x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations

Enrica Troiano*, **Laura Oberländer***, **Maximilian Wegge**, **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
{enrica.troiano,laura.oberlaender,maximilian.wegge,roman.klinger}@ims.uni-stuttgart.de

Abstract

Emotion classification is often formulated as the task to categorize texts into a predefined set of emotion classes. So far, this task has been the recognition of the emotion of writers and readers, as well as that of entities mentioned in the text. We argue that a classification setup for emotion analysis should be performed in an integrated manner, including the different semantic roles that participate in an emotion episode. Based on appraisal theories in psychology, which treat emotions as reactions to events, we compile an English corpus of written event descriptions. The descriptions depict emotion-eliciting circumstances, and they contain mentions of people who responded emotionally. We annotate all experiencers, including the original author, with the emotions they likely felt. In addition, we link them to the event they found salient (which can be different for different experiencers in a text) by annotating event properties, or *appraisals* (e.g., the perceived event undesirability, the uncertainty of its outcome). Our analysis reveals patterns in the co-occurrence of people's emotions in interaction. Hence, this richly-annotated resource provides useful data to study emotions and event evaluations from the perspective of different roles, and it enables the development of experiencer-specific emotion and appraisal classification systems.

Keywords: emotion analysis, corpus, affective computing, role labeling, emotion experiencer, appraisal theories, events

1. Introduction

Computational emotion analysis from text includes various subtasks, with the (arguably) most popular one being emotion classification – i.e., to categorize texts into a predefined set of emotion classes (Mohammad et al., 2018). The adopted categories often coincide with the list of "basic emotions" proposed by Ekman (1992), namely fear, joy, sadness, anger, disgust, surprise, or with inventories defined by other theorists, like Plutchik (2001). In addition to discrete labels, some works have been identifying structured information, namely emotion roles, which divides texts into spans that correspond to semantic roles. Depending on the considered domain, it can aim at distinguishing an emotion experiencer, a stimulus (i.e., the triggering event), a target towards which the emotion is directed, or a cue word evoking a specific affective state (Mohammad et al., 2014).

Structured knowledge of this sort has been proven informative for emotion classification (Oberländer et al., 2020), but to face one challenge at a time, the two strands of research are typically addressed separately. Hence, when the emotion experiencer is not known to a classification model, a preliminary decision has to be made: the classification should regard either the emotion expressed by the writer (Mohammad, 2012, i.a.), or the one triggered in the reader of the text (Haider et al., 2020, i.a.). Studies that include both perspectives are rare (Bostan et al., 2020; Buechel and Hahn, 2017b, i.a.), and so are those focused on the emotions of characters mentioned in the text (Kim and Klinger, 2019; Kim and Klinger, 2018). In fact, they are mainly dedicated to the analysis of emotion in literature – a task that comes with its own challenges, due to the artistic nature of the domain. Our paper sets a different focus: to account for a more complete understanding of an utterance's affective content, we are interested in the perspective of diverse emotion experiencers, both the writer of an event description and other in-text entities.

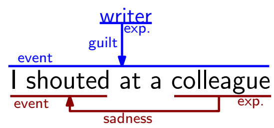

Figure 1: Stimulus and emotion annotation specific to experiencers.

We compile x-enVENT,[1] a corpus predominantly made of self-reported event descriptions. The texts were written by people who felt a particular emotion in response to such events, many of which involved third parties (Troiano et al., 2019). Hence, (1) we mark the textual spans corresponding to the emotion-triggering events and all of their experiencers (the writer being a special token), (2) we draw relations between the two, and (3) we specify what emotion results from their interplay. An example is shown in Figure 1, where the writer feels guilty for their own behaviour, while the person affected by it more likely feels sadness.

These ⟨event, experiencer, emotion⟩ tuples reflect the model of *appraisal* (Scherer, 1989), according to which emotions arise in response to the cognitive evaluation of events. This psychological theory places emotions in a multidimensional space: each dimension represents

*The first two authors contributed equally.

[1] We call the corpus x-enVENT, to indicate that it is English and annotated by trained eXperts, in contrast to a crowd-sourced resource under development (Crowd-enVENT).

a specific property of the event being evaluated, for instance its perceived *pleasantness* (likely to be high with joy and low with disgust), the mental or physical *effort* that it can be expected to cause in the experiencer (likely high with episodes of anger or fear), the *responsibility* held by the experiencer for what has happened (high for guilt) (Smith and Ellsworth, 1985). The annotation unit that we present in this work therefore consists of $\langle \text{ev}, \text{exp}, \text{emo}, \overrightarrow{\text{appraisal}} \rangle$ tuples, in which the event (ev) and the experiencer (exp) are token spans, the emotion (emo) is a single emotion label, and $\overrightarrow{\text{appraisal}}$ is a vector of numeric values for various appraisal dimensions.

Hence, we bring together work on appraisal theories for text analysis (Balahur et al., 2011; Hofmann et al., 2020; Hofmann et al., 2021), emotion role labeling (Oberländer et al., 2020; Mohammad et al., 2014; Kim and Klinger, 2018), and emotion classification. We combine the annotation layers, as exemplified in Figure 1, an example where the dimension of *responsibility* is scored with the highest degree for the writer and lowest for "a colleague", while *pleasantness* is low for both. Our corpus encompasses 720 event descriptions[2] and enables the development of experiencer-specific emotion and appraisal analysis systems. It further enables analyses of the interplay of emotions of people in interaction, as it emerges from text.

2. Previous Work

Resources for Emotion Recognition. The construction of emotion resources typically relies on psychological models of emotions. Following theories of basic emotions (Ekman, 1992; Plutchik, 2001), texts can be labelled with categorical classes. Alternatively, emotions can be described via continuous values in a vector space. Such is the basis of studies like Preoţiuc-Pietro et al. (2016),Yu et al. (2016) and Buechel and Hahn (2017a), which comprise the dimensions of valence, arousal and dominance motivated by Russell and Mehrabian (1977).

Usually, emotion classification and resource construction associate a text to one or more emotions from a specific perspective. Indeed, emotions arise in language whenever writers mention or evoke a mental state of their own or that of others (e.g., a character), as well as when they attempt to elicit a reaction in their readers. This has motivated the design of corpora with texts from various domains, like Reddit comments (Demszky et al., 2020), tales (Alm et al., 2005), blogposts (Aman and Szpakowicz, 2007), and labelled with the emotion of writers (Mohammad, 2012), of readers (Chang et al., 2015), or of both (Buechel and Hahn, 2017b).

As opposed to these studies, we annotate event descriptions from the perspective of each experiencer mentioned or presupposed (i.e., the writer) in the text.

Structured Emotion and Sentiment Analysis. Our setup is close to previous work in structured sentiment analysis. There, opinion holders are extracted (Toprak et al., 2010; Kessler et al., 2010), along with "aspects" and sentiment polarity values revealing the relation between aspect and holder. However, the linguistic variability of descriptions of emotion-inducing events is comparably richer than sentiment opinion expressions (Klinger and Cimiano, 2014): not only the experiencer needs to be situated in a given circumstance, but the link between such circumstance and the consequent emotion is to be grasped via world knowledge (e.g., that shouting at somebody, like in Figure 1, might be inappropriate).

Structured emotion analysis, on its part, has aimed at identifying segments of texts that mention emotion experiencers or stimuli (Wei et al., 2020; Neviarouskaya and Aono, 2013). Accordingly, the available resources contain labels at the sub-sentence level. Gao et al. (2017), for instance, built a corpus which marks emotion cause segments; Ghazi et al. (2015) did the same by leveraging emotion frames in FrameNet (Fillmore and others, 1976) that include a *stimulus* argument. Oberländer and Klinger (2020) compared clause-level and token-level stimulus detection.

Similar to corpora on emotion stimulus detection (Russo et al., 2011; Gui et al., 2016; Li and Xu, 2014; Xia and Ding, 2019; Chen et al., 2020; Kim and Klinger, 2018; Bostan et al., 2020; Mohammad et al., 2014), we consider emotion causes, or stimulus events, but we extend our definition of experiencers to both writers and third entities. We point out the ways in which they appraise events and the resulting emotion reaction, which is reconstructed from (but not felt by) the annotators.

Events and Appraisals. Appraisal annotations have enlivened some research efforts so far. Hofmann et al. (2020) exploited the idea that appraisals enable readers to interpret what others feel. They set up an annotation task in which event descriptions coming from enISEAR (Troiano et al., 2019), already labelled with the emotions of their writers, were associated to seven appraisal dimensions. Using the same dimensions and corpus, Hofmann et al. (2021) experimented with different annotation strategies. Compared to our work, they have disregarded the multitude of emotion perspectives available in text and only considered a limited number of appraisal dimensions.

Other than corpora, the knowledge base EmotiNet was motivated by appraisal theories (Balahur et al., 2011). It describes events with respect to their atomic elements, such as actors, actions and objects, as well as their properties, defined along the lines of appraisal criteria. Further, Cambria et al. (2020) presented a logical representation of events inspired by appraisal theories, but performed sentiment analysis, and Shaikh et al. (2009) used logical expressions to combine event properties with the goal to infer an emotion category.

3. Appraisals

Appraisal theories approach emotions as componential processes, that is, as "an episode of interrelated, syn-

[2] Available at `http://www.ims.uni-stuttgart.de/data/appraisalemotion`

Variable	Description	Values
Event	the most salient fact for the evaluation of an emotion experience	span
Experiencer(s)	the person(s) involved in the situation, and aware of it	span
Emotions	discrete names representing responses to events	disgust, joy, guilt, hope, sadness, surprise, shame, trust
Appr. Dimensions:		
- *suddenness*	the event was sudden or abrupt	1–5
- *familiarity*	the event was familiar to the experiencer	1–5
- *pleasantness*	the event was pleasant for the experiencer	1–5
- *understand*	the experiencer understood what was happening	1–5
- *goal relevance*	the event was important or relevant for experiencer's goals	1–5
- *self responsibility*	the event was caused by experiencer's own behaviour	1–5
- *other responsibility*	the event was caused by somebody else's behaviour	1–5
- *situational respons.*	the event was caused by chance or special circumstances	1–5
- *effort*	the situation required the experiencer a great deal of energy	1–5
- *exert*	the experiencer felt they needed to exert themselves to handle the event	1–5
- *attend*	the experiencer had to pay attention to the situation	1–5
- *consider*	the experiencer wanted to consider the situation	1–5
- *outcome probability*	the experiencer could anticipate the consequences of the event	1–5
- *expect. discrepancy*	the experiencer did not expect that the event would occur	1–5
- *goal conduciveness*	the event itself was positive or it had positive consequences for the experiencer	1–5
- *urgency*	the event required an immediate response from the experiencer	1–5
- *self control*	the experiencer had the capacity to affect the event	1–5
- *other control*	someone or something other than the experiencer was influencing what was going on	1–5
- *situational control*	the situation was the result of outside influences of which nobody had control	1–5
- *adjustment check*	the experiencer anticipated that they could live with the consequences of the event	1–5
- *internal check*	the event clashed with the experiencer's ideals and standards	1–5
- *external check*	the event violated laws or social norms	1–5

Table 1: The variables in our annotation task: event, experiencers, emotions, and 22 appraisal dimensions.

chronized changes in the states of all or most of the five organismic subsystems in response to the *evaluation* of a [...] stimulus-event" (Scherer, 2005). The five subsystems are cognitive (the appraisal), neurophysiological (bodily symptoms), motivational (action tendencies) and motor (facial and vocal expressions), as well as related to subjective feelings (the perceived emotional experience). All of them have corresponding linguistic realizations that evoke an emotion – e.g., bodily symptoms → "he couldn't stop shaking" → fear (Casel et al., 2021), but the cognitive component plays an additional role for emotion analysis, as it enables emotion decoding. Humans' empathy and ability to take the affective perspective of others is guided by the assessment of whether a certain event might have been important, threatening, or convenient for those who lived through it (Omdahl, 1995).

The change in appraisal hence consists in evaluating the situation with respect to the significance it holds for an individual: does the current event hamper my goals? can I predict what will happen next? do I care about it? While the criteria that are used to assess a situation can in principle be countless, appraisal theorists have come up with a number of criteria contributing to the development of an emotional episode (Ellsworth and Smith, 1988; Roseman et al., 1990; Tracy and Robins, 2006, among others). Most of them comprise the appraisal of *goal relevance*. Others include *pleasantness* and *novelty*. In our study, we use 22 appraisal dimensions (see Table 1), based on Smith and Ellsworth (1985), Scherer and Wallbott (1997) and Scherer and Fontaine (2013). These studies have developed a set of questions (e.g., "did you think that the event was pleasant?") in order to collect self-reports on appraisals. Yanchus (2006) raised concerns that this wording might bias the respondents: questions give people the chance to develop a theory in retrospect about their behaviour; instead, statements leave participants free to recall if the depicted behaviours (e.g., "I thought the event was pleasant"), applied to them or not. We abide by this idea and spell out each appraisal as an affirmation.

Appraisals hence reveal one's interaction with the environment. Two people with different goals, cultures and sets of beliefs might produce different evaluations of a stimulus. Consequently, specific appraisal combinations lead to different emotion reactions. Smith and Ellsworth (1985), for instance, qualify 15 emotions on the basis of *pleasantness*, *responsibility* of the emotion experiencer for triggering the event, *certainty* about what was happening, *attention* put on the emotion stimulus, *effort* expended to deal with it, and *situational control*, or the ability to influence the development of the situation. This can account for differences in how people respond to an event, as well as differences in the emotion inferred from (and chosen for) a text by annotators.

4. Corpus Creation

Our goal is to populate a corpus with appraisal dimension ratings for each experiencer mentioned in the text, quantifying the degree to which each property applies to

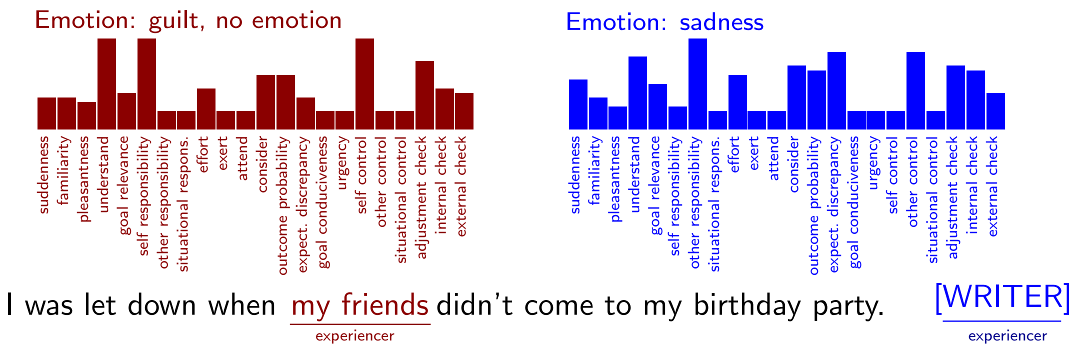

Figure 2: Annotated example. The event that is linked to both experiencers corresponds to the whole text here.

a given situation. Tracing back the original text authors to obtain first-hand appraisal self-recollections would be unfeasible. For this reason, we relied on the help of external readers. Their task consisted in assigning emotion labels, appraisal dimension ratings, and span annotations indicating experiencers and events in each text. Specifically, we asked our annotators to put themselves in the shoes of each experiencer to reconstruct both the emotion they felt and the way in which they might have appraised the event.

We conducted the annotation on the platform INCEPTION[3] (Klie et al., 2018). Annotators were four master students of computer science and computational linguistics, three male and one female, aged between 24 and 28. They were familiar with the field of emotion analysis and with appraisal theories, but the guidelines for the task still provided them with extensive examples and definitions for each concept to be annotated.

4.1. Data

The instances in our resource were sampled from various corpora that contain event descriptions and emotion annotations – the latter mostly provided by the authors of the texts themselves.

During the training phase of the annotators, we extracted data from the ISEAR corpus produced by psychologists (Scherer and Wallbott, 1994; Scherer and Wallbott, 1997), from its kin enISEAR (Troiano et al., 2019), EMPATHETIC-DIALOGUES (Rashkin et al., 2019) and Event2Mind (Rashkin et al., 2018). Later, we extracted texts only from enISEAR, which spans 1001 English sentences describing real-life events associated to 7 emotions.[4] Emotion names were manually masked by the authors of the corpus, such that follow-up emotion interpretation tasks based on those instances would not result in a trivial endeavour. Therefore, the texts coming from enISEAR are implicit emotion expressions, in which the affective meaning of texts is evoked (e.g., "I felt ... when my grandad passed away", "I felt ... when I first flew on a plane"), rather than spelled out.

[3] https://inception-project.github.io

[4] Our published corpus predominantly consists of data from enISEAR.

4.2. Annotation Guidelines

Annotators were presented with one instance at a time. The first step they had to accomplish was to assess whether the text contained an event. If they spotted one, they performed the following:

1. experiencer span identification, aimed at marking the textual span that contains the experiencer mention;

2. salient events span identification, i.e., marking the portion of text containing the event appraised by such experiencer[5];

3. emotion selection, to choose the reaction that the experiencer most likely had to the appraised event;

4. appraisal dimension rating, which consisted in scoring the value of each appraisal dimension with respect to the event.

Annotators repeated these steps for each event and experiencer. An example of the resulting annotation is provided in Figure 2.

4.2.1. Experiencer Span Identification

A text might mention different entities, but not all of them should automatically be deemed experiencers. The guidelines characterized an experiencer as the person who is involved in the situation, is aware of it, evaluates it, and is somewhat affected by what happened. In "Helen didn't notice that Julia lost her keys", only "Julia" would be the experiencer, while both would be annotated in "Julia lost Helen's keys, but Helen wasn't bothered and kept focusing on her homework", in which they likely react in different ways (e.g., Julia → guilt, Helen→no emotion). This example shows that (1) experiencers do not necessarily feel an emotion, (2) experiencers can include multiple entities, (3) each of them can be linked to different events (e.g., losing the keys, focusing on homework). Considering if experiencers are aware of the event is key to determine if they appraised the event but felt no emotion or were present in an event but did not assess it – irrelevant in our setup.

Experiencers could be separately marked in the text (i.e., "Julia", "Helen" individually), or be considered as a unique entity if the two were associated to the same emotion, elicited by the same event (e.g., "my friend and Helen passed the exam"). Annotators were instructed to

[5] Experiencer and event spans can overlap.

select the [WRITER] token, appended at the end of each instance, if they judged the text's author to be an experiencer. In "my daughter was building a snowman", "my daughter" would be annotated along with [WRITER], whose involvement can be recognized by the possessive "my", while in "Mary was building a snowman" only "Mary" would be taken as an experiencer.

4.2.2. Salient Events Span Identification

We gave a loose definition of "event", qualifying it as the occasion or the happening that is the most salient for the evaluation (appraisal) of the experience. For instance, despite sharing much lexical material, the sentences "I attended the funeral of my grandma" and "They started to yell at the funeral of my grandma" can be said to contain different salient events that took the focus of their experiencer.

We invited the annotators to include the arguments of predicates when marking an event (e.g., with transitive verbs, the event should include the object). Moreover, if a sentence contained different events and one experiencer (e.g., "I just bough a brand new car. I let my brother drive it even though he isn't a good driver"), only the focal event of the evaluation, which eventually elicited the main emotion, was annotated.

4.2.3. Emotion Selection

For an experiencer and an event, one label could be chosen among: *anger*, *disgust*, *fear*, *guilt*, *hope*, *joy*, *sadness*, *shame*, *surprise*, *trust*, *disappointment*, *frustration*, *anticipation*, *contentment*, or *pride*. Annotators could indicate that the inferred emotion did not fit any using the option "other", or signal the absence of an emotion reaction by picking the label "no emotion".[6]

Note that the data under consideration came with a prior emotion distribution and prevalent categories. We chose a richer set of emotions than in the enISEAR data, because we did not want to limit the possible emotions for other mentions of experiencers than the writer.

4.2.4. Appraisal Dimension Rating

Annotators aimed at reconstructing how events were assessed by experiencers relative to the 22 dimensions in Table 1. The rating was done on a 1-to-5 scale: the score given to a dimension represents the degree to which (according to the annotators) the event experiencer would agree with the statement describing the appraisal.

4.3. Data Aggregation

As each instance was labeled by four annotators, we aggregate their decisions into one final adjudicated annotation. The gold span-level annotations (experiencers and events) consist in the overlap between the majority of annotators' decisions, that is, in the shortest span appearing in all annotations. For instance, with individual annotations being "my friend", "my friend", "my friend and I", and "friend", the aggregated annotation would be "friend". In case the annotators' decisions differ considerably, or there is no token-level majority vote, we make use of a combination of automatic and heuristics-based manual checks, and aim at a high-recall approach including all entities involved in the emotion episode. Using the above example: if two annotators marked the whole phrase "my friend and I" as a single experiencer, another did not found any event experiencer, and the last only chose the token [WRITER], we would manually align the personal pronoun "I" to [WRITER], and propagate to it the emotion and appraisal annotations. Hence, we would end up with two experiencers: "my friend" and [WRITER]. 184 instances required this manual intervention.

Once the experiencers are defined, we aggregate the emotion annotations and appraisals. For the former, we include the disjunction of all emotion labels by all annotators who labeled such overlapping entities (e.g., all emotions associated to "my friend", "my friend and I", and "friend"). For the latter, we aggregate the ratings to a minimum, maximum, and average score. Note that in the published corpus, we provide the original individual annotations in addition.

[6]Events can be appraised without leading to an emotion.

Variable	Exact-F_1	Relaxed-F_1	Cohen's κ
Experiencer	0.86	0.88	0.84
Event	0.34	0.86	0.75
Emotion	–	–	0.62

Table 2: Inter-Annotator agreement for span annotations and emotion category. Scores for individual emotions together with frequencies are in Table 2.

5. x-enVENT Analysis

5.1. Inter-Annotator Agreement

Inter-annotator agreement results are in Table 2. We show three measures. In Exact-F_1, we take one annotator as the gold standard and the other as a prediction. An annotation span is accepted as a true positive if the whole span is exactly matching the other annotator's span. In the variant Relaxed-F_1, we accept a true positive if there is at least a one-token overlap between the two annotator's spans. We report averages across all annotator pairs. Cohen's κ (Cohen, 1960) refers to the average of a token-level pairwise assessments.

The F_1 measures show that the four annotators reached a high level of agreement for span annotations. As indicated by the Exact-F_1, their intuition were more consistent when labelling experiencers than events (F_1=.86 and .34, respectively). In fact, when considering the Relaxed variant, agreement relative to experiencers does not increase much (.88), while that concerning events is more than doubled (.86). This shows that annotators agree on the event with a token overlap, but do not exactly agree on the exact span. Also Cohen's κ points to a good agreement for the span annotations.

We further observe a moderate emotion agreement

Emotion Class	Cohen's κ	# Writer	# Other
Anger	0.53	204	132
Anticipation	0.58	0	2
Contentment	0.00	2	3
Disappointment	0.58	2	4
Disgust	0.80	66	21
Fear	0.71	134	86
Frustration	0.09	3	2
Guilt	0.68	164	95
Hope	0.17	9	30
Joy	0.84	116	146
Pride	0.00	0	1
Sadness	0.62	243	170
Shame	0.43	81	24
Surprise	0.02	48	21
Trust	0.00	0	4
No emotion	0.43	42	227
Other	0.17	4	3

Table 3: Inter-annotator agreement for separate emotions, calculated on all experiencer spans in which at least two annotators agreed by at least one token. # Writer and # Other denote the number of times in which either experiencer has such emotions.

Variable: Appraisals	Cohen's κ	Spearman's ρ
Suddenness	0.62	0.51
Familiarity	0.53	0.35
Pleasantness	0.82	0.69
Understand	0.88	0.34
Goal relevance	0.57	0.43
Self responsibility	0.80	0.79
other responsibility	0.77	0.77
Situational respons.	0.66	0.61
Effort	0.56	0.53
Exert	0.52	0.20
Attend	0.53	0.25
Consider	0.56	0.49
Outcome probability	0.63	0.48
Expectation discrepancy	0.68	0.55
Goal conduciveness	0.68	0.62
Urgency	0.45	0.24
Self control	0.61	0.52
Other control	0.67	0.60
Situational control	0.60	0.53
Adjustment check	0.69	0.56
Internal check	0.56	0.52
External check	0.67	0.65

Table 4: Agreement on appraisals. κ is based on a discretization at the threshold ≥ 4 to two classes.

(κ=.62), calculated on all experiencer annotations with an overlap between annotators. We break down emotion-class-specific agreement in Table 3. To give a better idea of their relative occurrences, we indicate the counts of writer (column # Writer) or another mentioned entity (# Other) being associated to each emotion. We see that the agreement is low only for very infrequent emotion classes.

As for appraisals, we analyze agreement based on the two measures in Table 4. The Cohen's κ score is based on a binarization based on the threshold of ≥ 4 for an appraisal category to hold. We do that under the assumption that some users of our corpus might prefer to perform discrete categorization instead of regression. In addition, we report the Spearman's correlation score on the original, non-binarized ratings. Note that not all annotators might have marked the same event participants. Similar to emotions, we compute their agreement only for sentences in which there is an experiencer overlap. We see that the agreement is moderate to good across nearly all dimensions. The judgments are positively correlated across all dimensions, with the weakest correlations holding for *effort*, *attention* and *urgency*.

5.2. Aggregated Corpus Analyses

The final corpus encompasses 720 texts (929 sentences).[7] It includes 912 event spans, on a total of 17.5k tokens (on average, 24.3 per sentence), and contains annotations for 1329 experiencers.[8] Experiencers are mostly a combination of writer and other entities within a sentence (in 270 texts, the experiencer is the writer only; in 8, only other entities). Further, the predominant emotions associated to the writers (*anger*, *guilt*, *fear*, *joy*, *shame*, *disgust*, *sadness*, see Table 3) correspond to the set of emotions for which the authors of texts in enISEAR self-labelled their own reactions.

5.2.1. Within Experiencer Analysis

We now turn to the analysis of the emotion and appraisal annotations for each experiencer in isolation, illustrated in Figure 3. The heatmap shows the average appraisal values for each emotion (limited to those that appear in the corpus more often than 15 times).

Some appraisal dimensions have similar average scores across emotion categories. This is the case for appraisals that resulted in acceptable κ agreement scores, but seemed particularly difficult to annotate during guidelines discussion. Columns that stand out are *adjustment check*, *expectation discrepancy* and *understand*, which tend to be higher than the others. For instance, while *understand* has lower averages for *fear* and *surprise*, it does not show high variability across cells, suggesting that evaluating in retrospect whether an experiencer understood what was going on during the event is difficult. A similar consideration holds for *effort* and *exert*.

By contrast, many appraisal dimensions hold more for specific emotions, like *self control* and *self responsibility*, which mainly concentrate on the rows from *guilt* to *no emotion*. Some intuitive features of emotions emerge there, indicating that certain categories are more strongly characterized by specific appraisal dimensions

[7] We release 180 extra texts that were not annotated by all annotators. For simplicity, we do not discuss them here.

[8] Word and sentence counts were obtained using the NLTK tokenizer. Prior to tokenization, we removed the "…" masks.

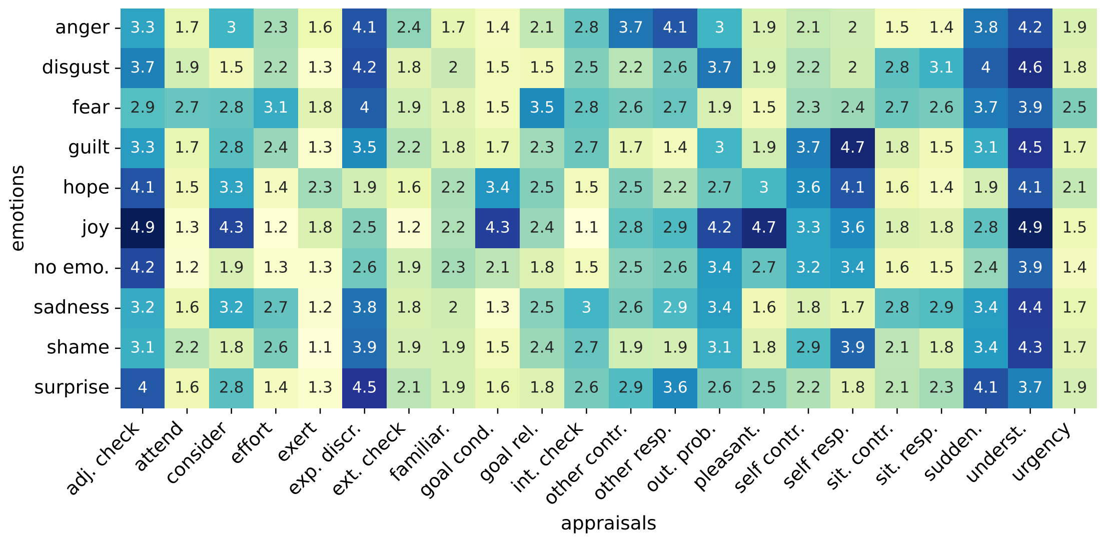

Figure 3: Analysis of appraisals and emotions for each experiencer.

than others. For instance, the label *surprise* seems to be chosen when events are appraised as sudden (*suddenness*) or as divergent from one's expectation (*expectation discrepancy*), and anger is more typical to events triggered by others (*other responsibility*) instead of the person reporting on it. The dimension of *self control* is particularly high for *guilt*, among the set of negatively-valenced emotions, but not for its kin *shame*. *Joy* and *hope*, the only positive emotions in the table, exhibit a greater score of *pleasantness*. They are also associated to a particularly high level of *adjustment check* (i.e., the idea that one can easily cope with the consequences of the event and *outcome probability* (the ability to predict the consequence of the event) as opposed to *fear*.

5.2.2. Between-Experiencer Analysis

The novelty of x-enVENT is that we can evaluate the relation of emotions and appraisals between different experiencers. We look into that in the following.

Different Experiencers, Different Reactions. If one experiencer feels a given emotion in response to an event, what emotions can be elicited in the other participants? We analyse the relation between the emotion of the writer vs. that of any other experiencer, and do that for the texts in which both appear – i.e., ignoring instances where only one of them is annotated. Results are shown in Table 4, where one cell represents the proportion of times any experiencer is associated to the emotion on the column, when the writer is annotated with the emotion on the row. We disregard infrequent classes in this depiction.

The class joy is standing out, as it seems shared by all participants in an event. The diagonal contains comparably high values for some other classes as well. This means that in some cases an emotion is common to different experiencers. However, for the majority of classes, the higher numbers are scattered off diagonal. That is, different emotion reactions can be inferred from

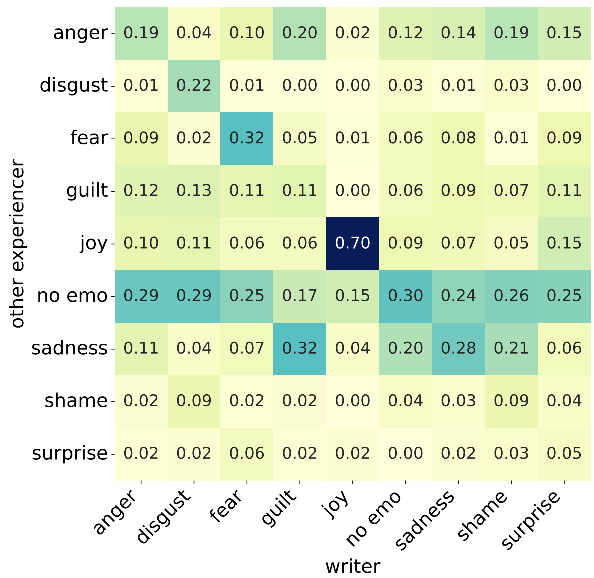

Figure 4: Averaged emotion co-occurrences between the writer (columns) and other experiencers (rows).

text with respect to different semantic roles. Interesting combinations are *guilt–anger* (.20) *no emotion–sadness* (.20). Next to that, the combination of *guilt* and *sadness* (.32) seems to suggest that the writer's sadness is often accompanied by another's guilt. Another interesting case is in the writer's emotion of *shame*, which often co-occurs with the *anger* of other entities (.19). Lastly, it is worth noticing that other experiencers often cannot attribute emotions to a situation that, instead, caused some in the writer. That happens for all of writer's emotions, with the lower cases being *joy* and *guilt* – i.e., two labels that, as discussed, likely co-occur with another emotion. The opposite is not true: non-emotional reactions of the writers are not uniformly distributed across the emotions of other entities.

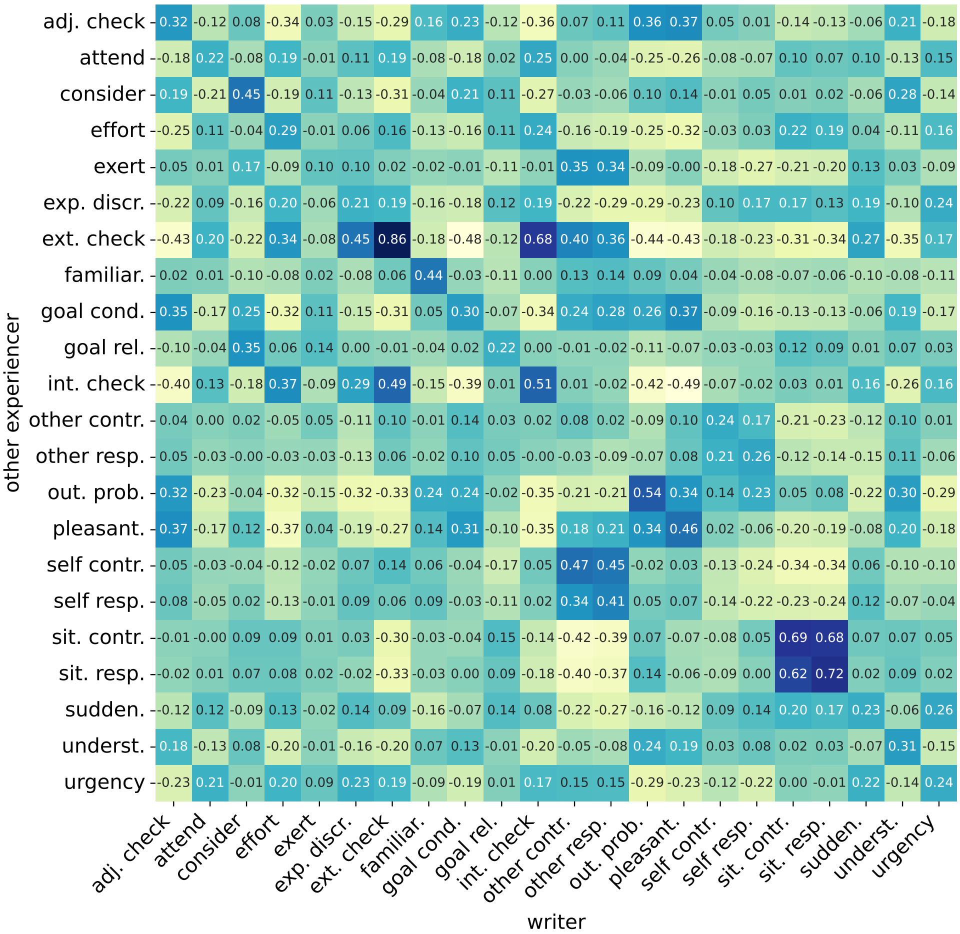

Figure 5: Spearman's correlations of the writer's (columns) and other experiencers' (rows) appraisal scores.

Appraisal Correlations. Next to emotions, we are interested in the relation between appraisals across different entities: if one perceives, for instance, *self responsibility* for an event, what is the appraisal of the other experiencer? Results are in Figure 5. For each (writer-other entity) pair in a text, we retrieve the scores of all appraisal pairs, where one element is the (averaged, see Section 4.3.) score assigned to the writer and the other is given to the mentioned entity. Hence, we calculate Spearman's correlation for such appraisal combinations corresponding to the cells of Figure 5.

Appraisals holding for the writers are positively correlated with the same dimensions for other entities (see diagonal). This, however, does not suggest that events are always similarly appraised by all participants, as many positive correlations can be found among diverse appraisal combinations. Examples are *internal check-external check* (ρ=.68), *expectation discrepancy-external check* (.45), and *other control* (or *responsibility*) - *self control* (*responsibility*). The latter pair indicates that, often, one participant triggers the event and the other is subject to it – but if an event is driven by external factors, it is so for both (see their *situational control/responsibility*). Among the negatively correlated, we notice *internal check-pleasantness*, *outcome probability-urgency*, *external check-adjustment check*.

6. Conclusion

This paper introduced x-enVENT, an English dataset motivated by cognitive appraisal theories of emotions and endowed with a multi-level set of annotations, including all participants in an emotion episode, their specific reaction, and their relation to the eliciting events. Our analysis shows that the text-level emotion annotation is typically not the same for all experiencers, hence, modelling emotion classification and role labeling together provides a more complete picture of emotion descriptions in text.

Our corpus is the fundament for future research to develop experiencer-specific emotion classification models. These will then enable a large-scale analysis of causal chains of emotions in context of multiple people.

Acknowledgements

This research is funded by the German Research Council (DFG), project "Computational Event Analysis based on Appraisal Theories for Emotion Analysis" (CEAT, project number KL 2869/1-2). We thank our annotators and Kai Sassenberg for fruitful discussions.

7. Bibliographical References

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Balahur, A., Hermida, J. M., and Montoyo, A. (2011). Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE transactions on affective computing*, 3(1):88–101.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA).

Buechel, S. and Hahn, U. (2017a). 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*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2017b). Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain, April. Association for Computational Linguistics.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., and Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 105–114, New York, NY, USA. Association for Computing Machinery.

Casel, F., Heindl, A., and Klinger, R. (2021). Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany, 6–9 September. KONVENS 2021 Organizers.

Chang, Y.-C., Chen, C.-C., Hsieh, Y.-L., Chen, C. C., and Hsu, W.-L. (2015). Linguistic template extraction for recognizing reader-emotion and emotional resonance writing assistance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 775–780, Beijing, China, July. Association for Computational Linguistics.

Chen, Y., Hou, W., Li, S., Wu, C., and Zhang, X. (2020). End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

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

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

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

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

Fillmore, C. J. et al. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, 280(1):20–32.

Gao, Q., Jiannan, H., Ruifeng, X., Lin, G., He, Y., Wong, K.-F., and Lu, Q. (2017). Overview of ntcir-13 eca task. In *Proceedings of the NTCIR-13 Conference*.

Ghazi, D., Inkpen, D., and Szpakowicz, S. (2015). Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Gui, L., Wu, D., Xu, R., Lu, Q., and Zhou, Y. (2016). Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November. Association for Computational Linguistics.

Haider, T., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France, May. European Language Resources Association (ELRA).

Hofmann, J., Troiano, E., Sassenberg, K., and Klinger, R. (2020). Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hofmann, J., Troiano, E., and Klinger, R. (2021). 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*, pages 160–170, Online. Association for Computational Linguistics.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA Sentment Corpus for the Automotive Domain. In *4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Kim, E. and Klinger, R. (2018). Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kim, E. and Klinger, R. (2019). Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, Juni.

Klinger, R. and Cimiano, P. (2014). The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Li, W. and Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749.

Mohammad, S., Zhu, X., and Martin, J. (2014). Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland, June. Association for Computational Linguistics.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad, S. (2012). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Neviarouskaya, A. and Aono, M. (2013). Extracting causes of emotions from text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 932–936, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Oatley, K. (2000). Emotion: Theories. *Encyclopedia of psychology*, 3:167–171.

Oberländer, L. A. M. and Klinger, R. (2020). Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Oberländer, L., Reich, K., and Klinger, R. (2020). Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, Barcelona, Spain, December. Association for Computational Linguistics.

Omdahl, B. L. (1995). *Cognitive Appraisal, Emotion, and Empathy*. Mahwah, NJ: Lawrence Erlbaum.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, 89(4):344–350.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Rashkin, H., Sap, M., Allaway, E., Smith, N. A., and Choi, Y. (2018). Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia, July. Association for Computational Linguistics.

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

Roseman, I. J., Spindel, M. S., and Jose, P. E. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. *Journal of Personality and Social Psychology*, 59(5):899–915.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Russo, I., Caselli, T., Rubino, F., Boldrini, E., and Martínez-Barco, P. (2011). EMOCause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon, June. Association for Computational Linguistics.

Scherer, K. R. and Fontaine, J. J. (2013). Driving the emotion process: The appraisal component. In J. J. R. Fontaine, et al., editors, *Series in affective science. Components of emotional meaning: A sourcebook*, chapter 12, pages 266–290. Oxford University Press, Oxford.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Scherer, K. R. and Wallbott, H. G. (1997). The ISEAR questionnaire and codebook. Geneva Emotion Research Group.

Scherer, K. R. (1989). Appraisal theory. *Handbook of Cognition and Emotion*.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

Shaikh, M. A. M., Prendinger, H., and Ishizuka, M. (2009). A linguistic interpretation of the occ emotion model for affect sensing from text. *Affective Information Processing*, pages 45–73.

Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):186–209.

Toprak, C., Jakob, N., and Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden, July. Association for Computational Linguistics.

Tracy, J. L. and Robins, R. W. (2006). Appraisal antecedents of shame and guilt: Support for a theoretical model. *Personality and social psychology bulletin*, 32(10):1339–1351.

Troiano, E., Padó, S., and Klinger, R. (2019). Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Wei, P., Zhao, J., and Mao, W. (2020). Effective interclause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online, July. Association for Computational Linguistics.

Xia, R. and Ding, Z. (2019). Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy, July. Association for Computational Linguistics.

Yanchus, N. J. (2006). *Development and validation of a self-report cognitive appraisal scale*. Ph.D. thesis, University of Georgia.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., and Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California, June. Association for Computational Linguistics.