Roman Klinger, Evgeny Kim, and Sebastian Padó

# Emotion Analysis for Literary Studies

## Corpus Creation and Computational Modelling

**Abstract:** Most approaches to emotion analysis in fictional texts focus on detecting the emotion class expressed over the course of a text, either with machine learning-based classification or with dictionaries. These approaches do not consider who experiences the emotion and what triggers it and therefore, as a necessary simplicifaction, aggregate across different characters and events. This constitutes a research gap, as emotions play a crucial role in the interaction between characters and the events they are involved in. We fill this gap with the development of two corpora and associated computational models which represent individual events together with their experiencers and stimuli. The first resource, REMAN (Relational EMotion ANnotation), aims at a fine-grained annotation of all these aspects on the text level. The second corpus, FANFIC, contains complete stories, annotated on the experiencer-stimulus level, i. e., focuses on emotional relations among characters. FANFIC is therefore a character relation corpus while REMAN considers event descriptions in addition. Our experiments show that computational stimuli detection is particularly challenging. Furthermore, predicting roles in joint models has the potential to perform better than separate predictions. These resources provide a starting point for future research on the recognition of emotions and associated entities in text. They support qualitative literary studies and digital humanities research. The corpora are freely available at http://www.ims.uni-stuttgart.de/data/emotion.

**Zusammenfassung:** Die meisten Ansätze zur Emotionsanalyse in fiktionalen Texten konzentrieren sich auf das Erkennen der in Text ausgedrückten Emotion, entweder mit maschinellem Lernen oder mit Wörterbüchern. Diese Ansätze berücksichtigen in der Regel nicht, wer die Emotion erlebt und warum. Dies stellt eine Vereinfachung dar, die dazu führt, dass über verschiedene Charaktere und Ereignisse aggregiert wird. Emotionen spielen aber eine entscheidende Rolle in der Interaktion zwischen den Charakteren und den Ereignissen, in die sie verwickelt sind. Wir füllen diese Lücke mit der Entwicklung von zwei Korpora und den zugehörigen Berechnungsmodellen, die einzelne Emotionen (Emotionsereignisse) und die dazugehörigen Charaktere bzw. Stimuli in Beziehung setzen. Die Ressource REMAN (Relationale EMotionsANnotation) zielt auf eine feinkörnige Annotati-

**Roman Klinger, Evgeny Kim, Sebastian Padó,** Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart


https://doi.org/10.1515/9783110693973-011

on all dieser Aspekte auf der Textebene. Unser FANFIC-Korpus enthält komplette Geschichten, annotiert auf der Erlebnis-Stimulus-Ebene, wobei aber Stimuli jeweils durch andere Figuren realisiert sind. Diese Ressource konzentriert sich deshalb auf die Relationen zwischen Figuren, während REMAN zusätzlich Ereignisbeschreibungen betrachtet. Unsere Modelle zeigen, dass insbesondere die automatische Erkennung von Stimuli eine Herausforderung ist. Weiterhin hat die gemeinsame Modellierung das Potential, besser zu funktionieren als getrennte Vorhersagen. Unsere Ressourcen bilden einen Ausgangspunkt für zukünftige Forschung zur Erkennung von Emotionen und assoziierten Entitäten im Text. Sie unterstützen qualitative Literaturwissenschaft und digitale geisteswissenschaftliche Forschung. Die Korpora sind frei verfügbar unter http://www.ims.uni-stuttgart.de/data/emotion.

# 1 Introduction

The analysis of affect in text in general became popular in computational linguistics as well as in application areas like social media mining or computational literary studies with the work by Wiebe (2000), who aimed at distinguishing subjective language from objective, factual statements. Based on this groundbreaking work, several subtasks have emerged, including sentiment analysis (classifying positive vs. negative statements). Another related subfield from this domain is to analyze emotions that are associated with text. This field recently attracted increasing attention, with the creation of corpora and automatic models. One focus is to analyze social media, as it is easy to access and process, and commonly full of relevant instances (Mohammad 2012a; Mohammad, Zhu, et al. 2014; Klinger, De Clercq, et al. 2018).

Emotions are also a crucial component of compelling narratives (Oatley 2002; Ingermanson and Economy 2009; Hogan 2015). Not only do emotions help readers to understand texts (Barton 1996; Robinson 2005) but they also improve readers' abilities of empathy and understanding of others' lives (Mar et al. 2009; Kidd and Castano 2013). This makes literature an interesting field for the study of emotions, as evidenced by the growing interest in emotion-oriented text analysis among digital humanities scholars.

Most research in this regard is based on annotated data, in different variations, either directly for the analysis of the text, the development of automatic systems, or the evaluation of such systems. Emotion annotation can be defined at different textual levels. For example, the corpus which originates from the ISEAR project (Scherer and Wallbott 1994) is annotated on the level of (short) documents, each of which contains the description of an emotionally charged situation. Ex-

amples of resources with sentence-level annotation include the work by Alm et al. (2005), a corpus of children stories, and Strapparava and Mihalcea (2007), who label news headlines. While these studies do not annotate any explicitly textual markers (also called cues) of emotion (Johnson-Laird and Oatley 1989), Aman and Szpakowicz (2007), who annotate blogposts, do include such textual markers. Wiebe et al. (2005) annotate a corpus of news articles with emotions at a word and phrase level. Mohammad, Zhu, et al. (2014) annotate emotion cues in a corpus of 4058 electoral tweets from US via crowdsourcing. Similar in annotation procedure, Liew et al. (2016) curate a corpus of 15 553 tweets and annotate it with 28 emotion categories, valence, arousal, and cues.

A number of studies, including the ones named in the previous paragraph, have explored automatic emotion analysis. In the context of literary studies, relatively simple setups have been used (see Kim and Klinger 2019), notably classification, where a single emotion label is assigned to a segment of text. This corresponds directly to the emotion annotation schemes sketched above. For instance, Kim, Padó, et al. (2017) show that emotions, recognized with dictionaries or bag-of-words models, can serve as features for genre classification in fiction. The predictive power of these models, however, remains generally limited.

We believe that the very simplicity of classification is one of the reasons for the limited performance: Such approaches ignore the semantic role-like structure of emotion, which are not textual categories but rather events. Obviously, the semantic roles in fiction should not be disconnected from their narratological embeddings: when there is an emotion, there is typically somebody who feels the emotion, a target for the motion, and a cause for it (Russell and Barrett 1999; Scarantino 2016). Consider the sentence "*Jack is afraid of John because John has a knife*". Following structural approaches to defining emotional episodes, the sentence can be rephrased as "emotion of fear is experienced by Jack (experiencer) because John (target) has a knife (cause)". Here, dictionary-based or bag-of-words approaches would probably capture that this sentence describes fear, but would fail in assigning the correct semantic roles to John and Jack. This could lead us to conclude, incorrectly, that their emotional experiences are the same.

Compared to classification approaches, there is a rather limited amount of both annotation and modelling work which considers emotions from a structured point of view. There are a few studies on English (Mohammad, Zhu, et al. 2014; Gao et al. 2015; Ghazi et al. 2015; Kim and Klinger 2018) and a considerable number on Mandarin Chinese (Gui, Yuan, et al. 2014; Li and Xu 2014; Gao et al. 2015; Gui, Wu, et al. 2016; Cheng et al. 2017; Gui, Hu, et al. 2017; Xu, Hu, et al. 2017; Chen et al. 2018; Ding et al. 2019; Xia and Ding 2019; Xia, Zhang, et al. 2019; Xu, Lin, et al. 2019). Notably, the corpus by Mohammad, Zhu, et al. (2014) considers experiencers, the stimuli, and targets. However, in the case of tweets, the experi-

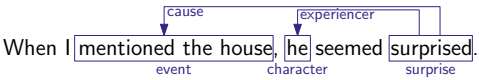


**Fig. 1:** Example annotation in REMAN for a sentence from Hugo (1885), with one character, an emotion word, and event and cause and experiencer annotations.

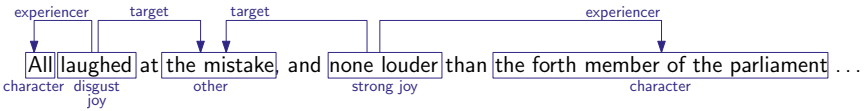


**Fig. 2:** Example annotation in REMAN for a sentence from Stimson (1943), with two characters who are experiencers of different emotions. Disgust and joy are annotated as a mixture of emotions. Both emotions have the same target.

encer is mostly the author of the tweet. Another recent resource of news headlines annotated via crowdsourcing is GoodNewsEveryone (Bostan et al. 2020). This corpus includes annotations of the perceived emotion of the reader in addition to the direct text realization of the emotion.

With this article, we present two corpora and modelling experiments which contribute to this situation, particularly for literature. In the REMAN corpus presented here (Relational Emotion Annotation), we aim for a more comprehensive analysis of emotion events in terms of the semantic roles that these events possess. Our work loosely follows the concept of directed emotions, as defined in FrameNet (Fillmore et al. 2003), and extends the work of Ghazi et al. (2015), who focus on detecting emotion stimuli in the FrameNet exemplary sentences annotated for emotions and causes. In REMAN, we annotate and extract who feels (*experiencer*) which emotion (*cue*, *class*), towards whom the emotion is expressed (*target*), and what is the event that caused the emotion (*stimulus*). Our study is the first one to apply this idea to literary texts. Figures 1 and 2 show examples of the more complex annotation in REMAN.

A second aspect which we believe to be understudied is the role of emotions in characterizing interpersonal relations. This direction links up emotion analysis with social network analysis, an important strand of research in computational literary studies (Agarwal et al. 2013; Nalisnick and Baird 2013; Piper et al. 2017, i. a.). The REMAN resource covers this direction to some extent, since some emotion stimuli happen to be characters, but does not do so in a focussed manner. Starting from the idea that structured emotion representations can serve as a basis for inferring relations between experiencers and stimulus characters, we create a second resource, the FANFIC corpus. In FANFIC, all emotion experiencers are annotated with the emotions they perceive and, if available, with the character which

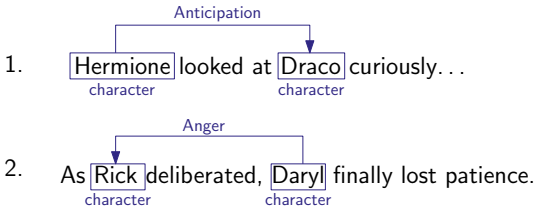


**Fig. 3:** Examples for emotional character interaction in the FANFIC corpus. Example (1) taken from Apryl_Zephyr (2016), example (2) from EmmyR (2014). The arrow starts at the experiencer and points at the causing character.

plays a role in causing the emotion. Figure 3 depicts two examples for emotional character interactions at the text level.

In the remainder of this chapter, we use a discrete set of emotions, based on fundamental emotions as proposed by Plutchik (2001), a common choice of emotion inventory. This model has previously been used in computational analysis of literature (Mohammad 2012b, i. a.). We refer the reader to social psychology literature for more details about alternative emotion theories (such as Ekman 1992) and on the emotional relationships among persons (Gaelick et al. 1985; Burkitt 1997).

Our work has potential to support literary scholars, for example, in analyzing differences and commonalities across texts. As an example, one may consider Goethe's *The Sorrows of Young Werther* (Goethe 1774), a book that gave rise to a plethora of imitations by other writers, who attempted to depict a similar love triangle between main characters found in the original book. The results of our study provides a computational methodology on the basis of which derivative works can be compared systematically with the original.

Our main contributions are therefore: (1) We discuss and make available two resources of fictional texts annotated for emotions, experiencers, causes, and targets as well as for emotional character relations; (2) We analyze the corpora and show which emotions are realized more often with stimuli than others. (3) We provide results of computational methods which automatize the annotation process of emotion words, roles and relations, and further (4) show that the prediction performance of all subtasks benefits from joint modelling, similar to the process of human reading, which is also not entirely linear but considers relations in the text to develop an understanding.

# 2 Annotation Task

We describe the creation and modelling of two resources, the REMAN corpus of emotion semantic role labeling and the FANFIC corpus of character relations.

## 2.1 REMAN: Semantic Role Labeling for Emotion Recognition in Literature

The REMAN corpus is a dataset of excerpts from fictional texts annotated for the phrases that evoke emotions, the experiencer of each emotion (a character in the text, if mentioned), the target and its cause, if mentioned (e. g., an entity, or event). An example of such an annotation is shown in Figures 1 and 2. Each annotation includes textual span labels such as emotions, characters, and events, as well as relational annotations that establish relations among text spans (viz., cause, experiencer, target). We now describe the conceptual background for each annotation layer in detail. The complete annotation guidelines are available online together with the corpus at http://www.ims.uni-stuttgart.de/data/emotion.

### 2.1.1 Conceptualization

We conceptualize ***emotions*** as an individual's experiences that fall in the categories in Plutchik's classification of emotions, namely *anger*, *fear*, *trust*, *disgust*, *joy*, *sadness*, *surprise*, and *anticipation*. In addition, we permit annotations with the class *other emotion* to capture cases when the emotion expressed in the text cannot be reliably categorized into one of the predefined eight classes. This emotion is not meant to extend the existing set of labels, but aims to cover ambiguous and vague emotion expressions. A list of the emotions along with example realizations can be found in Table 1.

Annotators are instructed to preferentially annotate individual key words (e. g., "afraid"), except in cases when emotions are expressed by complete phrases (e. g., "tense and frightened", "wholly absorbed") or by contextual realizations of emotion expression (e. g., "the corners of her mouth went down"). Additionally, emotion spans can be marked as intensified (i. e., amplified, "very happy"), diminished (i. e., downtoned, "a bit sad") and negated ("not afraid") without marking the modifier span or including the modifier word. Spans can be associated with one or more emotion labels (exemplified in Figure 2).

**Tab. 1:** Concepts used for the phrase annotation layer in REMAN together with examples.

| | Concept Value | Examples |
|---|---|---|
| Emotion | Anger | *angry, defend themselves by force, break your little finger, loss of my temper* |
| | Anticipation | *want, wish, wholly absorbed, looked listlessly round, wholly absorbed* |
| | Disgust | *repellent, cheap excitement, turn away from, beg never to hear again* |
| | Fear | *horrified, tense and frightened, shaking fingers* |
| | Joy | *cheerful, grateful, boisterous and hilarious, violins moved and touched him* |
| | Sadness | *failed, despair, the cloudy thoughts, staring at the floor* |
| | Surprise | *perplexing, suddenly, petrified with astonishment, loss for words, with his mouth open* |
| | Trust | *honor, true blue, immeasurable patience* |
| | Other | *careful, brave, had but a tongue, break in her voice, bit deeply into his thumb* |
| Modifier | strong | *I loved her the more* |
| | weak | *with a little pity* |
| | negated | *could not be content* |
| Entity | character | *the chairman of the board* |
| | event | *marry a man I did not love, because of his gold* |
| | other | *Lily's beauty* |

As a preparation for relation annotation, we annotate ***entities***, which are of a clear identity, for instance of a person, object, concept, state, or event (see Table 2). We only annotate them in the context of relations. The subtypes we are particularly interested in are:

**Character** An entity that acts as a character in the text. Character annotation should not omit important information (e. g., the annotation of "the man with two rings of the Royal Naval Reserve on his sleeve" is preferred over only annotating 'the man').

**Event** An event is an occasion or happening that plays a role in the text. Events can be expressed in many ways (see Table 2 for examples from the annotated dataset) and annotators are instructed to label the entire phrases including complementizers or determiners.

**Other** This is an umbrella concept for everything else that is neither a character nor an event, but participates in a relation.

Next, we annotate ***relations***, links between an emotion and other text spans and can be of type *experiencer*, *cause*, and *target*. They can be thought of as the roles that entities play with regard to specific emotions. These relations can only originate from the emotion annotations. In addition, we partially annotate *coreferences* to link personal pronouns to proper nouns.

**Tab. 2:** Typical linguistic realization of entities.

| Entity type | Linguistic realiz. | Examples |
|---|---|---|
| Character | noun phrase | *his son* |
| | adjectival phrase | *old man* |
| Event | verb phrase | *Mrs. Walton had got another baby.* |
| | adverbial phrase | *Jesus spoke unkindly to his mother when he said that to her.* |
| | prepositional phrase | [...] *giving her up.* |
| | clause | [...] *what she said to him* [...] |
| | noun phrase | *the journey* |
| Other | adjectival phrase | [...] *old age* [...] |
| | noun phrase | [...] *the heavens and the earth.* |
| | tense phrase | *She was the only treasure on the face of the Earth that my heart coveted.* |

Roles which are part of relations are:

**Experiencer** The experiencer relation links an emotion span and entity of type *character* who experiences the emotion. If the text contains multiple emotions with multiple experiencers, they all are subject to relation annotation.

**Target** The target relation links an emotion span and entity of any type towards which the emotion experienced by the experiencer is directed. If there are multiple targets of the emotion, then all of them should also be included in the relation annotation. See Figure 2 for the example of a target annotation.

**Cause** The cause relation links an emotion span and entity of any type, which serves as a stimulus, something that evokes the emotional response in the experiencer. If there are multiple causes for the emotion, then all of them are included in separate relation annotations.

**Coreference** The annotators are instructed to annotate as an experiencer the character that is the closest to the emotion phrase in terms of token distance. If the closest mention of the character is a pronoun and the text provides a referent that has a higher level of specificity than the pronoun (i. e., a proper noun or a noun denoting a group or class of objects), the annotators are asked to resolve the coreference. The coreference annotation can be used later to evaluate downstream task applications which associate emotions with a unique character instead of a pronoun.

### 2.1.2 Corpus Construction and Annotation

**Selection**

The corpus of 200 books is sampled from Project Gutenberg[1]. All books belong to the genre of fiction and were written by authors born after the year 1800[2]. We sample consecutive triples of sentences from this subsample of books. A triple is accepted for inclusion for annotation if the middle sentence includes a word that is known to be associated with an emotion, even in isolation. This increases the probability that actual emotion-role-relevant content is present in the instance. To realize this, we use the so-called NRC Emotion Dictionary, which consists of 14 183 linguistic units with an associated emotion (Mohammad and Turney 2013). We consider this middle sentence the target sentence and the annotators are instructed to label emotions in this second sentence only. Experiencers, causes and targets are annotated in the whole sentence triple if they refer to an emotion in the target sentence.

When selecting texts for annotation, there is a trade-of between short passages which are easy to parse but might not contain all relevant roles around an emotion expression, and longer passages which are more likely to contain all relevant relations but are more time-consuming to annotate. Ghazi et al. (2015), for instance, annotate only one sentence and speculate whether adding one sentence before and after will lead to better results. To check their hypothesis, we conduct a small pre-study experiment by extracting 100 random sentences from Project Gutenberg with the NRC dictionary and analyze how often the roles of experiencer, cause, and target are found in the target sentence and in the window of up to five sentences before and after. The analysis shows that 98 % of the snippets include the experiencer in the target sentence, while cause and target are found in the target sentence in 67 % of the texts. Another 29 % of the texts include cause and target in the window of one sentence before and after the target sentence. The remaining texts include cause and target in the window of two (2 %), three (1 %), and four (1 %) sentences around the target sentence. Evidently, three-sentence spans provide enough information regarding 'who feels what and why' without creating excessive annotation overhead (cumulatively, 96 % of cause and target are found in such sentence triples in the pre-study). We therefore opt for the anno-

**1** http://www.gutenberg.org/. Note that Project Gutenberg is currently not available in Germany due to an ongoing legal dispute. None of the texts under discussion regarding copyright are part of our corpus.

**2** We wanted to work with the texts from the nineteenth and twentieth centuries. However, metadata available to us does not include the book publication date, but specifies the birth year of the author.

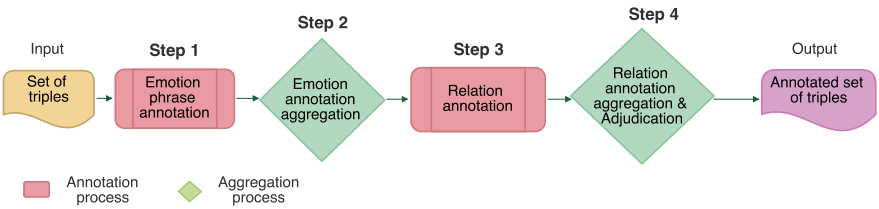


**Fig. 4:** A visualization of the multi-step annotation process.

tation of sentence triples with one sentence before and one after an emotion cue (preselected with the NRC dictionary).

### Annotation Procedure

The annotations are generated in a multistep process visualized in Figure 4. The people involved in the annotation were either *annotators* or *experts*, whose roles did not overlap. The annotations (of spans and relations) were performed by three graduate students of computational linguistics (two native English speakers, one non-native speaker) within a three-month period. Arising questions were discussed in weekly meetings with the experts (the two first authors of the paper) and the results documented in the annotation guidelines. Further, the experts perform manual adjudication in cases where automatic annotation aggregation is not possible (see below). We use WebAnno[3] (Yimam et al. 2013) as annotation environment. In the following, we discuss the four steps of generating the corpus.

**Step 1: Emotion phrase annotation** The *annotators* are asked to first decide whether the text to be annotated expresses an emotion and which emotion it is. If any exists, they label the phrase that led to their decision. The annotators are instructed to search for emotions that are expressed either as single words or phrases.

**Step 2: Emotion phrase aggregation** In the previous step, each annotator generates a set of annotations. In this step, the *expert* heuristically aggregates all spans that overlap between annotators in a semi-automatic process: Concrete emotions are preferred over the 'other-emotion' category, annotations with modifier are preferred over annotations without, and shorter spans are preferred over longer spans. Overlapping annotations with different emotion labels are all accepted.

**Step 3: Relation annotation** *Annotators* receive the texts that they annotated for emotions in Step 1, including the aggregated annotation from Step 2 based on

3 https://webanno.github.io/webanno/

all individual annotation. Thus, all annotators see the same texts and annotations in this step. For each emotion, the task is now to annotate entities that are experiencers, targets, or causes of the emotion and relate them to their respective emotions. The annotators are instructed to tag only those entities that have a role of an experiencer, cause, and target. The decision on the entity and relation annotation is made simultaneously: For each emotion, the annotators need to identify who experiences the emotion (which *character*) and why (because of an event, object, or other character).

**Step 4: Relation aggregation and adjudication** This final step is a manual *expert* step: Aggregate the relation annotations provided by the annotators. Heuristically, we prefer shorter spans for entities, but guide ourselves with common sense. For instance, consider the phrase "[...] *wishing rather to amuse and flatter himself by merely inspiring her with passion*". "*Wishing*" is labelled as emotion. One annotator tagged "*to amuse and flatter himself by merely inspiring her with passion*" as event, another tagged only "*by merely inspiring her with passion*", which is incomplete, as the target of the emotion is the act of amusing and flattering oneself.

Note that we do not discard the rejected annotations but publish all annotations of all annotators.

## 2.2 FANFIC: Character Relations in Fanfiction

The goal of the FANFIC corpus is different from REMAN but shares the goal of moving beyond identifying emotion labels for stretches of text. In REMAN, emotions are assigned to cues that are related to experiencers and stimuli. This is a detailed approach of modeling the structure of emotions in literature, but it might be too complex for computational modeling approaches to perform well. Further, we do not link the relational structure of emotions to the field of social network analysis.

This is what we aim at with the FANFIC corpus. On the one side, we opt for a simpler formulation of emotion structures, namely emotional relations between characters. On the other side, we aim at an evaluation on the document level, in the spirit of social network analysis.

### 2.2.1 Conceptualization

FANFIC is centered completely around interpersonal emotions: emotions are understood as relations between characters in a text. Formally, each emotion rela-

tion is a triple $(C_{\text{exp}}, e, C_{\text{cause}})$ in which the character $C_{\text{exp}}$ feels the emotion $e$ (mentioned in text explicitly or implicitly). The character $C_{\text{cause}}$ is part of an event which triggers the emotion $e$. We consider the eight fundamental emotions defined by Plutchik (2001) (anger, fear, joy, anticipation, trust, surprise, disgust, sadness). Each character corresponds to a token sequence for the relation extraction task. In a social network analysis setting, these characters correspond to normalized entities. Note that, in contrast to REMAN, we do not annotate the exact span that triggers the emotion (which is difficult in cases of implicit emotion descriptions), nor do we annotate causes that are events or objects.

### 2.2.2 Data Collection and Annotation

To be able to evaluate on the social network interaction level, the annotation of complete stories is required. We therefore annotate a sample of 19 complete English fan-fiction short stories, retrieved from the Archive of Our Own project[4] (due to availability, the legal possibility to process the texts and a modern language), and a single short story by Joyce (1914), "Counterparts". All fan-fiction stories were marked by the respective author as complete, are shorter than 1500 words, and include at least four different characters. They are tagged with the keywords 'emotion' and 'relationships' as metadata in the repository.

The annotators were instructed to mark every character mention with a canonical name and to decide if there is an emotional relationship between the character and another character (again, using WebAnno, as in the creation of the REMAN corpus). If so, they marked the corresponding emotion phrase with the emotion labels (as well as indicating if the emotion is amplified, downtoned or negated). Based on this phrase annotation, they marked two relations: from the emotion phrase to the experiencing character and from the emotion phrase to the causing character (if available, i. e., $C_{\text{cause}}$ can be empty). One character may be described as experiencing multiple emotions.

We generate a 'consensus' annotation by keeping all emotion labels by all annotators. This is motivated by the finding by Schuff et al. (2017) that aggregating with the goal of achieving a high recall leads to better performance for emotion prediction. As we focus here on emotion relationships, we retain all emotion labels from all annotators, providing for the richest possible emotion representation.

4 https://archiveofourown.org

# 3 Corpus Analyses

## 3.1 REMAN

In the following, we first discuss the annotation and then provide results of models trained on our resources.

### 3.1.1 Inter-Annotator Agreement and Consistency of the Annotations

The first step of our analysis is the evaluation of the quality of the annotations. For that, we make use of the Cohen's Kappa coefficient (κ, Artstein and Poesio 2008), a measure to measure the agreement of independent annotators. In this measure, the main ingredient is the the probability that two annotators agree, which is calculated by counting how often they made the same annotation choices. This value is the same as 'accuracy'. However, in contrast to this simple fraction of correct annotations, Cohen's kappa normalizes by the *expected agreement*.

We calculate this measure at the token level. In addition, we calculate $F_1$ score on the phrase level both with exact match (where all tokens need to be the same between annotators) and with fuzzy match (where one token overlap is sufficient such that the annotation counts as being the same). In this manner, we calculate agreement both for the phrase annotation and the relation annotation.

Table 3 reports the IAA agreement scores for emotion, entity, and relation annotations for each pair of annotators. Among all emotions, *joy* has the highest number of instances (336) and the highest agreement scores (average κ=0.35), followed by *fear* (κ=0.30) and *sadness* (κ=0.24). *Other emotion* has the lowest agreement with average κ=0.07 – not surprisingly, given the nature of this label as category for difficult cases. For entity annotation, especially for *character* annotation, the agreement is higher, with the highest agreement between two annotators being κ=0.63. The agreement on the *event* and *other* entities is low (κ=0.23 and 0.14 and $F_1$=25 and 14, respectively). This is presumably the case because event annotations are often comparably long which makes it hard to achieve exact match. This also holds, to a lesser extent, for *character* annotations. If we allow partial overlaps to count as a match, the average $F_1$ increases to 57 for *character* (an increase of 4 percentage points (pp)), 44 for *event* (increase by 19 pp), and 23 for *other* category (increase by 9 pp).

For relation annotations, fuzzy evaluation also leads to higher agreement scores ($F_1$ increase for *experiencer*, *cause* and *target* by 10 pp, 7 pp, and 12 pp respectively). These results are in line with previous studies on emotion cause an-

**Tab. 3:** Pairwise inter-annotator agreement for phrase annotation and relation annotation in REMAN (annotators a, b, c). $F_1$ is in %. Regarding the relation scores, in strict $F_1$, a TP holds if the relation label agrees and the entity it points to has the same label and span. In fuzzy $F_1$ ($\approx F_1$), a TP holds if the relation and the entity it points to have the same lavel, but the span boundary of the entity may differ.

| | | a vs. b | | | b vs. c | | | a vs. c | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Type** | **κ** | **strict $F_1$** | **$\approx F_1$** | **κ** | **strict $F_1$** | **$\approx F_1$** | **κ** | **strict $F_1$** | **$\approx F_1$** |
| Emotion | anger | .25 | 25 | 39 | .15 | 15 | 38 | .18 | 18 | 33 |
| | anticipation | .09 | 9 | 23 | .07 | 7 | 20 | .18 | 18 | 39 |
| | sadness | .32 | 32 | 41 | .22 | 23 | 41 | .19 | 20 | 29 |
| | joy | .38 | 39 | 50 | .40 | 40 | 55 | .28 | 28 | 44 |
| | surprise | .26 | 26 | 43 | .22 | 23 | 33 | .27 | 27 | 37 |
| | trust | .17 | 17 | 26 | .14 | 14 | 21 | .12 | 13 | 32 |
| | disgust | .23 | 23 | 41 | .10 | 10 | 26 | .19 | 19 | 31 |
| | other | .07 | 7 | 7 | .06 | 6 | 11 | .08 | 8 | 22 |
| | fear | .35 | 35 | 48 | .28 | 28 | 35 | .28 | 28 | 41 |
| Entity | character | .63 | 63 | 68 | .48 | 49 | 51 | .48 | 48 | 54 |
| | event | .29 | 31 | 60 | .09 | 10 | 30 | .32 | 34 | 44 |
| | other | .11 | 12 | 28 | .11 | 11 | 18 | .20 | 21 | 23 |
| Relation | experiencer | | 65 | 73 | | 48 | 57 | | 46 | 55 |
| | cause | | 20 | 28 | | 34 | 39 | | 26 | 32 |
| | target | | 27 | 36 | | 18 | 29 | | 14 | 28 |

notation (Russo et al. 2011), and show that disagreements mainly come from the different choices about the precise annotation spans, while the spans typically overlap.

### 3.1.2 Assessing Low Agreement

As we show in Section 3.1.1, the agreement across all annotation layers is relatively low, even for a semantic annotation task. There are several reasons. Indeed, emotion categorization is highly subjective and emotions often co-occur (Schuff et al. 2017). In addition, the cause and target of the emotion are not always clearly recognizable in the text and are also subjective categories (two annotators may find two different causes for the same emotion), problems that emotion role annotation inherits from general semantic role annotation (Ellsworth et al. 2004) – hence the low agreement scores across all categories. The only exception are *experiencer* annotations, which are the most reliable among all annotations and match the substantial agreement scores of character annotation (the only type of entities that can be involved in an experiencer relation).

**Tab. 4:** REMAN corpus statistics for emotions annotation. Columns indicate the frequency of each emotion.

| | | | | Modifier | | | Annotation Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Type** | **Total** | **Adjudic.** | **strong** | **weak** | **neg.** | **1 token** | | **≥ 2 token** | |
| Emotions | anger | 192 | 156 | 5 | 12 | 7 | 106 | 68% | 50 | 32% |
| | anticipation | 248 | 201 | 5 | 3 | 11 | 161 | 80% | 40 | 20% |
| | disgust | 242 | 190 | 2 | 7 | 14 | 144 | 76% | 46 | 24% |
| | fear | 254 | 183 | 11 | 16 | 17 | 145 | 79% | 38 | 21% |
| | joy | 434 | 336 | 31 | 20 | 28 | 289 | 86% | 47 | 14% |
| | sadness | 307 | 224 | 10 | 2 | 13 | 168 | 75% | 56 | 25% |
| | surprise | 243 | 196 | 12 | 4 | 7 | 156 | 80% | 40 | 20% |
| | trust | 264 | 232 | 3 | 3 | 33 | 191 | 82% | 41 | 18% |
| | other emotion | 432 | 207 | 4 | 4 | 4 | 133 | 64% | 41 | 36% |
| Entities | character | 2072 | 1715 | | | | 1288 | 75% | 427 | 25% |
| | event | 858 | 615 | | | | 38 | 6% | 577 | 94% |
| | other | 771 | 485 | | | | 114 | 24% | 371 | 76% |

We illustrate the difficulties the annotators face when annotating emotions with roles with the following example:

> They had never seen ... what was really hateful in his face; ... they could only express it by saying that the arched brows and the long emphatic chin gave it always a look of being lit from below ...'.

In our study, both annotators agree on the character ("they") and the emotion ('hateful' expressing disgust). They also agree that the disgust is related to properties of the face which is described. However, one annotator marks "his face" as target, the other marks the more specific but longer "the arched brows and the long emphatic chin gave it always a look of being lit from below" as cause.

If we abstract away from the text spans, both annotators agree that the emotion of disgust has something to do with "his face", however they disagree on the target annotation and the cause annotation. We take such cases to indicate that while the annotation task is indeed difficult, the surface-oriented inter-annotator agreement measures that we compute arguably underestimate the amount of conceptual agreement among annotators. It is on this basis that we consider our annotation to be meaningful despite the low agreement.

**Tab. 5:** REMAN corpus statistics for relation annotation. Rows indicate the total frequency of each relation and each relation-entity combination.

| | Entities involved | | |
|---|---|---|---|
| **Relation** | **char.** | **event** | **other** |
| experiencer | 1704 | | |
| cause | 87 | 398 | 343 |
| target | 444 | 315 | 257 |
| overall relations | 2238 | 717 | 601 |

### 3.1.3 Corpus Statistics

Tables 4 and 5 show the total number of annotations for each annotation category, broken down by different criteria. In Table 4, the *Total* column shows the overall number of annotations generated by all annotators, while the *Adjudic.* column shows the number of accepted annotations. The REMAN corpus consists of 1720 sentence triples, 1115 of which include an emotion. This is a comparably low number, given that we picked the triples based on words that are associated with an emotion, according to an emotion dictionary. But this also shows, that the annotators of our corpus do not agree with an emotion assignment only based on a dictionary, challenging the use of such simple approach for emotion detection. Still, our corpus is densely populated with emotions, with 64 % of triples having an emotion annotation.

*Joy* has the highest number of annotations, while *anger* has the lowest number of annotations. *Joy*, in addition, is modified as *strong* and *weak* more often than other emotions, while *trust* is negated more often compared to other emotions. In most cases, emotion phrases are single tokens (e. g., 'monster', 'irksome'), out of which 47 % on average are found in the NRC dictionary. *Other emotion* has the largest proportion of annotations that span more than one token (36 % out of all annotations in this category), which is in line with our expectation that lower levels of specificity for emotion annotation make it more difficult to find a single token that indicates an emotion.

In Table 5, we see that, based on the definition of the annotation task, the role of experiencers can only be filled by characters. Causes and targets can be filled by characters, events, and other entities. Interestingly, characters are more often the target of an emotion than the cause. Events are more likely to cause the emotion than being the target of it.

**Tab. 6:** FANFIC corpus: $F_1$ scores at different levels in % for agreement between annotators (a1, a2, a3).

| | a1–a2 | a1–a3 | a2–a3 |
|---|---|---|---|
| Instances labelled | 24 | 19 | 24 |
| Instances unlabelled | 33 | 27 | 29 |
| Graph labelled | 66 | 69 | 66 |
| Graph unlabelled | 90 | 93 | 92 |

## 3.2 FANFIC

### 3.2.1 Inter-Annotator Agreement

Recall that the goal of FANFIC is to use emotion annotation for the construction of social networks. From this perspective, it makes sense to define inter-annotator agreement not just in terms of the textual surface, but also at the level of the network computed from the annotations.

Therefore, we calculate the agreement along two dimensions, namely unlabelled vs. labeled and instance vs. graph-level. Table 6 reports the pairwise results for three annotators. In the *Instances labelled* setting, we accept an instance being labeled as true positive if both annotators marked the same span of text to label the characters as experiencer and cause of an emotion and classified their interaction with the same emotion. In the *Instances unlabelled* case, the emotion label is allowed to be different. On the graph level (*Graph labelled* and *Graph unlabelled*), the evaluation is performed on an aggregated graph of interacting characters, i. e., a relation is accepted by one annotator if the other annotator marked the same interaction somewhere in the text. We use the $F_1$ score to be able to measure the agreement between two annotators on the span levels. For that, we treat the annotations from one annotator in the pair as correct and the annotations from the other as predicted.

As Table 6 shows, agreement on the textual level is low with values between 19 and 33 % (depending on the annotator pair), which also motivated our previously mentioned aggregation strategy. The values for graph-labelled agreement, which arguably provide a more relevant picture for our use-case of network generation, are considerably higher (66 % to 93 %). This shows that annotators agree when it comes to detecting relationships, regardless of where exactly in the text they appear.

| Emotion | All | Rel. |
|---|---|---|
| anger | 258 | 197 |
| anticipation | 307 | 239 |
| disgust | 163 | 122 |
| fear | 182 | 120 |
| joy | 413 | 308 |
| sadness | 97 | 64 |
| surprise | 143 | 129 |
| trust | 179 | 156 |
| **total** | **1742** | **1335** |

**Tab. 7:** FANFIC corpus: Statistics of emotion and relation annotation. 'All' indicates the total number of emotion annotations. 'Rel.' indicates the number of emotional relationships (including a causing character) instantiated with the given emotion.

### 3.2.2 Corpus Statistics

Table 7 summarizes the aggregated results of the annotation. The column 'All' lists the number of experiencer annotations (with an emotion), the column 'Rel.' refers to the counts of emotion annotations with both experiencer and cause. In this sense, 'Rel.' column is a subset of 'All' column.

*Joy* has the highest number of annotated instances and the highest number of relationship instances (413 and 308 respectively). In contrast, *sadness* has the lowest number of annotations with a total count of instances and relations being 97 and 64 respectively. Overall, we obtain 1335 annotated instances, which we use to build and test our models.

## 4 Computational Modeling

We now come to the computational modeling part of our study, where we investigate how difficult the manually annotated emotion structures described above are to predict automatically with current NLP methods. Given the differences between the annotation schemes of REMAN and FANFIC, the models we develop differ to an extent. In the case of REMAN, we phrase the prediction of the emotional structure, including its arguments (experiencer, cause, emotion cue, target) as a sequence prediction task. We further analyze if the information about one of the roles is helpful to recognize another. In the case of FANFIC, we phrase the computational modeling as classification which relation exists between all pairs of characters from a given novel. This is a standard formulation for relation extraction.

**Tab. 8:** Experimental results for the REMAN corpus: Results for predicting Emotions and Roles (column *Predict.*) in Exp. 1–3 (column *Exp*).

| | | | | | Strict | | | Fuzzy | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Predict.** | **Exp** | **# Ann.** | **Model** | **Features** | **P** | **R** | **$F_1$** | **P** | **R** | **$F_1$** |
| | 1 | | Rule-based | dict | 19 | 83 | 31 | | | |
| | 1 | | MLP | BOW | 55 | 21 | 31 | | | |
| Emotion | 2 | 1925 | CRF | all + dictionary | 56 | 6 | 11 | 56 | 6 | 11 |
| | 3 | | CRF | all + dict + exp | 55 | 9 | 16 | 69 | 12 | 20 |
| | 2 | | biLSTM-CRF | embeddings | 57 | 35 | 43 | 62 | 39 | 48 |
| | 2 | | CRF | all + person | 0 | 0 | 0 | 0 | 0 | 0 |
| Cause | 2 | 1550 | biLSTM-CRF | embeddings | 0 | 0 | 0 | 0 | 0 | |
| | 2 | | CRF | all + person | 50 | 2 | 4 | 50 | 2 | 4 |
| Exp'cer | 3 | 1717 | CRF | all + person + emo. | 74 | 15 | 24 | 78 | 15 | 26 |
| | 2 | | biLSTM-CRF | embeddings | 49 | 21 | 30 | 49 | 21 | 30 |
| Target | 3 | 1017 | CRF | all + emo. | 50 | 3 | 6 | 50 | 3 | 6 |
| | 3 | | biLSTM-CRF | embeddings | 0 | 0 | 0 | 0 | 0 | 0 |

## 4.1 REMAN: Role Identification as Sequence Labeling

### 4.1.1 Experiment 1: Coarse-grained emotion classification

We start with our experiments by first studying how well we can identify the emotion in our sentence triples, without looking at the role labeling. This is therefore a standard approach to emotion analysis. The task is to assign one emotion to a sentence triple (target sentence plus two context sentences). We consider this task as the first step towards the full structured prediction tasks as defined above: It confirms that we can at least correctly predict the core of the emotion structure, namely the emotion itself.

We compare a dictionary-based approach and a bag-of-words-based classifier. For the dictionary-based classification, we take the intersection between the words in the triple and the NRC dictionary and assign the triple with the corresponding emotion labels. The $F_1$ score is calculated by comparing the set of labels predicted by dictionaries against the set of gold labels for each triple. The gold labels come from the annotation of words and phrases within each triple. For the BOW approach, we convert each triple into a sparse matrix using all words in the corpus as features. We then classify the triples with a multi-layer perceptron with three hidden layers, 128 neurons each, with an initial learning rate of 0.01 that is divided by 5 if the validation score does not increase after two consecutive epochs by at least 0.001.

The results of all experiments are summarized in Table 8. Experiment one corresponds to the first two rows (labeled with a '1' in the column 'Exp'). We evaluate our models in the same two ways as for inter-annotator agreement: Either by accepting a TP if it is exactly found (exact match) or if at least one token is overlapping with the annotation (fuzzy match).

Emotion classification with dictionaries and bag of words shows mediocre performance. The recall with the dictionary classification is comparably high ($F_1$=0.83), which is due to the fact that texts were sampled using these dictionaries. However, as we said earlier, annotators are free to label any words and phrases as emotion-bearing, hence low precision, recall, and consequently $F_1$ score. The MLP with BOW features does not perform better but shows increased precision at the cost of lower recall.

The experiments therefore show a comparably low performance for the classification setting. However, given that for emotion classification in social media (e. g., Schuff et al. 2017), the results are also typically around 0.60 $F_1$, this is a reasonable result, as literature can be considered a linguistically more complex genre.

#### 4.1.2 Experiment 2: Fine-grained emotion and role detection

We now turn to the setting of recognizing the words that correspond to the role and which trigger the emotion. We phrase this as a sequence labeling task, i. e., a sequence of input tokens is assigned a corresponding sequence of output labels. A classical example from NLP is part-of-speech tagging, where each word is assigned a part of speech. We can phrase emotion structure prediction on the REMAN data as a sequence prediction task, where the input is again the sentence, and each word is assigned either an emotion (if it is a cue for an emotion event), one of the labels *experiencer*, *target*, or *cause*, if it is part of a phrase that fills the corresponding role, or *none*, if it does not participate in the emotion structure. Note that we lose the explicit relation between emotion event and its roles; however, since few sentences contain multiple emotion events, we simply assume that all roles and emotion events within a sentence belong to one another.

Consider the example depicted in Figure 1: The phrase "I mentioned the house" is labelled as an event and is assigned a role of a *cause* for the emotion of *surprise*, and the word "he" is labelled as a character and is assigned a role of an *experiencer* of the same emotion. We represent these relationships by tagging "I

mentioned the house" as *cause* and "he" as *experiencer*, capturing the text spans that are linked by relations with an emotion.[5]

We use conditional random fields (CRF) (Lafferty et al. 2001) and bidirectional long short-term memory networks with a CRF layer (biLSTM-CRF), which are both known to provide generally good performance in sequence prediction tasks (Benikova et al. 2014; Huang et al. 2015). Conditional random fields can be considered an extension of hidden Markov models in the sense that they also are probabilistic and that they also model transition probabilities. However, they do that in the spirit of maximum entropy classifiers, which can deal with many correlated features (see Klinger and Tomanek 2007 for more details). BiLSTM-CRF are essentially conditional random fields, but extract feature representations with a deep neural network. Remarkably, in the biLSTM unit of this network, the Markov property is relaxed such that long distant relations can be considered.

We evaluate the performance of fine-grained emotion and role (experiencer, target, and cause) prediction in a sequence labelling fashion. We train separate CRF and biLSTM-CRF models for each relation, as some annotations overlap (e. g., experiencers can also be targets/causes). The CRF uses part-of-speech tags (detected with spaCy[6], Honnibal 2013), the head of the dependency, if it is capitalized, and offset conjunction with the features of previous and succeeding words as features. For the *emotion* category, we use the presence in the NRC dictionary in addition and, for *experiencer*, the presence in a list of English pronouns. We train for 500 iterations with L-BFGS (Liu and Nocedal 1989) and L1 regularization.

The biLSTM-CRF model uses a concatenated output of two biLSTM models (one trained on word embeddings with dimension 300, and one trained on character embeddings from the corpus with dimension 100) as an input to a CRF layer. The word embeddings that we use as input are pre-trained on Wikipedia[7] using *fastText*. We use Adam as activation function, a dropout value of 0.5, and train the model for 100 epochs with early stopping if no improvement is observed after ten consecutive epochs.

The results for this experiment are also shown in Table 8, with the corresponding rows marked with '2' in the 'Exp' column. As results of this experiment show, the recall is low for all predicted categories. Presumably, a major reason, discussed in Section 3, is that substantial numbers of emotion annotations are words or phrases that are not found in the NRC dictionary. On average, only 46 % of emotion annotations are single tokens that can be found in the NRC dictionary,

**5** In more detail, we use the inside-outside-beginning (IOB) encoding which is standard in sequence prediction (Ramshaw and Marcus 1995).

**6** https://spacy.io/

**7** As available at https://github.com/facebookresearch/fastText (Bojanowski et al. 2017).

but for some emotions this number is much lower (only 14 % of *anticipation* cues). For the categories cause and target, their realizations tend to be rather long spans of text (e. g., 94 % of target events are multiword expressions). Faced with the large amount of variability in the training data, the model often abstains from making any predictions whatsoever for these categories. This explains the zero $F_1$ score for cause prediction with CRF and biLSTM-CRF. We see a somewhat better performance for target prediction with CRFs, which is attributable to the fact that most target relations are triggered by characters, 75 % of which are single tokens.

The highest precision and $F_1$ across all categories is observed for the *emotion* category with biLSTM-CRF (strict $F_1$=43 and fuzzy $F_1$=48). The strict $F_1$ is by 12 pp higher than predicted with dictionaries and with BOW in text classification experiment.

The *experiencer* category is second best, even though the recall for this category is still very low. This can be explained by the fact that experiencers are expressed in the text mostly as personal pronouns. Since the number of personal pronouns in our texts is relatively low (13 % of all tokens in a sentence triple on average), and only a small fraction of them act as experiencers (<1 % of all tokens in a sentence triple on average), the classifier cannot learn when an entity is an experiencer or not.

#### 4.1.3 Experiment 3: Potential for joint modeling of emotion and role prediction

In the final experiment on REMAN, we analyze if there is a potential for *joint modeling* of relations to improve over learning each relation separately. Joint modeling means that the different parts of the emotion structure are not predicted individually, as is the case in simple models, but at the same time. In this manner, joint models can take into account interdependencies between different parts of the structure. They can be thought of as attempting to arrive at a global understanding, similar to human readers of a text.

To that end, we analyze the potential interactions between predictions with gold labels of all other predictions. Specifically, when training our models, we provide the classifier with the information which sequence of tokens is an experiencer (in the case of emotion phrase prediction) and which sequence of tokens is an emotion (in case of experiencer, cause, and target detection). Since this information is taken from the manual annotations, this does not constitute a 'real' joint model, but a so-called oracle: its results constitute an upper bound for the performance when more knowledge is available.

Recall that the goal of this experiment is to estimate if joint modeling of emotion and roles yields a benefit beyond individual prediction. Table 8 shows that for

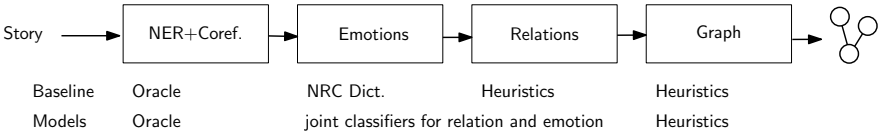


**Fig. 5:** FANFIC: Models for the emotional relationship prediction. Oracle: a set of character pairs from the gold data.

the *emotion* category, $F_1$ increases by 5 pp in strict and by 9 pp in fuzzy evaluation if we provide the classifier with the information which sequence of tokens is an *experiencer*. For *experiencer* prediction, $F_1$ increases by 20 pp in strict and by 22 pp in fuzzy evaluation if we tell the classifier which word or sequence is labelled as emotion.

These results indicate the complementarity of both categories. A qualitative study on a subsample of linguistic properties of emotions and experiencers shows that when the emotion expression and experiencer are parts of the same phrase (verb or adjectival phrase), the emotion word serves as a head to the word that represents an experiencer. Hence, the classifier is able to partially learn that any phrase that is a part of the emotion phrase, whose head is a personal pronoun or a proper name, is a potential *experiencer*. The same applies to *experiencer*: if the head of the governing phrase is an emotion, then the head of the current phrase is a potential *experiencer*. However, due to variability of emotion expressions, this cannot always be the case.

As we have seen, the task of predicting parts of the emotion structure, such as *experiencer*, *cause*, and *target* is a difficult one. In addition to previous observations that informing the classifier about an emotion simplifies the *experiencer* prediction, we have also observed that in many cases characters are experiencing emotions because of other characters. This observation is interesting on its own, as focusing on emotion character relationships potentially adds an interesting angle of analysis to the study of emotions in text. This motivates the focus of the FANFIC model below.

## 4.2 FANFIC: Emotional Character Relationships

As stated above, the ultimate goal in FANFIC is to predict graphs whose nodes are the characters of a text and whose edges are labeled with emotions. The creation of such graphs requires substantial processing beyond what we have seen for REMAN, as shown in Figure 5: First, references to characters have to be recognized and aggregated (Named Entity Recognition + Coreference). Since our focus

**Tab. 9:** FANFIC: Example of different indicator conditions. *No-Ind.*: no positional indicators are added. *Role*: uses tags <e> (experiencer) and <c> (cause). *Entity*: uses only tag <et> (entity). The *M-* conditions mask the lexical material (i. e., name) of the entity.

| Indicator condition | Example |
|---|---|
| No-Ind. | Alice is angry with Bob |
| Role | <e>Alice</e> is angry with <c>Bob</c> |
| MRole | <e/> is angry with <c/> |
| Entity | <et>Alice</et> is angry with <et>Bob</et> |
| MEntity | <et/> is angry with <et/> |

is on emotion prediction, we do not automate this step here, but instead rely on gold character annotations. Then, Emotions have to be recognized and mapped onto character pairs (Relations). Finally, all character–emotion relations have to be aggregated into a graph.

We cast the relation detection as a classification task in which each instance consists of two character mentions with up to *n* tokens context to the left and to the right of the character mentions and the classes are the set of emotions. We use a two-layer GRU neural network (Chung et al. 2014) with max and averaged pooling with different variations of encoding the character positions with indicators (inspired by Zhou et al. 2016, who propose the use of positional indicators for relation detection). Our variations are exemplified in Table 9. The goal of an indicator is to mark a character that is either an experiencer or a cause of the expressed emotion in text. We use different encodings of these roles: 'Role' and 'MRole' (masked role) indicators inform the classifier about these roles, while 'Entity' and 'MEntity' (masked entity) indicators do not (they only indicate that marked characters are entities in the relationship). Note that the prediction of directed relations is simpler in the 'Role' and 'MRole' cases, compared to 'Entity' and 'MEntity', as the model has access to gold information about the relation direction.

We obtain word vectors for the deep learning models from GloVe (pre-trained on Common Crawl, d=300, Pennington et al. 2014) and initialize out-of-vocabulary terms with zeros (including the position indicators).

Given that we have comparably limited data on the story-level, we perform cross-story validation, where each story is used as one separate test/validation source. For model selection and meta-parameter optimization, we use 50 % randomly sampled annotations from this respective test/validation instance as a validation set and the remainder as test data.

We evaluate on three different levels of granularity: Given two character mentions, in the instance-level evaluation, we only accept the prediction to be correct if exactly the same mention has the according emotion annotation. We then ag-

| Model | Instance | Story | Graph |
|---|---|---|---|
| NoInd | 26 | 25 | 35 |
| Role | 33 | 33 | 41 |
| MRole | 38 (38) | 39 (39) | 40 (42) |
| Entity | 23 | 22 | 39 |
| MEntity | 28 | 28 | 39 |

**Tab. 10:** FANFIC: Cross-validated results for different models as $F_1$ scores. 'Instance': aggregated over all instances in the dataset. 'Story': performance averaged over all stories. 'Graph': performance on graph level averaged over all stories. See Table 9 for the examples of the indicator implementation. Results on independent test data shown in brackets.

gregate the different true positive, false positive and false negative values across all stories before averaging to an aggregated score (similar to micro-averaging). On the story-level, we also accept a prediction to be a true positive the same way, but first calculate the result precision/recall/$F_1$ for the whole story before averaging (similar to macro-averaging). On the graph-level, we accept a prediction for a character pair to be correct without considering the exact position.

**Results**

Table 10 shows the results on development data and independent test data for the best models. The GRU+MRole model achieves the highest performance on the instance and story levels, and shows a clear improvement over the GRU+NoInd. model. GRU+Role achieves the highest performance on the graph level. As expected, we observe a better performance on a graph level for all models. The absolute numbers, however, are not very high, but the increase from the instance to the graph level shows that constructing the graph is somewhat 'forgiving': Not each individual prediction at the textual level needs to be correct for a correct interaction graph to emerge.

This is shown in practice on Figure 6, which illustrates a fully predicted network from a fan fiction story based on the *Star Wars* universe (Miralana 2015). The error analysis on the predicted network shows that the mistakes made by the model are not immediately obvious. One example is the *trust* relationship between Finn and Rey. Although the textual instance used to classify the interaction contains 'trust' vocabulary ("they could **help** and be **supportive**"), the overall tone suggests that Finn *anticipates* Rey asking for his help rather than directly imposing trust on her. However, as we do not take into account the exact positions, this mistake is still considered a true positive, as a *trust* relationship is present in the gold data. Another example is the *anticipation* relationship between Rey and Leia that is tagged with *sadness* in the gold data. Consider the following text that was used to classify the relationship: "She adored the older woman and enjoyed her company ..., there were certain things that she didn't want to share with her ...." The text implies that though Rey is pious towards Leia, some aspects of their rela-

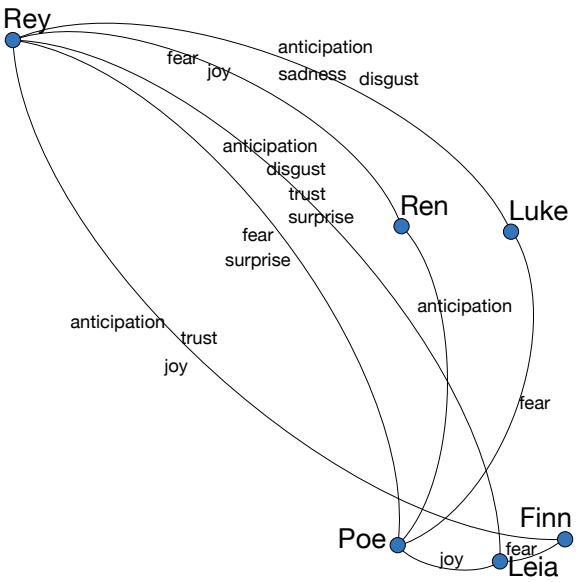


**Fig. 6:** Example of a predicted network.

tionship do not allow her to be fully open with that woman, hence sadness. The erroneous relationship assignment is then presumably triggered due to specific words such as "adored", "enjoyed" and "share", which often indicate *joy* and *anticipation*. This prediction does not count as true positive, as the gold data does not contain *anticipation* among correct relationship between Rey and Leia.

In general, we find that the sequential and embedding information captured by a GRU as well as additional positional information are all relevant for a substantial performance, at least on the fine-grained emotion prediction task. At the same time, we note that the results presented in this section are based on a setting where names of characters come directly from the annotation. This is an unrealistic scenario, as it is not possible to get character annotations for all books we might be interested in analyzing. We shall address this question in our future work.

# 5 Discussion, Conclusion, and Future Work

As both the inter-annotator agreement numbers and the results of our computational models show, the tasks of annotating emotions and corresponding roles manually and automatically are both difficult. Contributing factors are the high

lexical variability of emotion expressions (see Table 1) and of the linguistic form of cause and target expressions. At the same time, the resources we present provide useful and valuable insights in the language of emotion expression and, therefore, should be useful to the communities in linguistics, NLP, and literary studies who are interested in the study of textual expressions of emotion.

Developing such resources has its limitations: Due to the subjective nature of emotions, it is extremely challenging to develop annotation guidelines which would lead to annotations with less variation among annotators, in particular if the annotation includes complex structural choices. That is in line with previous research. For instance, both Schuff et al. (2017) and Russo et al. (2011) find that aggregating labels of multiple annotators not by majority vote, but by forming the union, leads to datasets that are, surprisingly, easier to model computationally.

In REMAN, we tackle this problem by employing a multi-step procedure that helps to improve the agreement of the relation annotation. This does not help in the emotion annotation itself, but helps in the role assignment. The introduction of our multi-step annotation procedure lead to an increased inter-annotator agreement for *experiencer* and *cause* annotations by 13 pp and 5 pp in strict evaluation. This indicates that the task seems easier to annotators if they perform role assignment with predefined emotion annotations.

Another difficulty arises from the nature of the texts we work with. Fictional texts are highly metaphoric and full of allusions, which requires thoughtful reading (often reading between the lines) and a global understanding. However, this is something that our annotators cannot develop in the REMAN case: they only have access to one sentence pre- and post-context each. Therefore, it is not always possible to annotate the cause, target, or even the experiencer. This is a trade-off: On the one side, we did not want to annotate full books to cover a range of sources with manageable annotation effort. On the other side, more context might have improved results. Future work will therefore aim at better understanding how to preselect the relevant context that is needed for reliable annotation and secondly use such knowledge for a follow-up annotation project.

Some of the challenges that are posed with the REMAN corpus are addressed in a different way with the FANFIC approach: Here, we formulated the task of emotional character network extraction from fictional texts. We argued that joining social network analysis of fiction with emotion analysis leverages simplifications that each approach makes when considered independently. However, it should be noted that these evaluations are hard to compare, as the tasks and data sets are different.

In ongoing work, we aim at the development of a a real-world application pipeline in which character pairs are not given by an oracle, but rather extracted from text automatically using named entity recognition. To better understand the

relation between instance and graph levels, we explore the best strategy for edge labeling either by a majority vote or accepting the edges with the highest confidence scores. Further, modeling the task in an end-to-end learning setting from text to directly predict the graph, in the spirit of multi-instance learning, is one of the next steps. To that end, we suggest obtaining more gold data with character relations and optimize the pipeline towards the best performance on additional data.

**Acknowledgment:** This research has been conducted within the CRETA project (http://www.creta.uni-stuttgart.de/) which is funded by the German Ministry for Education and Research (BMBF). This research was partially funded by the German Research Council (DFG), project SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1–1). We thank Laura Ana Maria Bostan and the CRETA consortium for fruitful discussions.

# References

Agarwal, Apoorv, Anup Kotalwar, and Owen Rambow (2013). "Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan, pp. 1202–1208.

Alm, Cecilia Ovesdotter, 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*. Vancouver, BC, pp. 579–586.

Aman, Saima and Stan Szpakowicz (2007). "Identifying Expressions of Emotion in Text". In: *Proceedings of Text, Speech and Dialogue*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 196–205.

Apryl_Zephyr (2016). *Friends*. URL: https://archiveofourown.org/works/8081986 (visited on June 1, 2020).

Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.

Barton, James (1996). "Interpreting character emotions for literature comprehension". In: *Journal of Adolescent & Adult Literacy* 40.1, pp. 22–28.

Benikova, Darina, Chris Biemann, Max Kisselew, and Sebastian Pado (2014). "GermEval 2014 Named Entity Recognition Shared Task: Companion Paper". In: *Workshop Proceedings of the 12th edition of the KONVENS conference*. Hildesheim, Germany, pp. 104–112.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Bostan, Laura Ana Maria, 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 International Conference on Language Resources and Evaluation*. Marseille, France, pp. 1547–1559. (Visited on June 1, 2020).

Burkitt, Ian (1997). "Social relationships and emotions". In: *Sociology* 31.1, pp. 37–55. URL: https://www.jstor.org/stable/42855768?seq=1 (visited on June 1, 2020).

Chen, Ying, 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*. Brussels, Belgium, pp. 646–651.

Cheng, Xiyao, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou (2017). "An emotion cause corpus for chinese microblogs with multiple-user structures". In: *ACM Transactions on Asian and Low-Resource Language Information Processing* 17.1, p. 6.

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *Proceedings of the Deep Learning and Representation Learning Workshop at NIPS 2014*. Montreal, Canada.

Ding, Zixiang, Huihui He, Mengran Zhang, and Rui Xia (2019). "From Independent Prediction to Reordered Prediction: Integrating Relative Position and Global Label Information to Emotion Cause Identification". In: *The Thirty-Third AAAI Conference on Artificial Intelligence*. New Orleans, LA, pp. 6343–6350.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition & Emotion* 6.3-4, pp. 169–200.

Ellsworth, Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó (2004). "PropBank, SALSA and FrameNet: How Design Determines Product". In: *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora at LREC*. Lisbon, Portugal.

EmmyR (2014). *PianoP*. URL: https://archiveofourown.org/works/2481311 (visited on June 1, 2020).

Fillmore, Charles J., Miriam R.L. Petruck, Josef Ruppenhofer, and Abby Wright (2003). "Framenet in Action: The Case of Attaching". In: *International Journal of Lexicography* 16.3, pp. 297–332.

Gaelick, Lisa, Galen V. Bodenhausen, and Robert S. Wyer (1985). "Emotional communication in close relationships." In: *Journal of Personality and Social Psychology* 49.5, p. 1246.

Gao, Kai, Hua Xu, and Jiushuo Wang (2015). "A rule-based approach to emotion cause detection for Chinese micro-blogs". In: *Expert Systems with Applications* 42.9, pp. 4517–4528.

Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz (2015). "Detecting emotion stimuli in emotion-bearing sentences". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. Cairo, Egypt, pp. 152–165.

Goethe, Johann Wolfgang von (1774). *Die Leiden des jungen Werthers*. URL: http://www.deutschestextarchiv.de/book/show/goethe_werther01_1774 (visited on June 1, 2020).

Gui, Lin, 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*. Copenhagen, Denmark, pp. 1593–1602.

Gui, Lin, 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*. Austin, Texas, pp. 1639–1649.

Gui, Lin, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou (2014). "Emotion cause detection with linguistic construction in chinese weibo text". In: *Natural Language Processing and Chinese Computing*. Springer, pp. 457–464.

Hogan, Patrick Colm (2015). "What Literature Teaches Us About Emotion: Synthesizing Affective Science and Literary Study". In: *The Oxford Handbook of Cognitive Literary Studies*. Ed. by Lisa Zunshine. Oxford University Press. Chap. 13, pp. 273–290.

Honnibal, Matthew (2013). *A Good Part-of-Speech Tagger in about 200 Lines of Python*. Online: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991*.

Hugo, Victor (1885). *Les Misérables*. Only accessible outside of Germany. Project Gutenberg: http://www.gutenberg.org/ebooks/135. (Visited on June 1, 2020).

Ingermanson, Randy and Peter Economy (2009). *Writing fiction for dummies*. Indianapolis, IN: John Wiley & Sons.

Johnson-Laird, Philip Nicholas and Keith Oatley (1989). "The language of emotions: An analysis of a semantic field". In: *Cognition & Emotion* 3.2, pp. 81–123.

Joyce, James (1914). *Dubliners*. London: Grant Richards.

Kidd, David Comer and Emanuele Castano (2013). "Reading literary fiction improves theory of mind". In: *Science* 342.6156, pp. 377–380.

Kim, Evgeny 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*. Santa Fe, New Mexico, USA, pp. 1345–1359.

Kim, Evgeny and Roman Klinger (2019). "A Survey on Sentiment and Emotion Analysis for Computational Literary Studies". In: *Zeitschrift für Digitale Geisteswissenschaften* 4. DOI: 10.17175/2019_008.

Kim, Evgeny, Sebastian Padó, and Roman Klinger (2017). "Investigating the Relationship between Literary Genres and Emotional Plot Development". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Vancouver, BC, pp. 17–26.

Klinger, Roman, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur (2018). "IEST: WASSA-2018 Implicit Emotions Shared Task". In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels, Belgium, pp. 31–42.

Klinger, Roman and Katrin Tomanek (2007). "Classical Probabilistic Models and Conditional Random Fields". Tech. rep. TR07-2-013. ISSN 1864-4503, Technical Report. Department of Computer Science, Dortmund University of Technology.

Lafferty, John, Andrew McCallum, and Fernando Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the International Conference on Machine Learning*. Williamstown, MA, pp. 282–289.

Li, Weiyuan and Hua Xu (2014). "Text-based emotion classification using emotion cause extraction". In: *Expert Systems with Applications* 41.4, pp. 1742–1749.

Liew, Jasy Suet Yan, Howard R. Turtle, and Elizabeth D. Liddy (2016). "EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 1149–1156.

Liu, Dong C. and Jorge Nocedal (1989). "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3, pp. 503–528.

Mar, Raymond A, Keith Oatley, and Jordan B Peterson (2009). "Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes". In: *Communications* 34.4, pp. 407–428.

Miralana (2015). *What's so strange about a down-home family romance?* URL: https : / / archiveofourown.org/works/5474927 (visited on June 1, 2020).

Mohammad, Saif (2012a). "# Emotional tweets". In: *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. Montréal, QC, pp. 246–255.

Mohammad, Saif (2012b). "From once upon a time to happily ever after: Tracking emotions in mail and books". In: *Decision Support Systems* 53.4, pp. 730–741.

Mohammad, Saif and Peter Turney (2013). "Crowdsourcing a word–emotion association lexicon". In: *Computational Intelligence* 29.3, pp. 436–465.

Mohammad, Saif, 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*. Baltimore, Maryland, pp. 32–41.

Nalisnick, Eric T and Henry S Baird (2013). "Extracting sentiment networks from Shakespeare's plays". In: *Proceedings of the 12th International Conference on Document Analysis and Recognition*. Washington, DC, pp. 758–762.

Oatley, Keith (2002). "Emotions and the story worlds of fiction". In: *Narrative impact: Social and cognitive foundations* 39, p. 69.

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

Piper, Andrew, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths, and Hardik Vala (2017). "Studying Literary Characters and Character Networks". In: *Digital Humanities 2017: Conference Abstracts*. Montreal, Canada, pp. 119–122.

Plutchik, R. (2001). "The Nature of Emotions". In: *American Scientist* 89.4, pp. 344–350.

Ramshaw, Lance and Mitch Marcus (1995). "Text Chunking using Transformation-Based Learning". In: *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, MA, pp. 82–95.

Robinson, Jenefer (2005). *Deeper than reason: Emotion and its role in literature, music, and art*. Oxford University Press on Demand.

Russell, James A and Lisa F Barrett (1999). "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." In: *Journal of Personality and Social Psychology* 76.5, pp. 805–819.

Russo, Irene, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco (2011). "Emocause: an easy-adaptable approach to emotion cause contexts". In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 153–160.

Scarantino, Andrea (2016). "The philosophy of emotions and its impact on affective science". In: *The handbook of emotions*, pp. 3–65.

Scherer, Klaus R. and Harald G. Wallbott (1994). "Evidence for universality and cultural variation of differential emotion response patterning." In: *Journal of Personality and Social Psychology* 66.2, p. 310.

Schuff, Hendrik, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger (2017). "Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus". In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Copenhagen, Denmark, pp. 13–23.

Stimson, Frederic Jesu (1943). *The King's Men: A Tale of Tomorrow*. Project Gutenberg: http : //www.gutenberg.org/ebooks/18960. (Visited on June 1, 2020).

Strapparava, Carlo and Rada Mihalcea (2007). "SemEval-2007 Task 14: Affective Text". In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic, pp. 70–74.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie (2005). "Annotating Expressions of Opinions and Emotions in Language". In: *Language Resources and Evaluation* 39.2, pp. 165–210.

Wiebe, Janyce M. (2000). "Learning Subjective Adjectives from Corpora". In: *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. Austin, TX, pp. 735–740.

Xia, Rui 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*. Florence, Italy, pp. 1003–1012.

Xia, Rui, Mengran Zhang, and Zixiang Ding (2019). "RTHN: A RNN-Transformer hierarchical network for emotion cause extraction". In: *arXiv preprint arXiv:1906.01236*.

Xu, Bo, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu (2019). "Extracting Emotion Causes Using Learning to Rank Methods From an Information Retrieval Perspective". In: *IEEE Access* 7, pp. 15573–15583.

Xu, Ruifeng, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui (2017). "An ensemble approach for emotion cause detection with event extraction and multi-kernel SVMs". In: *Tsinghua Science and Technology* 22.6, pp. 646–659.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann (2013). "WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria, pp. 1–6.

Zhou, Peng, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu (2016). "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany, pp. 207–212.