Secondary Publication

Wegge, Maximilian; Klinger, Roman

Automatic Emotion Experiencer Recognition

Date of secondary publication: 20.06.2024

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-959309

Primary publication

Wegge, Maximilian; Klinger, Roman (2023): „Automatic Emotion Experiencer Recognition“. In: Christopher Klamm, Gabriella Lapesa, Valentin Gold, Theresa Gessler, Simone Paolo Ponzetto (Ed.), Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, Ingolstadt: Association for Computational Lingustics, pp. 1–7, https://aclanthology.org/2023.cpss-1.1/.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

Automatic Emotion Experiencer Recognition

Maximilian Wegge and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

The most prominent subtask in emotion analysis is emotion classification; to assign a category to a textual unit, for instance a social media post. Many research questions from the social sciences do, however, not only require the detection of the emotion of an author of a post but to understand who is ascribed an emotion in text. This task is tackled by emotion role labeling which aims at extracting who is described in text to experience an emotion, why, and towards whom. This could, however, be considered overly sophisticated if the main question to answer is who feels which emotion. A targeted approach for such setup is to classify emotion experiencer mentions (aka "emoters") regarding the emotion they presumably perceive. This task is similar to named entity recognition of person names with the difference that not every mentioned entity name is an emoter. While, very recently, data with emoter annotations has been made available, no experiments have yet been performed to detect such mentions. With this paper, we provide baseline experiments to understand how challenging the task is. We further evaluate the impact on experiencer-specific emotion categorization and appraisal detection in a pipeline, when gold mentions are not available. We show that experiencer detection in text is a challenging task, with a precision of .82 and a recall of .56 (F_1 =.66). These results motivate future work of jointly modeling emoter spans and emotion/appraisal predictions.

1 Introduction

Computational emotion classification is among the most prominent tasks in the field of textual emotion analysis. It is typically formulated as either a classification or regression task, depending on the underlying emotion theory and intended application and domain: Texts can be classified into one or multiple discrete emotion categories, following the concept of basic emotions by Ekman (1992) or Plutchik (2001), as continuous values within the vector space of valence, arousal and dominance (Russell and Mehrabian, 1977) or based on the emoter's cognitive appraisal of the emotion-eliciting event (e.g., the level of *control* or *responsibility*; Smith and Ellsworth, 1985).

Recent work has emphasized the relevance of perspective, i.e., whose emotion is considered given an emotion-eliciting event. Typically, emotions are investigated from either the writer's or the reader's perspective, with only few approaches that consider both (e.g., Buechel and Hahn, 2017). Although not exclusively focused on it, perspective is also addressed in the context of semantic role labeling ("Who is feeling the emotion?"), besides the emotion target ("Who is the emotion directed towards?") and cause ("What is causing the emotion?") (Mohammad et al., 2014; Bostan et al., 2020a). Troiano et al. (2022) build upon this idea and extend the investigation to all potential emoters affected by an event. For each entity, they consider their emotions and the appraisal of the corresponding event, which allows to disambiguate the individual emotions.

Consider the example "Ken Paxton: Texas House votes to impeach Trump ally"[1]. Here, "Ken Paxton" could be attributed *guilt* because of the impeachment process following a potential appraisal of *self responsibility*. "Trump" being described as an ally might develop *anger* because he might evaluate the situation differently and assign an appraisal of *other responsibility*. "Texas House" could be considered a named entity, but does not represent an emoter. The writer's emotion is presumably irrelevant in such news headline. Experiencer-agnostic approaches can only assign emotions and appraisal to the entire text, thus oversimplifying the relations between individual experiencers.

Wegge et al. (2022) compare experiencer- and text-level emotion/appraisal predictors on self-

[1] https://www.bbc.com/news/world-us-canada-65736478

Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences, pages 1–7
September 22, 2023. ©2023 Association for Computational Linguistics

reported event descriptions. They find that an experiencer-specific predictor is able to capture the individual information, while a conventional classifier averages over all individual (potentially contradictory) information in the entire text. While they provide a computational approach for experiencer-specific emotion and appraisal classification, they rely on gold annotations of experiencer-spans. They do not investigate whether these spans can be predicted reliably and what consequences this would have on the classification task.

In this paper, we evaluate (i.) the performance of an automatic experiencer-detection model and (ii.) the impact of the imperfect automatic prediction on emotion and appraisal classification. We show that there is a substantial drop in the pipeline model in contrast to using gold annotations, which motivates future joint modeling work.

2 Related Work

Computational emotion classification is commonly grounded in theories of basic emotions, i.e., Ekman (1992) or Plutchik (2001), while regression models often handle emotions as tuples of continuous values within a vector space, for instance of valence, arousal, and dominance (Russell and Mehrabian, 1977). Emotion intensity prediction combines both classification and regression tasks by assigning not only an emotion category but a corresponding intensity score as well (Mohammad and Bravo-Marquez, 2017). In appraisal theories, emotions depend on the emoter's cognitive evaluation of the event (Smith and Ellsworth, 1985; Scherer et al., 2001) and are either defined by it directly or are understood to emerge out of it, depending on the respective theory (Scarantino, 2016).

This cognitive appraisal can be modeled with variables that represent the emoter's event evaluation, for instance whether the emoter could anticipate the consequences of the event (*outcome probability*) or whether the emoter is responsible for what is happening (*self responsibility*) rather than another entity (*other responsibility*). The appraisal theories make an obvious aspect explicit: the emotion is developed by an entity that is part of an emotional episode. This work therefore puts emphasis not only on a cause or expression of an emotion, but also by whom it is perceived.

Emotion classification received substantial attention in a variety of domains like social media posts (Mohammad and Bravo-Marquez, 2017; Stranisci et al., 2022; i.a.), news headlines (Bostan et al., 2020a) or literary texts (Alm et al., 2005). Most work focused on the emotions from a single perspective. Semantic role labeling does consider more than one perspective, but is primarily focused on the relations between experiencers, targets, and causes (Bostan et al., 2020a; Mohammad et al., 2014; Kim and Klinger, 2018a). The work on emotion experiencer detection is a more direct access to the emotion experiencer (Wegge et al., 2022; Troiano et al., 2022). In comparison to emotion role labeling, that is a simplification that enables a more straight-forward modeling. These modeling differences are similar to representing aspect-based sentiment analysis as an aspect classification task rather than finding full graph representations of evaluative phrases and mentioned aspects (compare the two shared task setups described by Barnes et al., 2022; Pontiki et al., 2014).

Appraisal theories already motivated some NLP research (Troiano et al., 2023; Hofmann et al., 2020; Stranisci et al., 2022), but only recently, Troiano et al. (2022) investigate all potential perspectives involved in an event with their x-enVENT corpus, based on self-reported event descriptions (Troiano et al., 2019). The corpus is annotated with potential emoters, their respective emotions and 22 appraisals (score from 0–5 for each dimension). Wegge et al. (2022) proposed first models to assign emotions and appraisals to experiencer mentions, but did rely on the experiencer annotations. Therefore, it is still an open research question what the challenges of emotion experiencer detection are; the gap that we aim at filling with this paper.

3 Methods

Our methods consists of a pipeline of (a) experiencer detection followed by (b) experiencer-aware emotion/appraisal detection. For the second step, we follow Wegge et al. (2022) who purely relied on gold annotations for the first step.

The experiencers consist of sequences of tokens within a text (we assume experiencer-spans to be non-overlapping). The writer's perspective is represented with such annotation on a special token prefix `writer`. One text can contain multiple experiencer spans. Each experiencer gets assigned a set of emotion labels (6 Ekman emotions + other, no emotion, and shame) and a set of up to 22 appraisal dimensions (see Table 3 for a list of classes).

Our pipeline consists of two steps: (i.) the detec-

tion of experiencers and (ii.) the prediction of emotions/appraisal dimensions for each experiencer.

Models. For detecting the experiencer-spans, we fine-tune a transition-based named entity recognition model (NER) from the spaCy library (Honnibal et al., 2020) on the x-enVENT corpus (Troiano et al., 2022). The data set consists of 720 instances which we split into 538 for training (of which we use 61 for validation) and 107 for testing. We omit 14 instances that contain overlapping spans.[2]

Our goal is to ensure comparability with previous work on experiencer-specific emotion and appraisal classification. Therefore, we apply the same models as Wegge et al. (2022), by fine-tuning Distil-RoBERTa (Liu et al., 2019, using Hugging Face's transformers library, Wolf et al., 2020) with a multi-output classification head to jointly predict all emotion labels (see their paper for implementation details). Experiencer-spans are encoded via positional indicators in the text (cf. Zhou et al., 2016). We differ from the previous approach in formulating the prediction of appraisal dimensions as classification instead of regression to have a straight-forward access to an evaluation of the overall pipeline in which additional experiencers might appear that are not available in the gold annotation. To this end, we use a threshold of 4 to discretize the continuous appraisal scores. The appraisal classification head is analogous to the one for emotions.[3]

Evaluation. We evaluate the performance of our pipeline by calculating the F_1 in two settings. In the *strict* evaluation, only exact matches of token spans make true positives. In the *relaxed* setting, we additionally accept partial matches with at least one token overlap as true positives.

We apply the experiencer-specific classifiers to the experiencer-spans detected in the first pipeline component instead of the gold spans. We consider this in the calculation of F_1 by treating every predicted emotion or appraisal label as a false positive if the associated experiencer-span has no correspondence in the gold data (we accept overlapping spans). Analogously, if a gold experiencer-span was not recognized by the experiencer-span detector, we consider each gold emotion and appraisal label that was associated with that span a false negative. We compare our results against the performance values on gold-annotated experiencer spans.

[2]We use the default spaCy configuration, learning rate 0.001, weight decay, dropout 0.1, Adam optimizer.

[3]Our code is available at https://www.ims.uni-stuttgart.de/data/appraisalemotion.

	P		R		F_1	
	s	r	s	r	s	r
incl. WRITER	90	93	77	80	83	86
excl. WRITER	74	82	50	56	60	66

Table 1: Span-prediction results (s: strict; r: relaxed).

	GOLD SPANS			PIPELINE			
Emotion	P	R	F_1	P	R	F_1	ΔF_1
anger	73	53	61	77	45	57	−4
disgust	76	81	79	64	56	60	−19
fear	82	60	69	68	57	62	−7
joy	48	82	60	49	69	57	−3
no emotion	54	79	64	47	47	47	−17
other	33	5	9	50	5	9	±0
sadness	61	77	68	57	65	61	−7
shame	57	73	64	54	59	56	−8
Macro avg.	49	66	56	40	62	49	−7
Micro avg.	55	72	62	43	67	52	−10

Table 2: The experiencer-specific emotion classifier is evaluated on expert-annotated (GOLD SPANS) and automatically detected (PIPELINE) experiencer-spans.

4 Results

We report results for both pipeline components.

4.1 Experiencer-Span Detection

Table 1 reports the precision, recall and F_1 of the span-detector for all non-writer experiencers (excl. WRITER) as well as to all experiencer-spans (incl. WRITER). Recognizing the `writer` token as an experiencer is trivial (F_1 =1.0).

As to be expected, the performance of the span-predictor is lower in the evaluation setup that considers only the non-writer experiencers. There is a considerable difference in the exact and relaxed evaluation setup, which shows that the model sometimes only finds a subset of the experiencer tokens. The task is challenging: while the precision is acceptable, only half of the experiencers are found. This is to some degree a result of the annotation of the data – the corpus authors tasked the annotators to only label the first occurrence of each mention of an experiencer in a text – a property that is challenging to be grasped automatically.

4.2 Emotion and Appraisal Classification

Table 2 reports the results of the emotion classifier applied to the automatically predicted experiencer-spans (PIPELINE setting) as well as the baseline results (GOLD SPANS) that were obtained on expert-annotated experiencer-spans. Across almost all

	GOLD SPANS			PIPELINE			
Appraisal	P	R	F_1	P	R	F_1	ΔF_1
suddenness	67	65	66	64	59	62	−3
familiarity	0	0	0	0	0	0	±0
pleasantness	8	87	83	78	78	78	−9
understand	80	100	89	77	82	80	−9
goal relev.	38	33	0.35	29	22	25	−10
self resp.	64	95	76	61	70	65	−11
other resp.	73	73	73	64	60	62	−11
sit. resp.	52	79	62	45	68	54	−8
effort	67	29	40	20	14	17	−23
exert	0	0	0	0	0	0	±0
attend	50	17	25	50	17	25	±0
consider	72	66	69	65	57	61	−8
outcome prob.	55	75	63	51	62	56	−7
expect. discrep.	72	63	67	67	56	61	−6
goal conduc.	59	62	60	60	57	59	−1
urgency	0	0	0	0	0	0	±0
self control	58	89	70	58	64	61	−9
other control	75	55	63	63	45	52	−11
sit. control	52	78	62	46	67	55	−7
adj. check	75	75	75	72	53	61	−14
int. check	33	12	18	25	12	17	−1
ext. check	0	0	0	0	0	0	±0
Macro avg.	46	64	54	42	48	45	−9
Micro avg.	58	86	69	54	69	61	−8

Table 3: Appraisal classification results of the appraisal classifier evaluated on expert-annotated (GOLD SPANS) and automatically detected (PIPELINE) experiencer-spans.

emotion categories, the PIPELINE classifier performs worse than the GOLD SPANS baseline, which is expected as the evaluation method penalizes erroneously detected experiencer-spans. However, the drop in performance differs between emotions. For *anger*, *joy*, *sadness*, *fear*, *shame* the difference is less than 10pp F_1– for these emotions, experiencers can be found more reliably than for *disgust* (19pp) or *no emotion* (17pp).

The notable decrease in performance for *no emotion* is in line with the observation that predicting non-writer spans is more challenging than predicting writer-spans. From all spans annotated with *no emotion*, 84% are non-writer spans. However, the classification performance also drops for emotion classes that are frequently annotated in writer-spans; The pipeline classifier shows its biggest decrease in performance (19pp) for *disgust*, although 76% of all spans annotated with *disgust* are writer-spans. This is due to the span-predictor's low recall: a low number of recognized spans leads to a higher number of false negatives for all emotion classes associated with these spans. The biggest increase in FN introduced by the span-predictor is observed for *disgust* (71%), the lowest for *other* (21%).

Analogous to the emotion classifier, we observe a decrease in performance for the appraisal predictor, reported in Table 3. Again, there is a substantial difference in the drop of performance, with *effort* and *adjustment check* showing the highest loss (23pp and 14pp, respectively) and *goal conduciveness*, *internal check*, *attend* being the lowest (1pp or no difference). Both *effort* and *adjustment check* appear only seldom in writer-spans (33% each), while *goal conduciveness*, *internal check* and *attend* appear more often in writer spans (between 39% and 44%) and are less prone to unrecognized spans (44%/40% of FN are introduced through missing spans for *goal conduciveness/attention*, 29% for *internal check*; cf. Table 7). However, the individual differences are less pronounced than for the emotion classification results, due to the sparseness of some appraisal dimensions.

We show more detailed emotion/appraisal-specific statistics of writer spans and false negatives in the appendix.

5 Discussion and Conclusion

In this paper, we presented the first evaluation of experiencer detection in text and the impact of these predictions on the emotion/appraisal classification. We found that experiencer detection is challenging but the results are promising.

The emotion/appraisal detection interacts with the span prediction task. This indicates that a joint model that can explore interactions between experiencer and emotion/appraisal dimensions might work better than the pipeline setting. Such model is however not trivial to be build, because the emotion/appraisal classification depends on a variable number of spans. Possible approaches include a purely token-level classification task or multiple sequence labeling setups. Such engineering attempts can also find inspiration in emotion–cause pair extraction models (e.g., Yuan et al., 2020).

Our work also motivates other follow-up studies, namely to extend the experiments to corpora that are fully annotated with emotion role graphs (Campagnano et al., 2022), from which some contain experiencer annotations (Bostan et al., 2020b; Kim and Klinger, 2018b; Mohammad et al., 2014). We expect our approach to show improvements over full graph predictions for the subtask of experiencer-specific emotion prediction due to fewer model parameters.

Acknowledgements

This research is funded by the German Research Council (DFG), project "Computational Event Analysis based on Appraisal Theories for Emotion Analysis" CEAT (project number KL 2869/1-2).

References

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.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020a. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020b. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

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

Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.

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

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

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

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:1907.11692.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

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

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

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. 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.

Andrea Scarantino. 2016. The philosophy of emotions and its impact on affective science. In *Handbook of emotions*, chapter 4, pages 3–48. Guilford Press New York, NY.

Klaus R Scherer, A Schorr, and T Johnstone. 2001. *Appraisal considered as a process of multi-level sequential checking*, volume 92. Oxford University Press.

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

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Enrica Troiano, Laura Oberländer, Maximilian Wegge, and Roman Klinger. 2022. x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of The 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72.

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.

Maximilian Wegge, Enrica Troiano, Laura Ana Maria Oberlaender, and Roman Klinger. 2022. Experiencer-specific emotion and appraisal prediction. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 25–32, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. 2020. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3568–3573, Online. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

A Distributions Emotion Spans and False Negatives

	Writer		Non-Writer	
Emotion	%	#	%	#
anger	.61	204	.39	132
disgust	.76	66	.24	21
fear	.61	135	.39	85
joy	.45	118	.55	147
no emotion	.16	43	.84	226
other	.50	59	.50	58
sadness	.59	249	.41	174
shame	.64	209	.36	116

Table 4: Frequency (absolute and relative) of writer and non-writer spans annotated with a given emotion.

	total	due to non-recogn. span	
Emotion	#	#	%
anger	28	7	.25
disgust	7	5	.71
fear	13	4	.31
joy	12	7	.58
no emotion	23	14	.61
other	19	4	.21
sadness	21	9	.43
shame	21	10	.48

Table 5: Number of false negative emotion predictions (relative and absolute) that were introduced due to the experiencer predictor not recognizing the span.

B Distributions Appraisal Spans and False Negatives

	Writer		Non-Writer	
Appraisal	%	#	%	#
suddenness	.62	333	.38	202
familiarity	.9	3	.91	30
pleasantness	.53	99	.47	87
understand	.58	642	.42	460
goal relev.	.47	40	.53	45
self resp.	.47	244	.53	273
other resp.	.50	256	.50	251
sit. resp.	.70	140	.30	59
effort	.33	25	.67	51
exert	.38	3	.62	5
attend	.44	18	.56	23
consider	.54	140	.46	119
outcome prob.	.54	211	.46	177
expect. discrep.	.60	380	.40	252
goal conduc.	.44	76	.56	96
urgency	.40	10	.60	15
self control	.39	136	.61	217
other control	.50	199	.50	203
sit. control	.67	135	.33	67
adj. check	.33	145	.67	301
int. check	.39	26	.61	41
ext. check	.21	9	.79	34

Table 6: Frequency (absolute and relative) of writer-/non-writer spans annotated with a given appraisal class.

	total	due to non-recogn. span	
Appraisal	#	#	%
suddenness	30	11	.37
familiarity	3	1	.33
pleasantness	5	3	.60
understand	28	28	1
goal relev.	7	2	.29
self resp.	25	23	.92
other resp.	28	13	.46
sit. resp.	6	3	.50
effort	6	3	.50
exert	2	1	.50
attend	5	2	.40
consider	15	5	.33
outcome prob.	20	13	.65
expect. discrep.	41	16	.39
goal conduc.	9	4	.44
urgency	3	1	.33
self control	23	19	.83
other control	33	12	.36
sit. control	6	3	.50
adj. check	36	20	.56
int. check	7	2	.29
ext. check	6	4	.67

Table 7: Number of false negative appraisal predictions (relative and absolute) that were introduced due to the experiencer predictor not recognizing the span.