

# Token Sequence Labeling vs. Clause Classification for English Emotion Stimulus Detection

Laura Oberländer and Roman Klinger

Institut für Maschinelle Sprachverarbeitung, University of Stuttgart

Pfaffenwaldring 5b, 70569 Stuttgart, Germany

{laura.oberlaender, roman.klinger}@ims.uni-stuttgart.de

## Abstract

Emotion stimulus detection is the task of finding the cause of an emotion in a textual description, similar to target or aspect detection for sentiment analysis. Previous work approached this in three ways, namely (1) as text classification into an inventory of predefined possible stimuli (“Is the stimulus category *A* or *B*?”), (2) as sequence labeling of tokens (“Which tokens describe the stimulus?”), and (3) as clause classification (“Does this clause contain the emotion stimulus?”). So far, setting (3) has been evaluated broadly on Mandarin and (2) on English, but no comparison has been performed. Therefore, we analyze whether clause classification or token sequence labeling is better suited for emotion stimulus detection in English. We propose an integrated framework which enables us to evaluate the two different approaches comparably, implement models inspired by state-of-the-art approaches in Mandarin, and test them on four English data sets from different domains. Our results show that token sequence labeling is superior on three out of four datasets, in both clause-based and token sequence-based evaluation. The only case in which clause classification performs better is one data set with a high density of clause annotations. Our error analysis further confirms quantitatively and qualitatively that clauses are not the appropriate stimulus unit in English.

## 1 Introduction

Research in emotion analysis from text focuses on classification, i.e., mapping sentences or documents to emotion categories based on psychological theories (e.g., Ekman (1992), Plutchik (2001)). While this task answers the question *which* emotion

is expressed in a text, it does not detect the textual unit, which reveals *why* the emotion has been developed. For instance, in the example “*Paul is angry because he lost his wallet.*” it remains hidden that *lost his wallet* is the reason for experiencing the emotion of anger. This stimulus, e.g., an event description, a person, a state of affairs, or an object enables deeper insight, similar to targeted or aspect-based sentiment analysis (Jakob and Gurevych, 2010; Yang and Cardie, 2013; Klinger and Cimiano, 2013; Pontiki et al., 2015, 2016, i.a.). This situation is dissatisfying for (at least) two reasons. First, detecting the emotions expressed in social media and their stimuli might play a role in understanding why different social groups change their attitude towards specific events and could help recognize specific issues in society. Second, understanding the relationship between stimuli and emotions is also compelling from a psychological point of view, given that emotions are commonly considered responses to relevant situations (Scherer, 2005).

Models which tackle the task of detecting the stimulus in a text have seen three different problem formulations in the past: (1) Classification into a predefined inventory of possible stimuli (Mohammad et al., 2014), similarly to previous work in sentiment analysis (Ganu et al., 2009), (2) classification of precalculated or annotated clauses as containing a stimulus or not (Gui et al., 2016, i.a.), and (3) detecting the tokens that describe the stimulus, e.g., with IOB labels (Ghazi et al., 2015, i.a.). We follow the two settings in which the stimuli are not predefined categories (2+3, cf. Figure 1).

These two settings have their advantages and disadvantages. The clause classification setting is more coarse-grained and, therefore, more likely to perform well than the token sequence labeling setting, but it might miss the exact starting and endpoints of a stimulus span and needs clause annotations or a syntactic parse with the risk of error

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

Clause-based Classification:	
No Stimulus	Stimulus
[ She's pleased at ]	[ how things have turned out . ]

Token Sequence Labeling:									
O	O	O	O	B	I	I	I	I	O
She	's	pleased	at	how	things	have	turned	out	.

Figure 1: Different formulations for emotion stimulus detection.

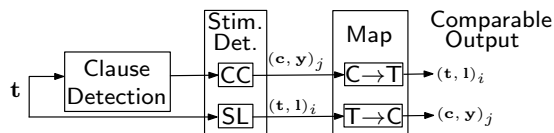


Figure 2: Framework for emotion stimulus detection. Tokens  $\mathbf{t}$  are split into clauses for clause class. Mapping ensures that both methods result in clause classifications  $(\mathbf{t}, \mathbf{l})_i$  and token sequences with labels  $(\mathbf{c}, \mathbf{y})_j$ .

propagation. The token sequence labeling setting might be more challenging, but has the potential to output more exactly which tokens belong to the stimulus. Further, sequence labeling is a more standard machine learning setting than a pipeline of clause detection and classification.

These two different formulations are naturally evaluated in two different ways and have not been compared before, to the best of our knowledge. Therefore, it remains unclear which task formulation is more appropriate for English. Further, the most recent approaches have been evaluated only on Mandarin Chinese, with the only exception being the *EmotionCauseAnalysis* dataset being considered by Fan et al. (2019), but not in comparison to token sequence labeling. No other English emotion stimulus data sets have been tackled with clause classification methods. We hypothesize that clauses are not appropriate units for English, as Ghazi et al. (2015) already noted that: “such granularity [is] too large to be considered an emotion stimulus in English”. A similar argument has been brought up during the development of semantic role labeling methods: Punyakanok et al. (2008) stated that “argument[s] may span over different parts of a sentence”.

Our contributions are as follows: (1) we develop an integrated framework that represents different formulations for the emotion stimulus detection task and evaluate these on four available English datasets; (2) as part of this framework, we propose a clause detector for English which is required to perform stimulus detection via clause classification in a real-world setting; (3) show that token

sequence labeling is indeed the preferred approach for stimulus detection in most available English datasets; (4) show in an error analysis that this is mostly because clauses are not the appropriate unit for stimuli in English. Finally, (5), we make our implementation and annotations for both clauses and tokens available at <http://www.ims.uni-stuttgart.de/data/emotion-stimulus-detection>.

The remainder of the paper is organized as follows. We first introduce our integrated framework of stimulus detection which enables us to evaluate clause classification and token sequence labeling in a comparable manner (Section 2). We then turn to the experiments (Section 3) in which we analyze results on four different English data sets. Section 4 discusses typical errors in detail, which leads to a better understanding of how stimuli are formulated in English. We conclude in Section 6.

## 2 An Integrated Framework for Stimulus Detection

The two approaches for open-domain stimulus detection, namely, clause classification and token sequence labeling, have not been compared on English. We propose an integrated framework (Figure 2) which takes tokens  $\mathbf{t}$  as input, splits this sequence into clauses and classifies them (clause detection can be bypassed if manual annotations of clauses are available). The token sequence labeling does not rely on clause annotations. The output, either clauses  $\mathbf{c}$  with classifications  $\mathbf{y}$  ( $\mathbf{y} \in \{\text{yes, no}\}^n$ ) or tokens  $\mathbf{t}$  with labels  $\mathbf{l}$  are then mapped to each other to enable a comparative evaluation. We explain these steps in the following subsections.

### 2.1 Clause Extraction

The clause classification methods rely on representing an instance as a sequence of clauses. Clauses in English grammar are defined as the smallest grammatical structures that contain a subject and a predicate, and can express a complete proposition (Kroeger, 2005). We show our algorithm to detect clauses in Algorithm 1.

To mark the segments that would potentially approximate clauses, we rely on the constituency parse tree of the token sequence (Line 2). For that reason, we use the Berkeley Neural Parser (Kitaev and Klein, 2018). As illustrated by Feng et al. (2012) and Tafreshi and Diab (2018) we also do that by segmenting the constituency parse tree of the instance (Line 9) at the borders of constituents

---

**Algorithm 1: Clause Extraction**

---

```
Input: text
Output: Clauses  $\mathbf{c}$ 
1  $\mathbf{t} \leftarrow \text{tokenize}(\text{text})$ 
2  $\text{tree} \leftarrow \text{parse}(\mathbf{t})$  // constituency parse
3  $\text{gaps} \leftarrow \{0, |\mathbf{t}|\}$  // potential clause bounds
4  $\text{segments} \leftarrow \emptyset$  // initial. set of segments
5 foreach node  $n$  in tree do
6   if  $\text{label}(n) \in \{S, \text{SBAR}, \text{SBARQ}, \text{INV}, \text{SQ}\}$ 
7      $\ell \leftarrow$  first token leaf that  $n$  governs
8      $r \leftarrow$  last token leaf that  $n$  governs
9      $\text{gaps} = \text{gaps} \cup \{\text{idx}_\ell, \text{idx}_r + 1\}$ 
10 foreach adjacent pair  $(i, j)$  in  $\text{sort}(\text{gaps})$  do
11    $\text{segments} = \text{segments} \cup \mathbf{t}[i : j]$ 
12 repeat
13   foreach  $s_i$  in  $\text{segments}$  do
14     if  $s_i \sim / \wedge [ \wedge \text{A-z a-z 0-9} ] + \$ /$ 
15        $s_{i-1} = s_{i-1} \parallel s_i$ 
16        $\text{segments} = \text{segments} \setminus s_i$ 
17     if  $|s_i| \leq 3$ 
18        $s_{i+1} = s_i \parallel s_{i+1}$ 
19        $\text{segments} = \text{segments} \setminus s_i$ 
20 until convergence
21 return  $\text{segments}$ 
```

---

labeled as clause-type (Bies et al., 1995). We then join the segments until convergence heuristically based on punctuation (Line 12). We illustrate the algorithm in the example in Figure 3.

## 2.2 Stimulus Detection

Our goal is to compare sequence labeling and clause classification. To attribute the performance of the model to the formulation of the task, we keep the differences between the models at a minimum. We therefore first discuss the model components and then how we put them together.

Our models are composed of four layers. As **Embedding Layer**, we use pretrained embeddings to embed each token in the instance  $s = t_1 \dots t_n$  to obtain  $\vec{e}_1, \dots, \vec{e}_n$ . For the **Encoding Layer**, we use a bidirectional LSTM which outputs a sequence of hidden states  $\vec{h}_1, \dots, \vec{h}_n$ . In an additional **Attention Layer**, each word or clause is represented as the concatenation of its embedding and a weighted average over other words or clauses in the instance:  $\vec{u}_i = [\vec{h}_i; \sum_{j=1}^n a_{i,j} \cdot \vec{h}_j]$ . The weights  $a_{i,j}$  are calculated as the dot-product between  $\vec{h}_i$  and every other word, and by normalizing the scores using softmax  $\vec{a}_i = \text{softmax}(\vec{h}_i^T \cdot \vec{h}_j)$ . We concatenate all representations to obtain the final representation vector  $\vec{s}$ . The **Output Layer** is different for the two different task formulations (sequence labeling vs. single softmax). For the case of the single softmax, the input to the classifier is the representation of the clause obtained on the previous layer and the

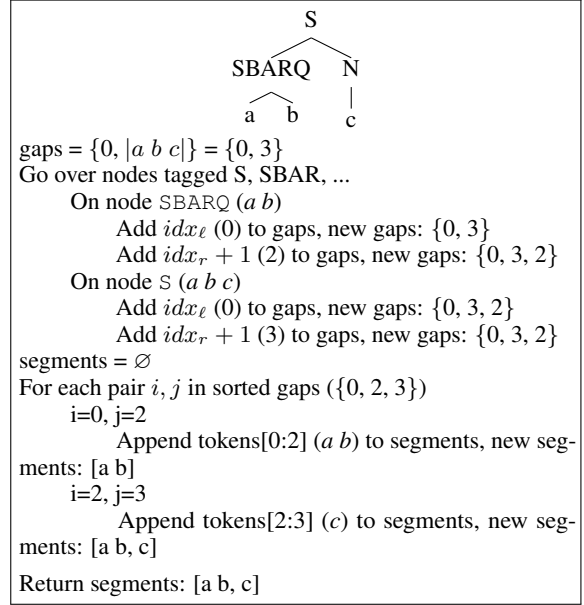


Figure 3: Example for the application of Algorithm 1.

classifier output is defined as  $\vec{o}_i = \text{softmax}(W \cdot \text{ReLU}(\text{Dropout}(h(\vec{s}))))$ . When labels are not predicted independently from each other but rather in a sequential manner, we use a linear-chain conditional random field (Lafferty et al., 2001). It takes the sequence of probability vectors from the previous layer  $\vec{u}_1, \vec{u}_2, \dots$  and outputs a sequence of labels  $y_1, y_2, \dots$ . The score of the labeled sequence is defined as the sum of the probabilities of individual labels and the transition probabilities:  $s(y_{1:n}) = \sum_{i=1}^n \vec{u}_i(y_i) + \sum_{i=2}^n T[y_{i-1}, y_i]$ , where the matrix  $T$  that contains the transition probabilities between one label and another (i.e.,  $T[i, j]$  represents the probability that a token labeled  $i$  is followed by a token labeled  $j$ ). At prediction time, the most likely sequence is chosen with the Viterbi algorithm (Viterbi, 1967).

With these components, we can now put together the actual models which we use for stimulus detection. We compare three different models, one for token sequence labeling (SL) and two for clause classification (CC). The model architectures are illustrated in Figure 4.

**Token Sequence Labeling (SL).** In this model, we formulate emotion stimulus detection as token sequence labeling with the IOB alphabet (Ramshaw and Marcus, 1995). As embeddings, we use word-level GloVe embeddings (Pennington et al., 2014). The sequence-to-sequence architecture comprises a bidirectional LSTM, an attention layer and the CRF output layer.

**Independent Clause Classification (ICC).** This

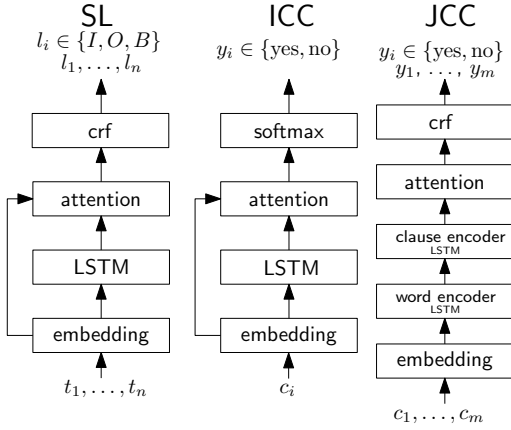


Figure 4: Comparable model architectures.

model, similarly proposed by Cheng et al. (2017), takes the clauses from the clause detector (or from annotated data) and classifies them as containing the stimulus or not. The model has a similar architecture to the one before, with the exception of the final classifier, which is a single softmax to output a single label. The training objective is to minimize the cross-entropy loss. This model does not have access to clauses other than the one it predicts for.

**Joint Clause Classification (JCC).** In this model, the neural architecture we employ is slightly different from before to enable it to make a prediction for clauses in the context of all clauses. It comprises multiple LSTM modules as word-level encoders, one for each clause. The LSTM at the word-level encodes the tokens of one clause into one representation. The next layer is a clause-level encoder based on two bidirectional LSTMs, where the clause representations are learned and updated by integrating the relations between multiple clauses. After we obtain the final clause representation for each clause, we perform sequence labeling with a CRF on the clause level. The training objective is to minimize the negative log-likelihood loss across all clauses. This implementation follows the architecture by Xia et al. (2019), with the change of the upper layer, which is, in our case, an LSTM clause encoder and not a transformer, to keep the architecture comparable across our different formulations. Therefore, this is comparable to all other hierarchical models proposed for the task (Ding et al., 2019; Xu et al., 2019; Xia and Ding, 2019).

### 2.3 Mapping between Task Formulations

The last component of our integrated framework maps the different representations of each formulation of emotion stimulus detection between each

other, namely clause classifications to token sequence labeling and vice versa. We obtain clause classifications from token label sequences ( $T \rightarrow C$  in Figure 2) by accepting any clause that has at least one token being labeled as B or I as a stimulus clause. The other way around, clause classes are mapped to tokens ( $C \rightarrow T$ ) in such a way that the first token of a stimulus clause is a B and all the remaining tokens in the respective clause are I. Tokens from clauses that do not correspond to a stimulus all receive O labels.

## 3 Experiments and Results

We now put the models to use to understand the differences between sequence labeling and clause classification for English emotion stimulus detection and the suitability of clauses as the unit of analysis.

### 3.1 Data Sets

We base our experiments on four data sets.<sup>1</sup> For each data set, we report the size, the number of stimulus annotations and statistics for tokens and clauses in Table 1.

**EmotionStimulus.** This data set proposed by Ghazi et al. (2015) is constructed based on FrameNet’s *emotion-directed* frame.<sup>2</sup> The authors used FrameNet’s annotated data for 173 emotion lexical units, grouped the lexical units into seven basic emotions using their synonyms and built a dataset manually annotated with both the emotion stimulus and the emotion. The corpus consists of 820 sentences with annotations of emotion categories and stimuli. The rest of 1,594 sentences only contain an emotion label. For this dataset, we see the lowest average number of clauses for which all tokens correspond to a stimulus ( $\mu$  w. all S/I in Table 1). This result shows that the stimuli annotations rarely align with the clause boundaries.

**ElectoralTweets.** Frame Semantics also inspires a dataset of social media posts (Mohammad et al., 2014). The corpus consists of 4,056 tweets of which 2,427 contain emotion stimulus annotations on the token level. The annotation was performed via crowdsourcing. The tweets are the shortest instance type in length and have a higher average of clauses per instance than the *GoodNewsEveryone* or the *EmotionStimulus* datasets. They also show

<sup>1</sup>Corpora which we do not consider for our experiments are discussed in the related work section.

<sup>2</sup>[https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Emotion\\_directed](https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Emotion_directed)

Data set	Size	Stimuli	Tokens				Clauses			
			$\mu$	$\sigma$	$\mu S/I$	$\mu S/C$	Total	w. S	$\mu I$	$\mu$ w. all S/I
<i>EmotionStimulus</i>	2,414	820	7.29	5.20	0.12	0.11	5,818	1,117	2.41	0.05
<i>ElectoralTweets</i>	4,056	2,427	6.22	4.00	0.20	0.17	13,612	3,295	3.36	0.34
<i>GoodNewsEveryone</i>	5,000	4,798	7.27	3.67	0.55	0.50	9,190	6,301	1.84	0.52
Emotion Cause Ana.	2,655	2,580	8.48	5.20	0.20	0.10	19,473	2,897	7.33	0.33

Table 1: Data sets available for the Emotion Stimulus Detection task in English. Size: number of annotated instances, Stimuli : number of instances with stimuli annotated;  $\mu$ ,  $\sigma$ : mean/standard deviation of length of stimuli in tokens;  $\mu S/I$ : mean number of stimulus tokens per instance;  $\mu S/C$ : mean number of stimulus tokens per clause; Total: total number of clauses, w. S: number of clauses that contain a stimulus;  $\mu I$ : average number of clauses per instance;  $\mu$  w. all S/I: average number of clauses in which all tokens correspond to annotated stimuli.

the same mean of stimulus tokens per instance as *EmotionCauseAnalysis* with a slightly higher mean for the number of clauses in which all tokens correspond to stimulus annotations.

**GoodNewsEveryone.** The data set by [Bostan et al. \(2020\)](#) consists of news headlines. From a total of 5000 instances, 4,798 contain a stimulus. The headlines have the shortest stimuli in token count. Similar to the *ElectoralTweets*, they also have a high average stimulus token density in clauses. This set has the lowest mean number of clauses per instance ( $\mu I$  in Table 1).

**EmotionCauseAnalysis** ([Gao et al., 2017](#)) comparably annotate English and Mandarin texts on the clause level and the token level. In our work, we use the English subset, which is the only English corpus annotated for stimuli both at the clause level and at the token level. This dataset has the fewest instances without stimuli among all the others. It also has the longest instances and stimuli. The mean of stimuli tokens annotated per clause is comparable to *EmotionStimulus* despite having a higher mean of stimuli tokens per instance. In the upcoming experiments, we use the clause annotations and not automatically recognized clauses with Algorithm 1 as input to our framework.

### 3.2 Clause Identification Evaluation

Before turning to the actual evaluation of the emotion stimulus detection methods, we evaluate the quality of the automatic clause detection. For an intrinsic evaluation, we annotate 50 instances from each test corpus in each data set with two annotators trained on the clause extraction task in two iterations. The two annotators are graduate students and have different scientific backgrounds: computational linguistics (A1) and computer science with a specialization in computer vision (A2). Each student annotated 50 instances of each dataset from

the datasets we use in the same order. As an environment for the annotation process, we used a simple spreadsheet application. We did this small annotation experiment as an inner check for our understanding of the clause extraction task. None of the annotators is a native English speaker; A1 is a native speaker of a Romance language, and A2 a German speaker. The inter-annotator agreement is shown in Table 2. We achieve an acceptable average agreement of  $\kappa=0.65$ .

We now turn to the question if annotated clauses (as an upper bound to an automatic system) align well with annotated stimuli (**Stimuli vs. Anno. Clauses** in Table 2). The evaluation is based on recall (i.e., measuring for how many stimuli a clause exists), either for the whole stimulus (exact), or for the left or the right boundary. We see that except for the corpus *EmotionStimulus*, the right boundaries match better than the left.

Turning to extracted clauses instead of annotated ones (**Extra. vs. Anno. Clauses**) we first evaluate the automatic extraction algorithm. We obtain  $F_1$  values between 0.76% and 0.80%, which we consider acceptable though they also show that error propagation could occur.

For the actual extrinsic evaluation, if clause boundaries are correctly found for annotated stimuli (**Stimuli vs. Extra. Clauses**), we see that the results are only slightly lower than for the gold annotations, except for *EmotionStimulus*. Therefore, we do not expect to see error propagation due to an imperfect extraction algorithm for most data sets.

These results suggest that clauses are not an appropriate unit for stimuli in English. Still, we do not know yet if the clause detection task’s simplicity outweighs these disadvantages in contrast to token sequence labeling. We turn to answer this in the following.

Dataset	Intrinsic							Extrinsic		
	IAA	Stimuli vs. Anno. Clauses			Extra. vs. Anno. Clauses			Stimuli vs. Extra. Clauses		
	$\kappa$	Exact	Left	Right	Precision	Recall	F1	Exact	Left	Right
<i>EmotionCauseAnalysis</i>	0.60	0.60	0.35	0.86	0.77	0.75	0.76	0.59	0.36	0.84
<i>GoodNewsEveryone</i>	0.77	0.62	0.29	0.90	0.87	0.76	0.80	0.61	0.27	0.89
<i>EmotionStimulus</i>	0.59	0.47	0.83	0.11	0.86	0.72	0.76	0.17	0.26	0.07
<i>ElectoralTweets</i>	0.63	0.56	0.39	0.63	0.82	0.78	0.80	0.54	0.43	0.60

Table 2: Evaluation of Clause Detection. Note that for *EmotionCauseAnalysis*, the clauses stem from the annotation provided in the original data and not from our automatic detection method.

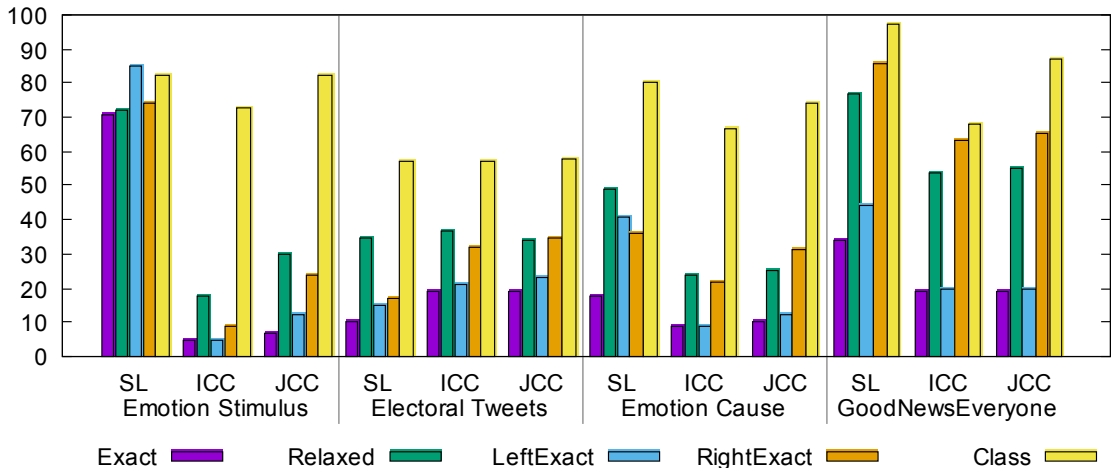


Figure 5: Results of the three different models across four different datasets

### 3.3 Stimulus Detection Evaluation

#### 3.3.1 Evaluation Procedure

We evaluate the quality of all models with five different measures. Motivated by the formulation of *clause classification*, we (1) evaluate the prediction on the clause level with precision, recall, and  $F_1$ . For the *sequence labeling* evaluation, we use four variations. (2) *Exact*, where we consider a consecutive token sequence to be correct if a gold annotation exists that exactly matches, (3) *Relaxed*, where an overlap of one token with a gold annotation is sufficient, (4) *Left-Exact* and (5) *Right-Exact*, where at least the most left/right token in the prediction needs to have a gold-annotated counterpart.

One might argue that sequence labeling evaluation is unfair for the clause classification, as it is more fine-grained than the actual prediction method. However, for transparency across methods and analysis of advantages and disadvantages of the different methods, we use this approach in addition to clause classification evaluation.

We split the data for each set randomly into three sets: 80% train, 10% dev, and 10% test. We use dropout with a probability of 0.5, train with Adam (Kingma and Ba, 2015) with a base learning rate of

0.003, and a batch size of 10. At test time, we select the model with the best validation accuracy after 50 epochs with a patience of 10 epochs. All models use embedding sizes of 300 and hidden state sizes of 100 (Pennington et al., 2014). We do not tune hyperparameters for any of the architectures and implement all models with the AllenNLP library (Gardner et al., 2018).

#### 3.3.2 Results

We now study the performance of the different models on the English data sets. Figure 5 summarizes the results. (Precision and recall values are available in Table 7 in Appendices.)

**Which of the modeling approaches performs best on English data?** If we only compare the absolute numbers in  $F_1$ , we see that the clause classification evaluation (Class) shows the highest result across all models and data set. The only exception is the *EmotionStimulus* data, in which the Left-Exact evaluation is slightly higher. When we rely on this evaluation score, we see that the token sequence labeling method shows a superior result to the classification methods in two data sets, namely *GoodNewsEveryone* and *EmotionCauseAnalysis*. On *ElectoralTweets* and *EmotionStimulus*, the re-












Ann. Pred.	Error types	SL				ICC				JCC				Sum
		ET	GNE	ES	ECA	ET	GNE	ES	ECA	ET	GNE	ES	ECA	
	Early stop	0	4	1	3	0	6	2	7	0	5	1	4	33
	Late stop	11	9	10	8	19	30	7	25	17	31	13	22	202
	Early start & stop	0	3	0	1	9	11	5	1	6	10	3	2	51
	Early start	152	16	0	6	192	73	9	164	220	58	3	159	1052
	Late start	28	3	0	1	3	8	1	0	2	7	1	0	54
	Late start & stop	2	1	0	0	0	2	0	1	0	1	0	1	8
	Contained	0	0	0	0	0	0	0	1	0	0	0	2	3
	Multiple	143	189	11	260	47	24	9	11	37	34	4	0	769
	Surrounded	9	10	0	5	19	31	28	43	22	33	26	28	254
	False Negative	231	160	59	228	126	112	10	97	85	50	1	81	1240
	False Positives	10	18	2	14	78	92	11	38	60	73	4	26	426
	All	586	413	83	526	493	389	82	388	449	302	56	325	4092

Table 3: Counts for each error type for each model across all data sets.

sults are *en par* across all methods with this evaluation measure. We find this surprising to some degree, as this evaluation is more natural for the classification tasks (ICC and JCC) than for sequence labeling (SL), which requires the mapping step.

As this suggests that clauses are not the appropriate unit, it is worth comparing these results with the Exact evaluation measure, which evaluates on the token-sequence level. We observe that token sequence labeling outperforms both clause classification methods on three of the four data sets, with *ElectoralTweets* being the only exception with the shortest textual instances and the highest number of clauses in which all tokens correspond to stimulus annotation (see Table 1). Therefore, we conclude that token sequence labeling is superior to clause classification on (most of our) English data sets.

**Do clause classification models perform better on the left or the right side of the stimulus clause?** Given the evaluation of the clause detection, we expect the right boundary to be better found for *GoodNewsEveryone* and *EmotionCauseAnalysis* and the left boundary for *EmotionStimulus*. Surprisingly, this is not entirely true – the right boundary is found with higher  $F_1$  on all data sets, not only on those where the clauses are better aligned with the stimulus’ right boundary. Nevertheless, the effect is more reliable for *GoodNewsEveryone*, as expected.

**Does token sequence labeling perform better on the left or the right side of the stimulus clause?** We can ask this similar question for token sequence labeling, though it might be harder to motivate than in the classification setting. Non-surprisingly,




such a clear pattern cannot be observed. For *ElectoralTweets* and *EmotionCauseAnalysis*, the difference between the left and right match is minimal. For *GoodNewsEveryone*, it can be observed to a lesser extent than for the classification approaches, and for *EmotionStimulus*, the left boundary is better found than the right boundary. It seems that for the longer sequences in *EmotionStimulus* and *EmotionCauseAnalysis*, the beginning of the stimulus span is easier to find than for shorter sequences.

#### Is joint prediction of clause labels beneficial?

This hypothesis can be confirmed; however, the differences are of a different magnitude depending on the data set. For *GoodNewsEveryone*, the effect is more substantial than for the other corpora. *ElectoralTweets* shows the smallest difference.

## 4 Error Analysis

In the following, we analyze the error types made by the different models on all data sets and investigate in which ways SL improves over the ICC and JCC models. We hypothesize that the higher flexibility of token-based sequence labeling leads to different types of errors than the clause-based classification models.

For quantitative analysis, we define different error types, illustrated in Table 3 with different symbols as abbreviations. The top bar illustrates the gold span, while the bottom corresponds to the predicted span. The error types illustrated with symbols  and  correspond to false positives;  are false negatives. All other error types correspond to either both false positive and false negative in a strict evaluation setting or true positives in one of





















Err	Example	Model	Data set
	Steve talked to me a lot   about being abandoned   and the pain   that caused.	JCC	ECA
	No what I told about the way   they treated you and me   made him angry.	SL	ECA
	Fuck Mitt Romney   and Fuck Barack Obama   ... God got me !!!!!	ICC	ET
	Maurice Mitchell wants   you to do more than vote.	ICC	GNE
	And he started   to despair that his exploration was going   to be entirely unsuccessful ...	ICC	ECA
	Deeply ashamed of my wayward notions.   I tried my best to contradict myself.	ICC	ES
	Anyone else find it weird   I get excited about stuff like the RNC tonight ?!   # polisciprbs	SL	ET
	Doesn't he do it well   said the girl following with admiring eyes,   every movement of him.	JCC	ECA
	If he feared   that some terrible secret might evaporate from them.   it was a mania with him.	SL	ECA
	I was furious   because the Mac XL wasn't real said Hoffman.	SL	ECA
	With such obvious delight in food, it 's hard   to see how Blanc remains so slim.	SL	ES
	Triad Thugs Use Clubs to Punish Hong Kong ' s Protesters.	JCC	GNE
	I'm glad to see you   so happy Lupin	ICC	ES




Figure 6: Examples for error types for different models and data sets. Extracted clauses are separate by |.

the relaxed evaluation settings.

**Do ICC and JCC particularly miss starting or end points of the stimulus annotation?** We see in Table 3 that for *Late stop* , CC models make considerably more mistakes across all datasets. ICC does so on ET and ECA, while JCC makes more mistakes on GNE and ES. For data sets in which stimulus annotations end with a clause, errors of this type are less likely. These results are more prominent for *Early start & stop* .

**Do all methods have similar issues with finding the whole consecutive stimulus?** We see this in the error type *Multiple* . When the CC models make this mistake, it can be attributed to the automatic fine-grained clause extraction, which can cause a small clause within a gold span to become a false negative. However, we see that SL shows higher numbers of this issue than CC. This result is also reflected in the surprisingly low number of *Contained*  – if the prediction is completely inside a gold annotation, the gold annotation tends to be long, and this increases the chance that it is (wrongly) split into multiple predictions.

**How do the error types differ across models?** The *Early Start (& Stop)* and *Surrounded* , , ) counts show differences across the different types of models. Presumably, the clause classification models do have difficulties in finding the left boundary, and they are more prone to “start early” than the token sequence labeling models. This might be due to gold spans starting in the middle of a clause which is predicted to contain the stimulus.

**How do the error types differ across data sets?** The results and error types differ across data sets (see particularly , , ). This points out

what we have seen in the evaluation already: The structure of a stimulus depends on the domain and annotation. The least challenging data set is *EmotionStimulus* with the lowest numbers of errors across all models. This result is caused by most sentences having similar syntactic trees, all stimuli are explicit and mostly introduced in a similar way.

For qualitative analyses, Figure 6 shows one example of each type of error described above. In the first example, the JCC model does not learn to include the second part of the coordination – “and the pain”. In the second example, similarly, the SL model misses the right part of the coordination. For most cases of independent clauses that we inspect, we see a common pattern for both types of models, which is that the prediction stops while encountering coordinating conjunctions. In the sixth example, the prediction span includes the emotion cue. This issue could be solved by doing sequence labeling instead or by informing the model of the presence of other semantic roles. These examples raise the following question: would improved clause segmentation lead to improvements for the clause-classification models across all data sets?

## 5 Related Work

The task of detecting the stimulus of an expressed emotion in text received relatively little attention.

Next to the corpora we mentioned so far, the *REMAN* corpus (Kim and Klinger, 2018) consists of English excerpts from literature, sampled from Project Gutenberg. The authors consider triples of sentences as a trade-off between longer passages and sentences. Further, Neviarouskaya and Aono (2013) annotated English sentences on the token level.



Besides English and Mandarin, [Russo et al. \(2011\)](#) developed a method for the identification of Italian sentences that contain an emotion cause phrase. [Yada et al. \(2017\)](#) annotate Japanese sentences on newspaper articles, web news articles, and Q&A sites. Table 8 in Appendices shows which corpora and methods have been used and compared in previous work for the available English and Chinese sets. We see that the methods applied on the Chinese sets are not evaluated on the English sets.

[Lee et al. \(2010\)](#) firstly investigated the interactions between emotions and the corresponding stimuli from a linguistic perspective. They publish a list of linguistic cues that help in identifying emotion stimuli and develop a rule-based approach. [Chen et al. \(2010\)](#) build on top of their work to develop a machine learning method. [Li and Xu \(2014\)](#) implement a rule-based system to detect the stimuli in Weibo posts and further inform an emotion classifier with the output of this system. Other approaches to develop rules include manual strategies ([Gao et al., 2015](#)), bootstrapping ([Yada et al., 2017](#)) and the use of constituency and dependency parsing ([Neviarouskaya and Aono, 2013](#)).

All recently published state-of-the-art methods for the task of emotion stimulus detection via clause classification are evaluated on the Mandarin data by [Gui et al. \(2016\)](#). They include multi-kernel learning ([Gui et al., 2016](#)) and long short-term memory networks (LSTM) ([Cheng et al., 2017](#)). [Gui et al. \(2017\)](#) propose a convolutional multiple-slot deep memory network (ConvMS-Memnet), and [Li et al. \(2018\)](#) a co-attention neural network model, which encodes the clauses with a co-attention based bi-directional long short-term memory into high-level input representations, which are further passed into a convolutional layer. [Ding et al. \(2019\)](#) proposed an architecture with components for “position augmented embedding” and “dynamic global label” which takes the relative position of the stimuli to the emotion keywords and use the predictions of previous clauses as features for predicting subsequent clauses. [Xia et al. \(2019\)](#) integrate the relative position of stimuli and evaluate a transformer-based model that classifies all clauses jointly within a text. Similarly, [Yu et al. \(2019\)](#) proposes a word-phrase-clause hierarchical network. The transformer-based model achieves state of the art, however, it is shown that the RNN based encoders are very close in performance ([Xia et al.,](#)

[2019](#)). Therefore, we use a comparable model that is grounded on the same concept of a hierarchical setup with LSTMs as encoders. Further, there is a strand of research which jointly predicts the clause that contains the emotion stimulus together with its emotion cue ([Wei et al., 2020](#); [Fan et al., 2020](#)). However, the comparability of methods across data sets has been limited in previous work, as Table 8 in the appendices shows.

## 6 Conclusion

We contributed to emotion stimulus detection in two ways. Firstly, we evaluated emotion stimulus detection across several English annotated data sets. Secondly, we analyzed if the current standard formulation for stimulus detection on Mandarin Chinese is also a good choice for English.

We find that the domain and annotation of the data sets have a large impact on the performance. The worst performance of the token sequence labeling approach is obtained on the crowdsourced data set *ElectoralTweets*. The well-formed sentences of *EmotionStimulus* pose fewer difficulties to our models than tweets and headlines. We see that the sequence labeling approaches are more appropriate for the phenomenon of stimulus mentions in English. This shows in the evaluation of the comparably coarse-grained clause level and is also backed by our error analysis.

For future work, we propose closer investigation of whether other smaller constituents might represent the stimulus better for English and a check of whether the strong results for the sequence labeling hold for other languages. Notably, the clause classification setup has its benefits, and this might lead to a promising setting as joint modeling or as a filtering step to finding parts of the text which might contain a stimulus mention. Another step is to investigate if the emotion stimulus and the emotion category classification benefit from joint modeling in English as it has been shown for Mandarin ([Chen et al., 2018](#)).

## Acknowledgments

This research has been conducted within the project SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1), funded by the German Research Council (DFG). We thank Enrica Troiano, Evgeny Kim, Gabriella Lapesa, and Sean Papay for fruitful discussions and feedback on earlier versions of the paper.

## References

- Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. [Bracketing guidelines for Treebank II style Penn Treebank project](http://languagelog ldc.upenn.edu/myl/PennTreebank1995.pdf). Online: <http://languagelog ldc.upenn.edu/myl/PennTreebank1995.pdf>.
- 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 *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.
- Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. [Joint learning for emotion classification and emotion cause detection](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651, Brussels, Belgium. Association for Computational Linguistics.
- Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. [Emotion cause detection with linguistic constructions](#). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China. Coling 2010 Organizing Committee.
- Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017. [An emotion cause corpus for chinese microblogs with multiple-user structures](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(1):6:1–6:19.
- Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. [From independent prediction to re-ordered prediction: Integrating relative position and global label information to emotion cause identification](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6343–6350.
- Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.
- Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. 2019. [A knowledge regularized hierarchical approach for emotion cause analysis](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5614–5624, Hong Kong, China. Association for Computational Linguistics.
- Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. [Transition-based directed graph construction for emotion-cause pair extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717, Online. Association for Computational Linguistics.
- Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. [Characterizing stylistic elements in syntactic structure](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533, Jeju Island, Korea. Association for Computational Linguistics.
- Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. [Beyond the stars: Improving rating predictions using review text content](#). In *Twelfth International Workshop on the Web and Databases (WebDB 2009)*.
- Kai Gao, Hua Xu, and Jiushuo Wang. 2015. [A rule-based approach to emotion cause detection for chinese micro-blogs](#). *Expert Systems with Applications*, 42(9):4517–4528.
- 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.
- 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. 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, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. [A question answering approach for emotion cause extraction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Copenhagen, Denmark. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Niklas Jakob and Iryna Gurevych. 2010. [Extracting opinion targets in a single and cross-domain setting with conditional random fields](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045. Association for 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.
- 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, Conference Track Proceedings*.
- Nikita Kitaev and Dan Klein. 2018. **Constituency parsing with a self-attentive encoder**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.
- Roman Klinger and Philipp Cimiano. 2013. **Bi-directional inter-dependencies of subjective expressions and targets and their value for a joint model**. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 848–854, Sofia, Bulgaria. Association for Computational Linguistics.
- Paul R. Kroeger. 2005. *Analyzing grammar: An introduction*. Cambridge University Press.
- John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. **Conditional random fields: Probabilistic models for segmenting and labeling sequence data**. In *International Conference on Machine Learning*, pages 282–289.
- Sophia Yat Mei Lee, Ying Cohen, Shoushan Li, and Chu-Ren Huang. 2010. **Emotion cause events: Corpus construction and analysis**. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pages 1121–1128, Valletta, Malta. European Language Resources Association (ELRA).
- Weiyuan Li and Hua Xu. 2014. **Text-based emotion classification using emotion cause extraction**. *Expert Systems with Applications*, 41(4):1742–1749.
- Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. **A co-attention neural network model for emotion cause analysis with emotional context awareness**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4752–4757, Brussels, Belgium. 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.
- Alena Neviarouskaya and Masaki Aono. 2013. **Extracting causes of emotions from text**. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 932–936, Nagoya, Japan. Asian Federation of Natural Language Processing.
- 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.
- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohamad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. **SemEval-2016 task 5: Aspect based sentiment analysis**. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.
- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. **SemEval-2015 task 12: Aspect based sentiment analysis**. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.
- Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. **The importance of syntactic parsing and inference in semantic role labeling**. *Computational Linguistics*, 34(2):257–287.
- Lance Ramshaw and Mitch Marcus. 1995. **Text chunking using transformation-based learning**. In *Third Workshop on Very Large Corpora*, pages 82–94.
- Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. **EMOCause: An easy-adaptable approach to extract emotion cause contexts**. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon. Association for Computational Linguistics.
- Klaus R. Scherer. 2005. **What are emotions? And how can they be measured?** *Social Science Information*, 44(4):695–729.
- Shabnam Tafreshi and Mona Diab. 2018. **Sentence and clause level emotion annotation, detection, and classification in a multi-genre corpus**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1246–1251, Miyazaki, Japan. European Language Resources Association (ELRA).

- Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.
- 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. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A RNN-Transformer Hierarchical Network for Emotion Cause Extraction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5285–5291, Macao, China. International Joint Conferences on Artificial Intelligence.
- Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Lian Yang, and Kan Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.
- Ruifeng Xu, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui. 2017. An ensemble approach for emotion cause detection with event extraction and multi-kernel svms. *Tsinghua Science and Technology*, 22(6):646–659.
- Shuntaro Yada, Kazushi Ikeda, Keiichiro Hoashi, and Kyo Kageura. 2017. A bootstrap method for automatic rule acquisition on emotion cause extraction. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 414–421.
- Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.
- Xinyi Yu, Wenge Rong, Zhuo Zhang, Yuanxin Ouyang, and Zhang Xiong. 2019. Multiple level hierarchical network-based clause selection for emotion cause extraction. *IEEE Access*, 7:9071–9079.

## A Appendix

Data	Model	SL Evaluation												CC Evaluation		
		Exact			Relaxed			Left-Exact			Right-Exact			Clause		
		P	R	F <sub>1</sub>	P	R	F <sub>1</sub>	P	R	F <sub>1</sub>	P	R	F <sub>1</sub>	P	R	F <sub>1</sub>
<i>EmotionStimulus</i>	SL	69	73	71	69	74	72	100	74	85	100	59	74	81	83	82
	ICC	03	26	05	10	100	18	03	26	05	05	44	09	82	70	73
	JCC	05	12	07	21	42	30	12	10	12	17	46	24	84	80	82
<i>ElectoralTweets</i>	SL	15	07	10	41	30	35	52	09	15	42	11	17	100	40	57
	ICC	12	47	19	22	100	37	13	47	21	21	74	32	59	59	59
	JCC	14	30	19	25	54	34	15	48	23	28	47	35	59	57	58
<i>EmotionCauseAnalysis</i>	SL	16	20	18	42	60	49	76	29	41	83	23	36	99	68	80
	ICC	05	35	09	14	100	24	05	35	09	13	82	22	79	64	67
	JCC	06	29	10	18	40	25	11	15	12	35	29	31	82	68	74
<i>GoodNewsEveryone</i>	SL	39	30	34	66	92	77	79	30	44	86	86	86	96	99	97
	ICC	15	29	19	37	100	54	15	29	20	48	92	63	71	67	68
	JCC	16	25	19	40	90	55	17	25	20	54	82	65	82	93	87

Figure 7: Results of the three different models across the five different datasets

Methods	Models	Papers	Data sets and Annotation Approach					
			Categorical Class. & Sequence Lab.		Sequence Labeling		Clause Class.	
			ET (en)	ES (en)	REMAN (en)	GNE (en)	ECA (en)	EDCE (zh)
	CRF	Ghazi et al. (2015)	–	+	–	–	–	–
	BiLSTM-CRF	Kim and Klinger (2018)	–	–	+	–	–	–
	BiLSTM-CRF	Bostan et al. (2020)	–	–	–	+	–	–
	SVM	Mohammad et al. (2014)	+	–	–	–	–	–
	CRF	Gao et al. (2017)	–	–	–	–	+	–
	LSTM	Cheng et al. (2017)	–	–	–	–	–	–
	JMECause	Chen et al. (2018)	–	–	–	–	–	–
	multi-kernel SVM	Xu et al. (2017)	–	–	–	–	–	+
	Multi-Kernel	Gui et al. (2016)	–	–	–	–	–	+
	ConvMS-Memnet	Gui et al. (2017)	–	–	–	–	–	+
	CANN	Li et al. (2018)	–	–	–	–	–	+
	PAE-DGL	Ding et al. (2019)	–	–	–	–	–	+
	HCS	Yu et al. (2019)	–	–	–	–	–	+
	Ranking	Xu et al. (2019)	–	–	–	–	–	+
	Hierarchical BiLSTM	Xia and Ding (2019)	–	–	–	–	–	+
	RTHN	Xia et al. (2019)	–	–	–	–	–	+
	<b>Our work</b>	<b>Ours (2020)</b>	+	+	–	+	+	–
	TransECPE	Fan et al. (2020)	–	–	–	–	–	+
	RankCP	Wei et al. (2020)	–	–	–	–	–	+

Figure 8: Mapping of previous state-of-the-art methods to data sets. + indicates that we are aware of a publication which reports on the method being evaluated on the respective data set and a – indicates our assumption that no reported results exist with the respective method being evaluated on the respective data set. ET corresponds to *ElectoralTweets*, ES to *EmotionStimulus*, GNE to *GoodNewsEveryone*, whereas the other data set are as being introduced above.