Experiencers, Stimuli, or Targets: Which Semantic Roles Enable Machine Learning to Infer the Emotions?

Laura Oberländer, Kevin Reich, and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Emotion recognition is predominantly formulated as text classification in which textual units are assigned to an emotion from a predefined inventory (e.g., fear, joy, anger, disgust, sadness, surprise, trust, anticipation). More recently, semantic role labeling approaches have been developed to extract structures from the text to answer questions like: "who is described to feel the emotion?" (experiencer), "what causes this emotion?" (stimulus), and at which entity is it directed?" (target). Though it has been shown that jointly modeling stimulus and emotion category prediction is beneficial for both subtasks, it remains unclear which of these semantic roles enables a classifier to infer the emotion. Is it the experiencer, because the identity of a person is biased towards a particular emotion (X is always happy)? Is it a particular target (everybody loves X) or a stimulus (doing X makes everybody sad)? We answer these questions by training emotion classification models on five available datasets annotated with at least one semantic role by masking the fillers of these roles in the text in a controlled manner and find that across multiple corpora, stimuli and targets carry emotion information, while the experiencer might be considered a confounder. Further, we analyze if informing the model about the position of the role improves the classification decision. Particularly on literature corpora we find that the role information improves the emotion classification.

1 Introduction

Emotion analysis is now an established research area which finds application in a variety of different fields, including social media analysis (Purver and Battersby, 2012; Wang et al., 2012; Mohammad and Bravo-Marquez, 2017; Ying et al., 2019, i.a.), opinion mining (Choi et al., 2006, i.a.), and computational literary studies (Alm et al., 2005; Kim and Klinger, 2019a; Haider et al., 2020; Zehe et al., 2020, i.a.). The most prominent task in emotion analysis is emotion categorization, where text receives assignments from a predefined emotion inventory, such as the fundamental emotions of *fear*, *anger*, *joy*, *anticipation*, *trust*, *surprise*, *disgust*, and *sadness* which follow theories by Ekman (1999) or Plutchik (2001). Other tasks include the recognition of affect values, namely valence or arousal (Posner et al., 2005) or analyses of event appraisal (Hofmann et al., 2020; Scherer, 2005).

More recently, categorization (or regression) tasks have been complemented by more fine-grained analyses, namely emotion stimulus detection and role labeling, to detect which words denote the experiencer of an emotion, the emotion cue description, or the target of an emotion. These efforts lead to computational approaches of detecting stimulus clauses (Xia and Ding, 2019; Wei et al., 2020; Gao et al., 2017) and emotion role labeling and sequence labeling (Mohammad et al., 2014; Bostan et al., 2020; Kim and Klinger, 2018; Ghazi et al., 2015; Zehe et al., 2020), with different advantages and disadvantages we discuss in Oberländer and Klinger (2020).

Further, this work led to a rich set of corpora with annotations of different subsets of roles. An example of a sentence annotated with semantic role labels for emotion is "$[\text{John}]_{\text{EXPERIENCER}}$ $[\text{hates}]_{\text{CUE}}$ $[\text{cars}]_{\text{TARGET}}$ because they $[\text{pollute the environment}]_{\text{STIMULUS}}$." A number of English-language resources are available: Ghazi et al. (2015)

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 119–128
Barcelona, Spain (Online), December 13, 2020.

	Whole Instance		Stimulus		Cue		Target		Exp.	
Dataset	#	∅len	#	∅len	#	∅len	#	∅len	#	∅len
Emotion-Stimulus, Ghazi et al. (2015)	2414	20.60	820	7.29	—	—	—	—	—	—
ElectoralTweets, Mohammad et al. (2014)	4056	19.14	2427	6.25	2930	5.08	2824	1.71	29	1.76
GoodNewsEveryone, Bostan et al. (2020)	5000	13.00	4798	7.29	4736	1.60	4474	4.86	3458	2.03
REMAN, Kim and Klinger (2018)	1720	72.03	609	9.33	1720	3.82	706	5.35	1050	2.04
Emotion Cause Analysis, Gao et al. (2017)	2558	62.24	2485	9.52	—	—	—	—	—	—

Table 1: Datasets with annotations of roles. # refers to the number of total instances. ∅len shows the average length of each role filler in each dataset in the number of tokens.

manually construct a dataset following FrameNet's emotion predicate and annotate the stimulus as its core argument. Mohammad et al. (2014) annotate Tweets for emotion cue phrases, emotion targets, and the emotion stimulus. In our previous work (Bostan et al., 2020) we publish news headlines annotated with the roles of emotion experiencer, cue, target, and stimulus. Kim and Klinger (2018) annotate sentence triples taken from literature for the same roles. A popular benchmark for emotion stimulus detection is the Mandarin corpus by Gui et al. (2016). Gao et al. (2017) annotate English and Mandarin texts in a comparable way on the clause level (*Emotion Cause Analysis*, ECA).

In this paper, we utilize role annotations to understand their influence on emotion classification. We evaluate which of the roles' contents enable an emotion classifier to infer the emotions. It is reasonable to assume that the roles' content carries different kinds of information regarding the emotion: One particular experiencer present in a corpus might always feel the same emotion; hence, be prone to a bias the model could pick up on. The target or stimulus might be independent of the experiencer and be sufficient to infer the emotion. The presence of a target might limit the set of emotions that can be triggered. Finally, as some of the corpora contain cue annotations, we assume that these are the most helpful to decide on the expressed emotion, as they typically have explicit references towards concrete emotion names.

2 Experimental Setting

In the following, we describe our experiments to understand which of the datasets' annotated roles contribute to the emotion classification performance.

Datasets. We base our experiments on five available datasets that are annotated for at least one of the roles of an experiencer, stimulus, target, or cue. The dataset by Ghazi et al. (2015) is one of the earliest we are aware of that contains stimulus annotations. They annotate based on FrameNet's *emotion-directed* frames that have a stimulus argument in the data (we refer to their corpus as *Emotion-Stimulus*, ES). Similarly early work is the Twitter corpus by Mohammad et al. (2014) (*ElectoralTweets*, ET). They also follow the emotion frame semantics definition but use data concerning the 2012 U.S. election. Therefore, their resource may be considered more diverse in language but more consistent in its domain than ES. More recently, Bostan et al. (2020) published an annotation of news headlines (*GoodNewsEveryone*, GNE). While they do not limit their corpus on a domain, they use a comparably narrow time window to retrieve the data and sample according to the inclusion of emotion words and popularity on social media. Kim and Klinger (2018, *REMAN*) and Gao et al. (2017, *Emotion Cause Analysis*, ECA) use literature data, which might be considered the most challenging for emotion analysis (for ECA, we use the English subset only).

As Table 1 shows, the literature data (REMAN, ECA) has the longest instances and also the longest stimulus annotations. The other resources have less than one third of their length in tokens, with GNE being the shortest. However, the overall annotation length does not differ dramatically. Cue, target, and experiencer annotations are only available in three out of five corpora (ET, REMAN, and GNE)[1].

Model Configuration. Our goal is to analyze the importance of different roles for the emotion classification. We use two different models, namely a bidirectional long short-term memory network (Hochreiter

[1]For ET, 90% of the annotated experiencers are the authors of the tweets without corresponding span annotation.

and Schmidhuber, 1997) with pretrained 300-dimensional GloVe embeddings[2] and a transformer-based model, RoBERTa (Liu et al., 2019). Both models take as input the text sequence and output the emotion class, where the concrete set of emotion labels depends on the dataset.

The models have different advantages and disadvantages in our experimental setting. The bi-LSTM with non-contextualized word embeddings might be more appropriate to be used in our setting in which we manipulate the input token sequence (see below). The transformer might benefit from the rich contextualized pretraining, which is particularly relevant given that the annotated corpora are of comparably limited size (in the context of deep learning)[3].

Setting and Hypotheses. We apply these models in several settings (illustrated in Table 2), which differ in the availability of information from the roles, namely (1), *As-Is*: This is the standard setting: The classifier has access to the whole text. (2), *Without* the text of the particular roles. (3), *Only* with the text of a particular role, masking the text that does not belong to it. Finally, (4), we keep the information available as is, but besides inform the model about the *Position* of the role. The latter is realized by adding positional indicators, inspired by Kim and Klinger (2019b) who showed the use of positional indicators for emotion relation classification[4].

Setting	Model Input							
As-Is	John	hates	cars	because	they	pollute	the	environment
Only Stim.	X	X	X	X	X	pollute	the	environment
Only Exp.	John	X	X	X	X	X	X	X
Only Tar.	X	X	cars	X	X	X	X	X
Without Stim.	John	hates	cars	because	they	X	X	X
Without Exp.	X	hates	cars	because	they	pollute	the	environment
Without Tar.	John	hates	X	because	they	pollute	the	environment
Pos. Stim.	John	hates	cars	because	they	⌊pollute	the	environment⌉
Pos. Exp.	⌊John⌉	hates	cars	because	they	pollute	the	environment
Pos. Tar.	John	hates	⌊cars⌉	because	they	pollute	the	environment

Table 2: Illustration of the experimental settings. X, ⌊, ⌉ denote special tokens added to the input according to each setting.

For roles that carry information relevant for emotion classification, we expect the *Without* setting to show a drop in performance compared to the *As-Is* setting. In such cases, the *Only* setting might show comparable performance, and the *Position* setting would show further improvements. When the role is a confounder, the performance in the *Without* setting is expected to be increased over the *As-Is* setting.

The label set depends on each of the datasets. For ES, we use the emotion labels *anger*, *disgust*, *fear*, *joy*, *no emotion*, *sadness*, and *surprise*; for ECA, we use *anger*, *sadness*, *disgust*, *joy*, *fear*, *surprise*, and *no emotion*. For GNE and ET, we merge the categories according to the rules described for ET by Bostan and Klinger (2018) and keep the primary emotions described in Plutchik's wheel. For REMAN, we group similarly and keep *anger*, *disgust*, *fear*, *joy*, *anticipation*, *surprise*, *sadness*, *trust*, and *no emotion*. ECA has a low number of instances annotated with multiple labels, which we ignore to keep all tasks as single-label classification. REMAN has emotion annotations only for the middle sentence in each triple. Thus we include only these middle segments in our experiments.

The results are based on a random split of each dataset into train, validation, and test (0.8, 0.1, 0.1). We report macro-averages across 10 runs for the bi-LSTM and 5 runs for RoBERTa.

3 Results

In the following, we discuss the results of the bi-LSTM model in detail and then point to differences to those of the transformer-based approach. Table 3 shows the results of our experiments for the bi-LSTM-

[2]We use 42B tokens, pretrained on CommonCrawl (Pennington et al., 2014), `https://nlp.stanford.edu/projects/glove/`

[3]The hyperparameters and details for the models are as follows. For the bi-LSTM, we set a dropout and recurrent dropout of 0.3 and optimize with Adam (Kingma and Ba, 2015), with a base learning rate of 0.0003, L2 regularization, on a batch size of 32, with early stopping with patience of 3, and initialization with Kaiming (He et al., 2015). We train for up to 100 epochs for the bi-LSTM model and 10 for the transformer-based model. Both models fine-tune their input representations during training. The hyperparameters of the model are optimized for ECA. For the bi-LSTM, we use AllenNLP (Gardner et al., 2018) and for the transformer the Hugging Face library (Wolf et al., 2019) (following the training procedure described by Devlin et al. (2019)). The code of our project is available at `http://www.ims.uni-stuttgart.de/data/emotion-classification-roles`.

[4]We experimented with adding two channels in the input embeddings which mark the tokens outside a role annotation with a 1 in one channel and the tokens which belong to the role annotation with a 1 in a second channel. The results were inferior to using positional indicators.

		As-Is			Without			Only			Position		
Dataset	Role	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ECA	Stimulus	41	39	39	**48**	**48**	**48**	30	25	23	**52**	**51**	**51**
ES	Stimulus	93	89	90	**94**	89	90	65	23	18	**95**	**90**	**92**
REMAN	Cue	47	27	25	**61**	14	8	**53**	14	8	42	23	19
	Stimulus				41	22	19	**91**	11	4	44	14	12
	Experiencer				29	23	19	**60**	11	6	32	25	21
	Target				19	12	9	**57**	10	3	31	23	21
ET	Cue	51	26	25	**63**	23	22	**79**	18	15	**62**	25	23
	Stimulus				50	23	21	**59**	15	11	**57**	**27**	**27**
	Experiencer				**53**	26	24	**80**	12	7	48	23	20
	Target				**56**	**27**	**26**	**64**	16	14	**65**	24	21
GNE	Cue	34	14	12	**62**	13	10	**93**	10	5	**64**	13	10
	Stimulus				**93**	10	5	**85**	11	7	**60**	13	9
	Experiencer				**55**	**18**	**15**	**93**	10	5	**63**	**15**	**13**
	Target				**86**	12	8	**93**	10	5	**62**	14	11

Table 3: Results of our bi-LSTM based model for emotion classification, with access to all tokens (*As-Is*), *Only* to the respective role, to all tokens *Without* the respective role, and all tokens together with the *Position*al indicators of the role added. All F_1 scores are macro averaged, the scores which are higher than in the As-Is setting are bold.

based model. Intuitively, we would expect the *As-Is* setting to outperform both the *Without* and *Only* settings because there is more information available to the model. Conversely, because information is added in *Position*, we expect it to outperform the *As-Is* setting. As we see in column *As-Is*, the scores for the emotion classification task differ substantially, even when all available information is shown to the model. In the *Without* setting, we see that removing information can sometimes help a model improve its decision. For instance, when we mask the labels of the respective role, we observe a performance increase for the experiencer role in GNE, which could potentially point to an unwanted bias for particular experiencers in this corpus. This is also the case for the stimulus role in ECA and the target role in ET.

As expected, an important role for emotion classification is the cue. In REMAN, the performance drops the most when the classifier does not see the cue span and gains the most when only the cue is available. For all other corpora, the cue role is not as important, but performance still shows a drop when it is not available (*Without*). Similarly, for all datasets except ECA, the performance drops when the stimulus is not shown. On the other hand, the stimulus alone is insufficient to infer the emotion with competitive performance. Noteworthy here is the corpus ES, in which the performance drop is particularly high.

These results show that the information contained in different roles is of varying importance and depends on the data's source and domain. In the setting *Position*, we leave all information accessible to the model but add positional indicators for the investigated role to the input for emotion classification. We see improvements in most cases, except REMAN, for which adding the positional information hurts the classification for all roles. This result could be because REMAN has very long annotation spans. Both ECA and ES show an improvement for their annotated role (stimulus). For ET, an increase in performance is shown when additional knowledge about the stimulus position is given, and for GNE, a slight improvement is shown when the model is given the experiencer's position information.

Table 4 shows the results of the transformer-based model evaluated in the same settings. As expected, the model shows performance improvements across all datasets in comparison to the bi-LSTM model. In the *As-Is* setting, we see a substantial increase in performance for REMAN. This result can be explained by the fact that the pretrained large language model has seen more literary English than the embeddings used as pretrained input to the bi-LSTM. GNE and ET scores are also improved across the roles. In the *Without* setting, we do not see the same patterns as for the bi-LSTM based model; the scores when hiding the stimulus for ECA, the target for ET, and experiencer for GNE do not increase over the scores of the *As-Is* setting.

This might have two reasons: On one hand, it is less likely to improve upon already high values

		As-Is			Without			Only			Position		
Dataset	Role	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ECA	Stimulus	68	70	68	4	17	7	4	17	7	**73**	**73**	**73**
ES	Stimulus	99	98	98	99	**99**	**99**	3	14	5	99	97	98
REMAN	Cue	67	60	66	3	12	5	3	12	5	**79**	**77**	**78**
	Stimulus				45	54	47	2	11	4	43	47	43
	Experiencer				60	60	56	2	11	4	62	56	56
	Target				46	42	42	2	11	3	44	45	42
ET	Cue	34	33	34	32	29	30	5	12	7	31	30	30
	Stimulus				**37**	33	34	9	15	11	33	32	32
	Experiencer				34	**34**	34	5	12	7	34	**34**	34
	Target				**35**	**34**	34	5	12	7	**35**	33	33
GNE	Cue	32	31	31	32	27	27	3	10	5	29	28	28
	Stimulus				7	11	7	24	23	23	**35**	**33**	**34**
	Experiencer				31	30	30	3	10	5	**35**	**32**	**33**
	Target				3	10	5	3	10	5	**35**	31	**32**

Table 4: Results of our transformer based model (RoBERTa) for emotion classification.

when changing the model configuration. On the other hand, and more interestingly, it might be that the contextualized embeddings compensate for missing information. Interestingly for the *Position* setting, the results are improving on all datasets, and REMAN gains from the cue's positional indicators. The dataset that stands out in this setting is ET, for which we see a slight decrease in performance across all roles available. The *Only* setting shows that the stimulus captures most of the emotion information for GNE and ET. The result for GNE is due to the particularly lengthy stimuli spans that sometimes stretch over the whole instance.

4 Conclusion and Future Work

Our experiments show that the importance of semantic roles for emotion classification differs between datasets and roles: The stimulus and cue are critical for classification, which correspond to the direct report of a feeling and the description that triggered an emotion. This result is shown in the drop in performance when removing these roles. This information is not redundantly available outside of these arguments.

It is particularly beneficial for the model's performance to have access to the position of cues and stimuli. This suggests that the classifier learns to tackle the problem differently when this information is available, especially so for ECA and ES – the cases in which literature has been annotated and the instances are comparably long.

The bi-LSTM model indicates that the experiencer role is a confounder in GNE. The performance can be increased when the model does not have access to its content. Similar results are observed for ET, in which the target role is a confounder. However, these results should be taken with a grain of salt given that they are not confirmed while switching to the transformer-based model. The differences in results between the bi-LSTM and the transformer also motivate further research, as they suggest that the contextualized representation might compensate for missing information, and is, therefore, more robust.

Finally, our results across both models and multiple datasets indicate that emotion classification approaches indeed benefit from semantic roles' information by adding the positional information. Similarly to targeted and aspect-based sentiment analysis, this motivates future work, in which emotion classification and role labeling should be modelled jointly. In this case, it can also be interesting to investigate what happens when the positional indicators are added to all roles jointly.

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft (project SEAT, KL 2869/1-1). We thank Enrica Troiano and Heike Adel for fruitful discussions and the anonymous reviewers for helpful comments.

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.

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

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

Yejin Choi, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney, Australia, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of NTCIR-13 ECA task. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pages 361–366, Tokyo, Japan, December.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, July. Association for Computational Linguistics.

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

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November. Association for Computational Linguistics.

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, May. European Language Resources Association.

K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

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*.

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

Evgeny Kim and Roman Klinger. 2019a. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy, August. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*.

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 preprint arXiv:1907.11692*.

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

Saif 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, June. Association for Computational Linguistics.

Laura Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for english emotion stimulus detection. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics (*SEM 2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

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

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

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

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491, Avignon, France, April. Association for Computational Linguistics.

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

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *SocialCom/PASSAT*, pages 587–592. IEEE.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online, July. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China, November. Association for Computational Linguistics.

Albin Zehe, Julia Arns, Lena Hettinger, and Andreas Hotho. 2020. Harrymotions – classifying relationships in harry potter based on emotion analysis. In *5th SwissText & 16th KONVENS Joint Conference*.

Appendix

Qualitative Discussion of Examples

We analyze a subset of interesting cases from the results section in the following, to better understand why removing stimuli from ECA improves the results and further why the same can be observed on ET for targets.

We show examples for these cases in Table 5. We observe in instances correctly classified in the *Without* setting that removing the stimulus makes the classification task easier by removing potential sources for overfitting: The remaining tokens contain the explicit cue, even though they are not explicitly annotated for ECA. For instance, in "[his angry outbreak]$_{\text{STIMULUS}}$ [saddened]$_{\text{CUE}}$ [me]$_{\text{EXPERIENCER}}$", we see that removing the stimulus which also contains a reference to another emotion, the task of picking the most dominant emotion from the remaining tokens is more straight-forward.

This holds similarly for other examples in ECA, in which the stimulus describes an event that could also be evaluated as scary; however, the experiencer mentions that he is surprised ("To my surprise").

		Label					
			without				
Dataset	Gold	All	Stim.	Exp.	Cue	Targ.	Text
GNE	**J**	Su	**J**	Su	Su	Su	[Djokovic]$_{\text{EXPERIENCER}}$ [happy]$_{\text{CUE}}$ [to carry on cruising]$_{\text{STIMULUS}}$
GNE	**J**	Su	**J**	Su	A	Su	[Trump]$_{\text{EXPERIENCER}}$ [upbeat]$_{\text{CUE}}$ [on potential for US-Japan trade deal.]$_{\text{STIMULUS}}$
ECA	**J**	F	**J**	–	–	–	["Michie Reetchie"]$_{\text{STIMULUS}}$, said Xavier, and again he burst into laughter that choked further speech. He controlled himself and laid his finger on his vein.
ECA	**Su**	F	**Su**	–	–	–	One morning Pop sent me down to the river to catch some fish for breakfast. To my surprise [there was a canoe in the water and there was no one in]$_{\text{STIMULUS}}$. Immediately I jumped into the river and brought the canoe to the side.
ECA	**F**	S	**F**	–	–	–	I did not answer, fearing [to tell him that I had been awake watching him]$_{\text{STIMULUS}}$
ECA	**A**	S	**A**	–	–	–	A massy stone and shook the ranks of Troy, as when in anger [against long - screaming cranes]$_{\text{STIMULUS}}$ a watcher of the field leaps from the ground in swift hand whirling round his head the sling and speeds the stone against them scattering.
ECA	**D**	A	**D**	–	–	–	[A year after being fired from his job]$_{\text{STIMULUS}}$ he has a lot of resentment towards his former boss.
ET	**D**	T	D	T	S	**D**	Three words to describe the entire [#GOP convention]$_{\text{TARGET}}$ [Mean and demeaning.]$_{\text{CUE}}$
ET	**A**	D	D	D	D	**A**	[#Republicans]$_{\text{TARGET}}$ are a joke . [Clint Eastwood]$_{\text{STIMULUS}}$ is their mascot ! America is in trouble if [these idiots]$_{\text{CUE}}$ win ! #RNC
ET	**J**	T	T	J	T	**J**	[Obama Voter]$_{\text{TARGET}}$ [Says Vote for Obama]$_{\text{STIMULUS}}$ [YES WE CAN AGAIN !]$_{\text{CUE}}$
ET	**J**	Ant	T	T	T	**J**	[So excited]$_{\text{CUE}}$ to vote this upcoming [election]$_{\text{TARGET}}$ [finally exercising my right to choose our next president]$_{\text{STIMULUS}}$ #Obama
ET	**D**	A	A	A	A	**D**	[Romney]$_{\text{TARGET}}$ is gonna put The Onion out of business . [#TheStench]$_{\text{CUE}}$
REMAN	**J**	noemo	noemo	**J**	noemo	–	And [she]$_{\text{EXPERIENCER}}$ returned the quiet but jubilant kiss that he laid upon her lips.

Table 5: Examples in which the prediction is incorrect when the model is applied on the whole instance, but it is correct when the respective role is removed. The correct prediction is marked in bold face. J: Joy, T: Trust, Su: Surprise, Ant: Anticipation, D: Disgust, F: Fear, A: Anger, S: Sadness

Detailed Results for Additional Positional Information

We have seen in the results that adding position information of the semantic roles increases the performance for both datasets which contain examples drawn from literature. This is particularly interesting for future research on jointly modelling roles and classification. Therefore, we show details per emotion class in Table 6 (only for the bi-LSTM model).

We see for the ECA dataset, that when the positional information is made accessible to the model, the classifier learns better to predict all emotion classes with a substantial improvement for anger and disgust. Similarly, ES improves over all emotions with the exception of disgust and sadness.

Data	Emotion	All			Stimulus Position		
		P	R	F_1	P	R	F_1
ECA	Anger	15	11	13	**36**	**44**	**40**
	Disgust	25	06	09	11	**11**	**11**
	Fear	56	56	56	**78**	**70**	**74**
	Joy	57	58	57	**65**	58	**61**
	Sadness	50	67	57	**57**	**72**	**64**
	Surprise	40	38	39	**63**	**53**	**58**
	Macro	40	39	38	**52**	**51**	**51**
ES	Anger	90	97	94	**92**	**98**	**95**
	Disgust	85	54	67	**100**	45	63
	Fear	97	88	93	95	**95**	**95**
	Joy	93	92	92	**100**	92	**96**
	Sadness	94	99	97	90	96	93
	Shame	100	94	97	100	**100**	**100**
	Surprise	91	95	93	88	**100**	**94**
	Macro	93	89	90	**95**	**90**	**91**

Table 6: Results per emotion for ECA and ES with and without positional stimuli information. Bold numbers indicate that their value is greater than in the As-Is setting.

Analysis of Content of Roles

Table 7 shows the most frequent tokens marked as *cue, stimulus, experiencer* or *target* over each dataset. They differ substantially per dataset and reflect well the respective source. The counts suggest a Zipfian distribution for *ElectoralTweets* (stimulus and target) and *GoodNewsEveryone* (experiencer, stimulus). This could explain the results obtained in the *Without* setting by the bi-LSTM-based model. The most common tokens annotated with the *target* role in *ElectoralTweets* also show the polarized nature of those who tweeted about the election.

Figure 1 shows the distribution of the most frequent tokens (across all roles) for the most frequent emotions of ET and GNE. The plots marked with "overall" show the prior distribution of emotions in the respective dataset. We see that for the emotion *admiration*, "president" stands out. Further we note that "Romney" is associated with *dislike* in this corpus.

For GNE we observe that the most frequent tokens are occurring less in instances annotated with *positive surprise* than overall, and more in instances annotated with *anger* (except for "Biden") showing that these tokens could be biased towards more negative emotions. This shows a bias of the dataset towards negative emotion when it comes to the most prominent tokens.

	Role	Tokens
ECA	Stim.	see (80), like (49), man (49), go (43), life (43), father (43), time (42), day (34), came (33), son (32)
ES	Stim.	see (36), way (12), find (11), left (9), people (9), prospect (8), thought (8), like (8), losing (8), work (7)
REMAN	Cue	love (32), suddenly (31), afraid (15), smile (12), beautiful (11), trust (11), pleasure (10), ugly (7), things (7), wish (6)
	Stim.	little (10), another (8), face (8), got (7), lord (7), left (7), great (7), wife (7), men (6), life (6)
	Exp.	man (23), woman (12), boy (7), old (7), isabel (6), people (6), god (5), father (5), heart (5), henry (5)
	Target	man (22), little (9), things (8), woman (8), see (8), old (7), god (6), wife (6), another (6), true (5)
ET	Cue	Obama (136), Romney (105), vote (89), like (65), Mitt (56), people (53), get (52), president (50), really (49), excited (49)
	Stim.	Obama (249), Romney (211), vote (108), Mitt (87), Barack (74), president (66), people (51), speech (40), like (40), get (35)
	Exp.	gop, anyone, presidency, clint
	Target	Obama (446), Romney (420), Mitt (146), Barack (112), People (53), president (40), election (20), debate (19), Michelle (19), Clinton (15)
GNE	Cue	killed (38), crisis (33), attacks (33), death (26), war (25), arrested (24), racist (24), help (22), new (20), fight (19)
	Stim.	Trump (279), border (68), Mueller (58), back (57), report (56), Iran (57), report (56), war (55), people (55), deal (55)
	Exp.	Trump (401), Donald (66), man (46), democrats (44), Biden (40), House (37), woman (36), police (35), Mueller (34), Sanders (33)
	Target	Trump (345), new (94), Mueller (54), House (44), border (43), people (42), democrats (41), deal (36), report (36), president (35)

Table 7: Most frequent 10 tokens with frequencies for each role and dataset.

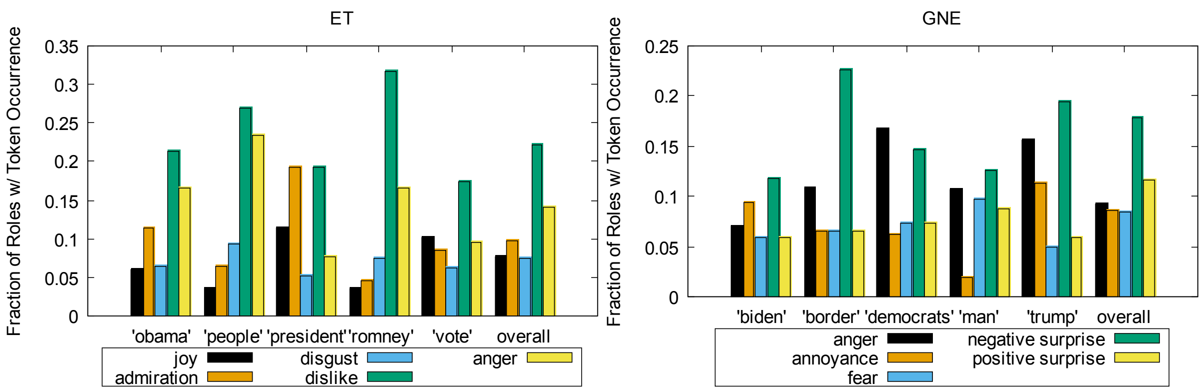

Figure 1: Emotion distribution of instances containing the respective tokens (% for the top-5 most frequent emotions for each dataset). "overall" represents the emotion distribution for those emotions across all instances.