Crowdsourcing and Validating
Event-focused Emotion Corpora for German and English

Enrica Troiano, Sebastian Padó and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
`{firstname.lastname}@ims.uni-stuttgart.de`

Abstract

Sentiment analysis has a range of corpora available across multiple languages. For emotion analysis, the situation is more limited, which hinders potential research on cross-lingual modeling and the development of predictive models for other languages. In this paper, we fill this gap for German by constructing deISEAR, a corpus designed in analogy to the well-established English ISEAR emotion dataset. Motivated by Scherer's appraisal theory, we implement a crowdsourcing experiment which consists of two steps. In step 1, participants create descriptions of emotional events for a given emotion. In step 2, five annotators assess the emotion expressed by the texts. We show that transferring an emotion classification model from the original English ISEAR to the German crowdsourced deISEAR via machine translation does not, on average, cause a performance drop.

1 Introduction

Feeling emotions is a central part of the "human condition" (Russell, 1945). While existing studies on automatic recognition of emotions in text have achieved promising results (Pool and Nissim (2016); Mohammad (2011), *i.a.*), we see two main shortcomings. First, there is shortage of resources for non-English languages, with few exceptions, like Chinese (Li et al., 2017; Odbal and Wang, 2014; Yuan et al., 2002). This hampers the data-driven modeling of emotion recognition that has unfolded, *e.g.*, for the related task of sentiment analysis. Second, emotions can be expressed in language with a wide variety of linguistic devices, from direct mentions (*e.g.*, "*I'm angry*") to evocative images (*e.g.*, "*He was petrified*") or prosody. Computational emotion recognition on English has mostly focused on explicit emotion expressions. Often, however, emotions are merely

inferable from world knowledge and experience. For instance, "*I finally found love*" presumably depicts a joyful circumstance, while fear probably ensued when "*She heard a sinister sound*". Attention to such *event-related emotions* is arguably important for wide-coverage emotion recognition and has motivated shared tasks (Klinger et al., 2018), structured resources (Balahur et al., 2011) and dedicated studies such as the "International Survey on Emotion Antecedents and Reactions" (ISEAR, Scherer and Wallbott, 1994). ISEAR, as one outcome, provides a corpus of English descriptions of emotional events for 7 emotions (anger, disgust, fear, guilt, joy, shame, sadness). Informants were asked in a classroom setting to describe emotional situations they experienced. This focus on private perspectives on events sets ISEAR apart. Even though from psychology, it is now established in natural language processing as a textual source of emotional events.

With this paper, we publish and analyze deISEAR, a German corpus of emotional event descriptions, and its English companion enISEAR, each containing 1001 instances. We move beyond the original ISEAR in two respects. (i), we move from on-site annotation to a two-step crowdsourcing procedure involving description generation and intersubjective interpretation; (ii), we analyze cross-lingual differences including a modelling experiment. Our corpus, available at https://www.ims.uni-stuttgart.de/data/emotion, supports the development of emotion classification models in German and English including multilingual aspects.

2 Previous Work

For the related but structurally simpler task of sentiment analysis, resources have been created in many languages. For German, this includes dictionaries

(Ruppenhofer et al., 2017, *i.a.*), corpora of newspaper comments (Schabus et al., 2017) and reviews (Klinger and Cimiano, 2014; Ruppenhofer et al., 2014; Boland et al., 2013). Nevertheless, the resource situation leaves much to be desired. The situation is even more difficult for emotion analysis. Emotion annotation is slower and more subjective (Schuff et al., 2017). Further, there is less agreement on the set of classes to use, stemming from alternative psychological theories. These include, *e.g.*, discrete classes vs. multiple continuous dimensions (Buechel and Hahn, 2016). Resources developed by one strand of research can be unusable for the other (Bostan and Klinger, 2018).

In German, a few dictionaries have been created for dimensional approaches. Among them is BAWL–R, a list of words rated with arousal, valence and imageability features (Vo et al., 2009; Briesemeister et al., 2011), where the nouns of the lexicon have been assigned to emotion intensities, amongst other values. Still, German resources are rare in comparison to English ones. To our knowledge, corpora with sentence-wise emotion annotations are not available for this language.

In particular, there is no German corpus with speakers' descriptions of emotionally intense events similar to the English ISEAR. ISEAR, the "International Survey on Emotion Antecedents and Reactions" (Scherer and Wallbott, 1997), was conducted by a group of psychologists who collected emotion data in the form of self-reports. The aim of the survey was to probe that emotions are invariant over cultures, and are characterized by patterns of bodily and behavioral changes (*e.g.*, change in breathing, felt temperature, speech behaviors). In order to investigate such view, they administered an anonymous questionnaire to 3000 students all over the world, in which participants were asked to reconstruct an emotion episode associated to one of seven basic emotions (anger, disgust, fear, guilt, joy, sadness, shame), and to recall both their evaluation of the stimulus and their reaction to it. For the final dataset, all the reports were translated to English, and accordingly, the responses of, *e.g.*, German speakers who took part in the survey are not available in their original language.

In this paper, we follow Scherer and Wallbott (1997) by re-using their set of seven basic emotions and recreating part of their questionnaire both in English and German. In contrast to ISEAR, we account for the fact that a description can be re-lated to different emotions by its writer and its readers. Affective analyses have rendered evidence that emotional standpoints affect the quality of annotation tasks (Buechel and Hahn, 2017). For instance, annotation results vary depending on whether workers are asked if a text is *associated* with an emotion and if it *evokes* an emotion, with the first phrasing downplaying the reader's perspective and inducing higher inter-annotator agreement (Mohammad and Turney, 2013). We take notice of these findings to design our annotation guidelines.

3 Crowdsourcing-based Corpus Creation

We developed a two-phase crowdsourcing experiment: one for generating descriptions, the other for rating the emotions of the descriptions. Phase 1 can be understood as sampling from $P(\text{description}|\text{emotion})$, obtaining likely descriptions for given emotions. Phase 2 estimates $P(\text{emotion}|\text{description})$, evaluating the association between a given description and all emotions. The participants' intuitions gathered this way are interpretable as a measure for the interpersonal validity of the descriptions, and as a point of comparison for our classification results.

The two crowdourcing phases targeted both German and English. This enabled us to tease apart the effects of the change of setup and change of language compared to the original ISEAR collection.

Phase 1: Generation. We used the Figure-Eight (https://www.figure-eight.com) crowdsourcing platform. Following the ISEAR questionnaire, we presented annotators with one of the seven emotions in Scherer and Wallbott's setup, and asked them to produce a textual description of an event in which they felt that emotion. The task of description generation was formulated as one of sentence completion (*e.g.*, "*Ich fühlte Freude, als/weil/...*", "*I felt joy when/because ...*"), after observing that this strategy made the job easier for laypersons, without inducing any restriction on sentence structure (for details, see Suppl. Mat., Section A). Further, we asked annotators to specify their gender (male, female, other), the temporal distance of the event (*i.e.*, whether the event took place days, weeks, months, or years before the time of text production), and the intensity and duration of the ensuing emotion (*i.e.*, whether the experience was not very intense, moderately intense, intense and very intense, and whether it lasted a few minutes, one hour, multiple hours, or more than one day). To obtain an

	Emotion	Statistics #tok	Temporal Distance				Intensity				Duration				Gender		
			D	W	M	Y	NV	M	I	VI	min	h	>h	≥d	M	F	O
German	Anger	15.1	46	25	31	41	3	25	67	48	23	29	39	52	112	31	–
	Disgust	13.1	38	38	42	25	12	52	48	31	95	37	8	3	110	33	–
	Fear	14.0	25	32	37	49	4	24	58	57	50	32	31	30	109	34	–
	Guilt	13.8	36	27	30	50	8	57	54	24	41	29	43	30	116	27	–
	Joy	11.6	40	30	29	44	2	18	60	63	14	18	42	69	107	35	1
	Sadness	11.5	29	26	42	46	3	31	43	66	16	9	27	91	113	30	–
	Shame	13.2	25	28	36	54	24	56	41	22	72	28	24	19	116	27	–
	Sum	13.2	239	206	247	309	56	263	371	311	311	182	214	294	783	217	1
English	Anger	28.3	45	29	25	44	9	34	48	52	30	23	36	54	62	81	–
	Disgust	22.4	57	25	21	40	12	51	37	43	66	27	24	26	57	86	–
	Fear	27.0	19	29	36	59	2	30	57	54	52	29	35	27	66	77	–
	Guilt	25.5	33	24	27	59	25	52	43	23	26	39	28	50	59	84	–
	Joy	23.6	32	24	31	56	2	27	48	66	14	13	43	73	60	83	–
	Sadness	21.6	40	24	31	48	10	45	38	50	17	21	23	82	62	81	–
	Shame	24.8	21	22	19	81	16	51	42	34	29	25	39	50	57	86	–
	Sum	24.7	247	177	190	387	76	290	313	322	234	177	228	362	423	578	–

Table 1: Statistics for prompting emotions across the average number of tokens (#tok) and the extra-linguistic labels of the descriptions. Temporal Distance, Intensity and Duration report the number of descriptions for events which took place days (D), weeks (W), months (M) or years (Y) ago, which caused an emotion of a specific intensity (NV: not very intense, M: moderate, I: intense, VI: very intense) and duration (min: a few minutes, one hour: h, multiple hours: >h, one or multiple days ≥d); Gender counts of the annotators are reported in the last column (male: M, female: F, other: O).

English equivalent to deISEAR, we crowd-sourced the same set of questions in English, creating a comparable English corpus (enISEAR). The generation task was published in two slices (Nov/Dec 2018 and Jan 2019). It was crucial for data quality to restrict the countries of origin (for German, DE/A; for English, UK/IR) – this prevented a substantial number of non-native participants who are proficient users of machine translation services from submitting answers. For each generated description, we paid 15 cents (see Suppl. Material, Section A for details).

Phase 2: Emotion Labeling. To verify to what extent the collected descriptions convey the emotions for which they were produced, we presented a new set of annotators with ten randomly sampled descriptions, omitting the emotion word (*e.g.*, "*I felt ... when/because ...*"), together with the list of seven emotions. The task was to choose the emotion the original author most likely felt during the described event. Each description was judged by 5 annotators. We paid 15 cents per task.

4 Corpus Analysis

Descriptive analysis. We include all descriptions from Phase 1 in the final resource and the upcoming discussion, regardless of the inter-annotator agreement from Phase 2. Both deISEAR and enISEAR comprise 1001 event-centered descrip-

tions: deISEAR includes 1084 sentences and 2613 distinct tokens, with a 0.19 type-token ratio; enISEAR contains 1366 sentences and a vocabulary of 3066 terms, with a type-token ratio of 0.12. Table 1 summarizes the Phase 1 annotation. For each prompting label[1], we report average description length, annotators' gender, duration, intensity and temporal distance of the emotional events.

The main difference between the two languages is description length: English instances are almost twice as long (24.7 tokens) as German ones (13.2 tokens). These differences may be related to the differences in gender distribution between languages.

Most patterns are similar across German and English. In both corpora, Anger and Sadness receive the longest and shortest descriptions, respectively. Enraging facts are usually depicted through the specific aspects that irritated their experiencers, like "*when a superior at work decided to make a huge issue out of something very petty just to [...] prove they have power over me*". In contrast, sad events are reported with fewer details, possibly because they are often conventionally associated with pain and require little elaboration, such as "*my grandmother had passed away*". Also the perceptual assessments of emotion episodes, as given by the extra-linguistic labels, are comparable between lan-

---

[1]Transl. de→en: Angst-Fear, Ekel-Disgust, Freude-Joy, Scham-Shame, Schuld-Guilt, Traurigkeit-Sadness, Wut-Anger

Emotion	German					English				
	≥1	≥2	≥3	≥4	$=5$	≥1	≥2	≥3	≥4	$=5$
Anger	135	125	107	81	52	137	129	112	89	59
Disgust	139	134	130	124	91	118	101	84	76	53
Fear	134	124	108	99	78	136	131	124	116	86
Guilt	137	126	102	67	31	137	130	124	89	44
Joy	142	142	142	140	136	143	143	143	143	137
Sadness	132	123	113	97	76	140	133	131	116	97
Shame	128	109	86	66	41	116	92	64	41	23
Sum	947	883	788	674	505	927	859	782	670	499

Table 2: Number of descriptions whose prompting label (column Emotion) agrees with the emotion labeled by all Phase-2 annotators ($=5$), by at least four (≥4), at least three (≥3), at least two (≥2), at least one (≥1).

guages. The majority of descriptions are located at the high end of the scale both for intensity and temporal distance, *i.e.*, they point to "milestone" events that are both remote and emotionally striking.

Agreement on emotions. We next analyze to what extent the emotions labelled in Phase 2 agree with the prompting emotion presented in Phase 1. Table 2 reports for how many descriptions (out of 143) the prompting emotion was selected one, two, three, four, or five (out of five) times in Phase 2. Agreement is similar between deISEAR and enISEAR. This indicates that the German items, although short, are sufficiently informative. In both languages, the agreement drops across the columns, yet half of the descriptions show perfect intersubjective validity ($=5$): 505 for German, 499 for English. We interpret this as a sign of quality.

Again, we find differences among emotions. Agreement is nearly perfect for Joy and rather low for Shame. These patterns can arise due to different processes. Certain emotions are easier to recognize from language (*e.g.*, "*when I saw someone else got stabbed near me*": Fear) than others (*e.g.* "*when my daughter was rude to my wife*": elicited for Shame, arguably also associated with Anger or Sadness). Patterns may also indicate closer conceptual similarity among specific emotions (Russell and Mehrabian, 1977, *cf.*).

To follow up on this observation, Figure 1 shows two confusion matrices for German and English which plot the frequency with which annotators selected emotion labels (Phase 2, rows) for prompting emotions (Phase 1, columns). The results in the diagonals correspond to the $=5$ columns in Table 2, mirroring the overall high level of validity of the descriptions, and spanning the range between Joy (very high agreement) and Shame (low agreement).
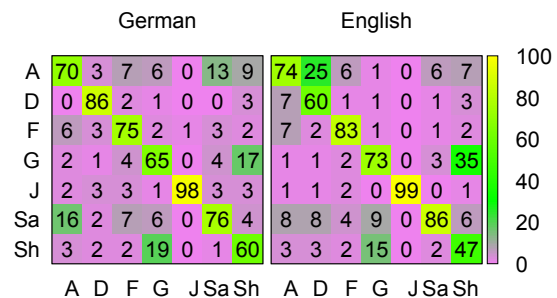

Figure 1: Confusion matrices for emotions. Columns: prompting emotions; rows: labeled emotions.

The off-diagonal cells indicate disagreements. In both languages, annotators perceive Shame descriptions as expressing Guilt, and vice versa (35% and 15% for English, 17% and 19% for German). In fact, Shame and Guilt "occur when events are attributed to internal causes" (Tracy and Robins, 2006), and thus they may appear overlapping.

We also see an interesting cross-lingual divergence. In deISEAR, Sadness is comparably often confused with Anger (13% of items), while in enISEAR it is Disgust that is regularly interpreted as Anger (25% of items). This might results from differences in the connotations of the prompting emotion words in the two languages. For Disgust ("*Ekel*"), German descriptions concentrate on physical repulsion, while the English descriptions also include metaphorical disgust which is more easily confounded with other emotions such as Anger.

Post-hoc Event type analysis. After the preceding analyses, we returned to the Phase 1 descriptions and performed a post-hoc annotation ourselves on a sample of 385 English and 385 German descriptions (balanced across emotions). We tagged them with dimensions motivated by Smith and Ellsworth (1985): whether the event was re-occurring (*general*), whether the event was in the *future* or in the *past*; whether it was a *prospective* emotion or actually felt; whether it had a *social* characteristic (involving other people or animals); whether the event had *self consequences* or *consequences for others*; and whether the author presumably had *situational control* or *responsibility*[2].

Table 3 shows the results. In both English and German, only a few units depict general and future events, in line with the annotation guidelines. Fear more often targets the future than other emotions. Most event descriptions involve other participants, especially in English. In general, events seem to

---

[2]One may be responsible, but not in control of the situation (*e.g.*, "*when I forgot to set an alarm*").

	Dimension	Anger	Disgust	Fear	Guilt	Joy	Sadness	Shame
German	General event	4	2	1	0	0	1	0
	Future event	0	0	1	0	0	0	0
	Past event	51	53	53	55	55	54	55
	Prospective	1	0	4	0	1	1	0
	Social	30	28	24	29	24	40	25
	Self conseq.	37	34	37	26	44	21	37
	Conseq. oth.	21	9	19	34	16	34	14
	Situat. control	2	5	4	24	9	3	19
	Responsible	20	31	17	51	26	23	40
English	General event	2	2	2	2	0	3	0
	Future event	0	0	0	0	0	0	0
	Past event	53	53	53	53	55	52	55
	Prospective	0	0	14	0	1	0	0
	Social	50	37	30	41	39	49	41
	Self conseq.	29	26	42	20	35	16	32
	Conseq. oth.	29	23	19	34	24	43	29
	Situat. control	3	7	8	31	15	2	24
	Responsible	13	29	34	53	34	16	43

Table 3: Event type analysis: Cells are counts of post-annotation out of 55 descriptions for each emotion.

Dataset	μF_1	An	Di	Fe	Gu	Jo	Sa	Sh
deISEAR	47	29	49	48	42	68	53	39
enISEAR	47	27	45	57	41	67	58	32

Table 4: Performance of ISEAR-trained classifier on our crowdsourced corpora, per emotion and micro-average F_1 (μF_1).

affect authors themselves more than other people, particularly in the case of Joy and Fear. Exceptions are Guilt and Sadness, for which there is a predominance of events whose effects bear down on others. Regarding the aspect of situational control, Shame and Guilt dominate. Guilt is particularly more frequent in descriptions in which the author is presumably responsible. These observations echo the findings by Tracy and Robins (2006).

Modeling. As a final analysis, we tested the compatibility of our created data with the original ISEAR corpus for emotion classification. We trained a maximum entropy classifier with L2 regularization with boolean unigram features on the original ISEAR corpus (7665 instances) and evaluated it on all instances collected in Phase 1 (with liblinear, Fan et al., 2008). We chose MaxEnt as a method as it constitutes are comparably strong baseline which is, in contrast to most neural classifiers, more easy to reproduce due to the convex optimization function and fewer hyper-parameters. We applied it to enISEAR and to a version of deISEAR translated with Google Translate[3], an effective baseline strategy for cross-lingual modeling (Barnes et al., 2016). In accord with the Phase 2 experiment, the emotion words present in the sentences were obscured. Table 4 shows a decent performance of the ISEAR model on our novel corpora, with similar scores and performance

differences between emotion classes to previous studies (Bostan and Klinger, 2018).

Modeling performance and inter-annotator disagreement are correlated: emotions that are difficult to annotate are also difficult to predict (Spearman's ρ between F_1 and the diagonal in Figure 1 is 0.85 for German, $p = .01$, and 0.75 for English, $p = .05$). It is notable that results for German are on a level with English despite the translation step and the shorter length of the German descriptions. That goes against our expectations, as previous studies showed that translation is only sentiment-preserving to some degree (Salameh et al., 2015; Lohar et al., 2018). We take this outcome as evidence for the cross-lingual comparability of deISEAR and enISEAR, and our general method.

5 Conclusion

We presented (a) deISEAR, a corpus of 1001 event descriptions in German, annotated with seven emotion classes; and (b) enISEAR, a companion English resource build analogously, to disentangle effects of annotation setup and English when comparing to the original ISEAR resource. Our two-phase annotation setup shows that perceived emotions can be different from expressed emotions in such event-focused corpus, which also affects classification performance.

Emotions vary substantially in their properties, both linguistic and extra-linguistic, which affects both annotation and modeling, while there is high consistency across the language pair English–German. Our modeling experiment shows that the straightforward application of machine translation for model transfer to another language does not lead to a drop in prediction performance.

---

[3]http://translate.google.com, applied on February 25, 2019

References

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 *Natural Language Processing and Information Systems*, pages 27–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1613–1623, Osaka, Japan. The COLING 2016 Organizing Committee.

Katarina Boland, Andias Wira-Alam, and Reinhard Messerschmidt. 2013. Creating an annotated corpus for sentiment analysis of German product reviews. Technical Report 2013/05, GESIS.

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

Benny B. Briesemeister, Lars Kuchinke, and Arthur M. Jacobs. 2011. Discrete emotion norms for nouns: Berlin affective word list (DENN–BAWL). *Behavior Research Methods*, 43(2):441.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem – dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, pages 1114–1122, The Hague, The Netherlands.

Sven Buechel and Udo Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. pages 1–12.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Roman Klinger and Philipp Cimiano. 2014. The US-AGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1656.

Roman Klinger, Orphee 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. Association for Computational Linguistics.

Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. 2017. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924.

Pintu Lohar, Haithem Afli, and Andy Way. 2018. Balancing translation quality and sentiment preservation (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 81–88, Boston, MA. Association for Machine Translation in the Americas.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.

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

Odbal and Zengfu Wang. 2014. Segment-based fine-grained emotion detection for chinese text. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 52–60, Wuhan, China. Association for Computational Linguistics.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39. The COLING 2016 Organizing Committee.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß Jonathan Sonntag, and Michael Wiegand. 2014. IGGSA Shared Tasks on German Sentiment Analysis. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, Hildesheim, Germany. University of Hildesheim.

Josef Ruppenhofer, Petra Steiner, and Michael Wiegand. 2017. Evaluating the morphological compositionality of polarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 625–633. INCOMA Ltd.

Bertrand Russell. 1945. *A History of Western Philosophy*. Routledge Classics.

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

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A

case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1241–1244, New York, NY, USA. ACM.

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.

Klaus R. Scherer and Harald G. Wallbott. 1997. The ISEAR questionnaire and codebook. *Geneva Emotion Research Group*.

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

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

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.

Melissa Vo, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41(2):534–538.

Jiahong Yuan, Liqin Shen, and Fangxin Chen. 2002. The acoustic realization of anger, fear, joy and sadness in chinese. In *Seventh International Conference on Spoken Language Processing*, pages 2025–2028, Denver, Colorado, USA.