Emotion-Aware, Emotion-Agnostic, or Automatic: Corpus Creation Strategies to Obtain Cognitive Event Appraisal Annotations

Jan Hofmann, Enrica Troiano, and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{jan.hofmann,enrica.troiano,roman.klinger}@ims.uni-stuttgart.de

Abstract

Appraisal theories explain how the cognitive evaluation of an event leads to a particular emotion. In contrast to theories of basic emotions or affect (valence/arousal), this theory has not received a lot of attention in natural language processing. Yet, in psychology it has been proven powerful: Smith and Ellsworth (1985) showed that the appraisal dimensions *attention*, *certainty*, *anticipated effort*, *pleasantness*, *responsibility/control* and *situational control* discriminate between (at least) 15 emotion classes. We study different annotation strategies for these dimensions, based on the event-focused enISEAR corpus (Troiano et al., 2019). We analyze two manual annotation settings: (1) showing the text to annotate while masking the experienced emotion label; (2) revealing the emotion associated with the text. Setting 2 enables the annotators to develop a more realistic intuition of the described event, while Setting 1 is a more standard annotation procedure, purely relying on text. We evaluate these strategies in two ways: by measuring inter-annotator agreement and by fine-tuning RoBERTa to predict appraisal variables. Our results show that knowledge of the emotion increases annotators' reliability. Further, we evaluate a purely automatic rule-based labeling strategy (inferring appraisal from annotated emotion classes). Training on automatically assigned labels leads to a competitive performance of our classifier, even when tested on manual annotations. This is an indicator that it might be possible to automatically create appraisal corpora for every domain for which emotion corpora already exist.

1 Introduction

Automatically detecting emotions in written texts consists of mapping textual units, like documents, paragraphs, or sentences, to a predefined set of emotions. Common sets of classes used for this purpose rely on psychological theories such as those proposed by Ekman (1992) (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) or Plutchik (2001). These theories are based on the assumption that there is a restricted number of emotions that have prototypical realizations. However, not all sets of emotions are appropriate for every domain. For instance, Dittrich and Zepf (2019) argue that some of the basic emotions are too strong for measuring how people feel when driving a car and, based on that, Cevher et al. (2019) resort to *joy*, *annoyance* (instead of *anger*), *insecurity* (instead of *fear*), *boredom*, and *relaxation* to classify in-car utterances. Haider et al. (2020) model the emotional perception of poetry and opt for the categories *beauty/joy*, *sadness*, *uneasiness*, *vitality*, *awe/sublime*, *suspense*, *humor*, *nostalgia*, and *annoyance*, following the definition of aesthetic emotions (Schindler et al., 2017; Menninghaus et al., 2019). Demszky et al. (2020) define a taxonomy of emotions, reaching a high coverage while maintaining inter-class relations.

An alternative to the use of categorical variables are the so-called "dimensional" approaches. The most popular of them models affective experiences along the variables of dominance, valence, and arousal (Russell and Mehrabian, 1977, VAD). Feldman Barrett (2006, 2017) theorizes that emotions are interpretations of continuous affective states experiencers find themselves in. Still, as Smith and Ellsworth (1985) note, not all emotions can be distinguished based on valence and arousal. One might argue that predicting three continuous variables instead of a richer set of categories is a simplification and can be limiting for downstream applications of emotion analysis models.

Smith and Ellsworth (1985) particularly argue that the VAD model does not capture all relevant aspects of an emotion in the context of an event. In a *fight or flight* situation (Cannon, 1929), for instance, the decision to take one of these two actions

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 160–170
April 19, 2021. ©2021 Association for Computational Linguistics

Emotion	Appraisals	Text
Joy	Attention, Certainty, Pleasant, Sit. Ctrl.	I felt ... when I knew that I was going back to Florida a year earlier than I thought I would.
Disgust	Attention, Certainty, Effort, Sit. Ctrl.	I felt ... when my kitten was sick and I had to clean it up.
Fear	Attention, Effort, Sit. Ctrl.	I felt ... when I was having a hard attach.
Guilt	Attention, Certainty, Effort, Respons., Control	I felt ... when I went on holiday and left our cat behind.
Sadness	Attention, Certainty, Sit. Ctrl.	I felt ... when I found out one of my favourite shops had shut down.

Table 1: Examples from the corpus of Hofmann et al. (2020).

is mostly made based on the effort that the emotion experiencer anticipates, but this is not represented by VAD. Therefore, Smith and Ellsworth (1985) propose a dimensional approach with the appraisal variables of how pleasant an event is (*pleasantness*, likely to be associated with *joy*, but unlikely to appear with *disgust*), how much effort an event can be expected to cause (*anticipated effort*, likely to be high when *anger* or *fear* is experienced), how certain the experiencer is in a specific situation (*certainty*, low, e.g., in the context of *hope* or *surprise*), how much attention is devoted to the event (*attention*, likely to be low, e.g., in the case of *boredom* or *disgust*), how much responsibility the experiencer of the emotion has for what has happened (*self-other responsibility/control*, high for feeling *guilt* or *pride*), and how much the experiencer feels to be controlled by the situation (*situational control*, high in *anger*).

As the cognitive appraisal is a fundamental subcomponent of emotions, we deem that appraisal dimensions are useful to perform emotion recognition, and that even the prediction of appraisals themselves can contribute to computational approaches to affective states. These appraisal dimensions have only recently found application in automatic emotion analysis in text: Hofmann et al. (2020) re-annotated a corpus of 1001 English emotional event descriptions (Troiano et al., 2019) for which the experienced emotion has been disclosed by the author of the description (Table 1 shows examples from their corpus). Their annotation is designed as a preliminary step for inferring discrete emotion categories. In contrast, we argue *that the prediction of appraisal dimensions is in itself valuable*. This intuition has an impact on our annotation strategy. While Hofmann et al. (2020) did not show any emotion label to the annotators, thus avoiding information leaks, we hypothesize that knowing such emotion helps understanding how the described events were originally appraised by their experiencers: at times, properly annotating appraisals as a third party might be unfeasible without having prior access to emotions.

We test this by comparing three annotation procedures: (1) we give the annotator access only to the text but not to its emotion label; (2) we give the annotator access to the text and the emotion, and evaluate if this additional information has an impact on annotation reliability and performance of a pretrained transformer-based classifier fine-tuned on these data; and (3) we automatically infer the appraisal dimensions from existing emotion annotations, investigating the hypothesis that manual annotation might not be necessary.

Our main contributions are that we show that (a) appraisal annotation is more reliable when annotators have access to the emotion label of the original experiencer, hence, *the event description itself does not carry sufficient information for annotation*. That also means that annotating appraisals for corpora in which the original emotion is not available might be particularly challenging. (b) Automatic, rule-based annotation of appraisals that leverages emotion labels is a viable alternative to human annotation, and therefore, appraisal corpora can automatically be created for domains for which emotion corpora are already available.

Our classifier further constitutes a novel state of the art for appraisal prediction on the data by Hofmann et al. (2020). The data is available at https://www.ims.uni-stuttgart.de/data/appraisalemotion.

2 Related Work

2.1 Resources for Emotion Analysis

There is a wealth of literature in psychology surrounding emotions, specifically regarding the way they are elicited, their universal validity, their number and stereotypical expressions, and their function (Scherer, 2000; Gendron and Feldman Barrett, 2009). The two prominent traditions which have dominated the field of emotion classification in natural language processing are discrete and dimensional models (Kim and Klinger, 2019).

Next to the creation of lexicons for emotion analysis (Pennebaker et al., 2001; Strapparava and Valitutti, 2004; Mohammad et al., 2013; Mohammad, 2018, i.a.), the annotation of text corpora received substantial attention (Bostan and Klinger, 2018). They vary across emotion categories and domains, with discrete classes being dominating – some exceptions focused on valence and arousal annotations are Buechel and Hahn (2017), Preoţiuc-Pietro et al. (2016), and Yu et al. (2016). For instance, the ISEAR study by Scherer and Wallbott (1994) led to self-reports of emotionally connotated events. Its creators aimed at understanding what aspects of emotions are universal and which are relative to culture. It was built by asking students to recall an emotion-inducing event and to describe it.

Other efforts focused more on creating corpora specifically for emotion analysis in NLP. Troiano et al. (2019) built enISEAR and deISEAR, whose 1001 event-descriptions were collected via crowdsourcing, with a questionnaire inspired by ISEAR, both in English and in German. TEC (Mohammad, 2012), another popular resource, is bigger in size (≈21k instances), contains tweets and was automatically annotated with hashtags. The Blogs corpus by Aman and Szpakowicz (2007) has sentence-level annotations for 5205 texts, annotated by multiple raters. While ISEAR, enISEAR and deISEAR are focused on describing specific emotion-inducing events, the Blogs corpus and TEC are more general.

This is also the set of corpora that we use in our study (a more comprehensive resource overview was made available by Hakak et al. (2017) and Bostan and Klinger (2018)).

2.2 Appraisal Theories

A richer perspective on emotions and their experience than affect models or fundamental emotion sets is provided by appraisal models (Scherer, 2009b), which did not receive a lot of attention from the NLP community so far. Appraisals are immediate evaluations of situations which guide the emotion felt by the experiencer (Scherer, 2009a). More precisely, an emotion is a synchronized change in five organismic subsystems (i.e., cognitive, peripheral efference, motivational, motor expression and subjective feeling) in response to the evaluation of a stimulus event important to an individual. Emotion states can be distinguished on the basis of their accompanying appraisals. For instance, fear emerges when an event is appraised as unforeseen and disagreeable, a frightening event is one appraised as an unforeseen, unpleasant, and contrary to one's goal (Mortillaro et al., 2012). The cognitive part of the emotion is the one guiding the evaluation of the stimulus along different dimensions. According to Scherer et al. (2001), they are relevance (i.e., the pleasantness of the event, and its relevance for one's goals), implication (i.e., its potential consequences), coping potential (i.e., one's ability to adjust to or control the situation) and normative significance (i.e. its congruity to one's values and beliefs). On a similar vein, Smith and Ellsworth (1985) argue that six cognitive appraisal dimensions can differentiate emotional experiences, as there is a relationship between the way situations are appraised along such dimensions and the experienced emotion. They are *pleasantness*, *anticipated effort*, *certainty*, *attention*, *responsibility/control* and *situational control*. We use their model to explore appraisals in text.

3 Experimental Setting

3.1 Annotation Guidelines

We adhere to the annotation guidelines and the appraisal dimensions of Hofmann et al. (2020), splitting the original *situational control* from Smith and Ellsworth (1985) into *control* and *circumstance*. Our judges take binary decisions with respect to seven appraisal dimensions. We ask them 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*)."

3.2 Experiment 1: Manual Annotation

To annotate the appraisal dimensions, judges need to make assumptions about the experienced situation. We believe this is possible, at times, purely from the textual description that needs to be judged. Other times, knowing which emotion a person developed might be necessary to understand how the overall experience was originally appraised.

To analyze this assumption and measure the importance of emotion labels to reliably assign appraisal dimensions, we build our experiment on top of the English enISEAR corpus by Troiano et al. (2019). Its authors asked workers on a crowdsourcing platform to complete sentences like “I felt [emotion name], when/that/because…”, where [emotion name] is replaced by a concrete emotion. In a later annotation round, other annotators had to infer the emotion of the text, and for this reason the creators of the corpus replaced emotion words with “…”. The resulting 1001 instances of enISEAR are labeled by the experiencers of the emotion themselves and have masked emotion words in the text.

We use these data to perform two annotation experiments on 210 instances, randomly sampled from enISEAR and stratified by emotion. Two annotators judge all of these instances in two different settings. Setting 1, EmoHide, replicates the study by Hofmann et al. (2020): the emotion label is not available to the annotator. In Setting 2, EmoVis, the emotion is presented along with the text. The two rounds of annotations (first EmoHide, later EmoVis which makes the emotion available) were distantiated by six months, to avoid a bias by recalling the previous round. We evaluate the reliability of the annotation via inter-annotator agreement with Cohen’s κ (1960), under the hypothesis that having knowledge of the emotion leads to more reliable human annotations.

Computational Modelling. One of the annotators annotated the full 1001 instances twice, that is, for the EmoHide and the EmoVis approaches as a basis to evaluate how well the realization of the appraisal concepts in the corpus can be modelled automatically. As we expect the annotations EmoVis to be more reliable, we also expect the model to perform better.[1]

We use RoBERTa (Liu et al., 2019) with the abstraction layer for tensorflow as provided by ktrain (Maiya, 2020), and choose the number of epochs to be 5, based on the appraisal prediction and emotion classification tasks in the data by Hofmann et al. (2020) (which we annotate in this paper). Only minor differences in performance can be seen between epochs 4–7. We keep this number of epochs fixed across all experiments and all other parameters at their default. The batch size is 5. More concretely, we opt for a 3×10-fold cross-validation setting, use the RoBERTa-base model in all experiments except for those on the German deISEAR (in the next Experiment 2), where we use XLM-R (Conneau et al., 2020).

Emotion	Attention	Certainty	Effort	Pleasant	Resp/Contr.	Sit. Control
Anger	1	1	1	0	0	0
Disgust	0	1	1	0	0	0
Fear	1	0	1	0	0	1
Guilt	0	1	1	0	1	0
Joy	1	1	0	1	1	0
Sadness	0	1	0	0	0	1
Shame	0	0	1	0	1	0
Surprise	1	0	0	1	0	1

Table 2: Discretized associations between appraisal dimensions and emotion categories, following Smith and Ellsworth (1985), as we use them for automatic annotation in Exp. 2.

3.3 Experiment 2: Automatic Annotation

As Smith and Ellsworth (1985) showed, appraisal dimensions are sufficient to discriminate emotion categories: this is knowledge which we can make use of, and we can leverage their findings to automatically assign discrete appraisal labels to enISEAR (see Table 2) in a rule-based manner. For comparability to the manual annotation setup, we opt for discrete labels which we infer from the continuous principle component analysis values from the original paper.

The question to be answered is if this rule-based annotation actually represents the same concepts as the manual annotation. To answer this, we compare the automatic annotation (AutoAppr) with both annotations that have been performed manually (EmoHide, EmoVis). Further, we train a model to predict these automatic annotations and evaluate on the manually annotated labels.

Since the automatic method relies on emotion labels, we expect its annotations to be more similar to EmoVis, where the annotators also have access to this information. For the same reason, we also assume that the model trained on automatically annotated labels performs better on EmoVis than on EmoHide. Finding that models trained on labels assigned in such rule-based manner show comparable performance to manual annotations (when tested on manual annotations) would suggest that

[1] We acknowledge that model performance on an annotated corpus can only to some degree be used to assess data quality. However, in combination with our inter-annotator agreement assessments, it serves as an indicator of the amount of noise.

	Inter Annotator Agreement			RoBERTa Modelling						
	EmoVis	EmoHide		EmoVis			EmoHide			
Appraisal	κ	κ	Δ	P	R	F_1	P	R	F_1	ΔF_1
Attentional Activity	.55	.30	+.25	.79	.84	.82	.84	.88	.86	−.04
Certainty	.71	.43	+.28	.94	.97	.96	.81	.93	.87	+.09
Anticipated Effort	.44	.38	+.06	.77	.83	.80	.66	.58	.62	+.18
Pleasantness	.93	.87	+.06	.92	.94	.93	.91	.92	.92	+.01
Responsibility	.80	.64	+.16	.85	.79	.82	.83	.81	.82	±.00
Control	.66	.71	−.05	.64	.49	.56	.74	.68	.71	−.15
Circumstance	.65	.54	+.11	.80	.72	.76	.76	.74	.75	+.01
Macro ∅	.68	.55	+.13	.82	.80	.80	.79	.79	.79	+.01
Micro ∅				.84	.85	.84	.80	.81	.81	+.03

Table 3: Experiment 1: Cohen's κ between annotators on EmoVis and EmoHide and modelling experiments. The model is separately trained and tested on EmoVis and EmoHide.

the latter might not be necessary to obtain appraisal prediction models.

In this automatic setup, we merge *responsibility* and *control*. While they are divided in the manually annotated corpora, this separation is not available in the results by Smith and Ellsworth (1985). This affects the comparability of the averages of performance measures between Exp. 1 and 2.

Further, under the assumption that automatic annotation shows competitive results on the manually annotated corpus enISEAR, we extend this analysis to other resources for corpus generalization. In addition to enISEAR, we use the original ISEAR dataset (Scherer and Wallbott, 1994), the German event corpus deISEAR (Troiano et al., 2019) and, as resources without focus on events, the Twitter Emotion Corpus (TEC) (Mohammad, 2012) and the Blogs corpus (Aman and Szpakowicz, 2007). Since these corpora are not manually annotated for appraisals, we only evaluate on automatic appraisal annotations.

4 Results

4.1 Experiment 1: Manual Annotation

Inter-Annotator Agreement. In Experiment 1, we compare the reliability of the annotation with and without access to the emotion label. We show the inter-annotator agreement results in Table 3. As we hypothesized, the agreement on EmoVis is clearly higher than on EmoHide, with .68 in comparison to .55 κ. The highest agreement increase is observed for *attention* (+.25) and *certainty* (+.28), followed by *responsibility* (+.16). The only decrease in agreement, for *control*, is comparably small (−.05).

Figure 1 (and Table 9 in the Appendix) shows the distribution of emotions for the different appraisal dimensions: for most dimensions, the annotation becomes more clearly connected to emotions with its availability, with *certainty* and *anticipated effort* being exceptions: here, the number of instances of a set of emotion classes partially increases. This confirms that knowledge of the emotion "denoises" the annotation.

Modelling. Table 3 also reports the prediction performance on appraisal classes using RoBERTa. We observe that the performances on EmoVis are higher than on EmoHide (+.03pp on micro F_1, .01 on macro F_1). This is in line with our assumptions, but the improvement is actually lower than we expected, given the more substantial difference in inter-annotator agreement. However, for *certainty* (+.09) and *anticipated effort* (+.18), the change is substantial. *Attentional Activity*, which shows a high increase in agreement, has a small decrease in modelling performance (−.04). *Control*, which does not improve in agreement, has a considerable loss in prediction performance (−.15).

From this experiment, we conclude that the annotation is more reliable with access to the emotion: this reflects on the modelling results, and it does so to different extents for different emotions.

4.2 Experiment 2: Automatic Annotation

Inter-Annotator Agreement. We now evaluate the rule-based annotation procedure, in which appraisal classes are purely assigned by the automatic procedure, shown in Table 2.

The agreement between the rule-based annotation AutoAppr and both manual annotations is shown in Table 4. As expected, we observe a higher agreement with EmoVis. Again, the differences

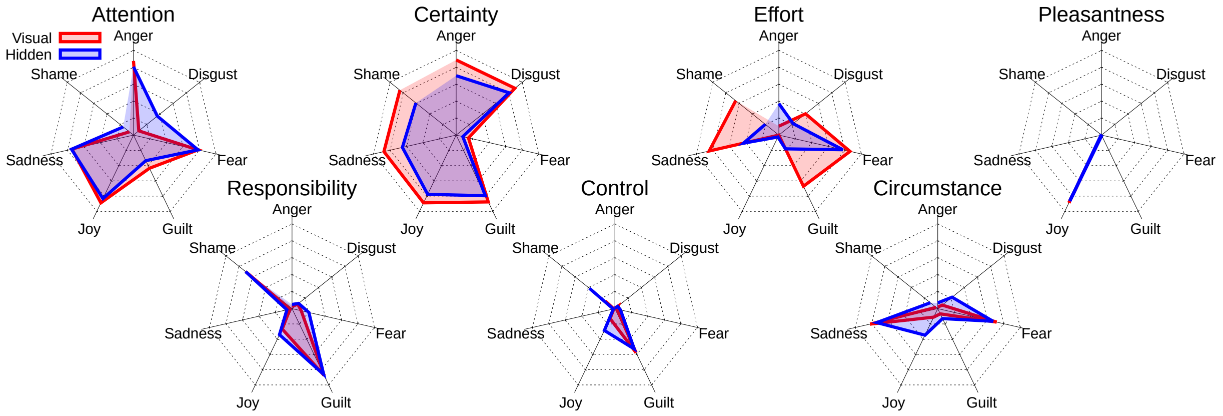

Figure 1: Frequency distribution of appraisals across emotions for EmoVis (Visual) and EmoHide (Hidden).

	Agreement to AutoAppr, κ				Modelling: RoBERTa trained on AutoAppr								
	EmoVis		EmoHide		EmoVis			EmoHide			AutoAppr		
Appraisal	A1	A2	A1	A2	P	R	F_1	P	R	F_1	P	R	F_1
Attentional Activity	.47	.63	.33	.48	.85	.59	.70	.88	.59	.71	.86	.85	.85
Certainty	.35	.53	.50	.51	.97	.81	.88	.81	.83	.82	.88	.90	.89
Anticipated Effort	.31	.08	.19	.16	.61	.77	.68	.39	.79	.52	.95	.96	.96
Pleasantness	.93	1.00	.91	.96	.90	.93	.92	.91	.95	.93	.96	.94	.95
Responsibility/Control	.39	.63	.39	.59	.66	.81	.72	.74	.78	.76	.91	.89	.90
Situational Control	.44	.64	.34	.59	.77	.74	.75	.78	.59	.67	.84	.83	.84
Macro ∅	.48	.58	.44	.54	.79	.77	.78	.75	.75	.74	.90	.89	.90
Micro ∅					.78	.75	.77	.70	.73	.72	.90	.90	.90

Table 4: Experiment 2 Main Results: Cohen's κ between annotators of EmoHide/EmoVis and AutoAppr, on the subset of 210 instances from enISEAR. The classifier is trained on the full set of 1001 instances annotated automatically (AutoAppr) and evaluated on all other annotations (cross-validation splits remain the same).

are not equally distributed across emotions and they resemble the changes in the other experiments, but agreement is lower between AutoAppr and the manual annotations than between the latters, suggesting that the automatic process does not lead to the same conceptual annotation[2].

Modelling. To answer the question how well one model trained on rule-based annotations performs on manual annotations, we test the model three times: on EmoVis, EmoHide, and AutoAppr. The right side of Table 4 reports the results. Note that *responsibility* and *control* have been merged, as explained in the experimental setting.

We see that the highest macro average F_1 is nonsurprisingly achieved when testing on AutoAppr (.90F_1). When testing the same model on EmoVis, the performance drops by 12pp (.78F_1), but is still substantially higher than for the corpus in which the emotions were not available to the annotators (.72F_1). Note that the performance of .78F_1 (EmoVis) and .74F_1 (EmoHide) are not too different from the model trained on manually annotated data, with .80F_1 and .79F_1. We therefore conclude that automatically labeling a corpus with appraisal dimensions leads to a meaningful model.

Corpus Generalization. Finally, we apply the automatic labeling procedure to other emotion corpora. Results are in Table 5.

Given the different nature of the domains and languages (German vs. English – deISEAR/enISEAR; tweets vs. blog texts – TEC vs. Blogs), these numbers cannot be directly compared, but we can observe that they are comparably high, similar to the other experiments. We carefully infer (without having compared the prediction on these corpora to manual annotations) that this is an indicator that automatic annotation of appraisal dimensions also works across different corpora and languages.

[2] These labels could be compared to humans' only on EmoVis, which turned out more reliable. We also consider EmoHide because it represents standard emotion annotation procedures, where judges assess texts without further information.

	deISEAR			ISEAR			TEC			Blogs		
Appraisal	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Attentional Activity	79	68	73	83	82	83	90	91	91	94	94	94
Certainty	79	90	84	89	92	90	87	89	88	97	97	97
Anticipated Effort	88	93	91	94	95	94	74	68	71	91	93	92
Pleasantness	80	69	77	91	90	91	85	86	86	97	97	97
Responsibility/Control	80	69	74	88	85	86	79	79	79	94	96	95
Situational Control	73	69	71	83	81	82	79	79	79	88	86	87
Macro ∅	81	76	78	88	87	88	82	82	82	94	94	94
Micro ∅	82	81	81	89	89	89	84	84	84	94	95	94

Table 5: Experiment 2, Generalization to other corpora. All results are averages across 3×10 cross validations. Note that the last three columns from Table 4 correspond to the same setting as it is in here.

4.3 Model Performance Notes and Comparison to Original Data Annotation

The data that we use was made available to support appraisal-based research in emotion analysis. It consists of the same instances we annotated in Hofmann et al. (2020). However, in this previous work, each instance was judged by three annotators, who did not have access to the emotion labels of the texts, and the experiments have been performed on labels obtained with the majority vote of the annotators. Instead, for the current experiments, the labels by only one annotator on all instances have been used. Therefore, the experiments of the two papers are not strictly comparable. In addition, Hofmann et al. (2020) adopted a CNN-based classifier. Brief, there are two sources for non-comparability in our experiments: different label sets and different models. We aimed at leveraging a more state-of-the-art transformer-based model, but at the same time, we needed different label sets to better understand the appraisal annotation processes.

For transparency reasons, we show the performance of our RoBERTA model on the original labels against the results by Hofmann et al. (2020). Table 6 compares the two studies with respect to appraisals and Table 7 with respect to emotion predictions. The emotion recognition models consist of a text-based model (T→E), a pipeline that first predicts the appraisal and then classifies the emotion without access to text with a two-layer dense neural network (T→A, A→E). To measure the complementarity of these two settings, a third model is an oracle ensemble (T→A→E + T→E) which accepts a prediction as true positive if one of the two models provides the correct prediction.

On this original data set by Hofmann et al. (2020), our model constitutes a new state of the art. The micro-averaged appraisal prediction with RoBERTa is 10pp higher than the original CNN-based model; the emotion classification has similar improvements, and the overall relation between the model configurations remains comparable.

	CNN			RoBERTa		
Appraisal	P	R	F_1	P	R	F_1
Attention	81	84	82	86	90	88
Certainty	84	86	85	87	94	91
Effort	68	68	68	79	77	78
Pleasantness	79	63	70	92	92	92
Responsibility	74	68	71	86	85	85
Control	63	49	55	81	73	77
Circumstance	65	58	61	74	69	71
Macro ∅	73	68	70	83	83	83
Micro ∅	77	74	75	84	85	85

Table 6: RoBERTA model performance on predicting appraisals on the original data by Hofmann et al. (2020), compared to their CNN results.

5 Qualitative Analysis

To better understand how revealing emotions affects the annotations in Experiment 1, we provide some concrete examples. Table 8 reports instances from enISEAR. We show for *which appraisal variables the agreement changes*, by marking the appraisal with + or −. For instance, −*attention* means that the annotators came to disagree on that appraisal dimension when the emotion was uncovered, while +*pleasantness* indicates that they came to agree thanks to the knowledge of the emotion label. Note that + does not mean that the dimension was marked as 1 by both annotators. The examples are sorted by the sum of changes in agreement.

In Example (1) an event is described in a way which leaves open if there is *responsibility*, *pleasantness*, *anticipated effort*, and even if the experiencer is entirely certain about what is happening.

	T→E						T→A,A→E						T→A→E + T→E					
	CNN			RoBERTa			CNN			RoBERTa			CNN			RoBERTa		
Emotion	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	51	52	52	62	62	62	34	62	44	44	72	54	66	81	73	72	85	78
Disgust	65	63	64	70	71	71	59	34	43	65	38	48	78	68	73	86	74	80
Fear	69	71	70	80	82	81	55	55	55	65	76	70	76	77	77	85	92	88
Guilt	47	42	44	60	60	60	38	50	43	50	68	58	60	63	62	72	80	76
Joy	74	80	77	92	96	94	77	69	72	92	95	93	79	80	80	94	98	96
Sadness	69	67	68	82	81	81	58	40	47	74	55	63	74	70	72	87	84	86
Shame	44	45	45	57	54	55	36	24	29	51	23	32	58	51	54	77	59	67
Macro ∅	60	60	60	72	72	72	51	48	48	63	61	60	70	70	70	82	82	82
Micro ∅			60			72			48			61			70			82

Table 7: Comparison of the CNN (Hofmann et al., 2020) and our RoBERTa model on the Text-to-Emotion baseline (T→E), the pipeline experiment (T→A,A→E) and the oracle ensemble experiment (T→A→E + T→E). These experiments follow the model configurations by Hofmann et al. (2020).

ID	Score	Emotion	Aggr. Change	Input Text
1	+4	Fear	+e +p +r +ci	I felt … when I was abseiling down a cliff-face.
2	+3	Joy	+ce +r +ci	I felt … when I got a new job.
3	±0	Guilt		I felt ... when I participated in gossip at work.
4	−2	Shame	−a −ce −r +ci	I felt … when I found out that my daughter had been having a difficult time and I didn't realise straight away what she was going through.
5	−2	Anger	−a −e	I felt … when we were charged by a care home for the three months after my father had died, even though we had emptied his room the day after his death.
6	−3	Fear	−a −ce −r	I felt … when cycling home after a long ride one evening, unaware how dark it had become, and thus relying on some very weak led lights that I'd never tested in complete darkness - I could barely see ten feet ahead of me.

Table 8: Examples of differences between annotations with masked and visible emotion labels. + and − indicate the agreement and the disagreement on a specific dimension which is reached after making the emotion visible. The score is the sum of agreement changes, either improvements (+1 for each dimension) or degradation (−1). a: *attention*, ce: *certainty*, e: *anticipated effort*, p: *pleasantness*, r: *responsibility*, ci: *circumstance*.

With the knowledge that the emotion is fear, it becomes clear that the situation does involve *anticipated effort*, is not *pleasant*, and that *circumstance* is not likely. The annotators also agree here that there is *responsibility* involved, which is likely an interpretation based on world-knowledge.

In the situation of getting a new job (Example (2)), knowing the emotion adds agreement regarding certainty, which is in line with added agreement that the person was responsible. Example (3) is an instance in which the situation was already entirely clear without knowing the emotion – the experiencer participated in gossip. That is a certain, non-pleasurable situation (when recognizing this) which is under their own control. Knowledge that guilt has been experienced does not add anything.

Example (4) shows that complex situations cause more disagreement in annotations which are not necessarily resolved by knowing the emotion. The described event is about a negative emotion felt because the experiencer did not recognize the bad mood of the daughter. Annotators come to disagree about *attention*, *certainty* and *responsibility*.

Example (5) describes another situation in which a negative event is discussed. Knowledge of the emotion puts a clear focus away from the sad part of the description (father dying) and puts it on something that causes anger. However, this shift does not resolve appraisal disagreements but indeed adds on top of them, with *attention* and *anticipated effort*.

Finally, Example (6) is another long description with annotators' focus on different aspects, one on the darkness (hence, no *responsibility*), the other on the cycling (*responsibility*). Here, knowledge of the emotion does not change the interpretability of the event. However, it informs on which part of the described situation the original author focused on.

These examples show that EmoVis helps solve ambiguities when events can be associated to multiple emotions, other times it helps people give more weight to specific portions of texts. In the first case agreement tends to be reached more easily.

6 Conclusion & Future Work

We analyzed how to build corpora of text annotated with appraisal variables and we evaluated how well such concepts can be modelled. By doing so, we brought together emotion analysis and a strand of emotion research in psychology which has received little attention from the computational linguistics community. We propose that in addition to well-established approaches to emotion analysis, like affect-oriented dimensional approaches or classification into predefined emotion inventories, psychological models of appraisals will be considered in future emotion-based studies, particularly those relying on event-oriented resources.

The use of appraisals is interesting from a *theoretical perspective*: motivated by psychology, we leveraged the cognitive mechanisms underlying emotions, thus accounting for many complex patterns in which humans appraise an event and emotionally react to it. In this light, it is interesting in itself that our annotators were able to empathically reconstruct the event appraisals experienced by others, even without knowing their emotion.

From a *practical perspective*, appraisal annotations are less prone to being poorly chosen for particular domains, in comparison to regular emotion classes, as the actual feeling develops based on the cognitive evaluation of an event. We also have shown that event descriptions alone might not be sufficient to properly annotate the hypothetical appraisal of the experiencer (which is, however, also an issue with traditional emotion analysis annotations and models – we cannot look into the feeler, we deal with private states). This shows and is presumably the reason that additional context (e.g., the emotion label) is required.

Some implications for future research and developments follow: ideally, appraisal (and emotion) annotations should stem directly from the experiencer. This is not doable in many NLP settings. For instance, when analyzing literature, it is impossible to ask fictional characters for their current event appraisal. However, we presume that settings on social media might be realistic, for instance by probing appropriate distant labeling methods, e.g., a careful choice of hashtags. If this is unfeasible, because text authors do not disclose their appraisals, the available emotion labels still represent a valuable source of information: as we have shown, they can guide the interpretation of the described events, and hence, the way in which these are post-assigned appraisal dimensions.

Finally, we provided evidence that even if a model is trained on automatically obtained appraisal labels, it is still capable of substantial performance. Therefore, we conclude that more corpora with appraisal dimensions from different languages and domains should be created from scratch. In the meantime, one can build on top of the rich set of available emotion corpora and automatically create appraisal-annotated resources out of them.

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft (projects SEAT, KL 2869/1-1 and CEAT, KL 2869/1-2) and has partially been conducted within the Leibniz WissenschaftsCampus Tübingen “Cognitive Interfaces”. We thank Kai Sassenberg and Laura Oberländer for inspiration and fruitful discussions.

References

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

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.

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.

Walter B. Cannon. 1929. *Bodily changes in pain, hunger, fear, and rage*. Appleton-Century, New York, US.

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.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 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. Association for Computational Linguistics.

Monique Dittrich and Sebastian Zepf. 2019. Exploring the validity of methods to track emotions behind the wheel. In *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems*, pages 115–127, Cham. Springer International Publishing.

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

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

Lisa Feldman Barrett. 2017. *How Emotions Are Made*. Houghton Mifflin Harcourt, New York, USA.

Maria Gendron and Lisa Feldman Barrett. 2009. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, 1(4):316–339.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.

Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, pages 397–402.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text.

Evgeny Kim and Roman Klinger. 2019. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift fuer Digitale Geisteswissenschaften*, 4.

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

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Ines Schindler, Julian Hanich, Thomas Jacobsen, and Stefan Koelsch. 2019. What are aesthetic emotions? *Psychological review*, 126(2):171.

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 Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marcello Mortillaro, Ben Meuleman, and Klaus R. Scherer. 2012. Advocating a componential appraisal model to guide emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 3(1):18–32.

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

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

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.

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

Klaus R. Scherer. 2000. Psychological models of emotion. In *The neuropsychology of emotion.*, Series in affective science., pages 137–162. Oxford University Press, New York, NY, US.

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

Klaus R. Scherer. 2009b. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1535):3459–3474. Publisher: The Royal Society.

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

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

Ines Schindler, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R. Scherer. 2017. Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS one*, 12(6):e0178899.

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 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).

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.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 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. Association for Computational Linguistics.

A Corpus Statistics for Manual Annotations of enISEAR

Table 9 shows the corpus statistics for the manually annotated corpora. In particular, it depicts how the availability of the emotion to the annotators influences the distribution of appraisal labels. The same numbers are also shown in a comparative manner in Figure 1. For most appraisal dimensions, the annotations are more specific, narrower, across the counts for emotions and mostly manifest in a lower number of those. An exception is *certainty*, which shows higher counts for all emotions, and *anticipated effort*, which receives higher counts for shame, sadness, fear, and guilt, but not for anger.

		Appraisal Dimension						
	Emotion	Attention	Certainty	Effort	Pleasant	Respons.	Control	Circum.
enISEAR EmoVis	Anger	141	143	17	0	4	1	3
	Disgust	13	143	65	0	14	8	11
	Fear	126	24	139	0	18	4	115
	Guilt	70	141	108	0	141	93	11
	Joy	143	143	0	143	43	21	18
	Sadness	120	141	136	0	4	2	132
	Shame	10	137	105	0	113	23	4
	Total	623	872	570	143	337	152	294
enISEAR EmoHide	Anger	130	113	60	0	8	1	11
	Disgust	58	129	35	1	16	7	35
	Fear	126	13	125	0	33	11	108
	Guilt	54	128	29	1	139	85	21
	Joy	134	125	4	139	55	46	56
	Sadness	121	105	69	1	10	3	119
	Shame	25	98	33	1	113	62	18
	Total	648	711	355	143	374	215	368

Table 9: Manual Annotation: Instance counts across emotions and appraisal annotations.