abrupt."*), but are likely to be irrelevant for dialogues (*"To the customer, the events in this conversation were sudden or abrupt."*), as we don't necessarily assume that any event referred to in these dialogues could be sudden. After considering all fourteen left dimension in this way, we decided to add three dimension:

- Event Predictability
- Other's Responsibility
- Familiarity

Event predictability is especially interesting for dreams, that are often erratic in nature, but also for customer service dialogues, as we assume customers often times learn more details about what went wrong during the interaction. *Others' responsibility*, is

Appraisal Dimension	Question Formulation
	To the customer/To the dreamer, the events in the dream/in the conversation...
Pleasantness	... were pleasant.
Familiarity	... were familiar.
Effort	... required a lot of energy to deal with (within the dream).
Own Responsibility	... were caused by their own behaviour (in the dream).
Others' Responsibility	... were caused by somebody else's behaviour (in the dream).
Chance Control	... were the result of outside influences (within the dream) of which nobody had control.
	(In the dream,)The dreamer/the customer...
Event Predictability	... could have predicted the occurrence of events (in the dream).
Attention	... paid attention to the events (in the dream).
Consequence Anticipation	... felt that they anticipated the consequences of the events (in the dream).
Own Control	... had the capacity to affect the events (in the dream).

Table 3.1.: The questions for each appraisal category. Formulations in parentheses only apply to dreams.

equally interesting for both domains. While with the details of what went wrong or what is being done to aid a situation, other agents - including the chat agent itself - are likely introduced over the course of customer service dialogues. For dreams, the Hall/van de Castle system for dream content analysis has a whole category dedicated to actors [25], implying the importance of others in the events of dreams. Lastly, *familiarity* may be more interesting for dreams - again because of their sometimes erratic natures - but also applies to customer service dialogues. We assume this, because the events mentioned in the dialogues range from ordinary initial claims, such as missing items, to more rare ones, such as receiving counterfeit items. With this range should come a range of responses and associated events that would be revealed or would unfold in response, which would also vary in regards to their familiarity.

This concluded the selection of appraisal dimensions. As we already formulated all questions to test whether they apply for each domain, we noticed that the questions can be naturally split into two types of formulations per domain. These are reflected in the final question setup depicted in fig. A.10 and fig. A.11. In addition, the questions are listed - with their corresponding appraisal category - in table 3.1. Other than the subject whose appraisals we are interested in (dreamer vs. customer), the question formulations stay the same between domains.

3.2.4. Study Design

This section is dedicated to the decisions and designs used to conduct the study, that have not been been discussed previously, such as participant payment, task size per participant and participant screening.

Time Estimate, Task Size and Payment

On Prolific, participants are payed for their participation on surveys. We decided to pay £9,00 per hour, the minimum hourly rate as recommended by Prolific, albeit not the minimum allowed rate.

In face-to-face trials, people took 35-60 minutes to complete the annotation task. We note, that self-annotations with the same tool take 10-15 minutes. While it is to be expected that new annotators will take longer, due to the size of the time gap we still assumed that those people have considered the task especially carefully, as they were asked to give feedback. To further gauge the required time and technical feasibility of the survey, we conducted a pilot study for both domains. In it, a set of three annotators rated the same instances. This confirmed our assumption: Annotators took between eleven and nineteen minutes to complete annotations - far less than the people we had asked previously. We chose to fix 20 minutes as an upper bound on the annotation task, henceforward paying each participant £3 to partake in the survey. All annotations collected during the pilot study are visualized in Appendix A.3. As we did not change the survey afterwards, they also serve to give an intuition of the gathered progression data at large.

With the time fixed, we needed to ensure the task each annotator was presented with was roughly equal in size and would take around 20 minutes. The number of instances is, by itself, not a suitable measure for this, as the their sizes vary from four to ten parts. We therefore decided to set an upper bound on the amount of *parts* each participant got to label, rather than on the number of instances they were presented with. Through a rough analysis of the time previous annotators had taken, we decided to fix 24 as an upper bound of parts per annotator. The problem of separating an amount of objects of certain sizes into the minimum number of bins of certain sizes is an NP-hard problem known as the bin-packing problem. We used available greedy algorithms to solve this problem and to split our data into such bins that later constituted the sets of instances participants got to annotate. This helped minimizing the number of required annotators while ensuring none of them was required to annotate more than the fixed maximum amount of parts overall.

The other formulation of the bin packing problem, which involves splitting objects of varying sizes into a fixed number of bins while trying to keep the sum sizes approximately equal between bins, was also used. As there is no downside to having participants label multiple instances, we decide to let the same annotator participate in multiple rounds of annotations if they so wished. This benefited us and the participants in two ways:

1. Only interested participants would return and they would already have a good grasp on the task.
2. The amount of possible annotators was higher for each instance.

Technically, we implemented this by releasing the survey bit-by-bit, updating the instances to be annotated in each round. This way, there was no chance of one annotator seeing the same instance twice. To split the data into multiple rounds, we used algorithms to solve the second formulation of the bin-packing problem. This ensured that each round yielded about the same amount of annotated parts, and therefore, that we required approximately the same amount of annotators for each round.

Data Acquisition for Inter-Annotator Comparison

As the data was split into small sections of about the same amount of parts, with each instance only being labeled by one participant, data had to be explicitly gathered for inter-annotator analyses. This means, that the instances that would be used to evaluate the data in chapter 4 would have to be labeled twice. To increase the comparability of each set of twice-labeled instances and to furthermore increase the statistical significance of IAA analyses, we chose to double the maximum amount of annotated parts per annotator for the collection of this subset of annotations. In consequence, we also increased the amount of time and the payment that annotators got by a factor of 2. This yielded an evaluation dataset comprised of a number of instances with two sets of annotations each and fixed annotator pairs.

Participant Screening

Prolific offers the option to filter possible survey participants through a number of filters. We made use of this option and filtered participants along a number of criteria. To ensure our annotations will be made by people with the highest knowledge of the English language, we set *Location* (USA or UK) *Nationality*, *Country of Birth*, *Immigration*, *Place of most time spent before turning 18* (UK, USA, Ireland, Australia, Canada, or New Zealand), *First Language*, *Fluent languages*, and *Primary Language* (English) to suitable options. As one of the customer service dataset is based on interactions with a chat bot, we wanted participants to be comfortable with the according medium. To increase these chances, we set filters for *Age* (18-55), *Devices with screens* (mobile phone, tablet reader, laptop, or desktop), *Weekly device usage* (2-6 times a week or every day), and *Chat/Messaging apps* (any). Furthermore, we screened out participants that had indicated any answer but no to the questions of *Neurodiversity*, *Autism Spectrum Disorder*, *Depression*, and *Anxiety*, as all these could change a participant's perception of emotion or judgement of presented scenarios.Motivated by the experiences of the original face-to-face trial, we als screened out participants who had not selected no when asked for *Literacy Difficulties*, *Dyslexia*, and *ADD/ADHD*, to avoid outliers in time. Finally, we set the *Approval Rate* screener to only allow participants whose contributions regularly get accepted (90%-100%).

3.3. Automatic Classification of Emotion Progressions

The gathered data served as the basis for the computational experiment, through which we sought to answer the third research question. In investigating how readily the gathered labels can be used for a sequence-to-sequence emotion classification task, we hoped to gain an insight into how readily the progressive nature of the novel annotations can be leveraged through machine learning. To this end, we defined a task and, which is detailed in section 3.3.1. Section 3.3.2 discusses the data preparation and implementation on a conceptual level. For further implementation details, we refer to section 4.2.1. Due to time constraints, we focused on the gathered categorical emotions for this task.

3.3.1. Task Design and Model Choice

The novelty of the gathered data consists mainly of its progressive nature, meaning the combination of considered context and sequential structure. We therefore intuit that any model which is able to make use of the prior information in a sequence should therefore benefit from the information encoded in the progressional labels and, hence, have an advantage over models that can not leverage this sequential and contextual information. With the computational experiment we seek to investigate if this intuition holds. In the following we detail different options of designing the However, since models vary in how they make use of sequential information, the choice of task and the choice of network are intertwined and should be considered in conjunction.

As explained in chapter 2, LSTMs are specifically designed to handle sequential data and propagate prior information and therefore were a clear candidate for any sequence-to-sequence classification task comparison we might choose to implement. While more recent and powerful transformer-based models are not inherently capable of taking into account the position information of each input, they are usually equipped with positional encodings, which add this information.

Before picking a model, we should considered how use them to answer the underlying question, meaning how evaluate if the progression information is being used by our network. A conceptually simple way of approaching this challenge is to alter or remove the sequential information: If a network does not train on emotional progressions, it should have a harder time predicting them than if it had trained on emotional progressions. This can broadly be achieved in two ways:

1. By altering the input data

2. By altering the model

The following discusses both options and how they can be implemented.

Disrupting Progressions through Data Alterations

While it is clear that the second option depends on the model choice, it may be less obvious that the first option also does. An easy way to illustrate this is by considering the difference between LSTMs and a variation of them called bidirectional LSTM (BiLSTM). While the former only propagates information in one direction, BiLSTMs employ two LSTMs that propagate information in opposite directions. This means that at any point in a sequence a BiLSTM has access to the entire sequence as context. If we now choose to shuffle the input sequence to disrupt the emotion progression, we can see how an LSTM would likely be affected by this alteration in the input data differently than a BiLSTM: While an LSTM can only consider what came before each item of the input sequence as context, the BiLSTM still has access to the entire sequence at any point. Therefore, while the context might have changed completely for the LSTM, only the order will have changed for the BiLSTM.

This is akin to the difference in the following labeling task for the third part of a text for humans:

1. Cut a text into parts, shuffle them, draw three, lay them out in a random order and, give them to a person. Let them read only the three parts you provided and ask them to judge the emotional content at the end.

2. Cut a text into parts, shuffle them, lay them all out in a random order, and give them to a person. Let them read the entire text, then ask them to judge the emotional content of the text up to the third part.

This example illustrates how the first setup is, besides the shuffling, the task human annotators were given to build the dataset. While we could introduce a similar effect for BiLSTMs, for example through masking, using LSTMs is arguably the more straightforward solution. Similarly, while the positional encoding of a transformer-based model would change for each entry in a shuffled input, the model would nonetheless consider the entire sequence. In this case, we could also employ strategies like masking or training on only one part as input at a time, as opposed to training on sequences, to mimic the loss of sequential information at training. Arguably, this again is less straightforward than shuffling the inputs during training for an LSTM.

Preventing Progressions from Being Leveraged through Model Alterations

An alternative to altering the input data that we considered consists of the second option introduced before: The utilized model architectures can be altered in a way that does not allow for context or sequential information to be leveraged by the system. When compared to the original architecture, this would also help give an insight into the extent to which the models learn to utilize the contextual information: If the altered models perform comparative to the original ones, this would indicate the classification of a part is largely based on just the information the current part provides. We ultimately decided against this approach, but still want to give a brief introduction of what this would have entailed and why we decided against it. For LSTMs, the alteration may be described as "cutting" the connections between cells, meaning to not allow propagation. Conceptually, this would mean not considering the inputs passed from the previous cell.

Mathematically, this could be achieved by setting those entries of the weight-matrices to zero which correspond to the previous output. Alternatively and equivalently, we may drop the corresponding dimensions altogether, reducing the dimensionality of the weight matrices W_f,W_o, W_i, and W_c in eq. (2.1), eq. (2.2), eq. (2.3), and eq. (2.5). As the forget gate, controlled by the linear layer parametrized through W_f and b_f, is used to control which parts of the previous memory will be kept and to which extent, we may drop them and replace them with a constant, pointwise zero multiplication, setting

$$f_t = 0$$

to circumvent the usage of C_{t-1} for sequence entry t. This further lowers the number of parameters, as one linear layer has been removed entirely. All other linear layers consider both the input of the current timestep x_t, as well as the output of the last cell h_{t-1}. In these cases, concatenation is used to combine both inputs. This means, that the impact of h_{t-1} can easily be negated by setting all entries of the corresponding weight matrices to 0. As before, we can reduce memory demands by simply dropping h_{t-1} from the computations and lowering the number conditionality of the weight matrices accordingly from $\mathbb{R}^{(n+m)\times u}$ to $\mathbb{R}^{x\times u}$, where $x_t \in \mathbb{R}^n$ and $h_t \in \mathbb{R}^m$ denote the dimensions of the embedding and cell state at any timestep t respectively, and u denotes the number of chosen units. Denoting those weight matrices of changed dimension with $\hat{W}$ yields the following set of adjusted equations that define the LSTM without recurrent connections:

$$
\begin{aligned}
i_t &= \sigma(\hat{W}_i x_i + b_i) && (3.1)\\
\tilde{C}_t &= \tanh(\hat{W}_c x_t + b_c) && (3.2)\\
C_t &= f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t && (3.3)\\
&= i_t \otimes \tilde{C}_t && (3.4)\\
o_t &= \sigma(\hat{W}_o x_t + b_o) && (3.5)\\
h_t &= o_t \otimes tanh(C_t) && (3.6)
\end{aligned}
$$

However, when training this model and a typical LSTM and comparing their performance, there is one key drawback: This reduction in trainable parameters and therefore in complexity may help the performance of the adapted LSTM over that of the vanilla LSTM, as complexity is lowered and training data is limited. To avoid this imbalance, we choose the shuffling of input data as the superior option. For the sake of simplicity, we will also avoid adjusting transformer-based models in a similar fashion, as they pose the additional challenge of adapting the positional encoding and the bi-directional context provided through the internal attention mechanism. Furthermore, while transformer-based models are more powerful than LSTMs, we are not interested in gauging how well our labels can potentially be predicted. Instead, we want to find out if and to what degree the progressional nature of our data is leveraged during training. Unless our base model does not end up performing above chance, a simpler model architecture such as LSTM is sufficient to answer this question.

Final Task Formulation

The consideration in this chapter led us to the following setup for the computational experiment: For each domain, we will train two LSTMs for a sequence-to-sequence emotion classification task. One time, we will provide text and label progressions in their original order, as found in the dataset. The second time around, we will randomize the order of the text and label sequences while maintaining the correct text and label correspondence. We will not alter the test sequences. To ensure the LSTMs are set up in a manner that allows them to learn and generalize from the training data, we perform a hyperparameter optimization for each input configuration. We then compare the best-performing representatives of the LSTMs trained on shuffled and intact emotion progressions. The intuition is that if the model learns to leverage the emotion progressions, we expect the best-performing system trained on intact progressions of a certain domain to outperform the best-performing system that was trained on disrupted progressions, as it has access to more meaningful information. We decide against comparing the same model for both training setups, as this introduces the need to choose between the best hyperparameters for the training on intact and disrupted progressions, which might introduce a bias.

3.3.2. Data Preparation and Implementation

Before training, a few preprocessing steps were required to generate suitable input representations of the gathered data. Both datasets contained instances that were labeled more than once, which is further discussed in section 4.1.4. Including the same instance multiple times might lead to biases, therefore we chose to only include each instance once. We were presented with two options for those instances that we collected more than one set of annotations for: We could either choose to aggregate the data in some way by computing one annotation that in some way takes into account multiple annotations, or we could pick one annotation for each instance and use it the way it was given. The vast majority of instances with more than one set of annotations was annotated by exactly two annotators. Therefore, there is no obvious way of aggregating the data, as methods like majority voting are not applicable. In addition, the progression information or underlying dynamics might be disturbed if we blindly aggregate annotations. Therefore, we chose to pick one annotation at random for each instance instead aggregating them. An alternative would be to utilize confidence scores to pick one annotation out of each set.

Next, we needed to obtain embeddings for each part. While simply using a pre-trained sentence transformers [75] was suitable for dreams, as a part consists of a sentence, the customer service data required further preparation and consideration. For one, each bi-turn usually consists of multiple sentences and we needed to decide how to aggregate these embeddings. The obvious choices were either concatenation or averaging. Averaging bears the advantage of keeping the input size manageable and the inputs comparable to the encoding used for dreams. In addition, averaging avoids an additional level of padding, as the turns tend to vary in length. On the other hand, averaging may lead to a significant loss in the informative value of the produced embedding, specifically if the encoded sentences greatly vary in their semantics and content. This poses an obvious problem for bi-turns: Averaging the embeddings generated over both turns would likely result in a large information loss, as the contents of the customer's and the agent's turn usually vary greatly. We, therefore, decided to combine both approaches: First, we split the conversations into sentences in the same automated fashion as we did for preprocessing dreams, as described in Chapter 3. In this step, the *==NEWMESSAGE==* tag is used as an additional end-of-sentence signifier. Furthermore, we decided to remove emojis before we pass the sentences to the encoder. We chose to do this for the sake of simplicity, as we were only interested in a baseline system to see if the progressional nature of our data could be leveraged. We note, however, that the research suggests that emojis carry valuable information for emotion prediction tasks which we are loosing with this step [30, 85, 38]. We then encoded each sentence using the same pre-trained sentence transformer as for dreams. From there, we either

1. averaged over all sentences, or
2. used the *==ADMIN==* and *==PART==* to split the bi-turns back into turns, averaged the embeddings within each turn, and finally concatenated the two embeddings obtained this way.

This leaves us with two sets of embeddings for customer service (CS) dialogues which we denote *CS Concat* and *CS Avg*, from this point forward.

4. Results

This chapter aims to analyze the results obtained with the methodology derived and described in the previous chapters with the main goal of answering the research questions posed in Chapter 1. To this end, we first analyze the gathered corpora to answer the first two questions, before detailing the results of the computational experiment, which is geared toward answering the third question.

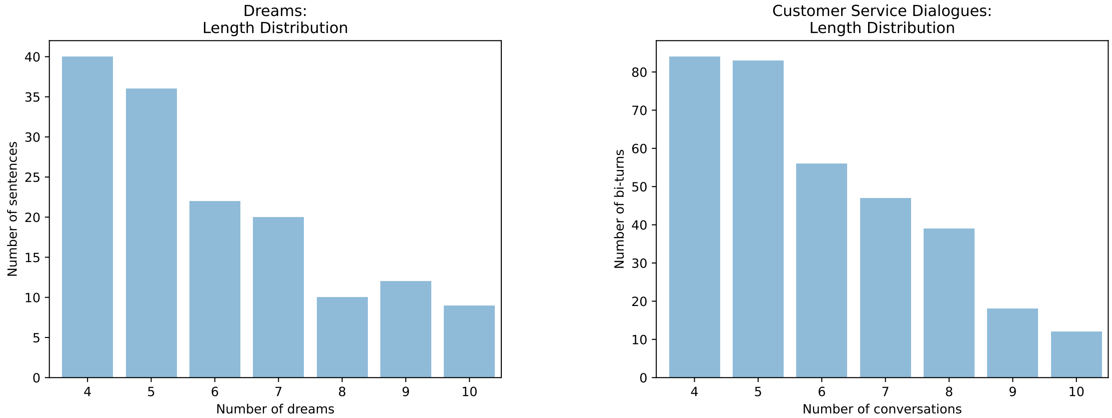
4.1. Data Analysis

In this section, we analyze the data we have gathered with the methodology described in Chapter 3. We first take a broad look at the gathered corpora in section 4.1.1, before considering the categorical labels and appraisal annotations in more detail in section 4.1.2 and section 4.1.3 respectively. In doing so, we hope to verify if the employed incremental task was suitable to gather progressional emotion annotation and into how these progressions build over the course of texts in the domains we considered.

4.1.1. Data Description

We decided to gather more data for the customer service dialogue corpus than for the dream corpus, as the raw data was substantially easier to gather and clean. The idea behind having two corpora of varying sizes was to have a bigger training dataset for one of the two domains, thereby increasing the chance of meaningful results for the machine learning experiment. In this spirit, the customer service dialogue dataset features 339 different annotated conversations, compared to the 149 dream reports featured in the dream corpus. Some of these texts were labeled more than once, either by mistake or by choice for later inter-annotator analyses, details on these instances can be found in section 4.1.4 and section 3.2.4 respectively. We included all annotations for instances labeled multiple times in the final dataset. This way anyone working on the data can freely decide how to integrate the annotations. Other options, such as aggregating or arbitrarily removing annotations, would no longer leave this choice. Counting multiple annotations of the same instances, the dream dataset features a total of 204 annotations, the customer dataset consists of 402 annotations. Both datasets have a mean length of around 5.9 parts per instance, the distribution of lengths is depicted in 4.1.

From this figure, we see that for both domains there is a bias towards shorter instances. This could be a result of the ranking algorithm, for which we normalized the sum of NRC-emotion word counts by the total word counts of a text. Shorter texts might be more to the point in their description of events and therefore relatively contain more



(a) The mean length of dream instances is 5.97 sentences.

(b) The mean length of customer service dialogue instances is 5.92 bi-turns.

Figure 4.1.: The distribution of instances of the number of parts for both corpora.

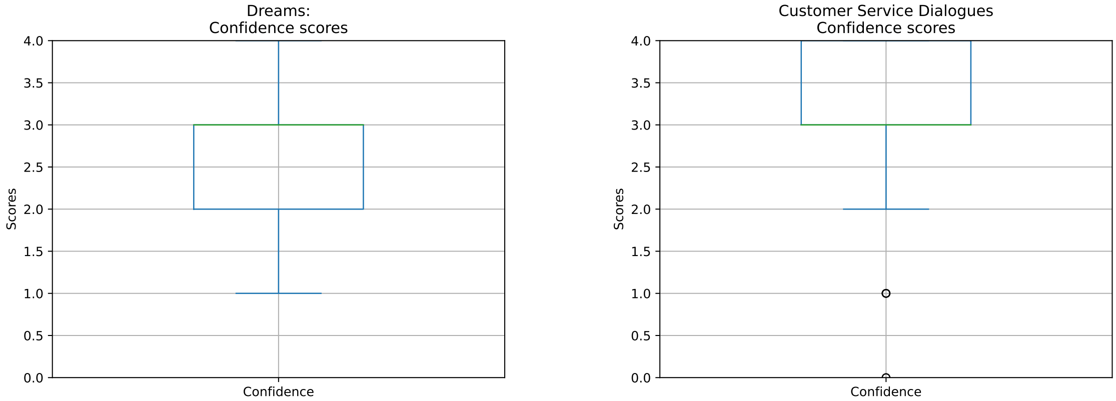
emotional words than longer descriptions. However, this may also reflect the underlying data: With a mean length of 6.7 parts, all reports in the considered dream series off DreamBank with a word count of up to 1000 and an adjusted sentence range from four to ten, also lean towards shorter dreams reports.

For each instance, the corpora contain:

- an instance identifier
- categorical emotion labels
- ordinal annotations for each appraisal dimension for each part
- an annotator ID to relate instances that were rated by the same person
- a confidence score

The exception to this rule is one customer service instance, which does not feature a confidence score, as the annotator seems to have skipped the question. We still include the annotation in the dataset. All other confidence scores contribute to the visualization of confidence scores in fig. 4.2, and Appendix A.2.

While for both domains, confidence values have a median of three on a scale from zero to 4, annotators were more confident in their annotations for customer service dialogues. Both distributions are asymmetric, with the confidence for customer service data reaching the highest value of four with its upper quartile, while the lower quartile of the confidence values for dream annotations reaches one point down, to two. The other quartile ends coincide with the mean in both cases. This means that annotators were relatively confident in their annotations, but more so when working with customer



(a) Dreams (b) Customer Service Dialogues

Figure 4.2.: Annotators' confidence in their annotations for both domains.

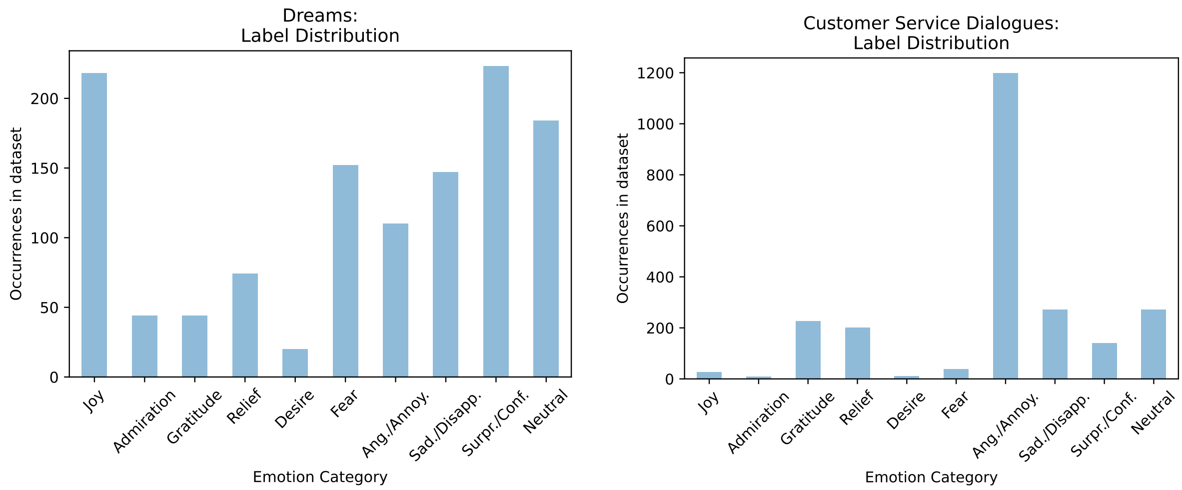
service data. One possible explanation for this can be found in the label distribution of the corpora at large, as we discuss in the following. A depiction of confidence scores by majority class and the last annotated class can be found in the Appendix in fig. A.13 and fig. A.14 respectively.

4.1.2. Categorical Label Analysis

We will now turn to the gathered categorical annotations and take a closer look at IAA and categorical emotion progressions. While the former analysis aims at answering the first question stated in Chapter 1 for categorical labels, the latter tackles the second question. To this end, we will focus on investigating if there are any overall trends in the distribution of the categories to answer questions like: Are customers less angry toward the end of the conversations? How often do certain emotions occur over the course of all instances? Is there a length difference between instances of different emotions?

Overall Category Occurences

Figure 4.3 shows the absolute frequency of categorical labels over all annotations included in the dataset. For dreams, we see the expected bias toward those labels that express a more negative sentiment, namely Fear, Anger/Annoyance, Sadness/Disappointment over those that express a more positive sentiment, namely Joy, Admiration, Gratitude, Relief, and arguably Desire. Nonetheless, Joy is the second most frequently annotated category, followed by Surprise/Confusion, which we also expected due to the at times erratic nature of dreams. Generally, we can observe that those labels that we either aggregated with Ekman emotion categories or that represented them from the start have been chosen far more frequently than those we copied from the customer service domain. This indicates that future annotations on dreams could benefit from focusing



(a) Dreams

(b) Customer Service Dialogues

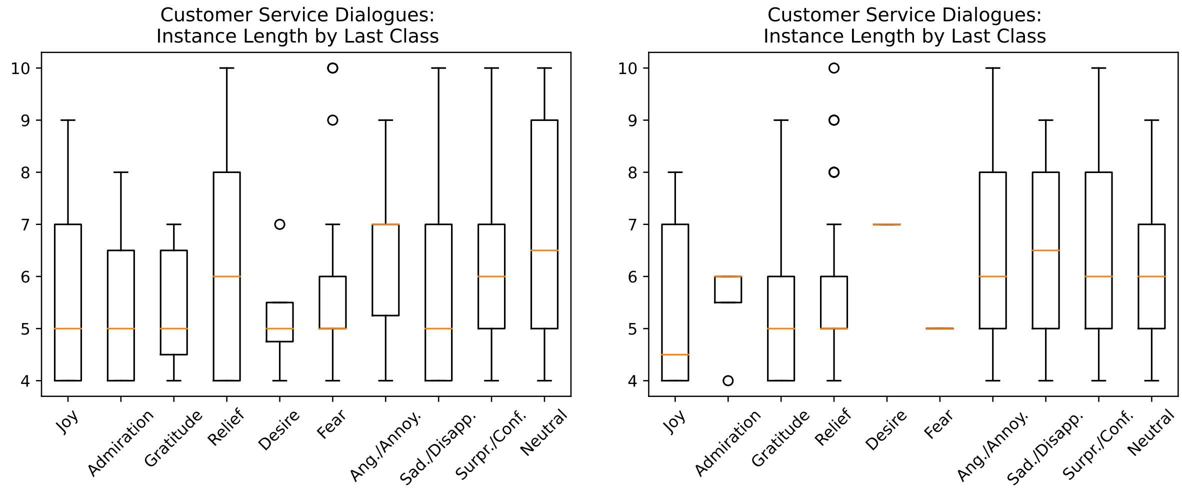
Figure 4.3.: Overall label distribution over all parts and annotations for both corpora

on basic emotion categories and that these seem to represent the experience of emotions in dreams well. In addition, surprise has been a less frequent label in previous corpora [11], making this distribution especially interesting. Since dream reports are structurally similar to event reports, adding them to the training set of a system that is trained on multiple datasets could help combat underlying data asymmetry found in other domains, specifically for surprise.

The customer service dataset, on the other hand, looks less balanced. The most striking observation we can make is that Anger/Annoyance has been chosen far more than any other category. A total of 58 annotations exclusively feature Anger/Annoyance as a label over all bi-turns. Removing them from the dataset does not alter any of the findings detailed in this section. We did expect a bias toward this category due to the nature of the domain, yet the magnitude of the difference is still surprising. Notably, only two out of the four categories we included for this domain specifically, namely Gratitude and Relief, are annotated frequently. Desire, Admiration, Fear, and Joy barely appear throughout the dataset. One possible reason is the ranking strategy. We ranked the instances based on their NRC word counts in the customer turns. It may be that customers who were angry communicated this more clearly than those who felt any other emotion. This might have led to a higher number of "angry" words in those instances that would later be labeled with anger, than the number of emotion words used to (not) communicate other emotions. These instances would then be selected to be annotated more often. One annotator explicitly commented on the prevalence of anger in their instances, their comment can be found in Appendix A.4. The distribution may also be a dataset or domain-inherent problem. A comparison to the label distribution previously gathered by Labat et al. [49] could help gain further insights but is outside the scope of this thesis. The computational experiments could help gauge to what extent this data can still be used to train machine learning models.

Additionally, we report the mean number of different emotion classes annotated over the progression of dreams to be 3.348, with a standard deviation of 1.307. This indicates that, on average, annotators changed the categorical emotion labels at least two times per dream, which implies the progressional information could not easily be reflected with one annotation, or even a start and end annotation. For customer service dialogues, the mean number of annotated classes per instance is 2.751, with a standard deviation of 1.134. From this, we conclude that, on average, the emotion category best describing the instances changes over their progression for both domains, though the progressions are more varied for dreams.

Instance Length and Emotion Category

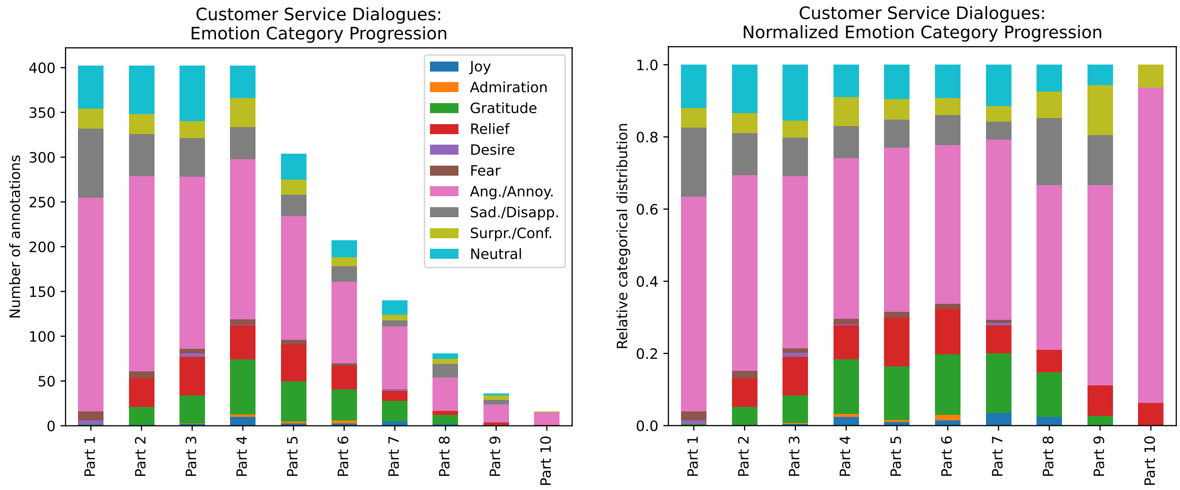


(a) Dreams (b) Customer Service Dialogues

Figure 4.4.: Instance lengths by emotion label for the last part.

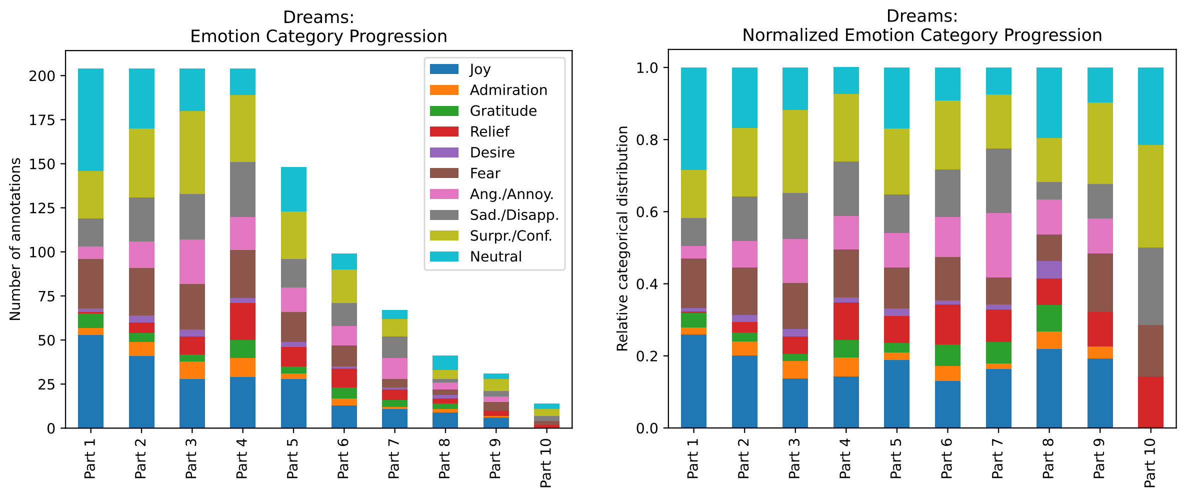
Figure 4.4 depicts the length distribution of instances depending on the category the annotators chose for the last part of each instance. This can be interpreted as the "overall" annotation of each text, as the annotators were to judge the emotional content of all parts up to and including the current one. For dreams, most classes feature reports of a wide variety of lengths, yet the median length stays at around five for six out of the ten classes. Exceptions to this are Relief, Anger/Annoyance, Sadness/Disappointment, Surprise/Confusion and Neutral, which all feature a higher median length. Out of the six classes with at least a hundred annotations overall, as depicted in fig. 4.3a, Joy and Anger/Annoyance are the only classes that do not feature any instances of length ten. Overall, there are more negatively valenced classes with a median clearly above 5 than positively valenced classes with the same property. For customer service dialogues we observe the same trend more clearly: The median length for negatively valenced classes and Neutral is higher than for positively valenced ones. The exceptions to this - Desire and Fear - do not feature enough samples to enable a reliable analysis. Joy in particular features the lowest overall median, indicating that overall joyous customer service instances may be shorter.

Progressions of Categorical Emotions



(a) Absolute frequency of labels per part. (b) Normalized frequency of labels per part.

Figure 4.5.: The label distribution for each part over the entire customer service dataset.



(a) Absolute frequency of labels per part. (b) Normalized frequency of labels per part.

Figure 4.6.: The label distribution for each part over the dream entire dataset.

Figures 4.5 and 4.6 allow for a more detailed look into how the emotion labels are distributed and change over the course of the instances. Both figures are based on all annotations in their respective corpus. There is no clear overall trend for either corpus, which is more surprising for the customer service dataset than for dreams. For the customer service domain, we may have expected to see anger play a greater role at the beginning of conversations, as the agent works to solve the problem. One possible explanation for this is that during the gathering of their underlying dataset, the agents were instructed to elicit a certain sentiment trajectory, some of them gearing towards a negative end-sentiment. Another one is the same, ranking-based argument we discussed previously. The lack of clear trends in the overall distribution makes following prediction tasks interesting, as models cannot simply learn to rely on the length of the previous text to make a prediction. For both domains, the number of different labels drops for the last parts, though this is very likely due to the reduced number of instances with this length. In regards to the first research question, we can gather from this that the underlying domain likely has a greater influence on the present emotions than where in a text we look for them.

Inter-Annotator Agreement for Categorical Emotions

The observed asymmetry in labels for customer service dialogues depicted in fig. 4.3 leads to a higher raw agreement in annotation for this domain for labels, as we can observe in fig. 4.7. This is an expected result, especially reflected in the micro F1 score in table 4.1b. All scores in this table were calculated only on those instances that were gathered for IAA evaluation purposes. Annotators were given double the number of instances to annotate for this task and two fixed annotators would each rate the same set of instances once. The same data underlies the confusion matrices in fig. 4.7, which further illustrate the data imbalance on the evaluation dataset.

Table 4.1 shows that while the micro F1 score is higher for customer service dialogues (0.517 for dialogues vs 0.426 for dreams) - due to the high agreement for Anger/Annoyance - the macro F1 score is higher for dreams (0.294 for dialogues vs 0.331 for dreams). These scores reflect the complexity of the annotation task. However, the scores can not be directly compared to any previously gathered datasets, as the processional nature of the annotations is new. From table 4.1a we can also observe that the F1 scores are significantly lower for those categories that were added for coherence with the previous label set for customer service dialogues, namely Admiration, Gratitude, Relief, and Desire. Excluding these categories and re-calculating the macro F1 score over only those categories that are either part of Ekman's emotion categories or aggregated with such plus Neutral yields a macro F1 of 0.449 - an increase of 0.118 over the reported macro F1 score including the customer service dialogue categories. We want to note that we cannot infer expected agreement of annotators on those instances that were now labeled with the customer service-specific categories from the data at hand, which means that excluding these categories for further annotation tasks may not lead to the same scores we obtain by excluding them post-hoc. In addition, this closes the

Micro F1	0.426
Macro F1	0.331

Category	**F1-Score**
Joy	0.48
Admiration	0.267
Gratitude	0.118
Relief	0.231
Desire	0.0
Fear	0.414
Ang./Annoy.	0.423
Sad./Disapp.	0.562
Surpr./Conf.	0.491
Neutral	0.323

(a) F1 scores for dreams.

Micro F1	0.517
Macro F1	0.294

Category	**F1-Score**
Joy	0.286
Admiration	0.0
Gratitude	0.449
Relief	0.5
Desire*	x
Fear	0.0
Ang./Annoy.	0.723
Sad./Disapp.	0.159
Surpr./Conf.	0.146
Neutral	0.383

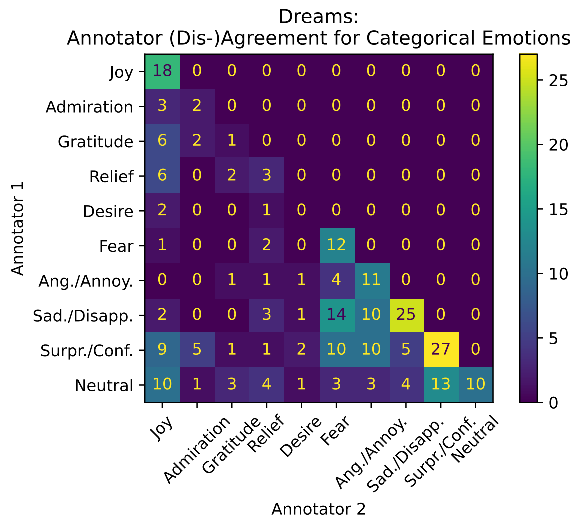
(b) F1 scores for customer service dialogues. * F1-score cannot be calculated, due to a lack of samples in the evaluation set, see fig. 4.7.

Table 4.1.: Inter-Annotator Agreement by category: F1 scores for emotion annotations. Class-specific F1 scores were calculated using the binary F1 metric by one-hot encoding the class in question.

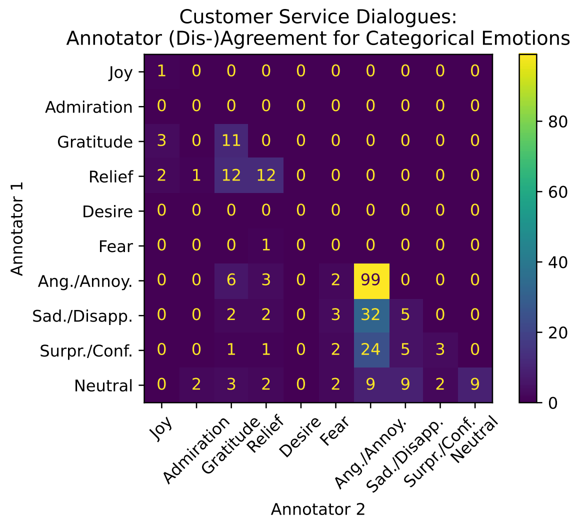
gap between the lowest (Neutral, 0.323) and highest (Sadness/Disappointment, 0.562) individual F1 scores significantly, which indicates that all categories are approximately equally hard to distinguish for annotators, with Sadness/Disappointment and Neutral being easier and harder respectively. This further supports the choice of basic emotion categories for future annotation tasks on dreams.

4.1.3. Appraisal Annotations Analysis

In this section, we analyze the appraisal annotations in the gathered datasets. As for categorical labels, we turn to inter-annotator measures to help answer the second question posed in Chapter 1. Since appraisal annotations are given as ordinal values instead of categories, we are also interested in checking how annotators' judgments vary over time: Can we observe convergence or divergence over the progression of the texts? For the second research question, we will consider the overall distribution of annotations overall instances, as well as changes in the mean score of appraisal annotations over the course of instances.



(a) Annotator agreement for dreams on the evaluation data.



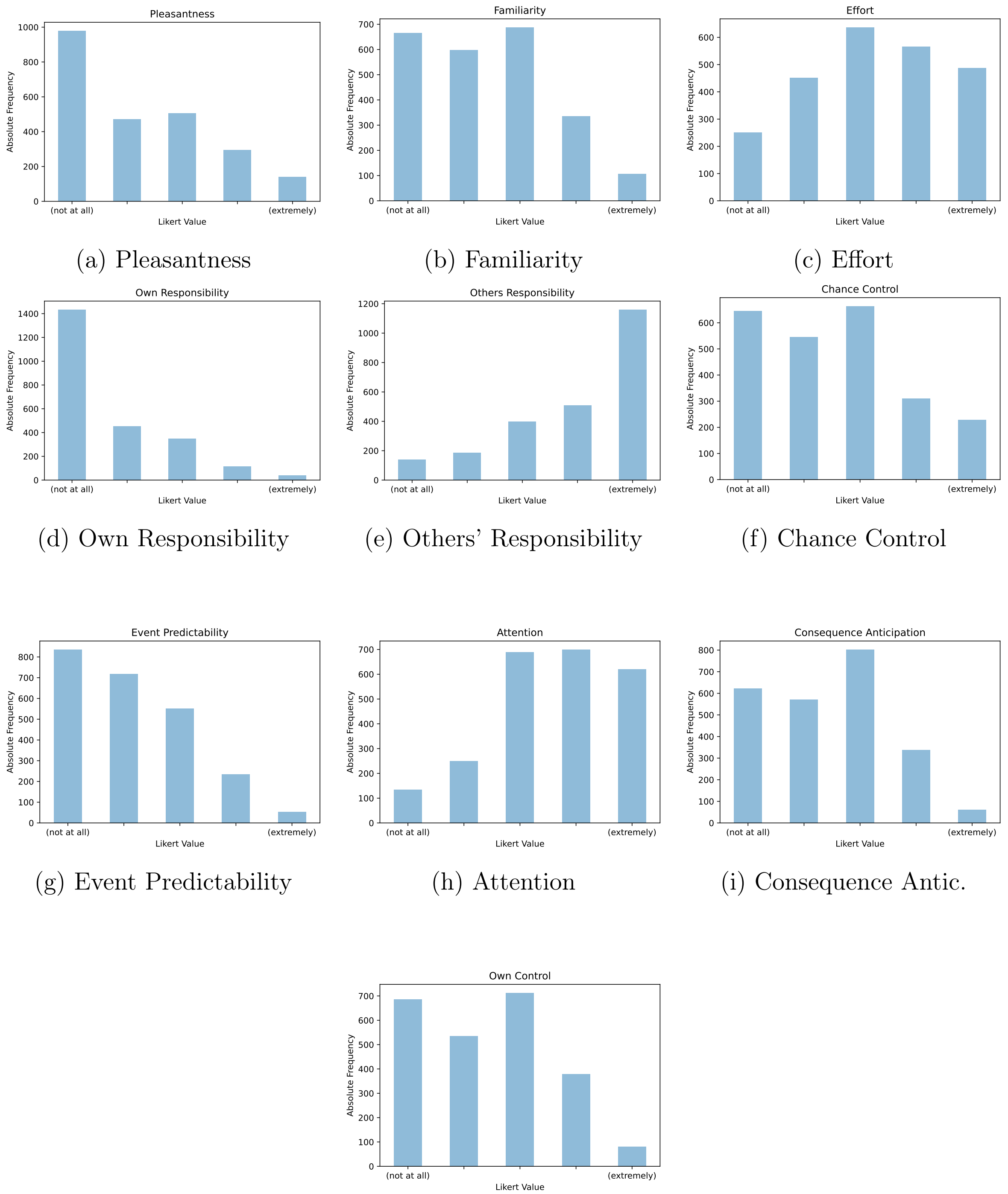
(b) Annotator agreement for customer service data on the evaluation data

Figure 4.7.: The label choices annotators made for evaluation instances for the same part. Entries on the main diagonal mean that the two annotators agreed on that label for the same part of the same instance.

Overall Appraisal Distributions

Figures 4.8 and 4.9 show the distribution of Likert scores representing appraisal ratings chosen by annotators over all annotations for customer service dialogues and dreams respectively. Analogously to 4.3, these statistics are based on all annotations in the respective dataset.

For customer service dialogues, we can see that the dimensions vary greatly in their distributions. Some, such as Own Responsibility, and Others' Responsibility, are heavily biased towards one end of the spectrum. These two dimensions, in particular, seem almost symmetrical - where Own Responsibility is heavily biased toward the "not at all option", Others' Responsibility is heavily biased toward "extremely". This is expected for the domain: Customers likely had little impact on the events and relied on the service agent and the company behind it to help them out. In addition, this is an indicator that the annotations we have gathered are to some degree consistent, at least for these two categories. Most categories show a moderate bias towards one of the two extremes, including Pleasantness (low), Event Predictability (low), Attention (high), Chance Control (low), Consequence Anticipation (low), and, to a lesser degree, Effort (high) and Own Control (low). This also mostly coincides with the findings of Smith and Ellsworth [86] and the high occurrence of anger in this corpus. According to their findings, anger is correlated with low Pleasantness and Own Control, moderately high Attention and Effort, as well as low Chance Control. The bias toward low Consequence Anticipation does not agree with Smith and Ellsworth's findings, as they found anger to be associ-



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

ated with moderately high Certainty, which maps to Consequence Anticipation. Further analyses of the correlations between the gathered appraisal and categorical annotations are outside the scope of this thesis and could be addressed in future work.

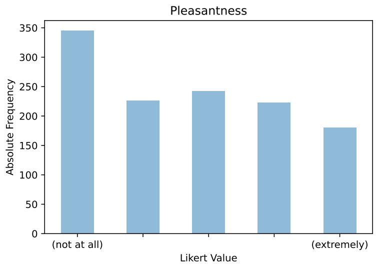
For the dream corpus, we see high biases for the dimensions Event Predictability (low) and Attention (high) in fig. 4.9. While the former might be due to the erratic nature of dreams, which is further highlighted by the moderate bias toward low Consequence Anticipation. The bias in Attention might reflect that most of the dreams do not revolve around mundane, everyday events. We can also observe a moderate bias toward low Pleasantness, and low Familiarity. Overall, the scores are more evenly distributed than for customer service dialogues. This further supports dreams as an interesting domain for further investigation, not just for categorical emotions but also for appraisals, as prediction tasks will both be non-trivial and data sources will likely contain a fair amount of samples for a wide variety of scores for a wide variety of dimensions.

Progressions of Appraisal Scores

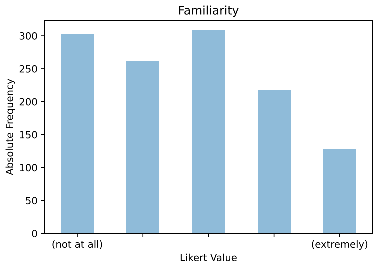
The annotations of appraisal scores allow for two central analyses. The first pertains to trends in appraisals over the course of all instances in a domain: Are there any trends for appraisal score at large over the progressions of instances in our corpora? The second set of central questions we can answer relates to the changes over the course of the text: How much do appraisals tend to change from one part to the next? Are there any indications that changes slow down or speed up over the progression of the text?

Figure 4.10 shows the mean values of appraisals over all annotated instances. These reveal two things: First, the biases we observed in our analysis of the overall distributions are also clearly visible when viewed over text progression. This is a direct consequence of the second observation: For most appraisal dimensions, the mean annotation values stay relatively constant when viewed over all instances. As for categorical emotion labels, this means that the distributions we observed earlier reflect the domain and underlying data at large, rather than being a product of a repeated, common progression between instances. For example, the bias toward high Attention does not reflect the sudden onset of events that require attention toward the middle of texts but rather reflects that the events reported in dreams are generally appraised as attention-worthy by the dreamer. However, we do observe an increase in the mean scores for attention over the course of texts in both domains. More generally, there is an increase for the mean scores for those appraisals that are biased toward high scores. The mean scores for Attention, Effort, and Others' Responsibility follow an overall upward trend up until at least the sixth sentence, the rounded mean length of instances in both corpora. We furthermore notice a sharper drop in those appraisal categories that displayed a low mean score for customer service dialogues over the last three parts. However, as the number of instances with this length is limited, this might reflect individual instances more than overall trends.

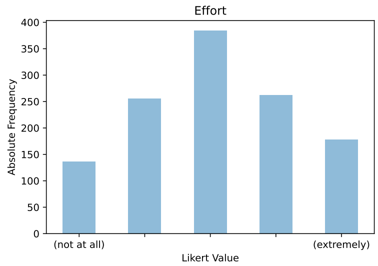
The stability of most mean scores over the progression of the texts begs the question



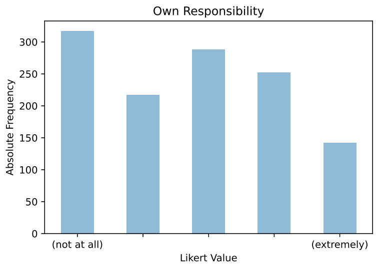
(a) Pleasantness



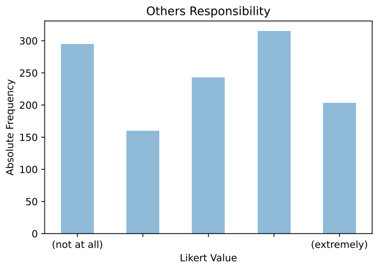
(b) Familiarity



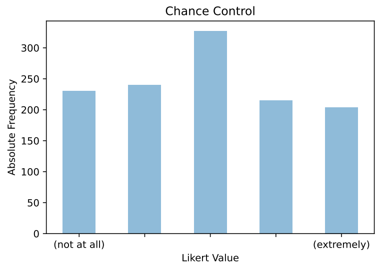
(c) Effort



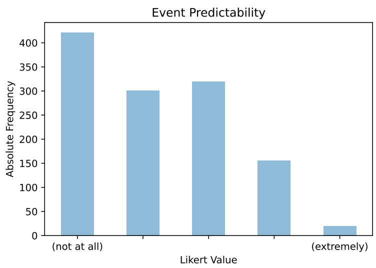
(d) Own Responsibility



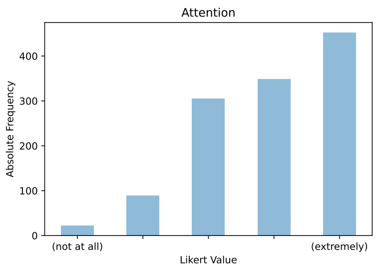
(e) Others' Responsibility



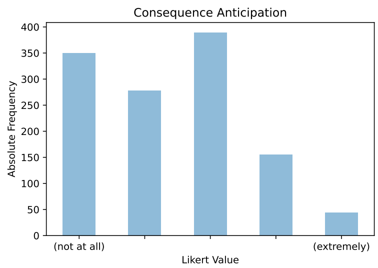
(f) Chance Control



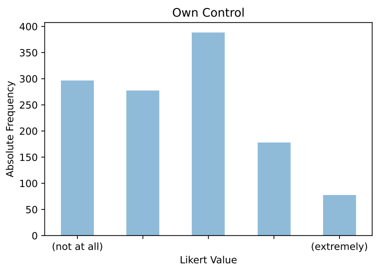
(g) Event Predictability



(h) Attention

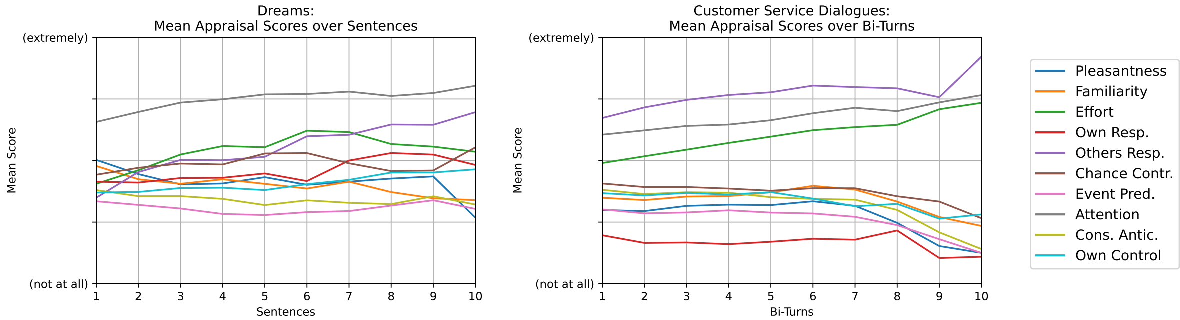


(i) Consequence Antic.



(j) Own Control

Figure 4.9.: Score distributions for each appraisal dimension over all annotations and parts for dreams.



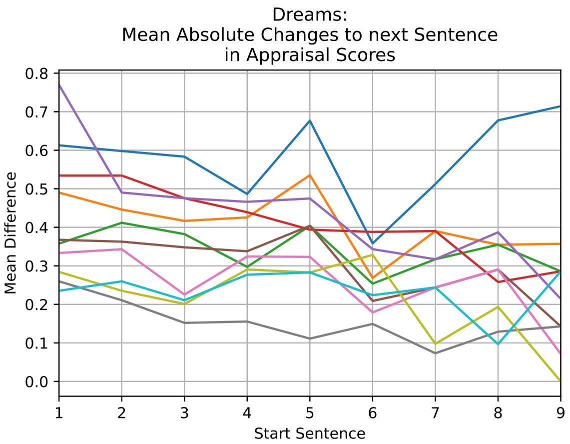
(a) Dreams (b) Customer Service Dialogues

Figure 4.10.: Mean appraisal scores over the progression of all instances.

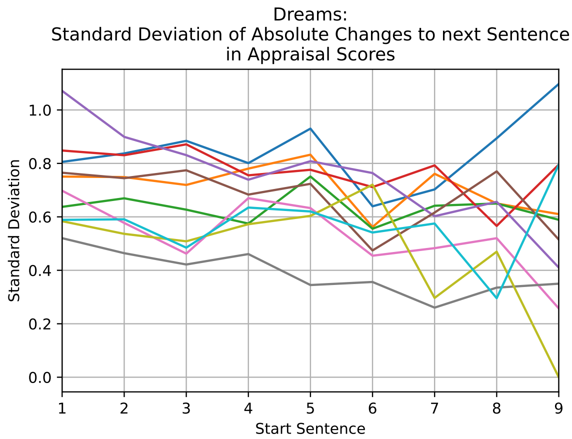
of whether this stability is due to a variety of different progressions, or if it is a product of appraisal annotations staying largely the same over the course of texts. In other words, we want to know if appraisals do progress, or if one annotation might ultimately sufficiently capture the emotional content of instances in terms of appraisal.

To gain insights into this question, we consider the mean absolute change between neighboring parts, as depicted in fig. 4.11. A mean absolute change at part one for a certain appraisal dimension would indicate that, on average, annotators changed their annotation by one point for the second part, when contrasted with the score they picked for the first. The mean absolute changes of all dimensions and for both corpora stay below one for all parts. Standard deviations are higher generally, but also stay below one by and large. Mean absolute changes are overall higher for dreams when compared to customer service dialogues for most appraisal categories, with most values staying below 0.3 for customer service dialogues. This means, that appraisal annotations for customer service dialogues were overall more static than for dreams. For the first six parts, we observe a particularly high change in annotations for Pleasantness, while Attention annotations stay the most constant. The mean changes staying mostly below one does, however, not indicate that annotations didn't change: For dreams, eight out of the ten appraisal categories display a sum of mean absolute changes of above one after four sentences - the minimum number of parts in our dataset, meaning that after four parts, on average, most appraisal annotations will have changed. The same is true for only three appraisal dimensions for customer service dialogues. After six parts, the rounded mean length of instances in both corpora, all appraisal dimensions for dreams sum up to more than one, and six out of ten appraisal dimensions exceed a sum of two. For customer service dialogues, eight out of ten summed up absolute mean changes exceed one, and they only sum up to a total of beyond two for Pleasantness. This indicates that dream annotations carry richer progression information than those found in customer service dialogues in terms of the number of score changes.

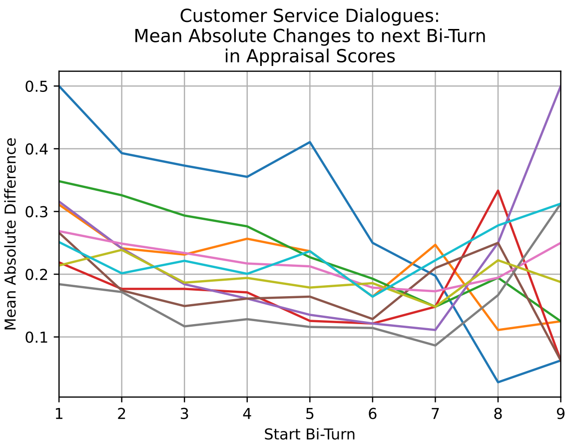
When analyzing these results it is noteworthy that during annotation, the scores participants checked for the last part where preselected for the current annotations. This could have introduced a bias toward lower changes between parts, as no change



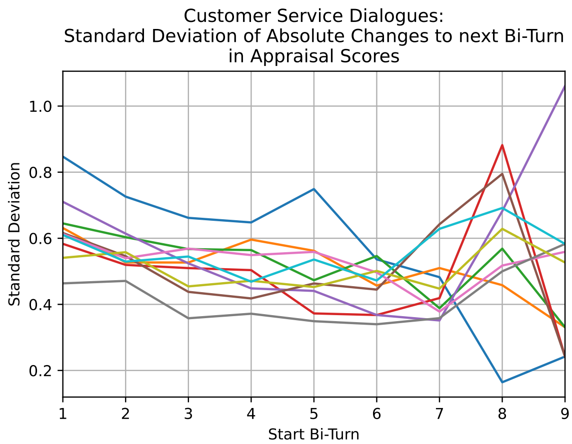
(a) Mean change in score between sentences for dreams.



(b) Standard deviation of change in score between sentence for dreams.



(c) Mean change in score between bi-turns for customer service dialogues.

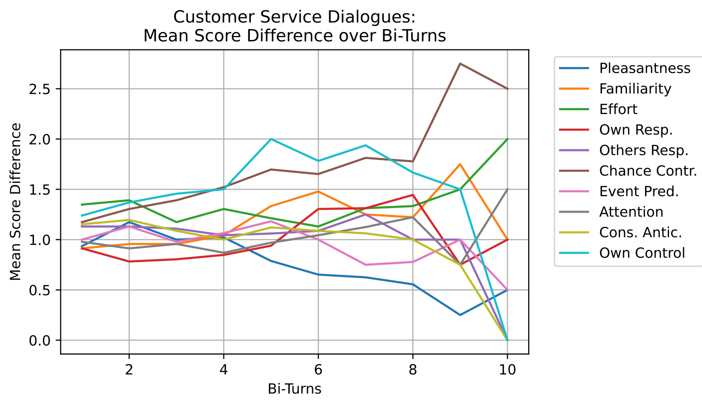


(d) Standard deviation of change in score between bi-turns for customer service dialogues.

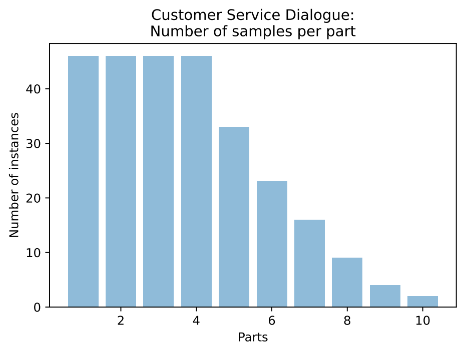
Figure 4.11.: The changes in appraisal scores between consecutive parts over the course of texts.

required the least action. If this is the case and how large this effect is could be analyzed by further annotations experiments without this preselection.

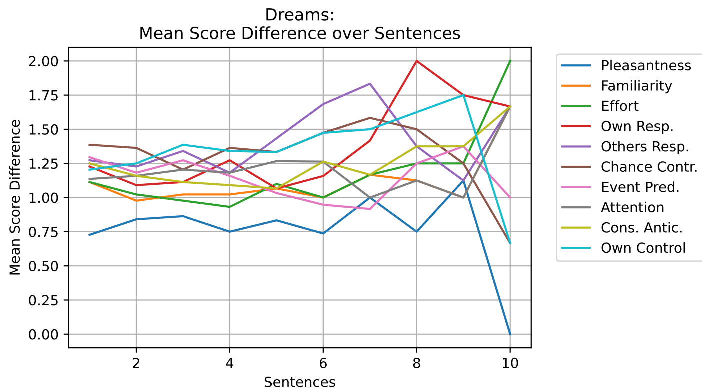
Inter-Annotator Correlation and Disagreement over Text Progression for Appraisal Dimensions



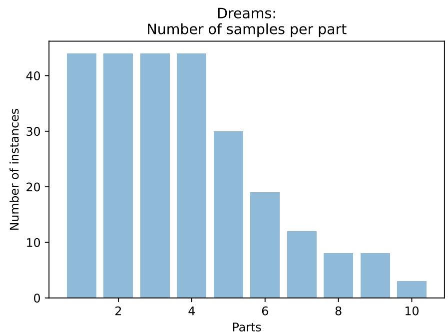
(a) The mean difference in scores for each appraisal category and part for customer service dialogues.



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

Figures 4.12c and 4.12a show the mean difference between the assigned Likert scores over the progression of the instances gathered for IAA analyses for customer service dialogues and dreams respectively. The visualizations are accompanied by displays of the number of instances in the evaluation set that reach any given number of parts in fig. 4.12d and fig. 4.12b respectively. The maximum possible difference between scores is four, as annotations were made on a five-point scale. A difference of two therefore can be interpreted as annotators no longer generally agreeing on whether a certain dimension is highly or lowly appraised. For both corpora, we can observe that all mean scores steadily stay between 0.5 and 1.6 for the first four parts. These also are the parts with the most annotations, as we set four as the minimum number of parts, hence every instance in the evaluation set contributes to these mean scores. Therefore, these are the most reliable scores. After five parts, means diverge between appraisal dimensions as the number of samples they are based on, and therefore their generalizability, decreases. Only the appraisals of Own Control and Chance Control for customer service dialogues ever display a mean score difference of above 1.5. From this, we gather that these appraisals might be hard to evaluate for customer service dialogues or that the wording of the questions might have been harder to understand in the context of this domain. All other mean differences consistently stay below this value. As this is still below the threshold of two, we deduce from this that annotators seemed to generally have agreed on their appraisal annotations. As a consequence, we furthermore deduce that the way of gathering annotations derived in Chapter 3 is generally suitable to gather progressions of appraisal annotations. Own Responsibility is consistently the appraisal category that annotations diverged for the least for the first four parts for customer service dialogues. This is not surprising since, as we have discussed before and as and easily visible from fig. 4.8d, annotations for this dimension were very biased toward the lowest score throughout the whole dataset. Far more interestingly, Pleasantness has the lowest mean difference almost throughout all parts for dreams. The margin is especially noticeable for the first six parts, and then again when it drops to zero for the few instances with a length of ten. From fig. 4.9a we gather that Pleasantness has a bias toward low scores, yet this bias is not as strong as for other dimensions. This could be either due to a very different distribution in the evaluation dataset, or due to Pleasantness being easier to spot and differentiate for annotators in dreams than other appraisals. Other than a spike in the mean difference for Others' Responsibility in dreams, these observations stay consistent for the next two parts, although, with fewer underlying instances, they start diverging more. As there are only a handful of annotations with more than six parts in the evaluation dataset, any behavior in the mean scores may speak more to the individual instances in that set than any dynamics of the overall data, we leave those without discussion.

Instead, we turn our attention to table 4.2. The annotations gathered by the pairs listed in the left columns of table 4.2a and 4.2b comprise the evaluation set. The annotated instances vary between pairs but are fixed within pairs. The numerical annotator ID is not of importance for the analyses carried out here, but are consistent with their IDs in the dataset and reported for reproducibility. All correlation analyses are carried out within annotator pairs only. Changes in values between annotator pairs can therefore be either due to the annotators themselves or due to the instances they were assigned. By doubling the number of parts annotated by each participant for the evaluation dataset, we tried to minimize the effects of the individual instances and increase the significance of our analyses. Yet, with only between six and nine instances annotated by each pair due to time constraints, we cannot guarantee significance, or rule out that the instances in one set were inherently harder to annotate than those in others. Table

Annotators	Pearson's r
25, 28	0.316
26, 29	0.22
27, 30	0.456
38, 41	0.465
39, 43	0.351
40, 42	0.466

(a) Dreams.

Annotators	Pearson's r
36, 39	0.359
35, 38	0.264
34, 37	0.446
44, 47	0.403
45, 49	0.346
46, 48	0.582

(b) Customer Service Dialogues.

Table 4.2.: Pearson's r for each annotator pair in the evaluation set of both domains. The annotator IDs are only used to group annotations made by the same person.

4.2 shows the Pearson correlation for each annotator pair over all appraisal annotations. Both domains show positive correlations for all pairs, with a mean correlation of 0.348 for dreams, and a slightly higher mean correlation of 0.4 for customer service dialogues. This means that, generally, when one annotator chose a higher value, the other annotator also tended to choose a higher value. All but two annotator pairs, one for each domain, show at least a moderate correlation of >0.3. While only one annotator pair reaches a high degree of correlation, with a Pearson score of >0.5, half of all annotator pairs reach scores of >0.4. We take this as another indicator that the data collection method derived in Chapter 3 is feasible in principle.

Taking a more detailed look at the correlations for each annotator pair by appraisal dimension, displayed in table 4.3, gives a little more insight into the data. This table reveals that, despite our best efforts, in 24 out of 60 cases, we do not reach statistical significance for dreams. For customer service dialogues, 29 out of 60 pairings do not yield a significant correlation coefficient. In addition, in three cases at least one of the annotators annotated one dimension with the exact same score over all parts of all instances. This happens once for own Responsibility, once for Attention, and once for Own Control. The same is true for customer service dialogues; as we can gather from table 4.3a, this happens twice for Own Responsibility and once for Attention. With this phenomenon occurring twice for Attention and three times for own Control over both datasets, it may also be an indicator that certain scores are generally more common for these domains. While Own Responsibility is heavily biased for customer service dialogues, as we gather from 4.8d, a bias is present, yet less pronounced for Attention, as we observe in fig. 4.8h, and 4.9h. As explained above, this could also be due to the individual instances in the evaluation sets. This effect is more easily visible when considering columns, as the number of statistically significant, well-defined correlation scores vary from one to seven between pairs. Yet, we only obtain a single significant Pearson score (0.402) for Attention for dream annotator pairs, which means the cause is more likely rooted in some interplay of that appraisal dimension, the domain, and the question formulation. The same is true for Event Predictability, with a correlation of -0.474 the annotations are heavily negatively correlated in this case. This could be a reflection on the question formulation for dreams, or of the general difficulty of annotating this dimension for dreams. The Pearson scores for the same dimension for customer service dialogues are significant four out of six times but vary heavily between -0.325 and 0.763. We observe a similar dynamic for the correlation scores of Own Control, with only one value reaching a significant correlation score (-0.517) for customer service dialogues and a wide range for dreams (-0.486 to 0.518). Pleasantness, Familiarity, Effort, Others' Responsibility, and Consequence Anticipation all display only positive correlations, where significant, over both domains. All correlation scores for these dimensions display at least a moderate correlation of >0.3. We can take interpret this as a sign that the question formulation of these dimensions was clear. Overall, the great number of missing significant values and wide ranges of correlation scores make further interpretation difficult. Should another dataset of this kind be gathered, a higher number of samples between annotator pairs could help get a clearer picture. Still, we conclude, that at least some appraisal categories lend themselves to being annotated in the way that we derived in Chapter 3.

4.1.4. Remarks

As mentioned before, during the data-gathering process some errors arose that led to superfluous annotations. Some of these, especially early on, were due to wrong setups on SoSci Survey, which discarded already annotated combinations on instances once the annotators were done instead of right after the combinations were assigned. This way, if an annotator started working while another one was already annotating a set of instances, there was a chance of the new annotator receiving the same set of instances to annotate. This was not caught during the pre-tests, as only one set of instances was annotated multiple times to test technical feasibility. It was caught and fixed in the first set of annotations.

In addition, a few annotators used the message functionality on Prolific or the feedback input box at the end of the survey to reach out and inform the author about error messages on SoSci Survey, depending on if they could reach the end of that survey or not. The feedback gathered through the survey can be found in Appendix A.4. In none of these instances, the error could be reproduced. In some of these cases, participants were still able to continue and their annotations showed no signs of corruption. However, one feedback given over the Prolific messaging functionality detailed that the reason was maintenance on SoSci Survey:

> "[...] I have received a message stating that the site is down for maintenance and will be available in three minutes. [...] after three minutes I tried to refresh the page but was given the following message: Questionnaire Error (page 3) [...] There is an error in the PHP code:"

Similar issues could have therefore also been caused by underlying problems in the platform, which could explain why they were not reproducible.

Annotators	36, 39	35, 38	34, 37	44, 47	45, 49	46, 48
Pleasantness	0.643	0.33	0.326	0.486	0.395	0.894
Familiarity	-	0.315	-	-	-	-
Effort	0.755	-	-	0.342	-	0.44
Own Responsibility	-	-	x	-	x	0.815
Others Responsibility	0.432	-	0.487	0.738	-	0.303
Chance Control	-	-	0.461	-0.297	-	0.635
Event Predictability	-0.319	0.347	0.763	-	-	0.689
Attention	-0.325	-	0.46	0.501	-	x
Consequence Anticipation	-	-	-	0.482	-	0.338
Own Control	-0.517	-	-	-	-	-

(a) Customer service dialogues

Annotators	25, 28	26, 29	27, 30	38, 41	39, 43	40, 42
Pleasantness	0.576	0.865	0.543	-	0.748	0.585
Familiarity	0.573	-	0.612	-	0.424	0.566
Effort	-	0.47	0.476	-	0.329	0.733
Own Responsibility	x	-0.657	0.601	0.401	0.536	0.429
Others Responsibility	-	0.608	0.352	0.383	-	-
Chance Control	-0.304	0.481	0.591	-	-	-
Event Predictability	-	-0.474	-	-	-	-
Attention	-	-	-	x	-	0.402
Consequence Anticipation	0.485	-	-	0.901	0.491	0.421
Own Control	-	-0.496	0.518	-	x	0.463

(b) Dreams

Table 4.3.: Pearson's r by appraisal dimension and annotator pair.
- denotes entries with scores that did not reach significance. x denotes entries where at least one annotator marked the same score for each part of each instance, leaving r undefined.

For three instances, annotations were corrupted. This means, that for some categories there were more annotations than instances. This, again, was not reproducible. All of these annotations were removed from the final dataset.

All of this led to some missing annotations. To maintain continuity in the dataset and make it reproducible in the manner described in Chapter 3, the author of this thesis annotated four instances herself: These instances carry the identifiers *d_vicky_20* (dreams), and *c315_6*, *c301_11*, *c171_9* (customer service dialogues). Two corrupted annotations were part of the evaluation set and the corruption only occurred for one of the two times these instances were annotated. We removed the corrupted versions from the dataset, kept the intact ones, and removed the instances from consideration for the evaluations. These instances are the dreams with identifiers *d_b_0259*, and *d_b2_3945*.

4.2. Computational Experiment

In this section, we first discuss the implementation details of the computational experiment described in 3.3.1. We then go on to analyze the data to see if there are any indications that the trained LSTMs were able to leverage the implicit progression information.

4.2.1. Experiment Setting and Implementation Details

The experiment implementation started with embedding the sentences in the corpora using the sentence transformer library by Huggingface [75]. More specifically, we employed the pre-trained model *all-MiniLM-L6-v2*, as it is described as a general-purpose model and because it offers a good trade-off between performance and speed [74]. This way, we obtained three sets of embeddings *Dreams*, *CS Concat*, and *CS Avg*: For any dream d, we obtained an embedding of size $p_d \times 384$, where $4 \leq p_d \leq 10$ denotes the number of parts in dream d. Analogously, the obtained embeddings in CS Avg that were computed as the mean of all sentence embeddings for each bi-turn of any given conversation c are of size $p_c \times 384$, where $4 \leq p_c \leq 10$ denotes the number of parts in c. In addition, we obtain one embedding of size $p_c \times 768$ for each conversation by computing the mean of each turn and concatenating these to obtain an embedding for the entire bi-turn. Together, these embeddings form the set of embeddings we denote as CS Concat.

Motivated by the relatively low count of parts labeled with Admiration, Gratitude, Relief, and Desire in dreams visible in fig. 4.3, we chose to aggregate these categories with Joy, yielding one category that is intuitively generally more positively valenced. The one exception to this is the class of Desire, as the valence value of this class could depend on the context. Yet, for the sake of obtaining only one aggregated class and arguing that, intuitively, Desire is not strongly negatively valenced, we still chose to aggregate it with the others. To ensure comparability, and since more positively valenced categories are generally rare for the customer service dataset, we chose the same aggregation strategy for customer service data. This yielded six target classes, that could be described as

1. Positive,
2. Fear,
3. Anger/Annoyance,
4. Sadness/Disappointment,
5. Surprise/Confusion, and

6. Neutral.

We point out that "Positive" is merely the descriptive name we choose to use in this thesis to denote the aggregated class. It does not imply that we mix emotion models and compare valence values to categorical labels. What stands out is how closely this label set resembles the set of emotions defined in the Hall/Van De Castle system of dream content analysis [25], which features Anger, Apprehension - a less intense form of fear according to Plutchik's wheel of emotion [69] - Sadness, Confusion, and a singular positive class, Happiness.

For the model implementation, we mainly made use of TensorFlow [55] and Keras [19]. We padded each of the previously computed embeddings and corresponding labels to size ten, the maximum instance length, and to this end introduced an additional padding class that we did not regard. We encoded all classes with a one-hot vector to enable the usage of the categorical cross-entropy function for the loss function. To aid training for all categories despite the biases in the label distribution, during each fold, we computed class weights from the training dataset using scikit-learn [65]. These were then passed through the *sample_weight* argument to the Keras API to make them usable with the one-hot encoded class labels.

Furthermore, we chose to evaluate the model through 10-fold cross-validation, as the datasets are limited in size, and trained for up to 150 epochs. For each fold, we split off an additional 10% of the training data as a validation set, that we used for early stopping. We set up early stopping with a minimum delta of 0.0001, meaning improvements are only registered as such if they were greater than 0.0001. Furthermore, we set up early stopping to restore the weights that produced the lowest loss on the validation dataset once training was stopped. For patience, meaning the number of epochs we continued to train despite a lack of improvement before stopping, we tried out the values 7 and 20. For systems trained on shuffled progressions, all training sequences were randomized independently of one another, so as to not introduce a new canonical ordering across all instances the system could learn to adapt to. Shuffling was not applied to masked timesteps, and labels and parts were shuffled in unison to maintain the proper structure of instances and the relationship between parts and assigned labels. To this end, we monitored the loss on this validation set.

The model itself was comprised of a masking layer, which masked all padded timesteps, followed by an LSTM layer. We set the number of units in the LSTM as another hyperparameter that we tried various values for, namely 5, 15, 25, 35, and 45. We fixed the recurrent dropout rate to 0.2 and tried out the parameters 0.1, 0.2, and 0.5 for the (input) dropout parameter. This was then followed by a dense layer, in which the number of nodes mirrored the number of output categories. The dense layer featured softmax as its activation function. We further varied the batch size, comparing results for batch sizes of 20 and 32.

After training, we picked the best hyperparameters for the early stopping patience, batch size, unit count, and dropout percentage by comparing mean macro F1 scores over all ten folds for the following analyses. We did this separately for each embedding set (Dreams, CS Concat, CS Avg) and each training scenario (intact progressions, shuffled progressions), obtaining six sets of hyperparameters in total. The choice to monitor macro f1 over other possible candidates, such as accuracy, was made due to the heavy label imbalance.

4.2.2. Experiment Results

In this section, we report and analyze the results of the computational experiment.

Overall Performance

	Dreams	**CS Concat**	**CS Avg**
Batch Size	20	32	32
Dropout	0.2	0.5	0.2
Units	35	35	15
Early Stopping Patience	7	20	20

(a) Best-performing hyperparameters when LSTMs were trained on intact progressions.

	Dreams	**CS Concat**	**CS Avg**
Batch Size	20	32	32
Dropout	0.2	0.1	0.5
Units	25	35	15
Early Stopping Patience	20	20	20

(b) Best-performing hyperparameters when LSTMs were trained on shuffled progressions.

Table 4.4.: The combination of the best-performing hyperparameters for all embeddings sets and training scenarios.

Table 4.4 summarizes the best hyperparameters found for each embedding set and training scenario with regard to the mean macro F1 over all folds, table 4.5 reports the performance scores of these systems in terms of mean macro F1 and accuracy. In the latter table, we first notice that all scores are relatively low. Due to the simplifications made in preprocessing and model choice, the limited training data and the complexity of the task, we did not the expect the classification performance to be exceptional. On first glance we also notice that the performance of all systems trained on intact sequences surpasses the performance of their counterparts trained on shuffled sequences; however, we point out that the according mean scores for both training scenarios lie within standard deviation of each other in all cases. Keeping this in mind, we still go on to describe and analyze the results in more detail.

Metric	Dreams		CS Concat		CS Avg	
	mean score	standard deviation	mean score	standard deviation	mean score	standard deviation
Macro F1	0.291	0.05	0.294	0.039	0.278	0.027
Accuracy	0.363	0.073	0.405	0.089	0.378	0.041

(a) Performance for the best-performing LSTMs trained on intact progressions.

Metric	Dreams		CS Concat		CS Avg	
	mean score	standard deviation	mean score	standard deviation	mean score	standard deviation
Macro F1	0.280	0.061	0.286	0.018	0.267	0.031
Accuracy	0.322	0.066	0.394	0.039	0.353	0.085

(b) Performance for the best-performing LSTMs trained on shuffled progressions.

Table 4.5.: Mean performance scores and standard deviation over all ten scores and their standard deviation.

As we had expected, the concatenated embeddings did apparently help retain more information, as the mean macro F1 score for CS Concat embeddings is slightly above the same score for CS Avg embeddings for both training scenarios. This effect is small, yet larger - 0.016 for the training on intact progressions, and 0.019 when trained on shuffled progressions - than the effect of the training scenario, which is 0.008 for the CS Concat embedding set and 0.011 for the CS Avg set. The consistency in performance increase for the intact training scenario over the shuffled one could imply that LSTMs could be somewhat able to leverage progression information, yet the minuscule difference between both training scenarios does not allow for a definitive answer. Taking into account the low overall performance, the core problem could lie in the simple approach we chose: A repetition of this task with a stronger baseline system, such as a transformer based architecture, more training data, and a less rigorous preprocessing of the data could help improve the overall performance, which might lead to a clearer result.

Performance by Class

Table 4.6 reports mean F1 scores per category for all trained systems. For dreams, the system based on the shuffled sequences outperforms its counterpart trained on intact sequences for all classes but Positive and Anger/Annoyance. The same is true for the results obtained on the CS Avg embedding set. Notably, for dreams the differences between the according F1 scores for the Positive and Anger/Annoyance class lies at 0.108 for both cases. This is by far the largest difference between according mean F1 scores for any class and embedding set, and it is also higher than any difference we observed in the mean macro F1 scores we compared previously. The next biggest difference we

Category	Dreams		CS Concat		CS Avg	
	intact	shuffled	intact	shuffled	intact	shuffled
Positive	0.532	0.426	0.605	0.576	0.582	0.521
Fear	0.332	0.379	0.015	0.067	0.0	0.075
Anger/Annoyance	0.267	0.159	0.496	0.501	0.447	0.43
Sadness/Disappointment	0.291	0.294	0.239	0.2	0.216	0.251
Surprise/Confusion	0.216	0.225	0.129	0.131	0.15	0.164
Neutral	0.304	0.314	0.313	0.299	0.317	0.272

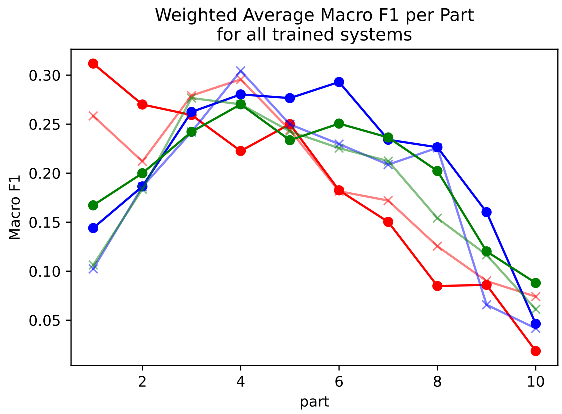
Table 4.6.: Mean F1 scores by category for all trained systems. Class-specific F1 scores were calculated using the binary F1 metric by one-hot encoding the class in question.

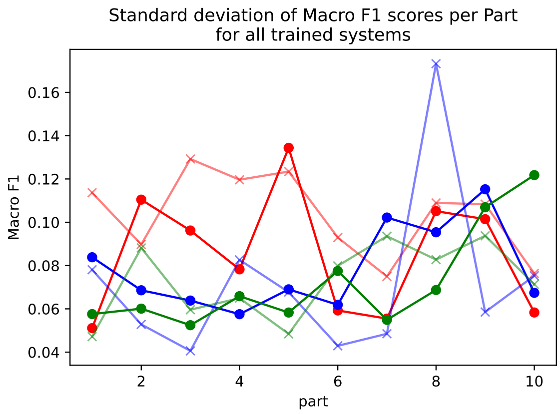
observe for dreams is for Fear, where the mean F1 scores differ by 0.047. This could indicate that progression information can be more easily leveraged for these classes in the case of dreams. For the CS Concat embedding set, differences in mean F1 scores across classes are small, with the largest difference of 0.052 being the one found for the classification of Fear. Each training scenario yields better results than its counterpart in three out of six cases, as we would expect by chance from two systems that largely have the same performance.

We also note that the F1 scores for the aggregated Positive class are higher than for any other class over all embedding sets and training scenarios. This is somewhat expected for dreams, as after the aggregation this class has by far the most samples and the system might therefore have a better chance of learning and generalizing from the distribution of the input data. Yet, the same is not true for customer service dialogues: Even after aggregation, there is far more data for labeled Anger/Annoyance than positive in the dataset. This is an indicator that the systems might have an easier time learning to distinguish between more positively valenced and negatively valenced classes, than they do learning to distinguish between the various negatively valenced classes. Yet, as this observation pertains to both training scenarios, it does not have direct implications for any of the research questions this thesis seeks to answer.

Performance Over Text Progression

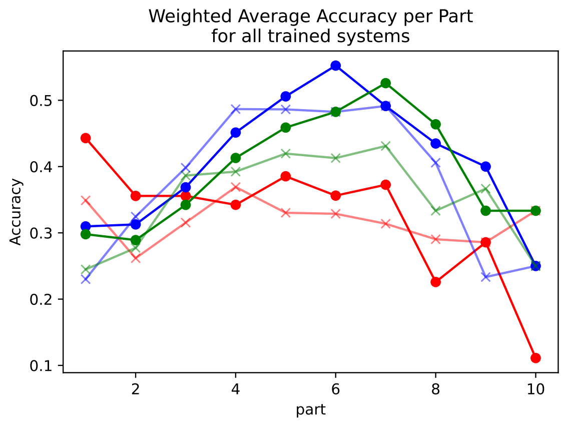
Figure 4.13 shows the development of mean macro F1 and mean accuracy for all best performing systems over parts. For dreams, the systems trained on intact and shuffled progressions both outperform their counterpart in five out of ten parts with respect to mean macro F1. The comparison for classification performance on customer service data is a little more interesting: On both embedding sets, the systems trained on intact progressions consistently outperform their counterparts from part six onward with only one exception. This is interesting because we would expect the classification of later parts to particularly benefit from the previous progression information, since more context can

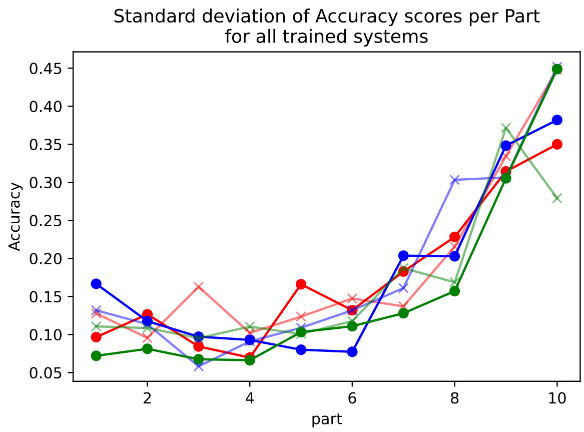




(a) Macro F1 scores for all trained systems over the course of instances.

(b) Standard deviation of Macro F1 scores for all trained systems per part.





(c) Accuracy scores for all trained systems over the course of instances.

(d) Standard deviation of accuracy scores for all trained systems over the text progression.

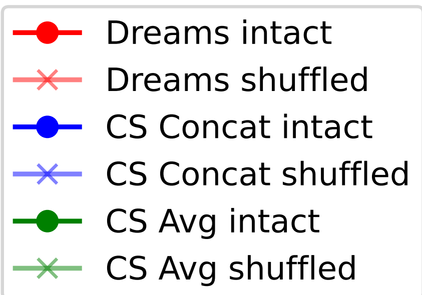


Figure 4.13.: Weighted average metrics over parts and the corresponding standard deviations. The contribution weight of the metric for each part and each fold was chosen proportional to the amount of instances that reached the required length.

be leveraged. While this might indicate that LSTMs learned to leverage the progressional nature of the data, there are multiple alternate explanations that could explain this behavior: For one, this could be due to chance, as differences generally remain small. The second reason could be an imbalance in the data that makes certain (majority) classes more likely later on. This is unlikely though, as we did not observe heavy imbalances in the emotion distribution over parts in the analysis of fig. 4.5. A third explanation could be that annotations for customer service labels stay largely static after a certain point and that the systems trained on progressions learn to stick with a class. However, this could also arguably be interpreted as the systems learning to leverage progressions in a sense. We did find a lower mean number of annotated classes for customer service dialogues than for dreams, yet this statistic alone can't support that customer service annotations are static, as they could also oscillate between two classes. We find a similar results for accuracy. Due to the quick uptake in standard deviation for accuracy after part six, the reported mean accuracy scores beyond this length should be considered with caution. Overall, the analysis over parts does not give a clear insight into whether or not progressions are leveraged by the systems that had accessed to them during training.

5. Conclusion and Future Work

In this thesis, we crowd-sourced two corpora containing emotion labels that describe emotion progressions throughout texts, both in terms of categorical emotion labels, as well as appraisal scores. The corpora varied in domain. One dataset was built on customer service dialogues, the other on dream reports. The novelty of this dataset consists of the progressional nature of the data: Annotations are sequential and each annotation is contextualized by the parts of the text that precede it, hence describing the emotional content of the entire text up and including the part they are linked to. Parts were defined with the domain in mind: While a part consisted of a sentence in the case of dreams, they were defined as the union of two consecutive turns - one by each party - for customer service dialogues. To this end, we designed a survey and asked participants to choose between ten emotion categories and evaluate ten appraisal dimensions on a five-point Likert scale for each part. Crucially, annotators were only ever shown each text up to the part they were currently annotating, which made the annotation task incremental and ensured the text following a part did not influence annotations. We then used this data to train LSTMs to evaluate if and to what extent these systems could leverage the progressional nature of the categorical emotion data.

Throughout this thesis, we aimed to answer three central questions: First, we were interested in finding out if the incremental annotation task design was suitable for gathering annotations that would describe the progressions of emotional content in text. Through the analysis of the gathered data, we have found plenty of evidence to support that this is the case. First, annotators reported high confidence in their annotations. In addition, the evaluated inter-annotator measurements for categorical emotions, F1, were moderate for all categories that were annotated a reasonable amount of times. We found similar results when analyzing Pearson's r for appraisal scores. As high inter-annotator agreement is generally hard to achieve for emotion annotations, we largely take this as a representation of the task's complexity, rather than a sign of the utilized manner of gathering data being fundamentally unsuitable for the task. We also found that Ekman's emotion categories seem to lend themselves well for the annotation of dreams, as labels that included one of these emotions were chosen more often and with greater inter-annotator agreement. An alternative could be the emotion set for dream content analysis as proposed by Hall/Van De Castle, as the final label set after aggregation closely resembled the five emotions proposed there. In addition, we analyzed the mean difference between appraisal scores over the progression of text. For shorter instances, we found a reasonable mean difference that indicated annotators did not generally fundamentally disagree on scores. However, we also found that the low number of long instances limited our ability to reliably analyze this metric for long texts. Paying special attention to the length distribution when gathering future datasets of this nature could help to resolve this. More generally, our evaluations were limited by the amount of data we gathered between pairs of annotators. We report F1 for categorical emotion labels instead of other, more common measures for inter-annotator agreement, such as κ-statistics to give insights into the reliability of our data, despite the sets of instances rated by the same two annotators being limited in all cases. Besides the novel structure of the data, this is a second reason that complicates direct comparison with other emotion corpora.

The second question aimed at getting insights into how emotion annotations build up over the course of texts. We found that while the emotion distributions vary greatly between domains, they have no clear dynamic over the course of texts that we uncovered. This is even true in the case of customer service dialogues, for which we would have expected certain changes in the distribution of classes over time, as we would have expected customers to get less angry over time. We can, however, not rule out that this specific finding for customer service dialogues is specific to our dataset, as the dialogues were gathered with certain sentiment trajectories in mind. In our dataset, we found a heavy bias toward anger for the customer service dialogues over all parts. The bias toward anger was also reflected in the associated appraisal dimensions. We found the annotations to be more balanced for dreams, over all but also over their progression, which could make them an interesting domain for future research in Emotion Analysis, and specifically for the analysis of emotion progressions. When analyzing the progressions of appraisal annotations over the course texts in both domains, we found evidence for more frequent changes in appraisal annotations for dreams than for customer service dialogues. This indicates that appraisal progressions in dreams are less static than those for customer service dialogues. The analysis of mean change between parts also indicates that progressional annotations capture information about the emotional content of dreams in terms of appraisal that is unlikely to be represented by a single annotation.

The third question was geared toward finding out if and to what extent machine learning systems could leverage the progressional nature of the gathered emotion annotations. For time reasons, we focused on the categorical labels for this task, which leaves any experiments on the gathered appraisal progressions to future research. The experiment comprised training LSTMs on both the original emotion progressions, as well as on manipulated instances, in which the individual parts were randomized during training. Both systems were then evaluated on instances with intact emotion progressions in a 10-fold cross-validation setup. Through this task, we hoped to get an insight into whether the trained systems learn to rely on the progressional nature of the annotations for prediction, or if the prediction is based largely on the input information of that part. Analyses of the classification performance in terms of macro F1 and accuracy showed that all trained systems performed poorly, which likely is a result of the simplicity of the chosen model architecture and the complexity of the task. Systems trained on intact label progressions consistently achieved slightly better performance scores than their counterparts that were trained on shuffled sequences. However, the increases were too small to conclude that the trained LSTMs were able to leverage the progressional nature of the gathered annotations. A performance comparison by class indicated, that progressions might have helped in the classification of positively valenced emotions and anger for dreams. Overall, this research question could not be conclusively answered in this thesis. Future experiments with a stronger machine learning architecture, more training data, or less rigorous preprocessing of the data could help get a clearer insight.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Druck-Exemplaren überein.

Datum und Unterschrift:

Declaration

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted hard copies.

Date and Signature:

6. Bibliography

[1] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen. "Transformer models for text-based emotion detection: a review of BERT-based approaches". In: *Artificial Intelligence Review* (2021), pp. 1–41. DOI: 10.1007/s10462-021-09958-2.

[2] L. von Ahn. *Offensive/Profane Word List*. URL: https://www.cs.cmu.edu/~biglou/resources/bad-words.txt (visited on 12/22/2022).

[3] *All emojis designed by OpenMoji – the open-source emoji and icon project. License: CC BY-SA 4.0*. URL: https://openmoji.org/.

[4] C. O. Alm, D. Roth, and R. Sproat. "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*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 579–586. URL: https://aclanthology.org/H05-1073.

[5] E. Altszyler, M. Sigman, S. Ribeiro, and D. F. Slezak. "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database". In: (2016). DOI: 10.48550/arXiv.1610.01520.

[6] S. Aman and S. Szpakowicz. "Identifying Expressions of Emotion in Text". In: *Text, Speech and Dialogue*. Ed. by V. Matoušek and P. Mautner. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 196–205.

[7] D. Barrett. "Dreams about COVID-19 versus normative dreams: Trends by gender". In: *Dreaming* 30.3 (2020), p. 216. DOI: 10.1037/drm0000149.

[8] V. Barriere, S. Tafreshi, J. Sedoc, and S. Alqahtani. "WASSA 2022 Shared Task: Predicting Empathy, Emotion and Personality in Reaction to News Stories". In: *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 214–227. DOI: 10.18653/v1/2022.wassa-1.20. URL: https://aclanthology.org/2022.wassa-1.20.

[9] K. Bartsch and H. M. Wellman. *Children talk about the mind*. Oxford University Press, Inc., 1995. ISBN: 0-19-508005-X.

[10] L. Bertolini, V. Elce, A. Michalak, G. Bernardi, and J. Weeds. "Automatic Scoring of Dream Reports' Emotional Content with Large Language Models". In: *arXiv preprint arXiv:2302.14828* (2023). DOI: 10.48550/arXiv.2302.14828.

[11] L.-A.-M. Bostan and R. Klinger. "An Analysis of Annotated Corpora for Emotion Classification in Text". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 2104–2119. URL: https://aclanthology.org/C18-1179.

[12] S. Buechel and U. Hahn. "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

[13] S. Buechel and U. Hahn. "Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation". In: *ECAI 2016*. IOS Press, 2016, pp. 1114–1122. DOI: 10.3233/978-1-61499-672-9-1114.

[14] S. Buechel and U. Hahn. "Emotion Representation Mapping for Automatic Lexicon Construction (Mostly) Performs on Human Level". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018.

[15] S. Buechel and U. Hahn. "Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation". In: *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–12. DOI: 10.18653/v1/W17-0801. URL: https://aclanthology.org/W17-0801.

[16] K. Bulkeley and M. Graves. "Using the LIWC program to study dreams". In: *Dreaming* 28.1 (2018), p. 43. DOI: 10.1037/drm0000071.

[17] L. M. Ceríaco, R. F. de Lima, M. Melo, and R. C. Bell. *Biodiversity of the Gulf of Guinea Oceanic Islands: Science and Conservation*. Springer Nature, 2022. ISBN: 978-3-031-06153-0. DOI: 10.1007/978-3-031-06153-0.

[18] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal. "SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 39–48. DOI: 10.18653/v1/S19-2005. URL: https://aclanthology.org/S19-2005.

[19] F. Chollet. *keras*. https://github.com/fchollet/keras. 2015.

[20] J. Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[21] A. Curci and B. Rimé. "Dreams, emotions, and social sharing of dreams". In: *Cognition and Emotion* 22.1 (2008), pp. 155–167. DOI: 10.1080/02699930701274102.

[22] C. Darwin. *The expression of the emotions in man and animals*. Ed. by J. Murray. 1872. URL: http://darwin-online.org.uk/content/frameset?pageseq=1&itemID=F1142&viewtype=text.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[24] G. W. Domhoff and A. Schneider. "Studying dream content using the archive and search engine on DreamBank.net". In: *Consciousness and Cognition* 17.4 (2008), pp. 1238–1247. DOI: 10.1016/j.concog.2008.06.010.

[25] G. W. Domhoff. "The Hall/Van de Castle System of Content Analysis". In: *Finding Meaning in Dreams: A Quantitative Approach*. Boston, MA: Springer US, 1996, pp. 9–37. ISBN: 978-1-4899-0298-6. DOI: 10.1007/978-1-4899-0298-6_2.

[26] K. Durand, M. Gallay, A. Seigneuric, F. Robichon, and J.-Y. Baudouin. "The development of facial emotion recognition: The role of configural information". In: *Journal of experimental child psychology* 97.1 (2007), pp. 14–27. DOI: 10.1016/j.jecp.2006.12.001.

[27] P. Ekman. "An argument for basic emotions". In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10.1080/02699939208411068.

[28] P. C. Ellsworth and K. R. Scherer. "Appraisal processes in emotion." In: *Handbook of affective sciences*. Oxford University Press, 2003, pp. 572–595.

[29] P. C. Ellsworth and C. A. Smith. "From appraisal to emotion: Differences among unpleasant feelings". In: *Motivation and emotion* 12.3 (1988), pp. 271–302. DOI: 10.1007/BF00993115.

[30] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Com-

putational Linguistics, 2017, pp. 1615–1625. DOI: 10.18653/v1/D17-1169. URL: https://aclanthology.org/D17-1169.

[31] J. L. Fleiss. "Measuring nominal scale agreement among many raters". In: *Psychological bulletin* 76.5 (1971), p. 378. DOI: 10.1037/h0031619.

[32] J. L. Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378. DOI: 10.1037/h0031619.

[33] A. Fogli, L. Maria Aiello, and D. Quercia. "Our dreams, our selves: automatic analysis of dream reports". In: *Royal Society open science* 7.8 (2020), p. 192080. DOI: 10.1098/rsos.192080.

[34] E. Frantova and S. Bergler. "Automatic emotion annotation of dream diaries". In: *Proceedings of the analyzing social media to represent collective knowledge workshop at K-CAP 2009, The fifth international conference on knowledge capture*. 2009.

[35] F. A. Gers, J. Schmidhuber, and F. Cummins. "Learning to forget: Continual prediction with LSTM". In: *Neural computation* 12.10 (2000), pp. 2451–2471. DOI: 10.1049/cp%3A19991218.

[36] S. L. Gilchrist, J. A. Davidson, and J. Shakespeare-Finch. "Dream emotions, waking emotions, personality characteristics and well-being–A positive psychology approach". In: *Dreaming* 17 (2007), pp. 172–185. DOI: 10.1037/1053-0797.17.3.172.

[37] M. Grandini, E. Bagli, and G. Visani. "Metrics for multi-class classification: an overview". In: *arXiv preprint arXiv:2008.05756* (2020).

[38] G. Guibon, M. Ochs, and P. Bellot. "From Emoji Usage to Categorical Emoji Prediction". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by A. Gelbukh. Cham: Springer Nature Switzerland, 2023, pp. 329–338. DOI: 10.1007/978-3-031-23804-8_26.

[39] E. Hartmann, M. Zborowski, and R. Kunzendorf. "The emotion pictured by a dream: An examination of emotions contextualized in dreams". In: *Sleep and Hypnosis* (2001).

[40] S. Hochreiter and J. Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[41] J. Hofmann, E. Troiano, and R. Klinger. "Emotion-Aware, Emotion-Agnostic, or Automatic: Corpus Creation Strategies to Obtain Cognitive Event Appraisal Annotations". In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Online: Association

for Computational Linguistics, 2021, pp. 160–170. URL: https://aclanthology.org/2021.wassa-1.17.

[42] J. Hofmann, E. Troiano, K. Sassenberg, and R. Klinger. "Appraisal Theories for Emotion Classification in Text". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 125–138. DOI: 10.18653/v1/2020.coling-main.11. URL: https://aclanthology.org/2020.coling-main.11.

[43] G. Hripcsak and A. S. Rothschild. "Agreement, the f-measure, and reliability in information retrieval". In: *Journal of the American medical informatics association* 12.3 (2005), pp. 296–298. DOI: 10.1197/jamia.M1733.

[44] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku. "EmotionLines: An Emotion Corpus of Multi-Party Conversations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. URL: https://aclanthology.org/L18-1252.

[45] R. Klinger, O. De Clercq, S. Mohammad, and A. Balahur. "IEST: WASSA-2018 Implicit Emotions Shared Task". In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 31–42. DOI: 10.18653/v1/W18-6206. URL: https://aclanthology.org/W18-6206.

[46] S. Labat, N. Ackaert, T. Demeester, and V. Hoste. "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*. Ed. by G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, and A. Uma. Marseille, France: European Language Resources Association (ELRA), 2022, pp. 66–72. ISBN: 9791095546986. URL: %7Bhttp://www.lrec-conf.org/proceedings/lrec2022/workshops/NLPerspectives/index.html%7D.

[47] S. Labat, T. Demeester, and V. Hoste. "Guidelines for annotating fine-grained emotion trajectories in customer service dialogues (version 1.0)". In: *LT3 Technical Report–LT3 20-01* (2020).

[48] S. Labat, T. Demeester, and V. Hoste. "(Accepted). EmoTwiCS: a corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter". In: *Language Resources and Evaluation* (2022).

[49] S. Labat, T. Demeester, and V. Hoste. "(Submitted). EmoWOZ-CS: A novel dataset to model emotion trajectories in text-based customer service dialogues". In:

[50] S. Labat, A. Hadifar, T. Demeester, and V. Hoste. "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)*. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022, pp. 106–112. URL: https://aclanthology.org/2022.wnut-1.12.

[51] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. ISBN: 1558607781. DOI: 10.5555/645530.655813.

[52] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551. DOI: 10.1162/neco.1989.

[53] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017, pp. 986–995. URL: https://aclanthology.org/I17-1099.

[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019). DOI: 10.48550/arXiv.1907.11692.

[55] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[56] S. M. Mohammad and P. D. Turney. "Crowdsourcing a word–emotion association lexicon". In: *Computational intelligence* 29.3 (2013), pp. 436–465. DOI: 10.1111/j.1467-8640.2012.00460.x.

[57] S. Mohammad and F. Bravo-Marquez. "WASSA-2017 Shared Task on Emotion Intensity". In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark:

Association for Computational Linguistics, 2017, pp. 34–49. DOI: 10.18653/v1/W17-5205. URL: https://aclanthology.org/W17-5205.

[58] A. Moors. "Appraisal Theory of Emotion". In: *Encyclopedia of Personality and Individual Differences*. Ed. by V. Zeigler-Hill and T. K. Shackelford. Cham: Springer International Publishing, 2017, pp. 1–9. ISBN: 978-3-319-28099-8. DOI: 10.1007/978-3-319-28099-8_493-1.

[59] D. Nadeau, C. Sabourin, J. De Koninck, S. Matwin, P. D. Turney, et al. "Automatic dream sentiment analysis". In: *Proceedings of the Workshop on Computational Aesthetics at the Twenty-First National Conference on Artificial Intelligence*. 2006.

[60] P. Nandwani and R. Verma. "A review on sentiment analysis and emotion detection from text". In: *Social Network Analysis and Mining* 11.1 (2021), p. 81. DOI: 10.1007/s13278-021-00776-6.

[61] T. A. Nielsen, D. Deslauriers, and G. W. Baylor. "Emotions in dream and waking event reports". In: *Dreaming* 1.4 (1991), p. 287. DOI: 10.1037/h0094340.

[62] L. A. M. Oberländer, E. Kim, and R. Klinger. "GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 1554–1566.

[63] C. Olah. *Understanding LSTM Networks*. 2015. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[64] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh. "Dimensional Emotion Detection from Categorical Emotion". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.358.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[66] J. W. Pennebaker, M. E. Francis, and R. J. Booth. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[67] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[68] *Plutchik-wheel.* 2011. URL: https://upload.wikimedia.org/wikipedia/commons/c/ce/Plutchik-wheel.svg (visited on 04/25/2023).

[69] R. Plutchik. “The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”. In: *American Scientist* 89.4 (2001), pp. 344–350. (Visited on 10/12/2022).

[70] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, 2019, pp. 527–536. DOI: 10.18653/v1/P19-1050. URL: https://aclanthology.org/P19-1050.

[71] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. “Emotion recognition in conversation: Research challenges, datasets, and recent advances”. In: *IEEE Access* 7 (2019), pp. 100943–100953. DOI: 10.1109/ACCESS.2019.2929050.

[72] J. Posner, J. A. Russell, and B. S. Peterson. “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology”. In: *Development and psychopathology* 17.3 (2005), pp. 715–734. DOI: 10.1017/S0954579405050340.

[73] A. H. Razavi, S. Matwin, J. De Koninck, and R. R. Amini. “Dream sentiment analysis using second order soft co-occurrences (SOSCO) and time course representations”. In: *Journal of Intelligent Information Systems* 42 (2014), pp. 393–413. DOI: 10.1007/s10844-013-0273-4.

[74] N. Reimers. *Pretrained Models.* URL: https://www.sbert.net/docs/pretrained_models.html.

[75] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

[76] A. Rogers, O. Kovaleva, and A. Rumshisky. “A primer in BERTology: What we know about how BERT works”. In: *Transactions of the Association for Computational Linguistics* 8 (2021), pp. 842–866. DOI: 10.1162/tacl_a_00349.

[77] I. J. Roseman. “Cognitive determinants of emotion: A structural theory”. In: *Review of personality & social psychology* (1984).

[78] I. J. Roseman and C. A. Smith. "Appraisal theory: Overview, assumptions, varieties, controversies". In: *Appraisal processes in emotion: Theory, methods, research* (2001), pp. 3–19.

[79] J. A. Russell and A. Mehrabian. "Evidence for a three-factor theory of emotions". In: *Journal of Research in Personality* 11.3 (1977), pp. 273–294. ISSN: 0092-6566. DOI: 10.1016/0092-6566(77)90037-X.

[80] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108* (2019). DOI: 10.48550/arXiv.1910.01108.

[81] K. R. Scherer. "The dynamic architecture of emotion: Evidence for the component process model". In: *Cognition and Emotion* 23.7 (2009), pp. 1307–1351. DOI: 10.1080/02699930902928969.

[82] M. Schredl. *Dream content analysis: Basic principles*. Vol. 3. 1. 2010, pp. 65–73. DOI: 10.11588/ijodr.2010.1.474.

[83] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger. "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*. 2017, pp. 13–23. DOI: 10.18653/v1/W17-5203.

[84] M. Schuster and K. K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093.

[85] A. A. M. Shoeb and G. de Melo. "EmoTag1200: Understanding the Association between Emojis and Emotions". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 8957–8967. DOI: 10.18653/v1/2020.emnlp-main.720.

[86] C. A. Smith and P. C. Ellsworth. "Patterns of cognitive appraisal in emotion". In: *Journal of personality and social psychology* 48.4 (1985), p. 813. DOI: 10.1037/0022-3514.48.4.813.

[87] M. A. Stranisci, S. Frenda, E. Ceccaldi, V. Basile, R. Damiano, and V. Patti. "APPReddit: a Corpus of Reddit Posts Annotated for Appraisal". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 3809–3818. URL: https://aclanthology.org/2022.lrec-1.406.

[88] E. Troiano, L. A. M. Oberlaender, M. Wegge, and R. Klinger. "x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022, pp. 1365–1375. URL: https://aclanthology.org/2022.lrec-1.146.

[89] E. Troiano, L. Oberländer, and R. Klinger. "Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction". In: *Computational Linguistics* (2022), pp. 1–71. DOI: 10.1162/coli_a_00461.

[90] E. Troiano, S. Padó, and R. Klinger. "Emotion ratings: How intensity, annotation confidence and agreements are entangled". In: *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2021.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[92] Y. Wang, J. Zhang, J. Ma, S. Wang, and J. Xiao. "Contextualized Emotion Recognition in Conversation as Sequence Tagging". In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, 2020, pp. 186–195. URL: https://aclanthology.org/2020.sigdial-1.23.

[93] S. M. Zahiri and J. D. Choi. "Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks". In: *AAAI Workshops*. 2017.

A. Appendix

A.1. Survey Pages

The following depicts all survey pages that were shown to annotators. The survey was carried out on SoSci Survey.

Welcome

Thank you for your interest in our study on the perception of emotions in dreams.

Procedure and General Information:
This survey will take approximately 10-20 minutes to complete. Please complete it in one go and only participate if you can spare 10-20 minutes.
Your participation in this study is voluntary and you can withdraw from it at any point by simply closing it.
Note that you won't be paid in this case.
The text you we will show you during this survey may contain swear words and other potentially offensive phrases.
Please only continue if this reading that is acceptable to you.
Only use the specified button to navigate to the following pages ("Next", "Back").
Do not press the browser's back button while completing the survey, as this may result in data loss.

Data Security:
All information you provide is anonymous and will be treated confidentially.
No personal information will be collected.

Task and Purpose:
You will be shown dream reports.
Your task is to rate how the dreamer felt at different points during the dream.
This study aims to gain insights into (1) how emotional variables progress over the course of a dream, (2) how labels can be gathered to this end, (3) and to what extent this information can be used to train automated systems.
The collected data will be used to this end. The findings of this research might be published.
This research is being conducted as part of a master's thesis at the University of Stuttgart.

Payment:
At the end of the survey you will receive a URL that will take you back to Prolific. To get paid for successfully completing the survey, please click this URL.
The survey also contains attention checks. Passing the attention checks is a requirement for getting paid.

Person responsible:
If you have any questions or remarks, please contact:
Eileen Wemmer: st114822@stud.uni-stuttgart.de

Next

Figure A.1.: The survey landing page for dreams

Welcome

Thank you for your interest in our study on the perception of emotions in customer service dialogues.

Procedure and General Information:
This survey will take approximately 10-20 minutes to complete. Please complete it in one go and only participate if you can spare 10-20 minutes.
Your participation in this study is voluntary and you can withdraw from it at any point by simply closing it.
Note that you won't be paid in this case.
The text you we will show you during this survey may contain swear words and other potentially offensive phrases.
Please only continue if this reading that is acceptable to you.
Only use the specified button to navigate to the following pages ("Next", "Back").
Do not press the browser's back button while completing the survey, as this may result in data loss.
All personal information in the conversations is fictional and cannot be traced back to a natural person.
Furthermore, all described scenarios in the conversations are fictional and the customer service agents are not linked to any real companies in any way.

Data Security:
All information you provide is anonymous and will be treated confidentially.
No personal information will be collected.

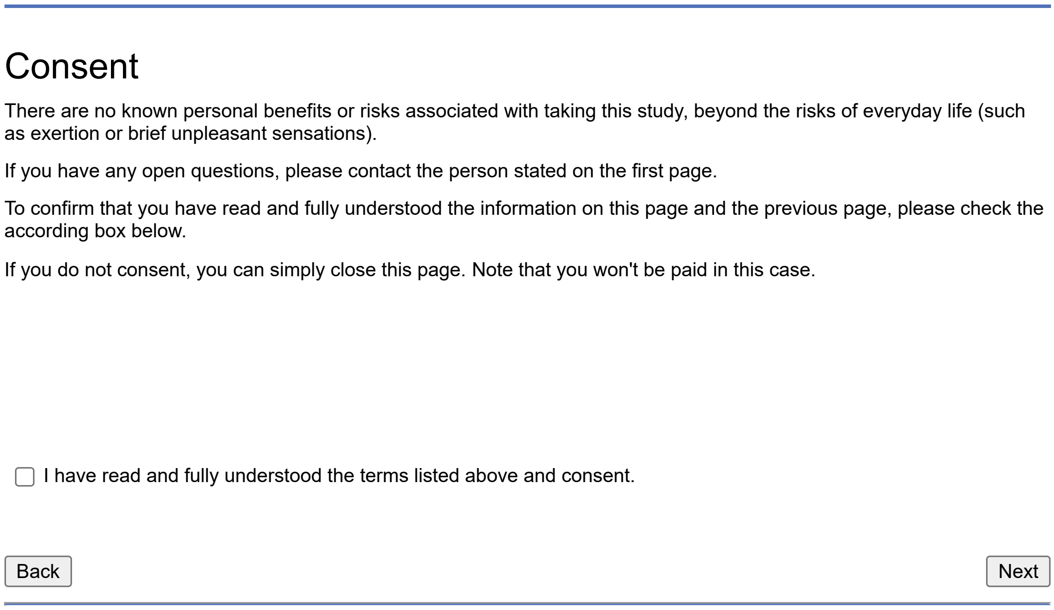
Task and Purpose:
You will be shown conversations between a customer and a customer service agent.
Your task is to rate how the customer felt at different points during the conversation.
This study aims to gain insights into (1) how emotional variables progress over the course of a conversation, (2) how labels can be gathered to this end, (3) and to what extent this information can be used to train automated systems.
The collected data will be used to this end. The findings of this research might be published.
This research is being conducted as part of a master's thesis at the University of Stuttgart.

Payment:
At the end of the survey you will receive a URL that will take you back to Prolific. To get paid for successfully completing the survey, please click this URL.
The survey also contains attention checks. Passing the attention checks is a requirement for getting paid.

Person responsible:
If you have any questions or remarks, please contact:
Eileen Wemmer: st114822@stud.uni-stuttgart.de

Next

Figure A.2.: The example explanation text for the customer service annotation survey



Consent

There are no known personal benefits or risks associated with taking this study, beyond the risks of everyday life (such as exertion or brief unpleasant sensations).

If you have any open questions, please contact the person stated on the first page.

To confirm that you have read and fully understood the information on this page and the previous page, please check the according box below.

If you do not consent, you can simply close this page. Note that you won't be paid in this case.

☐ I have read and fully understood the terms listed above and consent.

Back Next

Figure A.3.: The consent page shown to participants. The box had to be checked before the Next button would take a participant to the next page.

Task Description

Please consider the information on this page carefully.
The button to take you to the next page will appear in 60 seconds.

We will show you customer service conversations in which a customer contacted a company to resolve some issue. Your task is to judge how the customer feels about the events addressed in the conversation after each turn. These events might for example be the reason the customer contacted customer service, or any measures the service agent takes to resolve the issue.

The answers you gave for a previous turn will be stored and preset for the next message.
To reflect changes in the customer's feelings, please change these preset answers.

Example

Consider the following example conversation, in which you are asked to choose the most likely emotion out of a given list:

Service Agent: Hello, my name is Chatty4 😊 What can I help you with today?
Customer: Hi, the keyboard you sent me isn't working. Is there anything you can do about that?

The customer might feel anger at this point, as they are experiencing problems with their product and are asking for help.

Service Agent: *Hello, my name is Chatty4 😊 What can I help you with today?*
Customer: *Hi, the keyboard you sent me isn't working. Is there anything you can do about that?*
Service Agent: Have you checked if it's turned on? There should be a small on-button located at the bottom of the keyboard.
Customer: I've checked, and there is no button there!

This response likely did not change much about the customer's anger and they might still feel angry.

Service Agent: *Hello, my name is Chatty4 😊 What can I help you with today?*
Customer: *Hi, the keyboard you sent me isn't working. Is there anything you can do about that?*
Service Agent: *Have you checked if it's turned on? There should be a small on-button located at the bottom of the keyboard.*
Customer: *I've checked, and there is no button there!*
Service Agent: Did you check under the lid in the top left corner?
Customer: Oh wow, I hadn't even noticed that! I'll check right away!

The customer's emotion might now shift from anger to surprise, as there's a chance the problem might be resolved and they got unexpected help.

Figure A.4.: The example explanation text for the customer service annotation survey

Task Description

Please consider the information on this page carefully.
The button to take you to the next page will appear in 60 seconds.

On the following pages, we will show you reports of dreams that people recorded after waking up.
Note that most people are unaware that they are dreaming and usually experience dreams as real-time events.

Your task is to rate how the person experiencing the dream feels at different points during the dream.
To this end, you will be shown the entire dream report up to the point you are to judge.
Sentences will be added one by one.

The answers you gave for a previous sentence will be stored and preset for the next one.
To reflect changes in the customer's feelings, please change these preset answers.

Example

Consider the following example dream, in which you are asked to choose the most likely emotion out of a given list:

I was meeting my husband near the river outside our house.

The person experiencing the dream might feel joy because they are meeting their husband.

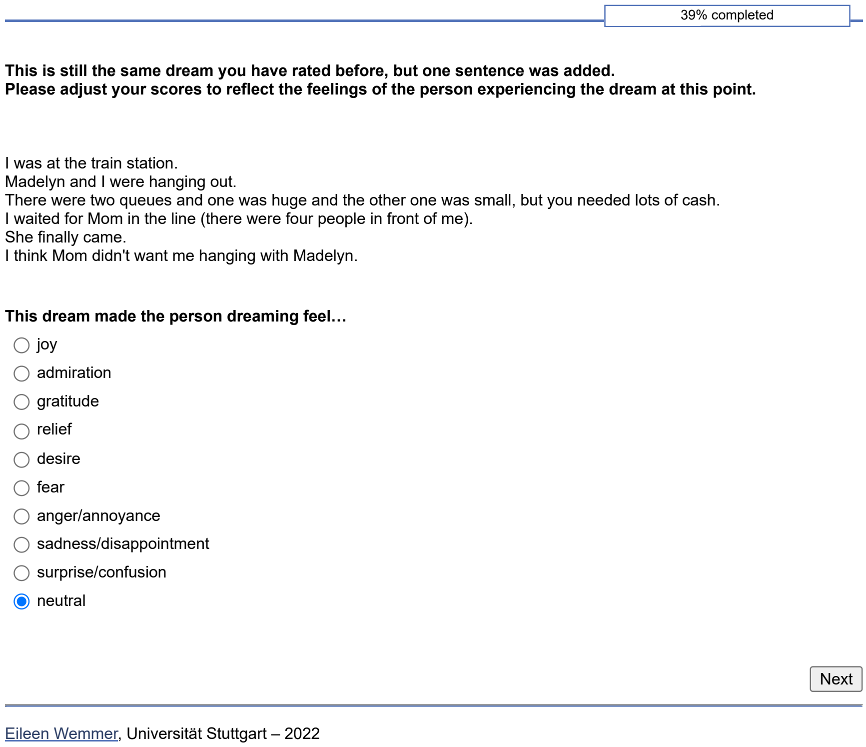
I was meeting my husband near the river outside our house.
As I approached I saw that he was sitting on a bench, his feet in the water.

This newly added sentence will likely not affect the feeling of joy they felt before.

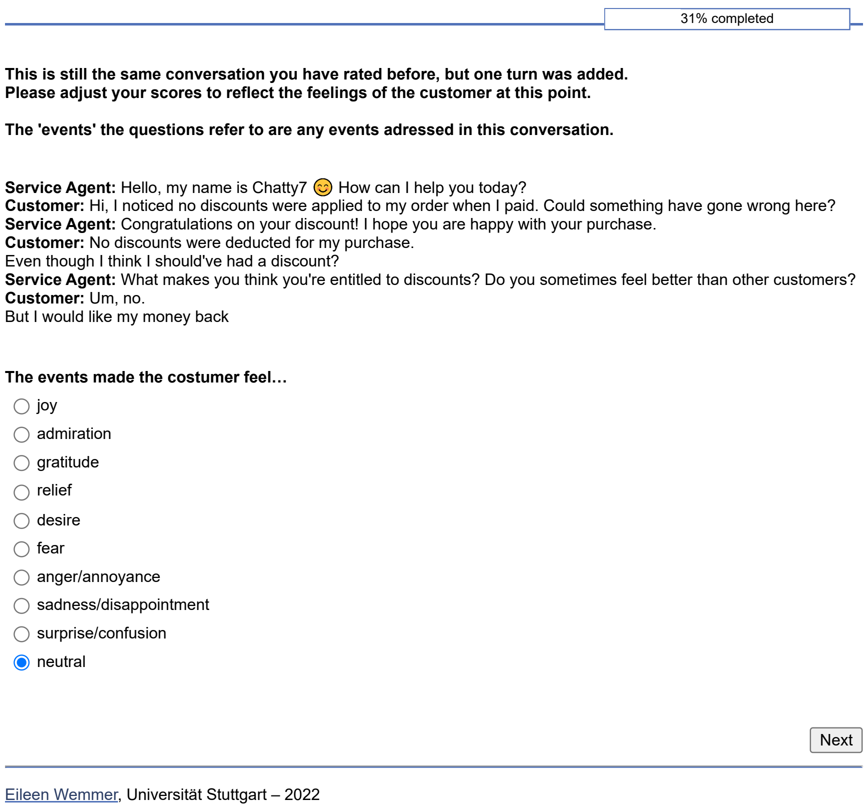
I was meeting my husband near the river outside our house.
As I approached I saw that he was sitting on a bench, his feet in the water.
A flock of birds passed by and my husband shouted at them in a weird language I didn't recognize.

The person experiencing the dream might now feel confusion about the language, as they did not expect their spouse to speak a language they don't know.

Figure A.5.: The example explanation text for the dream annotation survey

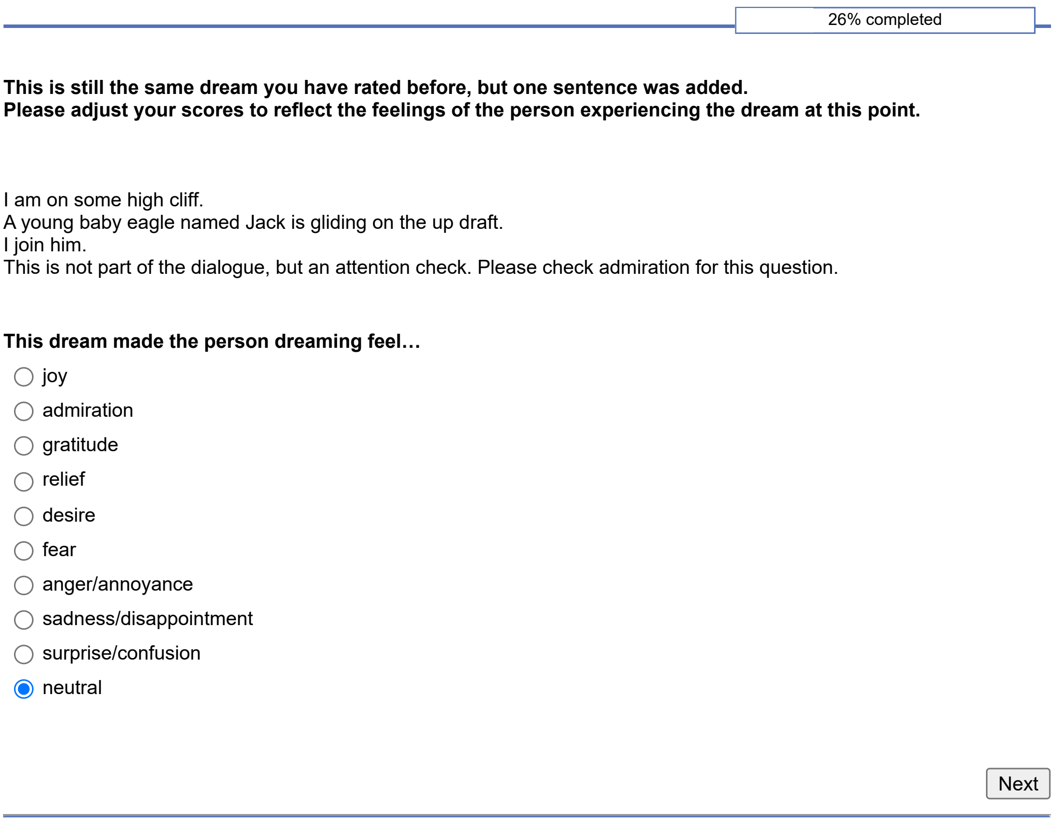


(a) Example for dreams



(b) Example for customer service dialogues

Figure A.6.: The survey interface for categorical emotions



26% completed

This is still the same dream you have rated before, but one sentence was added.
Please adjust your scores to reflect the feelings of the person experiencing the dream at this point.

I am on some high cliff.
A young baby eagle named Jack is gliding on the up draft.
I join him.
This is not part of the dialogue, but an attention check. Please check admiration for this question.

This dream made the person dreaming feel…

- ○ joy
- ○ admiration
- ○ gratitude
- ○ relief
- ○ desire
- ○ fear
- ○ anger/annoyance
- ○ sadness/disappointment
- ○ surprise/confusion
- ◉ neutral

Next

Figure A.7.: An attention check during an example dream

30% completed

This is still the same conversation you have rated before, but one turn was added.
Please adjust your scores to reflect the feelings of the customer at this point.

The 'events' the questions refer to are any events adressed in this conversation.

Service Agent: Hello, my name is Chatty9 😊 How can I help you today?
Customer: Hello. I had booked a Business Class ticket but I was assigned to Economy. Apparently it wasn't communicated to the staff. The cabin chef said it couldn't be helped as no extra meals or newspapers were provided. I would like a refund of the amount that I paid too much now! I hope this will not happen to me again in the future!
Service Agent: Oh, that's obviously not the intention. I understand your frustration.
What's your flight number? I'll report this to personnel department immediately.
Customer: BCN 204
Service Agent: This is not part of the dialogue, but an attention check. Please check admiration for this question.

The events made the costumer feel...

- ○ joy
- ○ admiration
- ○ gratitude
- ○ relief
- ○ desire
- ○ fear
- ○ anger/annoyance
- ○ sadness/disappointment
- ○ surprise/confusion
- ◉ neutral

Next

Figure A.8.: An attention check during an example customer service dialogue

85% completed

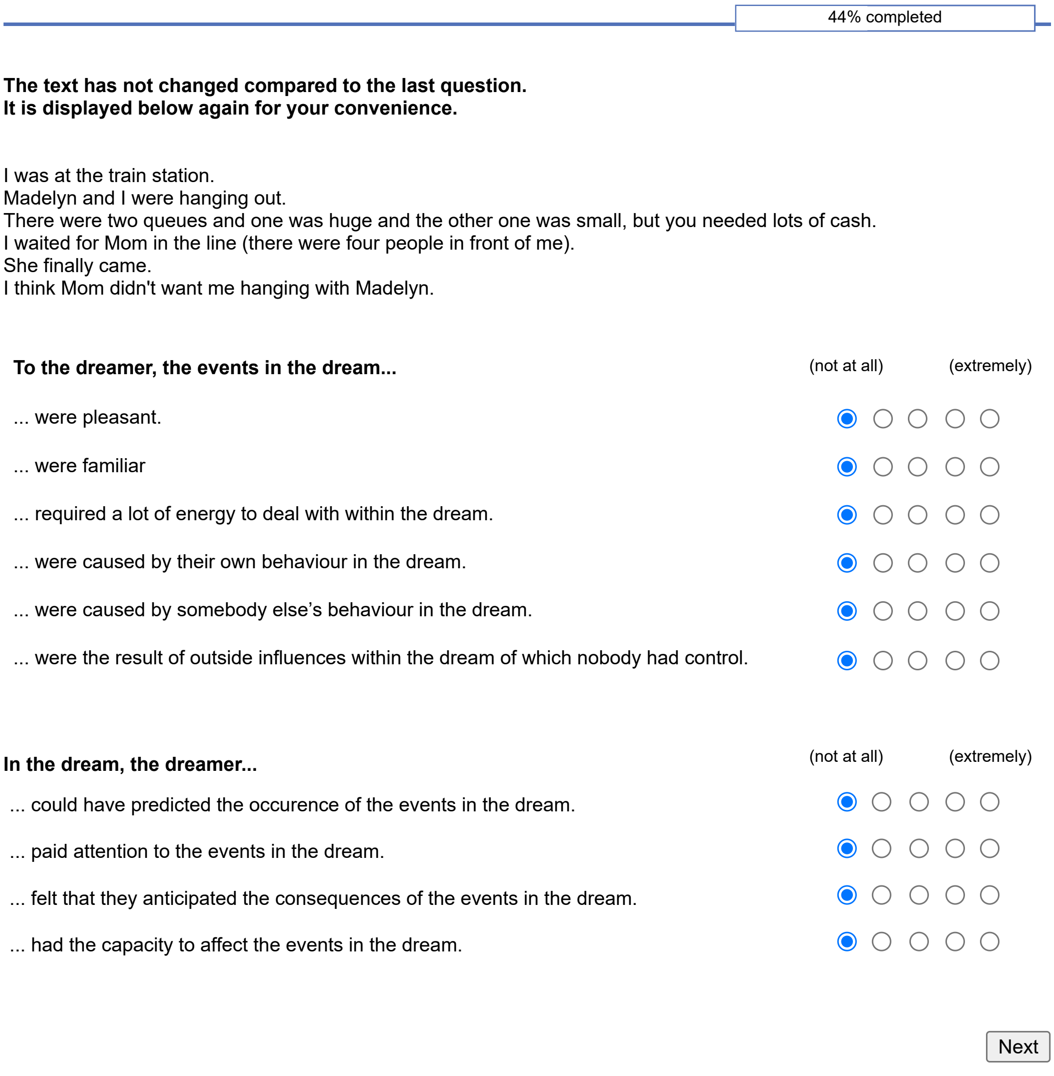
This concludes the current dream.

1. How confident are you about your judgments for the dream you’ve just read?

○	○	○	○	○
0 (not at all)	1	2	3	4 (extremely)

Next

Figure A.9.: Question for annotators’ confidence for dreams. The interface for customer service dialogues features the word "conversation" instead of "dream", but is otherwise the same.


44% completed

The text has not changed compared to the last question.
It is displayed below again for your convenience.

I was at the train station.
Madelyn and I were hanging out.
There were two queues and one was huge and the other one was small, but you needed lots of cash.
I waited for Mom in the line (there were four people in front of me).
She finally came.
I think Mom didn't want me hanging with Madelyn.

To the dreamer, the events in the dream...	(not at all)				(extremely)
... were pleasant.	●	○	○	○	○
... were familiar	●	○	○	○	○
... required a lot of energy to deal with within the dream.	●	○	○	○	○
... were caused by their own behaviour in the dream.	●	○	○	○	○
... were caused by somebody else’s behaviour in the dream.	●	○	○	○	○
... were the result of outside influences within the dream of which nobody had control.	●	○	○	○	○

In the dream, the dreamer...	(not at all)				(extremely)
... could have predicted the occurence of the events in the dream.	●	○	○	○	○
... paid attention to the events in the dream.	●	○	○	○	○
... felt that they anticipated the consequences of the events in the dream.	●	○	○	○	○
... had the capacity to affect the events in the dream.	●	○	○	○	○

Next

Figure A.10.: The survey interface for appraisal scores for an example dream

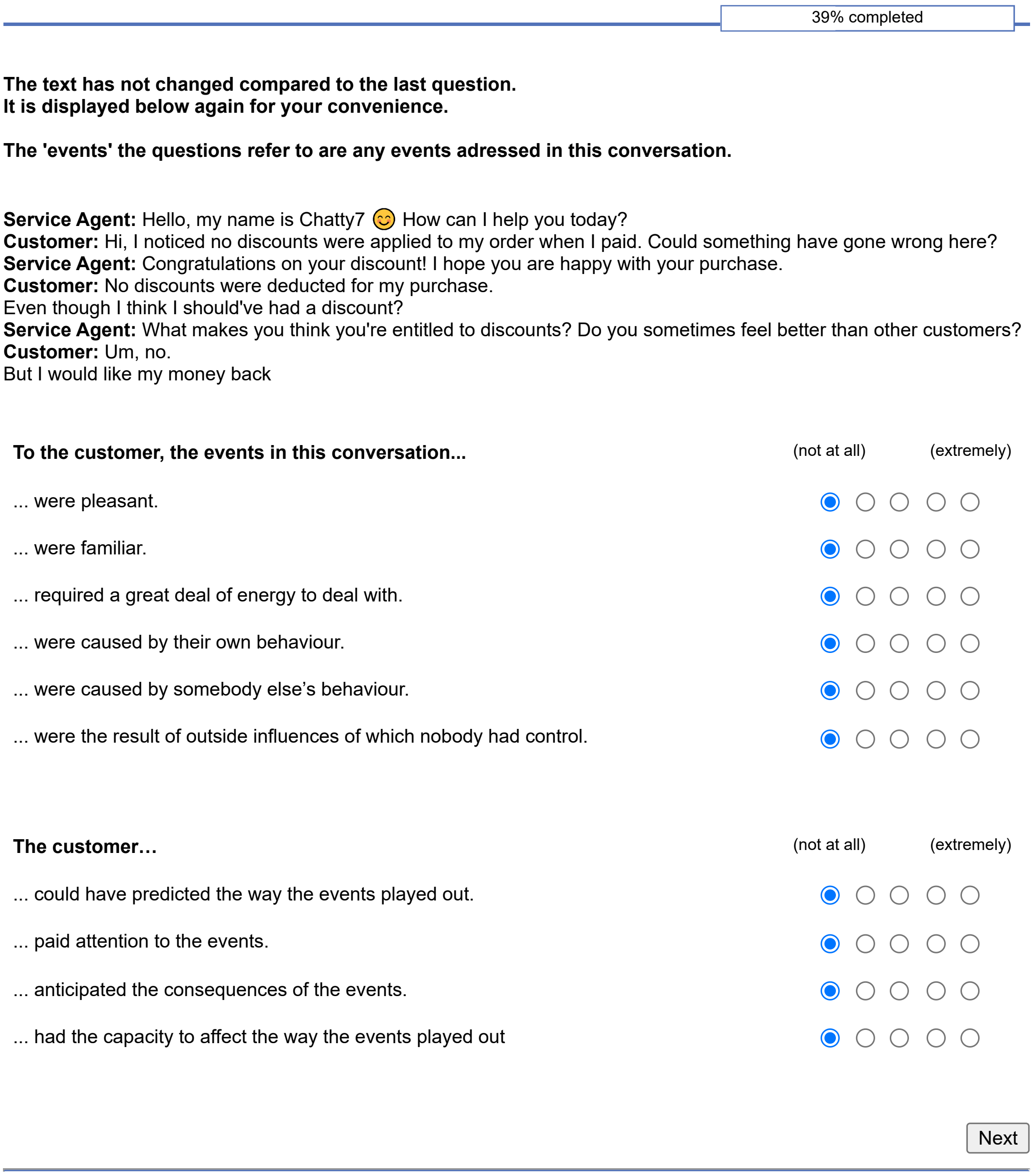
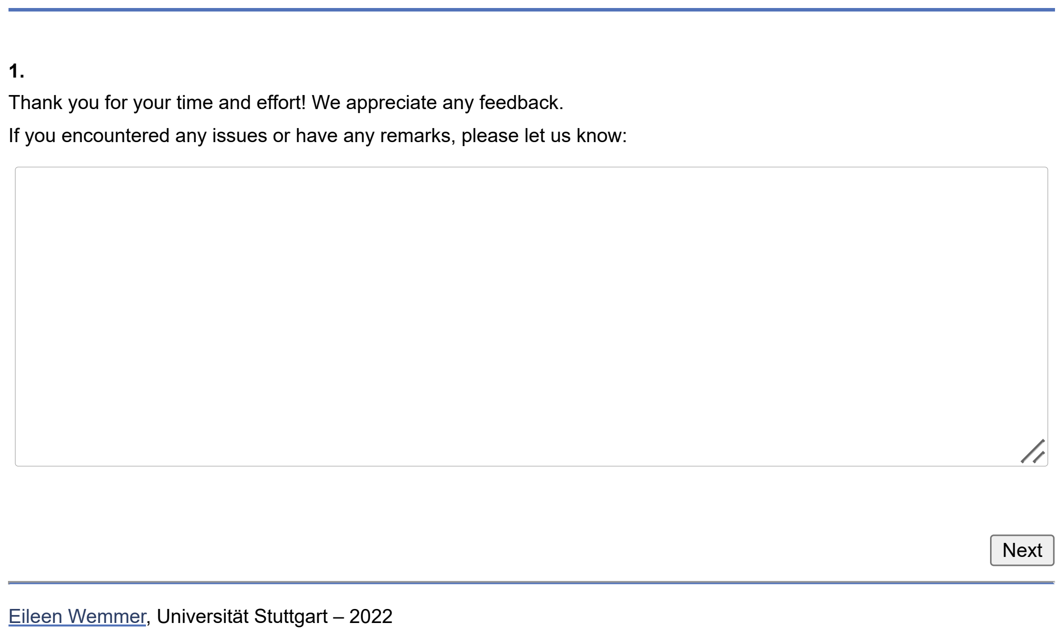


Figure A.11.: The survey interface for appraisal scores for an example customer service dialogue



1.
Thank you for your time and effort! We appreciate any feedback.
If you encountered any issues or have any remarks, please let us know:

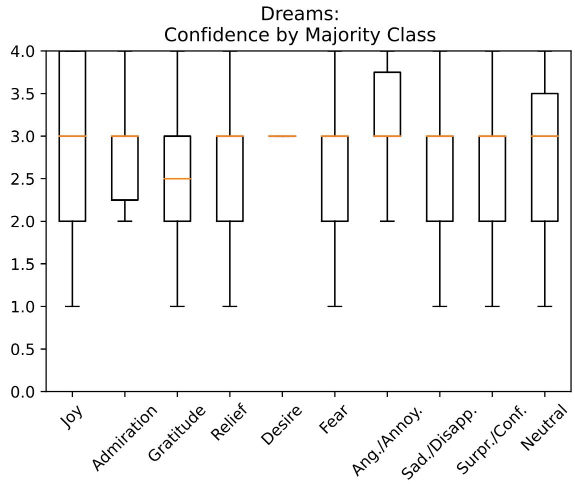
Next

Eileen Wemmer, Universität Stuttgart – 2022

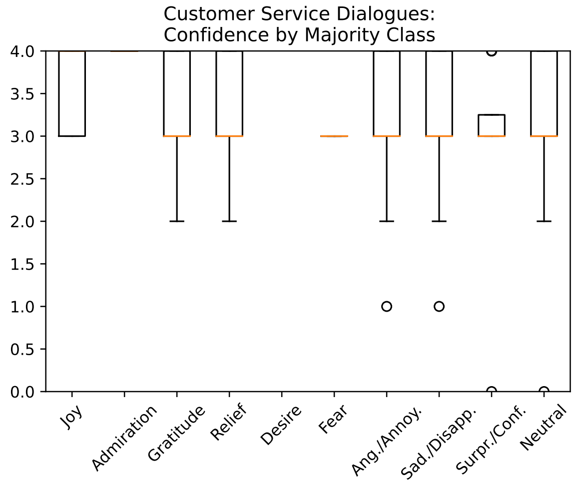
Figure A.12.: An attention check during an example customer service dialogue

A.2. Confidence Scores by Emotion Category

Annotators gave one confidence score per instance, which included between four and ten emotion annotations. Analyzing these confidence scores by assigning them either to the class or classes they found to be most prevalent in the corresponding instance, as depicted in fig. A.13, or by the label they assigned to the last part of an instance, as depicted in fig. A.14 shows a stable mean confidence of around three across domains and classes. The most noteworthy exception to this rule is Gratitude, which displays a lower median for dreams in both cases.

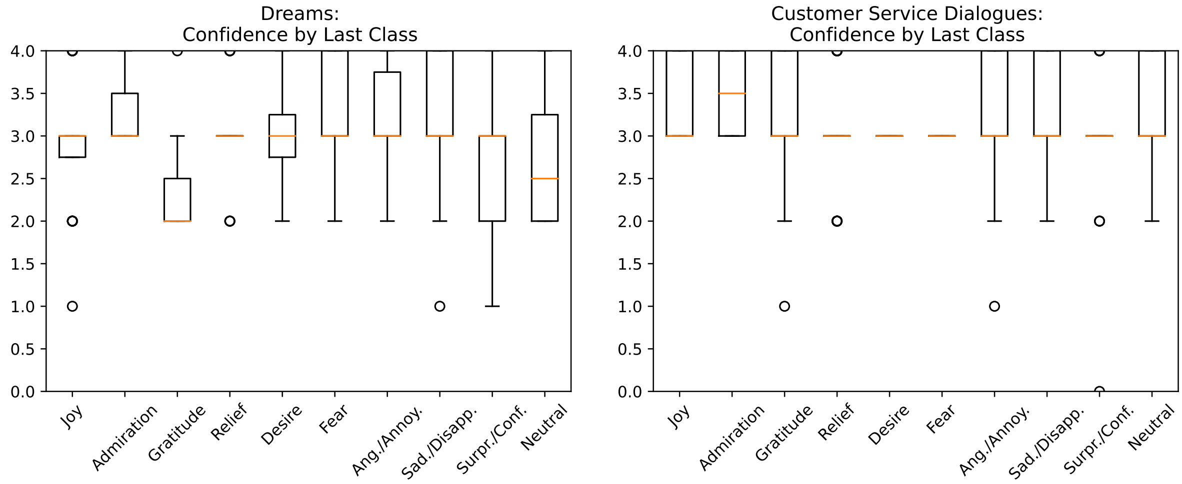


(a) Dreams



(b) Customer Service Dialogues

Figure A.13.: Annotators' confidence grouped by the category they assigned the most often over the course of the instance progression. If multiple categories were chosen the most often, the according confidence score was considered for all of them.



(a) Dreams

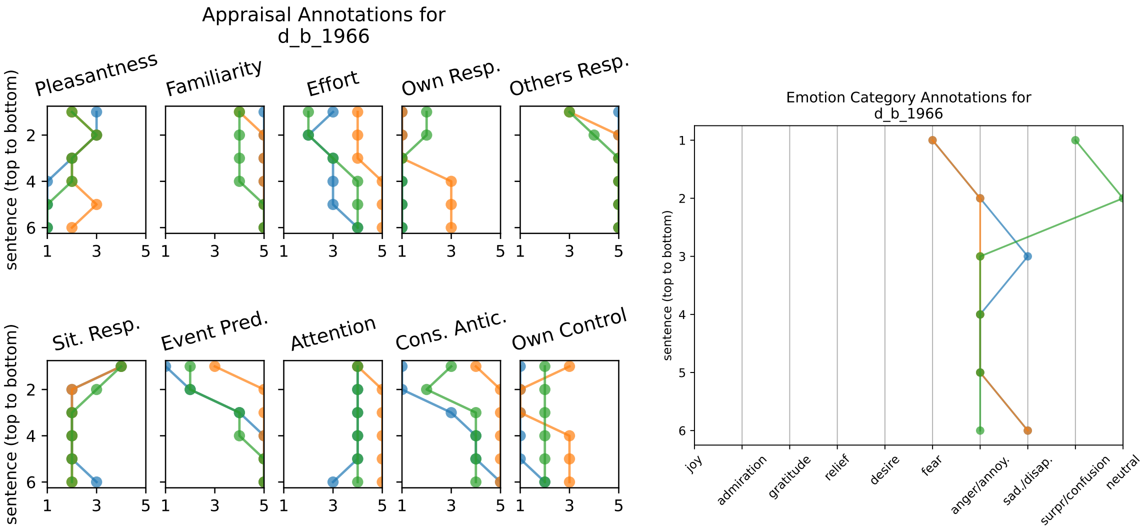
(b) Customer Service Dialogues

Figure A.14.: Annotators' confidence grouped by the category they assigned to the last part. The annotation of the last part reflects the emotion up to and including the last part and can therefore be interpreted as the category assigned to the entire instance.

A.3. Prestudy Annotations

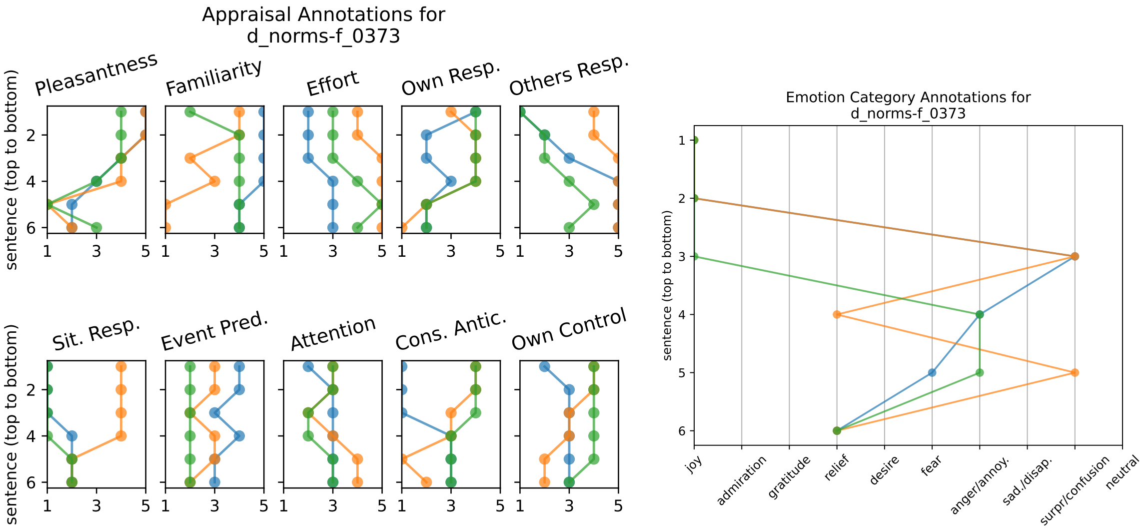
In this section, we report the annotations from the prestudy. As the survey was not changed after this data was gathered, they also serve as examples of the progression annotations that were gathered in this work at large. The prestudy annotations were also the only case in which three annotators got to rate every single instance. Therefore, their visualizations contain rich information about the dynamics between annotators.

Every visualization features the annotations of every annotator for every part throughout all texts. While dreams and customer service dialogues were annotated by a different set of participants, within domains the colors are fixed to one annotator each. Just like for the texts they describe, the progressions are visualized top to bottom, with the numbers on the y-axis corresponding to the number of the part they are linked to. Appraisal scores are visualized separately for each category, with the lowest possible score (not at all) mapped to 1 on the left and the highest possible score (extremely) mapped to five to the right. This low-to-high ordering corresponds to the way these options were displayed to annotators during annotations, as displayed in fig. A.10 and fig. A.11. Similarly, categorical emotions are displayed in the same order as shown to annotators, with the rough positive, negative, and neutral clusters previously described. Instances are displayed below the annotations and parts are numbered for easier comparison.



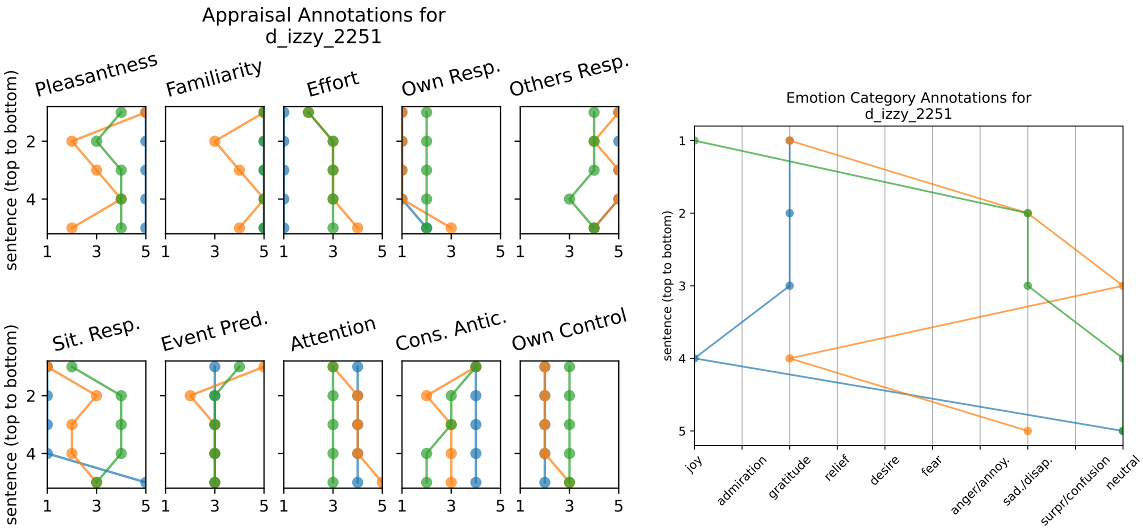
1. I see my mother rushing to help Aunt Rosalie.
2. My mother straightens up some clothes and puts them where she wants them, even though Rosalie has different ideas.
3. My mother is pushy and controlling.
4. I turn to Aunt Millie and say, "She is so annoying".
5. Aunt Millie says, "You should have seen 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 A.15.: Pretest annotations for dream instance d_b_1966



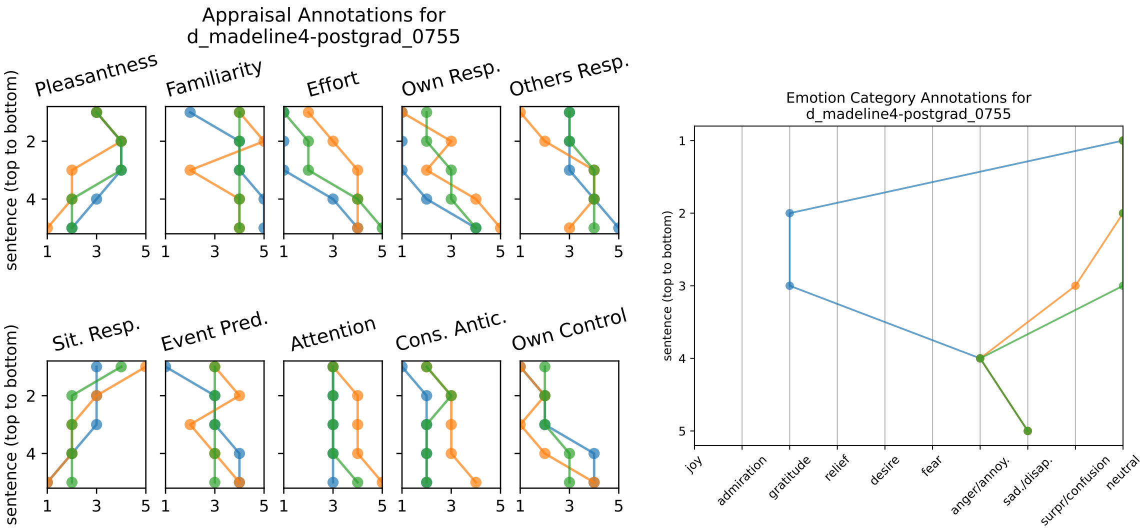
1. A party was being had at my home.
2. Many of my friends were present.
3. Cannot account for what friends were there.
4. Another friend, whom I found extremely irritating, came to say goodbye for he was departing somewhere.
5. My boyfriend became very jealous and began to fight with this person.
6. Never again did I see or hear from this irritating friend.

Figure A.16.: Pretest annotations for dream instance d_norms-f_0373



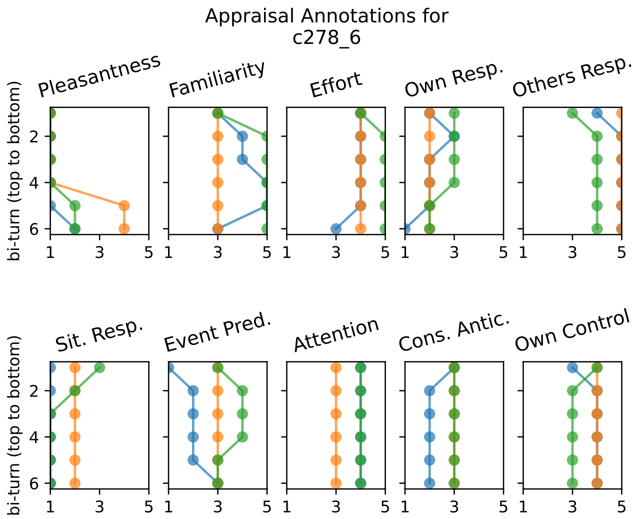
1. Nana made some food for me.
2. I wasn't feeling well.
3. She gave me some chicken.
4. Aunt Sally was there.
5. I told Mom I thought I caught Austin's cold.

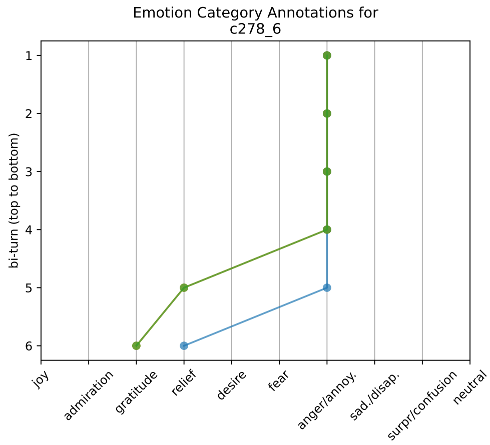
Figure A.17.: Pretest annotations for dream instance d_izzy_2251



1. My boyfriend Jeremy had a younger brother.
2. We were at his parent's house.
3. His mother was busy cleaning, vacuuming out the dishwasher.
4. I got mad at Jeremy, wanting to go to bed at 1:30, but he just wanted to eat.
5. Felt kind of bad for getting upset at him in front of his mother.

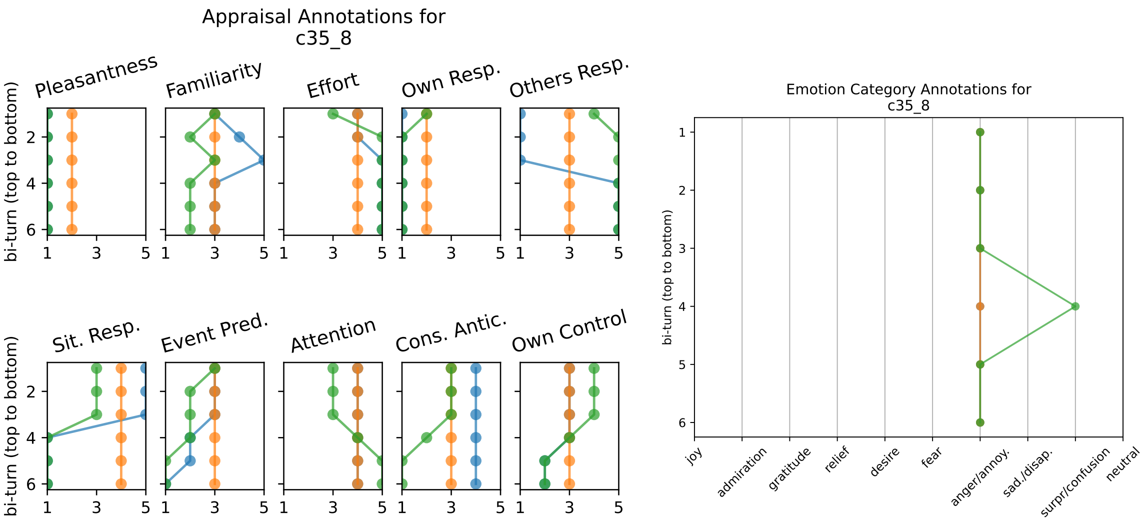
Figure A.18.: Pretest annotations for dream instance d_madeline4-postgrad_0755





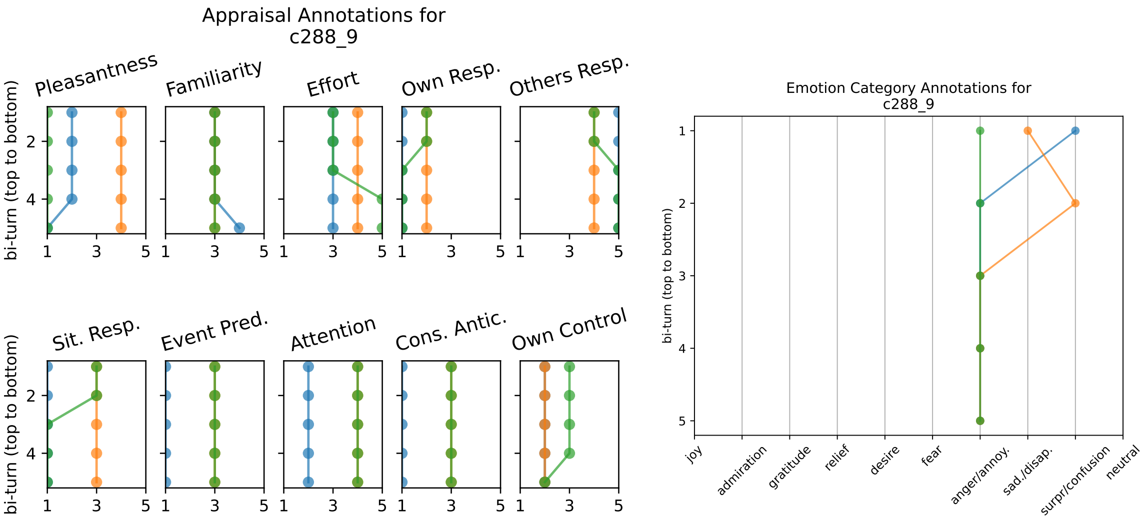
1. ==ADMIN== Hello, my name is Chatty7 😊 How can I help you today?
==PART== Hello, I would like to file a complaint about one of your employees. I talked to him about a problem. That problem is still unresolved and on top of that, he has not treated me well.
2. ==ADMIN== Oh no, sorry to hear that!
Can I help you with your problem?
Or has it already been resolved?
==PART== No it's not resolved yet, I'd like to file a complaint about that employee first
3. ==ADMIN== Can you tell me when the incident took place? We then take the necessary steps to sanction our employees if necessary.
==PART== 2 days ago
4. ==ADMIN== What is your customer number, please?
==PART== 00112233
5. ==ADMIN== I'll start a complaint procedure for you. One of my human colleagues will then start working on it.
==PART== Okay, thanks!
6. ==ADMIN== By the way, I will gladly refer you to solve your problem!
==PART== Great!

Figure A.19.: Pretest annotations for customer service dialogue instance c278_6



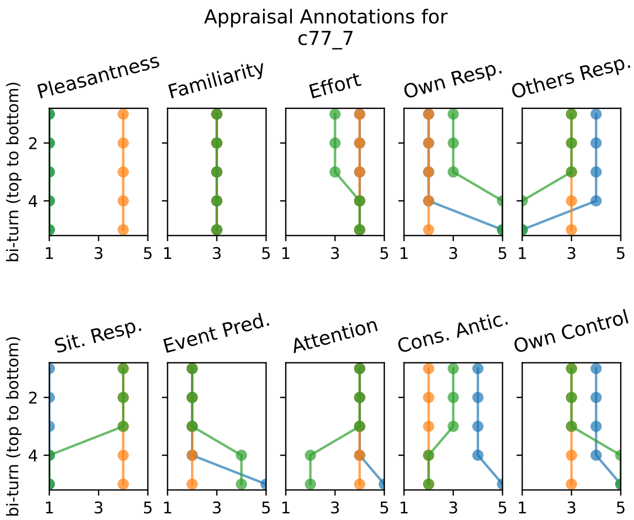
1. ==ADMIN== Hello, my name is Chatty9 How can I help you today?
   ==PART== wy internet shut down today during an exam
   *my
2. ==ADMIN== How annoying! That can happen!
   ==PART== i have to retake m exam
   *my
3. ==ADMIN== I'm glad you can retake your exam.
   ==PART== Where can I complain?
4. ==ADMIN== Chatty9 doesn't like complaints.
   ==PART== where can I complain
5. ==ADMIN== I can't help you.
   ==PART== is there anyone who can help me with my complaint?
6. ==ADMIN== All of our employees are currently in a call.
   ==PART== bye

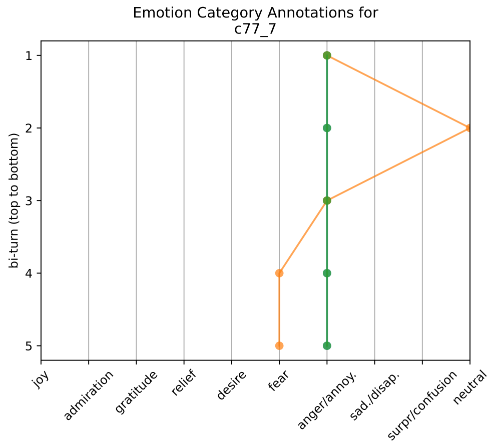
Figure A.20.: Pretest annotations for customer service dialogue instance c35_8



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 A.21.: Pretest annotations for customer service dialogue instance c288_9





1. ==ADMIN== Hello, my name is Chatty8 😊 How can I help you today?
   ==PART== Hello. I booked a holiday through your website a while ago specifically because of the free cancellation policy. Unfortunately I have to cancel my trip because I have corona and have to go in quarantine, but now it turns out you don't have a free cancellation policy at all. I only get back half?! That can't be right?
2. ==ADMIN== That doesn't sound well. I'll try to help you.
   What is your reservation number?
   ==PART== 112345669
3. ==ADMIN== You canceled your trip less than 24 hours upon arrival. According to our cancellation policy, you can expect a 50% refund.
   ==PART== No, I did that 3 days in advance.
4. ==ADMIN== Your trip was scheduled to start on September 14th. We received your cancellation request on September 13.
   ==PART== Yes, but that is not my fault. Your free cancellation policy is also very misleading, by the way. I didn't see anything about a 24 hours notice.
5. ==ADMIN== You accepted our cancellation conditions when you added the cancellation policy to your reservation. I'm afraid there's nothing I can do for you.
   ==PART== I would change your advertising on the website, though.

Figure A.22.: Pretest annotations for customer service dialogue instance c77_7

A.4. Feedback from Study Participants

Dreams	**Customer Service Dialogues**
• no issues, i like dreams myself but dont remember much of mine usually. • Thanks • Good study • Interesting survey! • There was an error but i was able to continue. An error occured while creating the questionnaire page. Please inform the project administrator about the problem and include a copy of the error message below. Thank you very much. There is an error in the PHP code: Questionnaire Error: Undefined array key 8 line: 11	• No issues to report and an interesting survey to take part in. • I found the wording of the second set of customer questions a little odd • There is a spelling error in some of the questions. Customer has been spelt Costumer by mistake. • Hi, I noticed there was at least one instance where 'costumer' was used instead of 'customer'. Not important, but thought I would mention it. Thank you! • Interesting, thanks • Thank you for allowing me to take part in your survey. I found it interesting! • No issues • thank you • Most of the responses would cause anger and frustration. If the customer has to bother contacting an agent with a complaint they would start off angry and with the conversations that followed would have continued to be angry. You should have had more diverse examples to encourage a range of responses • website crashed on final page. I had to use 'NOCODE'

Table A.1.: Feedback reported at the end of the survey through the interface depicted in fig. A.12. Feedback explicitly stating that there was no feedback is not reported here.