Appraisal Theories for Emotion Classification in Text

Jan Hofmann[1], **Enrica Troiano**[1], **Kai Sassenberg**[2,3], and **Roman Klinger**[1]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Leibniz-Institut für Wissensmedien, Tübingen, Germany
[3]University of Tübingen, Germany
{jan.hofmann,enrica.troiano,roman.klinger}@ims.uni-stuttgart.de
k.sassenberg@iwm-tuebingen.de

Abstract

Automatic emotion categorization has been predominantly formulated as text classification in which textual units are assigned to an emotion from a predefined inventory, for instance following the fundamental emotion classes proposed by Paul Ekman (fear, joy, anger, disgust, sadness, surprise) or Robert Plutchik (adding trust, anticipation). This approach ignores existing psychological theories to some degree, which provide explanations regarding the perception of events (for instance, that somebody experiences fear when they discover a snake because of the appraisal as being an unpleasant and non-controllable situation), even without having access to explicit reports what an experiencer of an emotion is feeling (for instance expressing this with the words "I am afraid."). Automatic classification approaches therefore need to learn properties of events as latent variables (for instance that the uncertainty and effort associated with discovering the snake leads to fear). With this paper, we propose to make such interpretations of events explicit, following theories of cognitive appraisal of events and show their potential for emotion classification when being encoded in classification models. Our results show that high quality appraisal dimension assignments in event descriptions lead to an improvement in the classification of discrete emotion categories.

1 Introduction

The task of emotion analysis is commonly formulated as classification or regression in which textual units (documents, paragraphs, sentences, words) are mapped to a predefined reference system, for instance the sets of fundamental emotions *fear*, *anger*, *joy*, *surprise*, *disgust*, and *sadness* proposed by Ekman (1999), or by Plutchik (2001), including also *trust* and *anticipation*. Machine learning-based models need to figure out which words point to a particular emotion which is experienced by a reader, the author, or a character of a text. Depending on the resource which has been annotated, the description of an emotion experience can vary. On Twitter, for instance, other than direct reports of an emotion state ("I feel depressed"), hashtags are used as emotion labels to enrich the description of events and stances ("I just got my exam result #sad"). In news articles, emotional events are sometimes explicitly mentioned ("couple infuriate officials", Bostan et al. (2020)) and other times require world knowledge ("Tom Cruise and Katie Holmes set wedding date", labeled as surprise, Strapparava and Mihalcea (2007)). In literature, a sequence of events which forms the narrative leads to an emotion in the reader. In this paper, we focus on text which communicates emotions without an explicit emotion word but rather describes events for which an emotion association is evident.

Such textual examples became popular in natural language processing research with the use of the data which has been generated in the ISEAR project (Scherer and Wallbott, 1997). It lead to a data set of descriptions of events which triggered specific affective states, originally to study event interpretations with a psychological focus. In text analysis, to infer the emotion felt by the writers of those reports, an event interpretation needs to be accomplished. For instance in the text "When a car is overtaking another and I am forced to drive off the road", the model needs to associate the event with *fear*.

However, nearly all computational approaches to associate text with emotions are agnostic to the way how emotions are communicated, they do "not know" how to interpret events, but, presumably, they purely learn word associations instead of actual event interpretations. One might argue that those approaches which predict fine-grained dimensions of affect, namely arousal and valence, tackle this problem (Buechel and Hahn, 2017; Preoţiuc-Pietro et al., 2016). However, these typically do not infer downstream emotion categories. Further, particularly regarding events, psychological theories offer more detailed information. As an example, the emotion component model (Scherer, 2005) advocates that cognitive appraisal dimensions underly discrete emotion classes (Smith and Ellsworth, 1985). These appraisal dimensions[1] evaluate (1) how pleasant an event is (*pleasantness*, likely to be associated with *joy*, but unlikely to appear with *disgust*), (2) how much effort an event can be expected to cause (*anticipated effort*, likely to be high when *anger* or *fear* is experienced), (3) how certain the experiencer is in a specific situation (*certainty*, low, e.g., in the context of *hope* or *surprise*), (4) how much attention is devoted to the event (*attention*, likely to be low, e.g., in the case of *boredom* or *frustration*), (5) how much responsibility the experiencer of the emotion has for what has happened (*self-other responsibility/control*, high for feeling *challenged* or *pride*), and (6) how much the experiencer has control over the situation (*situational control*, low in the case of *anger*).

Despite its richness, cognitive theories of appraisal and their empirical results have not been exploited for emotion prediction in text yet. We fill this gap with this paper and analyze the relation between appraisal dimensions and emotion categories in a text classification setting. We post-annotate an English emotion corpus of self-reports of emotion events (Troiano et al., 2019), which already contains annotations related to the emotions of *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness*, and *shame*, and add the appraisal dimensions by Smith and Ellsworth (1985) mentioned above. Further, we analyze if an automatic prediction of these dimensions from text is possible with standard neural methods, and if these predictions contribute to emotion classifications. Our main contributions are: (1) the first event-centered corpus annotated with appraisal dimensions; (2) the evaluation how well text classification models can recognize these appraisal dimensions; (3) we show emotion classification benefits from the information of appraisal dimensions, when high quality predictions of these are available. Further, (4), we replicate the study by Smith and Ellsworth (1985) from a computational linguistics perspective, based on textual event descriptions.

The remainder of the paper is organized as follows. In Section 2, we review emotion theories and how they are used in automatic text analysis methods in detail. In Section 3, we present the procedure how we annotated our corpus. Section 4 finally shows how well text classification to recognize appraisal dimensions works and how this can be used to predict emotions. Finally, we conclude in Section 5.

2 Background on Emotion Psychology and Analysis

2.1 Emotion and Affect Theories

As a component of humans' life, emotions have been thoroughly studied in the field of psychology, where they are generally deemed responses to salient events. The debates surrounding their definition, however, has never come to a consensus, producing a varied literature on the topic. This has a clear implication for computational emotion analyses, for they must choose and follow one of the available psychological theories in order to motivate the emotion phenomenon that they research in language.

Some of such theories focus on the evolutionary function of emotions, and accordingly, on their link to actions (Izard, 1971; Tooby and Cosmides, 2008). The core idea is that emotions help humans accomplish every-day life tasks and communicate socially relevant information by triggering specific physiological symptoms. In particular, there are patterns of behaviour (e.g., smiling) that reflect discrete emotion terms (e.g., joy), which suggests that emotional states can be grouped based on a few natural language categories. One of the most popular sources for a set of fundamental emotions is the theory by Ekman (1992), who studied the relation between emotions and culture, as well as their facial expressions. Though doubted these days (Gendron et al., 2014), he claimed that the set of fundamental emotions, namely, *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* can be distinguished by facial muscular movements across cultures. As an addition to this model, Plutchik (2001) makes the assumption explicit that different fundamental emotions can occur together, for instance *trust* and *joy*, which is the case when *love* is experienced. Such emotion mixtures as well as an opposing property between *anger* and *fear*, *joy* and *sadness*, *surprise* and

[1]The examples follow the results by Smith and Ellsworth (1985), an excerpt is shown in Table 1.

anticipation, as well as *trust* and *disgust*, has been included in this model. In natural language processing, mostly a set of four to eight fundamental emotions is used, where anger, fear, joy, and sadness are shared by most approaches (an exception with 24 emotion classes is Abdul-Mageed and Ungar (2017)).

A diametrically opposite view is held by the constructivist tradition (Averill, 1980; Oatley, 1993; Barrett and Russell, 2015), in which actions and physiological changes are the building blocks that construct emotions, rather than their direct effect (Feldman Barrett, 2006). Feeling an emotion means categorizing the fluctuations of an affect system along some components. For instance, the affect components *valence* (degree of polarity), *arousal* (degree of excitement), and *dominance* (degree of control over a situation) (Posner et al., 2005) are used as dimensions to describe affect experiences in a 3-dimensional space, which can then be mapped to discrete emotion categories.

2.2 Theories of Cognitive Appraisal

An extension to this model in terms of underlying components is the work of Scherer (1982), Smith and Ellsworth (1985) and Oatley and Johnson-Laird (1987), who qualified emotions as *component* processes that arise to face salient circumstances: an emotion is an "episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus-event as relevant to major concerns of the organism" (Scherer et al., 2001). According to this view, there is an appraisal, that is, an information processing component, which enables people to determine the significance of a situation with respect to their needs and values. In the context of this appraisal (e.g., judging a snake as dangerous), the resources of four other components are mobilitated to deal with the situation. These are, next to the *cognitive component (appraisal)*, a *neurophysiological component (bodily symptoms)*, a *motivational component (action tendencies)*, a *motor expression (facial and vocal expression)*, and a *subjective feeling component (emotional experience)* (Scherer, 2005, Table 1).

While the notions of subjective experience and bodily symptoms were common to other emotion theories, appraisal represents a novelty that fills in some shortcomings of basic models. First, it explains how emotions are elicited. The origin of emotions is to be seen in the stimulus as appraised rather than in the stimulus as such. Second, appraisals provide a structured account for the differences among emotions. For instance, *anger* and *fear* are experienced when the evaluation of a negative event attributes it to external factors, whereas *guilt* and *shame* are felt if the causes of such event are identified in the self, as stable and uncontrollable personality traits, like in the case of *shame* (e.g., "I'm dumb"), or unstable and controllable behaviours for *guilt* (e.g., "I did not observe the speed limit") (Tracy and Robins, 2006).

We argue in this paper that this makes appraisals particularly useful for natural language processing, because they both provide a framework for research and represent a way of enriching existing data. As a matter of fact, few dimensions are sufficient to explain emotions based on cognitive appraisal. Smith and Ellsworth (1985) explain 15 emotions based on *pleasantness* (polarity), *self-other responsibility/control* (for initiating the situation), *certainty* (about what is going on), *attention* (whether the emotion stimulus is worth attending), *anticipated effort* (the amount of physical or mental activation before the stimulus), and *situational control* (the ability to cope with the situation). Compared to the valence-arousal-dominance model, where it is left unclear if the polarity dimension refers to a quality of the emotion stimulus or a quality of the feeling (Scherer, 2005), all these dimensions are unambiguously event-directed. In this paper, we focus on modelling the cognitive components described by Smith and Ellsworth (1985). We show their main findings in Table 1, limited to the emotions that we consider here.

2.3 Automatic Emotion Classification

Previous work on emotion analysis in natural language processing focuses either on resource creation or on emotion classification for a specific task and domain. On the side of resource creation, the early and influential work of Pennebaker et al. (2001) is a dictionary of words being associated with different psychologically relevant categories, including a subset of emotions. Later, Strapparava and Valitutti (2004) made WordNet Affect available to target word classes and differences regarding their emotional connotation, Mohammad and Turney (2012) released the NRC dictionary with more than 14,000 words for a set of discrete emotion classes, and a valence-arousal-dominance dictionary was provided by Mohammad (2018). Buechel et al. (2016) have developed a methodological framework to adapt existing affect lexicons

Emotion	Unpleasant	Responsibility	Uncertainty	Attention	Effort	Control
Happiness	−1.46	0.09	−0.46	0.15	−0.33	−0.21
Sadness	0.87	−0.36	0.00	−0.21	−0.14	1.15
Anger	0.85	−0.94	−0.29	0.12	0.53	−0.96
Fear	0.44	−0.17	0.73	0.03	0.63	0.59
Disgust	0.38	−0.50	−0.39	−0.96	0.06	−0.19
Shame	0.73	1.31	0.21	−0.11	0.07	−0.07
Guilt	0.60	1.31	−0.15	−0.36	0.00	−0.29
Boredom	0.34	−0.19	−0.35	−1.27	−1.19	0.12
Challenge	−0.37	0.44	−0.01	0.52	1.19	−0.20
Hope	−0.50	0.15	0.46	0.31	−0.18	0.35
Interest	−1.05	−0.13	−0.07	0.70	−0.07	−0.63
Contempt	0.89	−0.50	−0.12	0.08	−0.07	−0.63
Frustration	0.88	−0.37	−0.08	0.60	0.48	0.22
Surprise	−1.35	−0.94	0.73	0.40	−0.66	0.15
Pride	−1.25	0.81	−0.32	0.02	−0.31	−0.46

Table 1: The locations of emotions along appraisal PCA dimensions, as published by Smith and Ellsworth (1985), Table 6. The top part corresponds to those emotions we consider in our work here.

to specific use cases. Other than dictionaries, emotion analysis relies on labeled corpora. Some of them include information relative to valence and arousal (Buechel and Hahn, 2017; Preoţiuc-Pietro et al., 2016), but the majority of resources use discrete emotion classes, for instance to label fairy tales (Alm et al., 2005), blogs (Aman and Szpakowicz, 2007), tweets (Mohammad et al., 2017; Schuff et al., 2017; Mohammad, 2012; Mohammad and Bravo-Marquez, 2017; Klinger et al., 2018), Facebook posts (Preoţiuc-Pietro et al., 2016), news headlines (Strapparava and Mihalcea, 2007), dialogues (Li et al., 2017), literary texts (Kim et al., 2017), or self reports on emotion events (Scherer and Wallbott, 1997; Troiano et al., 2019). We point the reader to the survey by Bostan and Klinger (2018) for a more comprehensive overview on emotion datasets.

Most automatic methods to assign labels to text rely on machine learning (Alm et al., 2005; Aman and Szpakowicz, 2007; Schuff et al., 2017, i.a.). Recent shared tasks showed an increase in transfer learning from generic representations (Klinger et al., 2018; Mohammad et al., 2018; Mohammad and Bravo-Marquez, 2017). Felbo et al. (2017) proposed to use emoji representations for pretraining and Cevher et al. (2019) performed pretraining on existing emotion corpora followed by fine-tuning for a specific domain for which only little training data was available.

All previous machine learning-based approaches used models to predict emotions or affect values directly from text, without any access to appraisal dimensions. Only a couple of works incorporated cognitive components, for instance those coming from the OCC model (named after the authors Ortony, Clore and Collins' initials), which sees every appraisal as an evaluation of the pleasantness of events, objects, or actions with respect to one's goals, tastes or behavioural and moral standards (Clore and Ortony, 2013). Based on the OCC model, Shaikh et al. (2009) devised a rule-based approach to interpret text. They did not explicitly formulate their model following appraisal theories, but they moved towards a cognitively-motivated interpretation of events and interpersonal descriptions. Others have adopted patterns of appraisal to predict the emotions triggered by actions, as described in a text. Specifically, Balahur et al. (2011) and Balahur et al. (2012) have created EmotiNet, a knowledge base of action chains that includes information about the elements on which the appraisal is performed within an affective situation, namely, the agent, the action and the object involved in a chain. We share their motivation to delve into event representations based on the descriptions of their experiencers. Unlike their work, ours explicitly encodes appraisal dimensions and uses the classification into these categories for emotion prediction.

3 Corpus

The main objective of this study is to understand the relation between appraisal dimensions and emotion categories. Therefore, we build appraisal annotations on top of enISEAR, an existing corpus of 1001 English event descriptions which are already labeled with the discrete categories of *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness*, and *shame* (Troiano et al., 2019). Each instance has been generated by a crowdworker on the platform FigureEight by completing the sentence "I feel [emotion name], when …". This corpus has an advantage over the original ISEAR resource because it has a German counterpart which can be used in further studies; moreover, its emotion labels have been intersubjectively validated. Our corpus is available at `http://www.romanklinger.de/data-sets/appraisalEnISEAR.zip`.

3.1 Annotation

One presumable challenge in the post-annotation of events regarding the appraisal dimensions is that the annotators in our study do not have access to the private state of the experiencer of the event. However, under the assumption that events are perceived similarly in subjective feeling and evaluated comparably based on cognitive appraisal, we assume that this is not a major flaw in the design of the study. An alternative would have been to perform the text generation task as Troiano et al. (2019) did, but asking the authors of event descriptions for their appraisal in addition. We opted against such procedure as it would have meant to reproduce an existing study in addition to our research goal.

For the post-labeling of enISEAR, we aimed at formulating unambiguous and intuitive descriptions of appraisal dimensions, which would be faithful to those in Smith and Ellsworth (1985). As opposed to the subjects of their study, however, our annotators had to judge events that they did not personally experience. For this reason, we simplified our annotation guidelines in two respects. First, we opted for a binary setting, while Smith and Ellsworth (1985) used continuous scales to rate discrete emotion categories on the appraisal dimensions. Second, we split *control* into *Control* and *Circumstance* (i.e., *self* and *situational control*), in line with the discussion of this variable by Smith and Ellsworth (1985, p. 824f.), while retaining the category of *responsibility*. This was motivated by a series of discussions that revealed the difficulty for annotators to separate the concepts of responsibility and self control. Then, the annotators were instructed to read an event description, without having access to the emotion label, and to answer the following questions:

Most probably, at the time when the event happened, the writer…

- … wanted to devote further attention to the event. (*Attention*)
- … was certain about what was happening. (*Certainty*)
- … had to expend mental or physical effort to deal with the situation. (*Effort*)
- … found that the event was pleasant. (*Pleasantness*)
- … was responsible for the situation. (*Responsibility*)
- … found that he/she was in control of the situation. (*Control*)
- … found that the event could not have been changed or influenced by anyone. (*Circumstance*)

Each event description from enISEAR was judged by three annotators between the age of 26 and 29. One of them is a female Ph.D. student of computational linguistics, the others are male graduate students of software engineering. Two of the annotators are co-authors of this paper. The judges familiarised themselves with their task through four training iterations. At every iteration, we hand-picked 15–20 samples from the ISEAR dataset (Scherer and Wallbott, 1997), such that instances used for training would not be seen during the actual annotation, but had a comparable structure. Dissimilarities in the annotation were discussed in face-to-face meetings and the annotation guideline was refined.

The agreement improved from κ=0.62 to 0.67 in the four iterations. In one of them, we experimented with giving access to the emotion label, which lead to a large improvement in agreement (κ=0.83). Nevertheless, we decided to continue without this information, in order to evaluate the annotator's performance in a similar setting as we evaluate the automatic model – to predict appraisal for emotion classification. We show the pairwise inter-annotator scores of the final set in Table 2. The agreement

	Cohen's κ							
	between annotators				annotator–majority			
Appraisal Dimension	A1/A2	A1/A3	A2/A3	∅	A1	A2	A3	∅
Attentional Activity	.28	.24	.41	.31	.50	.76	.66	.64
Certainty	.41	.23	.29	.31	.62	.77	.46	.62
Anticipated Effort	.38	.33	.26	.32	.69	.67	.62	.66
Pleasantness	.89	.88	.90	.89	.93	.96	.94	.94
Responsibility	.68	.57	.63	.63	.80	.88	.76	.81
Control	.65	.56	.52	.58	.84	.81	.70	.78
Circumstance	.52	.32	.28	.37	.80	.69	.49	.66
Average	.59	.48	.52	.53	.77	.82	.70	.76

Table 2: Cohen's κ between all annotator pairs and between each annotator and the majority vote.

	Appraisal Dimension													
Emotion	Attention		Certainty		Effort		Pleasant		Respons.		Control		Circum.	
Anger	129	.90	119	.83	60	.42	0	.00	9	.06	1	.01	5	.03
Disgust	67	.47	134	.94	40	.28	2	.01	14	.10	11	.08	24	.17
Fear	129	.90	13	.09	121	.85	4	.03	43	.30	18	.13	66	.46
Guilt	55	.38	132	.92	36	.25	0	.00	133	.93	88	.62	11	.08
Joy	139	.97	140	.98	4	.03	141	.99	65	.45	41	.29	25	.17
Sadness	122	.85	112	.78	88	.62	1	.01	7	.05	2	.01	97	.68
Shame	32	.22	111	.78	51	.36	1	.01	106	.74	67	.47	12	.08
Total	673		761		400		149		377		228		240	

Table 3: Instance counts and ratios across emotions and appraisal annotations.

scores between the different annotator pairs are comparable.

These scores tell that rating appraisal dimensions for given events is challenging, and its difficulty varies depending on the categories. Given the comparably low agreement obtained for a subset of dimensions, we opt for a "crowd-sourcing"-like aggregation by taking the majority vote to form the final annotation, included in Table 2, on the right side of the table. We observe that the agreement between majority vote and each annotator is constantly above κ=.62, which is an acceptable agreement (∅κ=.76).

3.2 Analysis

In Table 3 are the cooccurrence counts across emotion and appraisal dimension pairs, as well as the relative counts normalized by emotion (enISEAR provides 143 descriptions per emotion). The most frequently annotated class is *certainty*, followed by *attention*. Appraisal dimensions are differently distributed across emotions: *anger* and *fear* require *attention*, *guilt* and *shame* do not; *disgust* and *anger* show the highest association with *certainty*, in opposition to *fear*. *Responsibility* and *control* play the biggest role in *guilt* and *shame*, while *joy*, non-surprisingly, strongly relates to *pleasantness*. *Fear* has a clear link with *anticipated effort* and, together with *sadness*, is characterized by the inability to control the circumstance.

These numbers are particularly interesting in comparison with the findings of Smith and Ellsworth (1985), who report the average scores along the PCA appraisal dimensions for each emotion[2]. Results are consistent in most cases. For instance, *joy* (or *happiness* in Table 1) stands out as highly pleasant and barely related to *anticipated effort*. Self *responsibility* is lowest in *anger*, an emotion that arises when blame is externalized, and mostly present in *shame* and *guilt*, which derive from blaming the self (Tracy and Robins, 2006). These two are also the emotions that annotators associated with *control* more than

[2]We report the subset of emotions that overlap with ours. Also note that their "Control" corresponds to our "Circum.".

others. *Attention* is prominent for events that elicited *anger* and which were under the control of others, as suggested by the low *situational control*. The highest *situational control*, on the contrary, appears with the data points labeled as *fear*, also characterized by a strong feeling of *uncertainty* and *anticipated effort*. There are also dissimilarities between the two tables, like the level of attention, reaching the lowest score for disgust in their study and not in ours. They also find that situational control is a stronger indicator for *shame* than for *guilt*, while effort is more marked in our sadness-related events than theirs.

These differences may partly be data-specific, partly due to the type of metrics shown in the tables. Most importantly, they can be traced to the annotation setup: while their subjects recalled and appraised personal events, our annotators evaluated the descriptions of events that are foreign to them.

It should be noted that for a reader/annotator it is challenging to impersonate in the writer: although some events have a shared understanding (e.g., "I passed the exam" is most likely appraised as pleasantness), others are tied to one's personal background, values and preferences. This may represent a source of disaccord both between the tables, among the annotators, and with the emotion labels themselves (e.g., "I felt ... when my mom offered me curry" has a *pleasant* gold label, while the original author meant it as a negative emotion, namely *disgust*).

4 Experiments

We now move to our evaluation if automatic methods to recognize emotions can benefit from being informed about appraisal dimensions explicitly. We first describe our models and how we address our research questions and then turn to the results.

4.1 Model Configuration

Figure 1 illustrates the four different tasks addressed by our models. **Task T→E** is the prediction of *emotions from text*, namely the standard setting in nearly all previous work of emotion analysis. We use a convolutional neural network (CNN) inspired by Kim (2014), with pretrained GloVe (Glove840B) as a 300-dimensional embedding layer (Pennington et al., 2014)[3] with convolution filter sizes of 2, 3, and 4 with a ReLu activation function (Nair and Hinton, 2010), followed by a max pooling layer of length 2 and a dropout of 0.5 followed by another dense layer.

As another model to predict emotions, we use a pipeline based on two steps, one to detect the appraisal from text, and the second to assign the appropriate emotion to the appraisal. We refer to **Task T→A** as the step of identifying *appraisal from text*. We use the model configuration of Task T→E, except for the sigmoid activation function and binary cross entropy loss instead of softmax and cross-entropy loss. As a second step, **Task A→E** predicts *emotion from appraisal*. The features are seven boolean variables. We use a neural network with two hidden layers with ReLU activation, followed by a dropout of 0.5.[4]

A disadvantage of the pipeline setting could be that the emotion prediction needs to handle propagated errors from the first step, and that the first step cannot benefit from what the second model learns. Therefore, we compare the pipeline setting (T→A, A→E) with a multi-task learning setting (**T→A/E**). The model is similar to Task T→E. The convolutional layer is shared by the tasks of predicting *emotions from text* and predicting *appraisal from text* and we use two output layers, one for emotion predictions with softmax activation and one for appraisal predictions with sigmoid activation.

4.2 Results

We perform each experiment in a repeated 10×10-fold cross-validatation setting and report average results. All partitions of training and test sets are the same across all experiments.

Experiment 1 (Appraisal Prediction, T→A) aims at understanding how well appraisal dimensions can be predicted from text. Emotion classification is an established task, and one might have some intuition on the expected performance with a given data set, but the prediction of appraisal dimensions has never been performed before. Hence, we report precision, recall, and F_1 for each appraisal component

[3]https://nlp.stanford.edu/projects/glove/

[4]We do not perform any further hyperparameter search. We also compared all model configurations to MaxEnt models to ensure that we do not suffer from overfitting due to a large number of parameters, in comparison to the limited training data. In all settings, the neural models were superiour. We therefore limit our explanations to those.

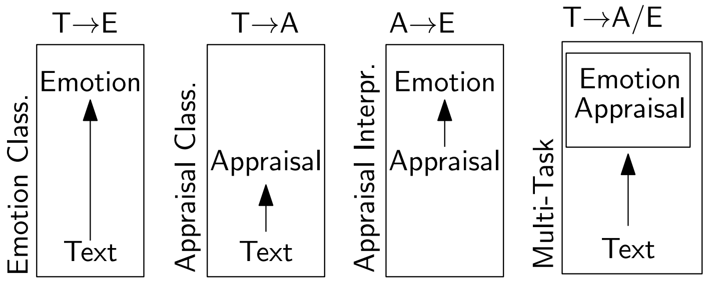

Figure 1: Tasks investigated in experiments on appraisal-driven emotion analysis.

	T→A		
Appraisal	P	R	F_1
Attention	81	84	82
Certainty	84	86	85
Effort	68	68	68
Pleasantness	79	63	70
Responsibility	74	68	71
Control	63	49	55
Circumstance	65	58	61
Macro ∅	73	68	70
Micro ∅	77	74	75

Table 4: Classifier performance on predicting appraisal dimensions.

considered in Table 4. The prediction of *certainty* works best (85%F_1) followed by *attention* (82%F_1). The lowest performance is seen for *control* (55%F_1) and *circumstance* (61%F_1). These results are only partially in line with the inter-annotator agreement scores. We obtain a 75 micro average F_1 score.

Experiment 2 (Appraisal Interpretation, A→E) aims at understanding how well emotions can be predicted from appraisals. We compare the baseline text-to-emotion setting (T→E) to the pipeline setting that first predicts the appraisal and then, from those, the emotion. In the pipeline setting we train the second step (A→E) on the gold appraisal annotations, not on the predictions.[5] We compare this setting to the performance of the appraisal-to-emotion model (A→E), when applied on gold appraisal annotations. This serves as an upper bound which can be reached with the best-performing appraisal prediction model.

We first turn to the results of the model which predicts the emotion based on annotated (gold) appraisal dimensions (A→E (gold)). Here, we observe a clear improvement in contrast to the emotion classification which has access to the text (T→E). *Anger* increases from .52 to .62, *disgust* decreases from .64 to .51, *fear* increases from .70 to .78, *guilt* from .44 to .63, *joy* from .77 to .96, *sadness* decreases from .68 to .66 and *shame* decreased from .45 to .43. On micro average, the performance increases from .60 to .66F_1. These results are an upper-bound for the performance that can be achieved with the pipeline model, under the assumption of having access to perfect appraisal predictions.

When moving to the real-world setting of first predicting the appraisal dimensions and then, based on those, predicting the emotion, the performance scores drop from 66 to 48%F_1. This is an indicator that the performance of our current appraisal prediction, though comparably reasonable with 75%F_1, is not yet sufficient to support emotion predictions, at least partially. The clear improvement in emotion prediction based on perfect appraisal annotations and the performance drop in the real-world setting for emotion prediction suggest that annotating more data with appraisal dimensions is necessary to further develop our approach.

[5]We also tested if training on the prediction leads to better results, but it constantly underperformed.

	T→E			T→A,A→E			A→E (Gold)			T→A/E			Oracle Ensemble		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	51	52	52	34	62	44	55	71	62	51	52	52	66	81	73
Disgust	65	63	64	59	34	43	53	48	51	64	64	64	78	68	73
Fear	69	71	70	55	55	55	79	78	78	70	68	69	76	77	77
Guilt	47	42	44	38	50	43	57	70	63	45	42	44	60	63	62
Joy	74	80	77	77	69	72	94	98	96	77	77	77	79	80	80
Sadness	69	67	68	58	40	47	69	63	66	68	68	68	74	70	72
Shame	44	45	45	36	24	29	56	35	43	43	43	43	58	51	54
Macro ∅	60	60	60	51	48	48	66	66	65	60	59	59	70	70	70
Micro ∅			60			48			66			59			70

Table 5: Comparison of the Text-to-Emotion baseline (T→E) with the performance of first prediction appraisal followed by emotion analysis (T→A,A→E) and the multi-task setting (T→A/E). The oracle consists of an ensemble of T→A,A→E and T→E, which is informed about which model is more likely to make the correct prediction.

Finally, **Experiment 3 (Multi-Task Learning, T→A/E and Oracle Ensemble)** evaluates if the model which learns appraisal and emotion jointly performs better than the pipeline model. Also these results are shown in Table 5. We see that the multi-task learning model cannot improve over the text-only setting. The question remains if the real-world pipeline (T→A, A→E) learns the same things as the text-based approach (T→E). To evaluate this, we design an ensemble, in which an oracle predicts which of the two models will obtain the correct result. In this experiment, we therefore accept a prediction as true positive if either the pipeline or the text-based prediction is correct. This result is shown in Table 5 in the columns "Oracle Ensemble". We see a clear improvement of 10pp over the text-based model and by 22 points over the pipeline. Therefore, we conclude that the model that is informed about the appraisal has the potential to contribute to the correct emotion labelling of a textual instance.

4.3 Discussion and Analysis

We have seen in the experiments and the results that the approach of predicting emotions based on appraisal shows a clear potential for a better performance. Though we have not been able to reach a substantial improvement in a real-world setting, in which appraisal dimensions are first predicted as a basis for the emotion prediction or in the multi-task setting, we observe that text- and appraisal-based models behave differently. Table 6 shows examples for the prediction in the real-world setting (T→A, A→E). In the top block, one example is shown for each emotion in which both appraisal and emotion are correctly predicted. This does not only include cases in which clear emotion indicators exist. The second block reports instances in which the appraisal is correct, but the emotion prediction is not. Here, the first sentence (when someone drove into my car…) is an example in which a flip of certainty would have changed the emotion. Similarly, the change of attention in "when I saw a group…" would have lead to the correct emotion prediction. These are therefore untypical cases of appraisal assignment. The last block shows examples where the wrong appraisal prediction leads to wrong emotion assignment.

It is further interesting to look into those cases which are wrongly predicted from text, but correctly predicted based on the gold appraisal annotations. We show examples for such cases in Table 7. Several of these cases are examples in which a word seems to indicate a particular emotion, which is actually not relevant to infer the emotion in the first place (e.g., *animal, vomiting, kids, high school*). Shame is particularly often predicted wrongly when the event is about the self. This is particularly problematic if the actual word pointing to an emotion appears to be non-typical (e.g., *crossword, anaesthetic*).

		Appraisal							
	Emotion (G/P)	A	Ce	E	P	R	Co	Ci	Text
Appr+Emo correct	Anger	1	1	0	0	0	0	0	when my neighbour started to throw rubbish in my garden for no reason.
	Disgust	0	1	0	0	0	0	0	to watch someone eat insects on television.
	Fear	1	0	1	0	0	0	1	when our kitten escaped in the late evening and we thought he was lost.
	Guilt	0	1	0	0	1	1	0	when I took something without paying.
	Joy	1	1	0	1	1	0	0	when I found a rare item I had wanted for a long time.
	Sadness	1	1	1	0	0	0	1	when my dog died. He was ill for a while. Still miss him.
	Shame	0	1	0	0	1	0	0	when I remember an embarrassing social faux pas from my teenage years.
Emo incorrect	Anger/Fear	1	0	1	0	0	0	0	when someone drove into my car causing damage and fear to myself – then drove off before exchanging insurance details.
	Disgust/Anger	1	1	0	0	0	0	0	when I saw a bird being mistreated when on holiday.
	Fear/Sadness	1	1	1	0	0	0	1	a huge spider just plopped on down on the sofa besides me, staring me out.
	Guilt/Disgust	0	1	0	0	0	0	0	when I watched a documentary that showed footage of farms of pigs and chickens and as a meat eater I felt awful guilt at how they are treated.
	Sadness/Anger	1	1	0	0	0	0	0	when I saw a group of homeless people and it was cold outside.
	Shame/Guilt	0	1	0	0	1	1	0	because I did something silly.
Ap+Emo incorrect	Anger/Shame	**0**	1	0	0	**1**	0	0	I feel ... because I can't stand when people lie.
	Disgust/Anger	**1**	1	0	0	0	**0**	0	when I saw a medical operation on a TV show.
	Fear/Guilt	1	0	**0**	0	**1**	0	**0**	when I was on a flight as I am ... of flying.
	Guilt/Shame	0	1	**1**	0	1	**1**	0	when I lost my sister's necklace that I had borrowed.
	Joy/Anger	1	1	0	**0**	0	0	**0**	when I saw bees coming back to my garden after few years of absence.
	Sadness/Guilt	**1**	1	0	0	**1**	0	**0**	when I watched some of the sad cases of children in need.
	Shame/Guilt	0	1	0	0	1	**1**	0	when I forgot a hairdressers appointment.

Table 6: Examples for the prediction of the pipeline setting (T→A, A→E) . A: Attention, Ce: Certainty, E: Effort, P: Pleasantness, R: Responsibility, Co: Control, Ci: Circumstance. First emotion mention is gold, second is prediction. Appraisal shown is prediction with errors shown in bold.

5 Conclusions and Future Work

We investigated the hypothesis that informing an emotion classification model about the cognitive appraisal regarding a situation is beneficial for the model performance. We were able to show that emotion classification performs better than text-based classification, under the assumption that perfect appraisal predictions are possible and shows complementary correct predictions. Yet, neither in a multi-task learning nor a real-world pipeline, in which the appraisal was predicted as a basis, we could show an improvement in emotion classification. This shows that, though our appraisal predictor is of reasonable performance, the model suffers from error propagation. This is still an encouraging result, suggesting that future work should further investigate the combination of appraisal information with emotion prediction, particularly in the light of our oracle ensemble that showed a clear improvement.

This first study on the topic raises a couple of research questions: Would there be other neural architectures which are better suited for including the appraisal information? Will more annotated data improve the prediction quality sufficiently? Finally, it should be analyzed if giving the annotators access to the emotion label when making the appraisal annotation could have changed the result.

Gold Emotion	A→E	T→E	Text
Anger	Anger	Fear	because I was overlooked at work.
Anger	Anger	Disgust	when I saw someone mistreating an animal.
Anger	Anger	Fear	when someone overtook my car on a blind bend and nearly caused an accident.
Disgust	Disgust	Shame	because I ate a sausage that was horrible.
Disgust	Disgust	Fear	when I was on a ferry in a storm and lots of people were vomiting.
Disgust	Disgust	Shame	because the milk I put in my coffee had lumps in it.
Fear	Fear	Shame	because I had to have a general anaesthetic for an operation.
Fear	Fear	Sadness	when my 2 year old broke her leg, and we felt helpless to assist her.
Fear	Fear	Anger	because we were driving fast in the rain in order to get somewhere before it shut, and the driver was going over the speed limit.
Guilt	Guilt	Shame	when I took something without paying.
Guilt	Guilt	Joy	for denying to offer my kids what they demanded of me.
Guilt	Guilt	Anger	when I had not done a job for a friend that I had promised to do.
Joy	Joy	Sadness	when witnessing the joy on my children's face on Christmas morning.
Joy	Joy	Shame	when I managed to complete a cryptic crossword.
Joy	Joy	Disgust	when I found a twenty pound note on the ground outside.
Sadness	Sadness	Fear	when it was raining this morning as I been planning to go on a camping trip.
Sadness	Sadness	Joy	I feel ... when I see the Christmas decorations come down, and know they won't be up again for another year.
Sadness	Sadness	Shame	when my friend's eye was watering after an injection into it and I could do nothing to help.
Shame	Shame	Joy	when I failed my ninth year at high school.
Shame	Shame	Guilt	when I had too much to drink in a pub, fell over and had to go to hospital.
Shame	Shame	Anger	when my mom caught me lying.

Table 7: Examples in which the appraisal model (on gold appraisal annotation) predicts the correct emotion and the baseline system does not.

Acknowledgements

This work was supported by Leibniz WissenschaftsCampus Tübingen "Cognitive Interfaces" and Deutsche Forschungsgemeinschaft (project SEAT, KL 2869/1-1). We thank Laura Bostan for inspiration and fruitful discussions and Valentino Sabbatino for his annotation work. Thanks to Sven Büchel for comments on an earlier version of this manuscript.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July. Association for Computational Linguistics.

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

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

James R Averill. 1980. A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier.

Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. 2011. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In Rafael Muñoz, Andrés Montoyo, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pages 27–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexandra Balahur, Jesus M. Hermida, and Andrew Montoyo. 2012. 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.

Lisa Feldman Barrett and James A Russell. 2015. *The psychological construction of emotion*. Guilford Publications, New York.

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

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France. European Language Resources Association (ELRA).

Sven Buechel and Udo Hahn. 2017. 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.

Sven Buechel, Johannes Hellrich, and Udo Hahn. 2016. Feelings from the Past—Adapting affective lexicons for historical emotion analysis. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 54–61, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Deniz Cevher, Sebastian Zepf, and Roman Klinger. 2019. Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 79–90, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Gerald. L. Clore and Andrew Ortony. 2013. Psychological Construction in the OCC Model of Emotion. *Emotion Review*, 5(4):335–343.

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

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *Handbook of Cognition and Emotion*. John Wiley & Sons, Sussex, UK.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. 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*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Lisa Feldman Barrett. 2006. Solving the emotion paradox: categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46.

Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251–262.

Carroll E. Izard. 1971. *The face of emotion*. Appleton-Century-Crofts, New York.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

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

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2012. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3).

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 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.

Saif Mohammad. 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. Association for Computational Linguistics.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, Haifa, Israel. Omnipress.

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

Keith Oatley. 1993. Social construction in emotions. In M. Lewis and J. M. Haviland, editors, *Handbook of emotions*, pages 341–352. Guilford Press, New York.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

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

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 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.

Klaus R. Scherer and Harald G. Wallbott. 1997. The ISEAR questionnaire and codebook. Geneva Emotion Research Group. `https://www.unige.ch/cisa/research/materials-and-online-research/research-material/`.

Klaus R. Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Klaus R. Scherer. 1982. Emotion as a process: Function, origin and regulation. *Social Science Information*, 48:555–70.

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

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

Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. 2009. A linguistic interpretation of the occ emotion model for affect sensing from text. *Affective Information Processing*, pages 45–73.

Craig. A. Smith and Phoebe. C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

John Tooby and Leda Cosmides. 2008. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In J.M. Haviland-Jones M. Lewis and L. F. Barrett, editors, *Handbook of emotions*, pages 114–137. Guilford Press, New York.

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

Enrica Troiano, Sebastian Padó, and Roman Klinger. 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.