Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1554–1566
Marseille, 11–16 May 2020

GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception

Laura Bostan, Evgeny Kim, Roman Klinger
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{laura.bostan, evgeny.kim, roman.klinger}@ims.uni-stuttgart.de

Abstract

Most research on emotion analysis from text focuses on the task of emotion classification or emotion intensity regression. Fewer works address emotions as a phenomenon to be tackled with structured learning, which can be explained by the lack of relevant datasets. We fill this gap by releasing a dataset of 5000 English news headlines annotated via crowdsourcing with their associated emotions, the corresponding emotion experiencers and textual cues, related emotion causes and targets, as well as the reader's perception of the emotion of the headline. This annotation task is comparably challenging, given the large number of classes and roles to be identified. We therefore propose a multiphase annotation procedure in which we first find relevant instances with emotional content and then annotate the more fine-grained aspects. Finally, we develop a baseline for the task of automatic prediction of semantic role structures and discuss the results. The corpus we release enables further research on emotion classification, emotion intensity prediction, emotion cause detection, and supports further qualitative studies.

Keywords: emotion, structured learning, role labeling

1. Introduction

Research in emotion analysis from text focuses on mapping words, sentences, or documents to emotion categories based on the models of Ekman (1992) or Plutchik (2001), which propose the emotion classes of *joy, sadness, anger, fear, trust, disgust, anticipation* and *surprise*. Emotion analysis has been applied to a variety of tasks including large scale social media mining (Stieglitz and Dang-Xuan, 2013), literature analysis (Reagan et al., 2016; Kim and Klinger, 2019), lyrics and music analysis (Mihalcea and Strapparava, 2012; Dodds and Danforth, 2010), and the analysis of the development of emotions over time (Hellrich et al., 2019).

There are at least two types of questions that cannot yet be answered by these emotion analysis systems. Firstly, such systems do not often explicitly model the perspective of understanding the written discourse (reader, writer, or the text's point of view). For example, the headline "Djokovic happy to carry on cruising" (Herman, 2019) contains an explicit mention of *joy* carried by the word "happy". However, it may evoke different emotions in a reader (*e. g.*, when the reader is a supporter of Roger Federer), and the same applies to the author of the headline. To the best of our knowledge, only one work considers this point (Buechel and Hahn, 2017c). Secondly, the structure that can be associated with the emotion description in text is not uncovered. Questions like "Who feels a particular emotion?" or "What causes that emotion?" remain unaddressed. There has been almost no work in this direction, with only a few exceptions in English (Kim and Klinger, 2018; Mohammad et al., 2014) and Mandarin (Xu et al., 2019; Ding et al., 2019).

With this work, we argue that emotion analysis would benefit from a more fine-grained analysis that considers the full structure of an emotion, similar to the research in aspect-based sentiment analysis (Wang et al., 2016; Ma et al., 2018; Xue and Li, 2018; Sun et al., 2019). Consider the headline: "A couple infuriated officials by landing their helicopter in the middle of a nature reserve" (Kenton, 2019) depicted in Figure 1. One could mark "officials" as the experiencer, "a couple" as the target, and "landing their helicopter in the middle of a nature reserve" as the cause of *anger*. Now let us imagine that the headline starts with "A *cheerful* couple" instead of "A couple". A simple approach to emotion detection based on cue words will capture that this sentence contains descriptions of *anger* ("infuriated") and *joy* ("cheerful"). It would, however, fail in attributing correct roles to the couple and the officials. Thus, the distinction between their emotional experiences would remain hidden from us.

In this study, we focus on an annotation task to develop a dataset that would enable addressing the issues raised above. Specifically, we introduce the corpus *GoodNewsEveryone*, a novel dataset of English news headlines collected from 82 different sources most of which are analyzed in the Media Bias Chart (Otero, 2018) annotated for emotion class, emotion intensity, semantic roles (experiencer, cause, target, cue), and reader perspective. We use semantic roles, since identifying who feels what and why is essentially a semantic role labeling task (Gildea and Jurafsky, 2000). The roles we consider are a subset of those defined for the semantic frame for "Emotion" in FrameNet (Baker et al., 1998).

We focus on news headlines due to their brevity and density of contained information. Headlines often appeal to a reader's emotions and hence are a potentially good source for emotion analysis. Besides, news headlines are easy-to-obtain data across many languages, void of data privacy issues associated with social media and microblogging.

Further, we opt for a crowdsourcing setting in contrast to an expert-based setting to obtain data annotated that is to a lesser extend influenced by individual opinions of a low number of annotators. Besides, our previous work showed that it is comparably hard to reach an acceptable agreement in such tasks even under close supervision (Kim and Klinger, 2018).

To summarize, our main contributions in this paper are, (1),

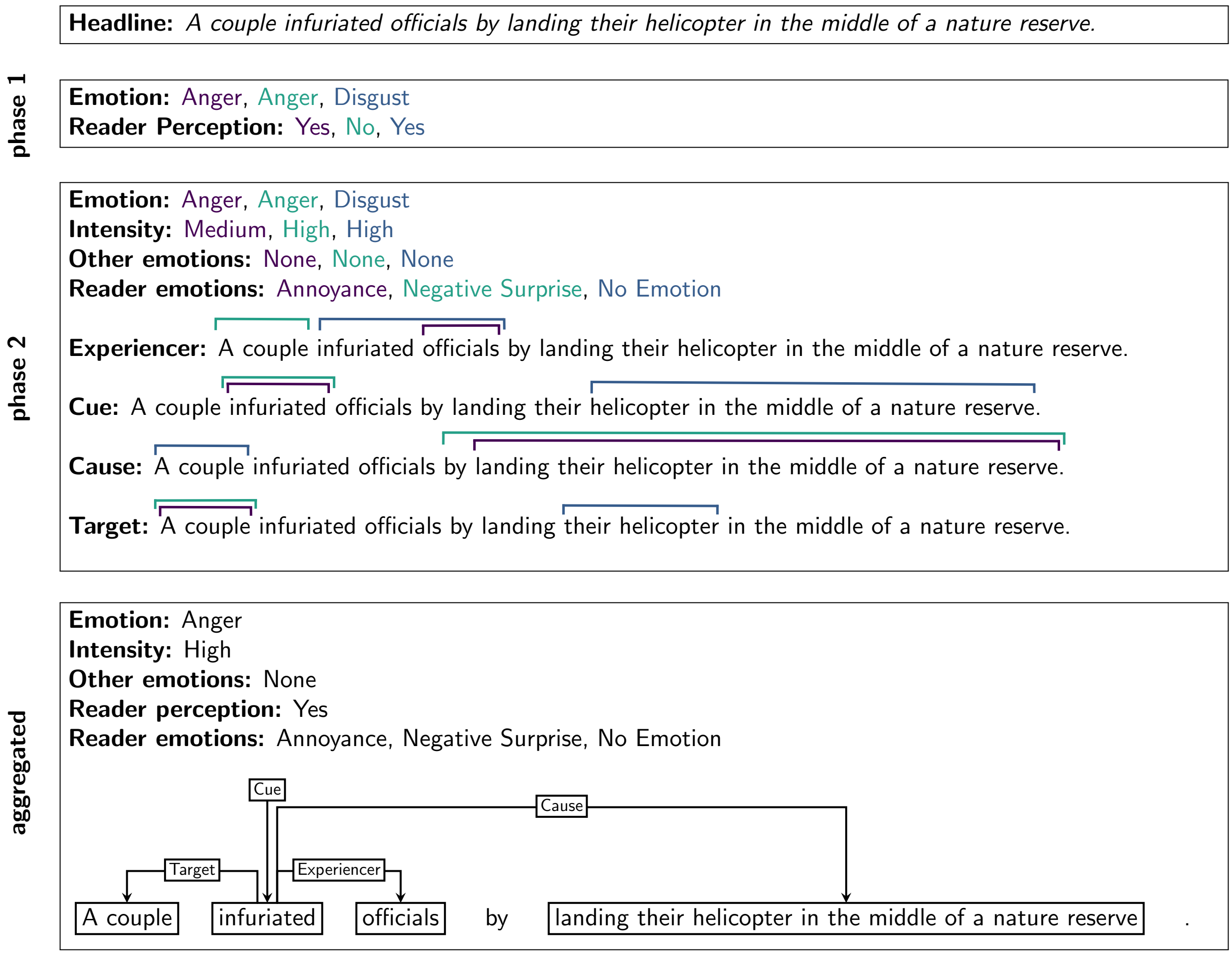

Figure 1: Example of an annotated headline from our dataset. Each color represents an annotator.

that we present the first resource of news headlines annotated for emotions, cues, intensities, experiencers, causes, targets, and reader emotion, (2), design a two-phase annotation procedure for emotion structures via crowdsourcing, and, (3), provide results of a baseline model to predict such roles in a sequence labeling setting. We provide our annotation guidelines and annotations at `http://www.ims.uni-stuttgart.de/data/goodnewseveryone`.

2. Related Work

Our annotation and modelling project is inspired by emotion classification and intensity prediction as well as role labeling and resources which were prepared for these tasks. We therefore look into each of these subtasks and explain how they are related to our new corpus.

2.1. Emotion Classification

Emotion classification deals with mapping words, sentences, or documents to a set of emotions following psychological models such as those proposed by Ekman (1992) (*anger, disgust, fear, joy, sadness*, and *surprise*) or Plutchik (2001); or continuous values of *valence*, *arousal* and *dominance* (Russell, 1980).
Datasets for those tasks can be created in different ways. One way to create annotated datasets is via *expert annotation* (Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007; Ghazi et al., 2015; Schuff et al., 2017; Buechel and Hahn, 2017c). A special case of this procedure has been proposed by the creators of the ISEAR dataset who make use of self-reporting instead, where subjects are asked to describe situations associated with a specific emotion (Scherer and Wallbott, 1994).

Crowdsourcing is another popular way to acquire human judgments (Mohammad, 2012; Mohammad et al., 2014; Mohammad et al., 2014; Abdul-Mageed and Ungar, 2017; Mohammad et al., 2018), for instance on Amazon Mechanical Turk or Figure Eight (previously known as Crowdflower). Troiano et al. (2019) recently published a data set which combines the idea of requesting self-reports (by experts in a lab setting) with the idea of using crowdsourcing. They extend their data to German reports (next to English) and validate each instance, again, via crowdsourcing.

Lastly, social network platforms play a central role in data acquisition with distant supervision, because they provide a cheap way to obtain large amounts of noisy data (Mohammad, 2012; Mohammad et al., 2014; Mohammad and Kiritchenko, 2015; Liu et al., 2017).

We show an overview of available resources in Table 1. Further, more details on previous work can for instance be found in Bostan and Klinger (2018).

	Dataset	Emotion Annotation	Int.	Cue	Exp.	Cause	Target	Size	Source
Emotion & Intensity Classification	ISEAR	Ekman + {shame, guilt}	✗	✗	✗	✗	✗	7,665	Scherer et al. (1994)
	Tales	Ekman	✗	✗	✗	✗	✗	15,302	Alm et al. (2005)
	AffectiveText	Ekman + {valence}	✗	✗	✗	✗	✗	1,250	Strapparava et al. (2007)
	TEC	Ekman + {±surprise}	✗	✗	✗	✗	✗	21,051	Mohammad et al. (2015)
	fb-valence-arousal	VA	✗	✗	✗	✗	✗	2,895	Preoţiuc-Pietro et al. (2016)
	EmoBank	VAD	✗	✗	✗	✗	✗	10,548	Buechel and Hahn (2017a)
	DailyDialogs	Ekman	✗	✗	✗	✗	✗	13,118	Li et al. (2017)
	Grounded-Emotions	Joy & Sadness	✗	✗	✗	✗	✗	2,585	Liu et al. (2017)
	SSEC	Plutchik	✗	✗	✗	✗	✗	4,868	Schuff et al. (2017)
	EmoInt	Ekman − {disgust, surprise}	✓	✗	✗	✗	✗	7,097	Mohammad et al. (2017)
	Multigenre	Plutchik	✗	✗	✗	✗	✗	17,321	Tafreshi and Diab (2018)
	The Affect in Tweets	Others	✓	✗	✗	✗	✗	11,288	Mohammad (2018)
	EmoContext	Joy, Sadness, Anger & Others	✗	✗	✗	✗	✗	30,159	Chatterjee et al. (2019)
	MELD	Ekman + Neutral	✗	✗	✗	✗	✗	13,000	Poria et al. (2019)
	enISEAR	Ekman + {shame, guilt}	✗	✗	✗	✗	✗	1,001	Troiano et al. (2019)
Roles	Blogs	Ekman + {mixed, noemo}	✓	✓	✗	✗	✗	5,025	Aman et al. (2007)
	Emotion-Stimulus	Ekman + {shame}	✗	✗	✗	✓	✗	2,414	Ghazi et al. (2015)
	EmoCues	28 emo categories	✗	✓	✗	✗	✗	15,553	Liew et al. (2016)
	Electoral-Tweets	Plutchik	✗	✓	✓	✓	✓	4,058	Mohammad et al. (2014)
	REMAN	Plutchik + {other}	✗	✓	✓	✓	✓	1,720	Kim and Klinger (2018)
	GoodNewsEveryone	**extended Plutchik**	✓	✓	✓	✓	✓	**5,000**	**Bostan et. al (2020)**

Table 1: Related resources for emotion analysis in English.

2.2. Emotion Intensity

In emotion intensity prediction, the term *intensity* refers to the *degree* an emotion is experienced. For this task, there are only a few datasets available. To our knowledge, the first dataset annotated for emotion intensity is by Aman and Szpakowicz (2007), who ask experts to map textual spans to a set of predefined categories of emotion intensity (*high, moderate*, and *low*). Recently, new datasets were released for the EmoInt shared tasks (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018), both annotated via crowdsourcing through best-worst scaling.

2.3. Cue or Trigger Words

The task of finding a function that segments a textual input and finds the span indicating an emotion category is less researched. First work that annotated cues was done manually by one expert and three annotators on the domain of blog posts (Aman and Szpakowicz, 2007). Mohammad et al. (2014) annotate the cues of emotions in a corpus of 4,058 electoral tweets from the 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.

To the best of our knowledge, there is only one work (Kim and Klinger, 2018) that leverages the annotations for cues and considers the task of emotion detection where the exact spans that represent the cues need to be predicted.

2.4. Emotion Cause Detection

Detecting the cause of an expressed emotion in text received relatively little attention, compared to emotion detection. There are only few works on English that focus on creating resources to tackle this task (Ghazi et al., 2015; Mohammad et al., 2014; Kim and Klinger, 2018; Gao et al., 2015). The task can be formulated in different ways. One is to define a closed set of potential causes after annotation. Then, cause detection is a classification task (Mohammad et al., 2014). Another setting is to find the cause in the text without sticking to clause boundaries. This is formulated as segmentation or clause classification on the token level (Ghazi et al., 2015; Kim and Klinger, 2018). Finding the cause of an emotion is widely researched on Mandarin in both resource creation and methods. Early works build on rule-based systems (Lee, 2010; Lee et al., 2010; Chen et al., 2010), which examine correlations between emotions and cause events in terms of linguistic cues. The works that follow up focus on both methods and corpus construction, showing large improvements over the early works (Li and Xu, 2014; Gui et al., 2014; Gao et al., 2015; Gui et al., 2016; Gui et al., 2017; Xu et al., 2017; Cheng et al., 2017; Chen et al., 2018; Ding et al., 2019). The most recent work on cause extraction is being done on Mandarin and formulates the task jointly with emotion detection (Xu et al., 2019; Xia and Ding, 2019; Xia et al., 2019). With the exception of Mohammad et al. (2014) who are annotating via crowdsourcing, all other datasets are manually labeled by experts, usually using the W3C Emotion Markup Language[1].

2.5. Semantic Role Labeling of Emotions

Semantic role labeling in the context of emotion analysis deals with extracting who feels (*experiencer*) which emotion (*cue*, *class*), towards whom the emotion is directed (*target*), and what is the event that caused the emotion (*stimulus*). The relations are defined akin to FrameNet's Emotion frame (Baker et al., 1998).

[1] `https://www.w3.org/TR/emotionml/`, last accessed Nov 27 2019

There are two works that work on annotation of semantic roles in the context of emotion. Firstly, Mohammad et al. (2014) annotate a dataset of 4,058 tweets via crowdsourcing. The tweets were published before the U.S. presidential elections in 2012. The semantic roles considered are the experiencer, the stimulus, and the target. However, in the case of tweets, the experiencer is mostly the author of the tweet. Secondly, Kim and Klinger (2018) annotate and release REMAN (Relational EMotion ANnotation), a corpus of 1,720 paragraphs based on Project Gutenberg. REMAN was manually annotated for spans which correspond to emotion cues and entities/events in the roles of experiencers, targets, and causes of the emotion. They also provide baseline results for the automatic prediction of these structures and show that their models benefit from joint modeling of emotions with its roles in all subtasks. Our work follows in motivation Kim and Klinger (2018) and in procedure Mohammad et al. (2014).

2.6. Reader vs. Writer vs. Text Perspective

Studying the impact of different annotation perspectives is another little explored area. There are few exceptions in sentiment analysis which investigate the relation between sentiment of a blog post and the sentiment of their comments (Tang and Chen, 2012) or model the emotion of a news reader jointly with the emotion of a comment writer (Liu et al., 2013).

Yang et al. (2009) deal with writer's and reader's emotions on online blogs and find that positive reader emotions tend to be linked to positive writer emotions. Buechel and Hahn (2017c) and Buechel and Hahn (2017b) look into the effects of different perspectives on annotation quality and find that the reader perspective yields better inter-annotator agreement values.

Haider et al. (2020) create an annotated corpus of poetry, in which they make the task explicit that they care about the emotion perceived by the reader, and not an emotion that is expressed by the author or a character. They further propose that for the perception of art, the commonly used set of fundamental emotions is not appropriate but should be extended to a set of aesthetic emotions.

3. Data Collection & Annotation

We gather the data in three steps: (1) collecting the news and the reactions they elicit in social media, (2) filtering the resulting set to retain relevant items, and (3) sampling the final selection using various metrics.

The headlines are then annotated via crowdsourcing in two phases by three annotators each in the first phase and by five annotators each in the second phase. As a last step, the annotations are adjudicated to form the gold standard. We describe each step in detail below.

3.1. Collecting Headlines

The first step consists of retrieving news headlines from the news publishers. We further retrieve content related to a news item from social media: tweets mentioning the headlines together with replies and Reddit posts that link to the headlines. We use this additional information for subsampling described later.

Emotion	Random	Entities	NRC	Reddit	Twitter	Total
Anger	257	350	377	150	144	1278
Annoyance	94	752	228	2	42	1118
Disgust	125	98	89	31	50	392
Fear	255	251	255	100	149	1010
Guilt	218	221	188	51	83	761
Joy	122	104	95	70	68	459
Love	6	51	20	0	4	81
Pessimism	29	79	67	20	58	253
Neg. Surprise	351	352	412	216	367	1698
Optimism	38	196	114	36	47	431
Pos. Surprise	179	332	276	103	83	973
Pride	17	111	42	12	17	199
Sadness	186	251	281	203	158	1079
Shame	112	154	140	44	114	564
Trust	32	97	42	2	6	179
Total	2021	3399	2626	1040	1390	10470

Table 2: Sampling methods counts per adjudicated emotion.

We manually select all news sources available as RSS feeds (82 out of 124) from the Media Bias Chart (Otero, 2019), a project that analyzes reliability (from *original fact reporting* to *containing inaccurate/fabricated information*) and political bias (from *most extreme left* to *most extreme right*) of U.S. news sources. To have a source with a focus on more positive emotions, we include Positive.News in addition.

Our news crawler retrieved daily headlines from the feeds, together with the attached metadata (title, link, and summary of the news article) from March 2019 until October 2019. Every day, after the news collection finished, Twitter was queried for 50 valid tweets for each headline[2]. In addition to that, for each collected tweet, we collect all valid replies and counts of being favorited, retweeted and replied to in the first 24 hours after its publication.

The last step in the pipeline is aquiring the top ("hot") submissions in the `/r/news`[3], `/r/worldnews`[4] subreddits, and their metadata, including the number of up- and downvotes, upvote ratio, number of comments, and the comments themselves.

3.2. Filtering & Postprocessing

We remove headlines that have less than 6 tokens (*e. g.*, "Small or nothing", "But Her Emails", "Red for Higher Ed"), as well as those starting with certain phrases, such as "Ep.","Watch Live:", "Playlist:", "Guide to", and "Ten Things". We also filter-out headlines that contain a date (*e. g.*, "Headlines for March 15, 2019") and words from the headlines which refer to visual content (*e. g.* "video", "photo", "image", "graphic", "watch").

3.3. Sampling Headlines

To aquire data across a wide political and stylistic spectrum, we stratify the remaining headlines by source (150 headlines

[2]A tweet is considered valid if it consists of more than 4 tokens which are not URLs, hashtags, or user mentions.

[3]`https://reddit.com/r/news`

[4]`https://reddit.com/r/worldnews`

	Question	Type	Variable	Codes
Phase 1	1. Which emotion is most dominant in the given headline?	closed, single	Emotion	Emotions + None
	2. Do you think the headline would stir up an emotion in readers?	closed, single	Emotion	Yes, No
Phase 2	1. Which emotion is most dominant in the given headline?	closed, single	Emotion	Emotions
	2. How intensely is the emotion expressed?	closed, single	Intensity	Low, Med., High
	3. Which words helped you in identifying the emotion?	open	Cue	String
	4. Is the experiencer of the emotion mentioned?	close	Experiencer	Yes, No
	5. Who is the experiencer of the emotion?	open	Experiencer	String
	6. Who or what is the emotion directed at?	open	Target	String
	7. Select the words that explain what happened that caused the expressed emotion.	open	Cause	String
	8. Which other emotions are expressed in the given headline?	closed, multiple	Other Emotions	Emotions
	9. Which emotion(s) did you feel while reading this headline?	closed, multiple	Reader Emotions	Emotions

Table 3: Questionnaires for the two annotation phases. Emotions are Anger, Annoyance, Disgust, Fear, Guilt, Joy, Love, Pessimism, Neg. Surprise, Optimism, Negative Surprise, Optimism, Positive Surprise, Pride, Sadness, Shame, and Trust.

from each source). We further subsample according to a set of different strategies. From each strategy, we use the same number of headlines. These are: 1) randomly select headlines, 2) select headlines with high count of emotion terms, 3) select headlines that contain named entities, and 4) select the headlines with high impact on social media.

Random Sampling. The goal of the first sampling method is to collect a random sample of headlines that is representative and not biased towards any source or content type. Note that the sample produced using this strategy might not be as rich with emotional content as the other samples.

Sampling via NRC. For the second sampling strategy, we hypothesize that headlines containing emotionally charged words are also likely to contain the structures we aim to annotate. This strategy selects headlines whose words are in the NRC dictionary (Mohammad and Turney, 2013).

Sampling Entities. We further hypothesize that headlines that mention named entities may also contain experiencers or targets of emotions, and therefore, they are likely to present a complete emotion structure. This sampling method yields headlines that contain at least one entity name, according to the recognition from spaCy that is trained on OntoNotes 5 and Wikipedia corpus.[5] We consider organization names, persons, nationalities, religious, political groups, buildings, countries, and other locations.

Sampling based on Reddit & Twitter. The last sampling strategy involves Twitter and Reddit metadata. This enables us to select and sample headlines based on their impact on social media (under the assumption that this correlates with the emotional connotation of the headline). This strategy chooses them equally from the most favorited tweets, most retweeted headlines on Twitter, most replied to tweets on Twitter, as well as most upvoted and most commented on posts on Reddit.

Table 2 on the previous page shows how many headlines are selected by each sampling method in relation to the most dominant emotion, which is the first of our annotation steps described in Section 3.4.1.

3.4. Annotation Procedure

Using these sampling and filtering methods, we select 9,932 headlines. Next, we set up two questionnaires (see Table 3) for the two annotation phases that we describe below. We use Figure Eight[6].

3.4.1. Phase 1: Selecting Emotional Headlines

The first questionnaire is meant to determine the dominant emotion of a headline if that exists, and whether the headline triggers an emotion in a reader. We hypothesize that these two questions help us to retain only relevant headlines for the next, more expensive, annotation phase.

During this phase, 9,932 headlines were annotated each by three annotators. The first question of the first phase (P1Q1) is: "Which emotion is most dominant in the given headline?" and annotators are provided a closed list of 15 emotion categories to which the category *No emotion* was added. The second question (P1Q2) aims to answer whether a given headline would stir up an emotion in most readers. The annotators could choose one from only two possible answers (*yes* or *no*, see Table 3 and Figure 1 for details).

Our set of 15 emotion categories is an extended set over Plutchik's emotion classes and comprises *anger*, *annoyance*, *disgust*, *fear*, *guilt*, *joy*, *love*, *pessimism*, *negative surprise*, *optimism*, *positive surprise*, *pride*, *sadness*, *shame*, and *trust*. Such a diverse set of emotion labels is meant to provide a more fine-grained analysis and equip the annotators with a wider range of answer choices.

3.4.2. Phase 2: Emotion and Role Annotation

The annotations collected during the first phase are automatically ranked, and the ranking is used to decide which headlines are further annotated in the second phase. Ranking consists of sorting by agreement on P1Q1, considering P1Q2 in the case of ties.

The top 5,000 ranked headlines are annotated by five annotators for emotion class, intensity, reader emotion, and other emotions in case there is not only one emotion. Along with these closed annotation tasks, the annotators are asked to answer several open questions, namely (1) who is the

[5] `https://spacy.io/api/annotation`, last accessed 27 Nov 2019

[6] `https://figure-eight.com`, last accessed 27 Nov 2019

Rule	Cue	Exp.	Cause	Target	Example
1. Majority	3,872	4,820	3,678	3,308	$(\text{span}_1; \text{span}_1; \text{span}_2) \rightarrow \text{span}_1$
2. Most common subsequence	163	70	1,114	1,163	$\{w_2, w_3\}; \{w_1, w_2, w_3\}; \{w_2, w_3, w_4\} \rightarrow \{w_2, w_3\}$
3. Longest common subsequ.	349	74	170	419	$\{w_1, w_2, w_3\}; \{w_1, w_2, w_3, w_4\}; \{w_3, w_4\} \rightarrow \{w_1, w_2, w_3\}$
4. Noun Chunks	0	11	0	0	
5. Manual	611	25	38	110	

Table 4: Heuristics used in adjudicating gold corpus in the order of application on the questions of the type *open* and their counts. w_i refers to the the word with the index i in the headline, each set of words represents an annotation.

experiencer of the emotion (if mentioned), (2) what event triggered the annotated emotion (if mentioned), (3) if the emotion had a target, and (4) who or what is the target. The annotators are free to select multiple instances related to the dominant emotion by copy-paste into the answer field. For more details on the exact questions and examples of answers, see Table 3. Figure 1 shows a depiction of the procedure.

3.4.3. Quality Control and Results

To control the quality, we ensured that a single annotator annotates a maximum of 120 headlines (this protects the annotators from reading too many news headlines and from dominating the annotations). Secondly, we let only annotators who geographically reside in the U.S. contribute to the task.

We test the annotators on a set of 1,100 test questions for the first phase (about 10% of the data) and 500 for the second phase. Annotators were required to pass 95%. The questions were generated based on hand-picked non-ambiguous real headlines through swapping out relevant words from the headline in order to obtain a different annotation, for instance, for "Djokovic happy to carry on cruising", we would swap "Djokovic" with a different entity, the cue "happy" to a different emotion expression.

Further, we exclude Phase 1 annotations that were done in less than 10 seconds and Phase 2 annotations that were done in less than 70 seconds.

After we collected all annotations, we found unreliable annotators for both phases in the following way: for each annotator and for each question, we compute the probability with which the annotator agrees with the response chosen by the majority. If the computed probability is more than two standard deviations away from the mean, we discard all annotations done by that annotator.

On average, 310 distinct annotators needed 15 seconds in the first phase. We followed the guidelines of the platform regarding payment and decided to pay for each judgment \$0.02 for Phase 1 (total of \$816.00). For the second phase, 331 distinct annotators needed on average $\approx$1:17 minutes to perform one judgment. Each judgment was paid with \$0.08 (total \$2,720.00).

3.5. Adjudication of Annotations

In this section, we describe the adjudication process we undertook to create the gold dataset and the difficulties we faced in creating a gold set out of the collected annotations. The first step was to discard wrong annotations for open questions, such as annotations in languages other than English, or annotations of spans that were not part of the headline. In the next step, we incrementally apply a set of rules to the annotated instances in a one-or-nothing fashion. Specifically, we incrementally test each instance for several criteria in such a way that if at least one criterium is satisfied, the instance is accepted and its adjudication is finalized. Instances that do not satisfy at least one criterium are adjudicated manually by us.

Relative Majority Rule. This filter is applied to all questions regardless of their type. Effectively, whenever an entire annotation is agreed upon by at least two annotators, we use all parts of this annotation as the gold annotation. Given the headline depicted in Figure 1 with the following target role annotations by different annotators: *"A couple", "None", "A couple", "officials", "their helicopter"*. The resulting gold annotation is *"A couple"* and the adjudication process for the target ends.

Most Common Subsequence Rule. This rule is only applied to open text questions. It takes the most common smallest string intersection of all annotations. In the headline above, the experiencer annotations *"A couple", "infuriated officials", "officials", "officials", "infuriated officials"* would lead to *"officials"*.

Longest Common Subsequence Rule. This rule is only applied if two different intersections are the most common (previous rule), and these two intersect. We then accept the longest common subsequence. Revisiting the example for deciding on the *cause* role with the annotations *"by landing their helicopter in the nature reserve", "by landing their helicopter", "landing their helicopter in the nature reserve", "a couple infuriated officials", "infuriated"* the adjudicated gold is *"landing their helicopter in the nature reserve"*.

Table 4 shows through examples of how each rule works and how many instances are "solved" by each adjudication rule.

Noun Chunks For the role of experiencer, we accept only the most-common noun-chunk(s)[7].

The annotations that are left after being processed by all the rules described above are being adjudicated manually by the authors of the paper. We show examples for all roles in Table 5.

[7] We used spaCy's named entity recognition model: `https://spacy.io/api/annotation\#named-entities`, last accessed Nov 25, 2019

Role	Chunk	Examples
Exp	NP	cops, David Beckham, Florida National Park, Democrats, El Salvador's President, former Trump associate
	AdjP	illegal immigrant, muslim women from Sri Lanka, indian farmers, syrian woman, western media, dutch doctor
Cue	NP	life lessons, scandal, no plans to stop, rebellion, record, sex assault
	AdjP	holy guacamole!, traumatized
	VP	infuriates, fires, blasts, pushing, doing drugs, will shock
Cause	VP	escaping the dictatorship of the dollar, giving birth in the wake of a storm
	Clause	pensioners being forced to sell their home to pay for care
	NP	trump tax law, trade war, theory of change at first democratic debate, two armed men
Target	AdvP	lazy students
	NP	nebraska flood victims, immigrant detention centers, measles crisis

Table 5: Example of linguistic realizations of the different roles.

Agreement	Emo./Non-Emo.	Reader Percep.	Dominant Emo.	Intensity	Other Emotions	Reader Emotions
κ	0.34	0.09	0.09	0.22	0.06	0.05
%	0.71	0.69	0.17	0.92	0.80	0.80
H (in bits)	0.40	0.42	1.74	0.13	0.36	0.37

Table 6: Agreement statistics on closed questions. Comparing with the questions in Table 3, Emotional/Non-Emotional uses the annotations of Phase 1 Question 1 (P1Q1). In the same way, Reader perception refers to P1Q2, Dominant Emotion is P2Q1, Intensity is linked to P2Q2, Other Emotions to P2Q8, and Reader Emotions to P2Q9.

	# of annotators agreeing			
Emotion	≥ 2	≥ 3	≥ 4	≥ 5
Anger	1.00	0.74	0.33	0.15
Annoyance	1.00	0.71	0.22	0.05
Disgust	1.00	0.78	0.21	0.08
Fear	1.00	0.83	0.44	0.23
Guilt	1.00	0.82	0.37	0.14
Joy	1.00	0.84	0.43	0.17
Love	1.00	0.90	0.62	0.48
Pessimism	1.00	0.76	0.24	0.07
Neg. Surprise	1.00	0.81	0.32	0.11
Optimism	1.00	0.69	0.31	0.12
Pos. Surprise	1.00	0.82	0.38	0.14
Pride	1.00	0.70	0.30	0.26
Sadness	1.00	0.86	0.50	0.24
Shame	1.00	0.63	0.24	0.13
Trust	1.00	0.43	0.05	0.05
Micro Average	1.00	0.75	0.33	0.16

Table 7: Percentage agreement per emotion category on most dominant emotion (second phase). Each column shows the percentage of emotions for which the # of annotators agreeing is greater than 2, 3, 4, and 5

Type	κ	F_1	%	Tok.	MASI	H
Experiencer	.40	.43	.36	.56	.35	.72
Cue	.31	.39	.30	.73	.55	.94
Cause	.28	.60	.16	.58	.47	.58
Target	.15	.36	.12	.45	.54	.04

Table 8: Pairwise inter-annotator agreement (mean) for the open questions annotations. We report for each role the following scores: Fleiss's κ, Accuracy, F_1 score, Proportional Token Overlap, MASI and Entropy

4. Analysis

4.1. Inter-Annotator Agreement

We calculate the agreement on the full set of annotations from each phase for the two question types, namely *open* vs. *closed*, where the first deal with emotion classification and second with the roles *cue*, *experiencer*, *cause*, and *target*.

4.1.1. Emotion

We use Fleiss' Kappa (κ) to measure the inter-annotator agreement for closed questions (Artstein and Poesio, 2008; Fleiss et al., 2013). Besides, we report the average percentage of overlaps between all pairs of annotators (%) and the mean entropy of annotations in bits. Higher agreement correlates with lower entropy. As Table 6 shows, the agreement on the question whether a headline is emotional or not obtains the highest agreement (.34), followed by the question on intensity (.22). The lowest agreement is on the question to find the most dominant emotion (.09).

All metrics show comparably low agreement on the closed questions, especially on the question of the most dominant emotion. This is reasonable, given that emotion annotation is an ambiguous, subjective, and difficult task. This aspect lead to the decision of not purely calculating a majority vote label but to consider the diversity in human interpretation of emotion categories and publish the annotations by all annotators.

Table 7 shows the counts of annotators agreeing on a particular emotion. We observe that *Love*, *Pride*, and *Sadness* show highest intersubjectivity followed closely by *Fear* and *Joy*. *Anger* and *Annoyance* show, given their similarity, lower scores. Note that the micro average of the basic emotions (+ love) is .21 for when more than five annotators agree.

4.1.2. Roles

Table 8 presents the mean of pair-wise inter-annotator agreement for each role. We report average pair-wise Fleiss' κ, span-based exact F_1 over the annotated spans, accuracy, proportional token overlap, and the measure of agreement on set-valued items, MASI (Passonneau, 2004). We observe a fair agreement on the open annotation tasks. The highest agreement is for the role of the *Experiencer*, followed by *Cue*, *Cause*, and *Target*.

This seems to correlate with the length of the annotated

	Dominant Emotion																Anno.	
Role	Anger	Annoyance	Disgust	Fear	Guilt	Joy	Love	Pessimism	Neg. Surprise	Optimism	Pos. Surprise	Pride	Sadness	Shame	Trust	Total	Mean Tok.	Std. Dev Tok.
Experiencer	371	214	292	294	144	176	39	231	628	212	391	52	238	89	95	3466	1.96	1.00
Cue	454	342	371	410	175	256	62	315	873	307	569	60	383	117	120	4814	1.45	1.10
Cause	449	341	375	408	171	260	58	315	871	310	562	65	376	118	119	4798	7.21	3.81
Target	428	319	356	383	164	227	54	297	805	289	529	60	338	111	117	4477	4.67	3.56
Overall	1702	1216	1394	1495	654	919	213	1158	3177	1118	2051	237	1335	435	451	17555	3.94	3.64

Table 9: Corpus statistics for role annotations. Columns indicate how frequent the respective emotions are in relation to the annotated role and annotation length.

spans (see Table 9). This finding is consistent with Kim and Klinger (2018). Presumably, *Experiencers* are easier to annotate as they often are noun phrases whereas causes can be convoluted relative clauses.

4.2. General Corpus Statistics

In the following, we report numbers of the adjudicated data set for simplicity of discussion. Please note that we publish all annotations by all annotators and suggest that computational models should consider the distribution of annotations instead of one adjudicated gold. The latter would be a simplification which we consider to not be appropriate.

GoodNewsEveryone contains 5,000 headlines from various news sources. Overall, the corpus is composed of 56,612 words (354,173 characters), out of which 17,513 are unique. The headline length is short, with 11 words on average. The shortest headline contains six words, while the longest headline contains 32 words. The length of a headline in characters ranges from 24 the shortest to 199 the longest.

Table 9 presents the total number of adjudicated annotations for each role in relation to the dominant emotion. *GoodNewsEveryone* consists of 5,000 headlines, 3,312 of which have an annotated dominant emotion via majority vote. The rest of the 1,688 headlines (up to 5,000) ended in ties for the most dominant emotion category and were adjudicated manually. The emotion category *Negative Surprise* has the highest number of annotations, while *Love* has the lowest number of annotations. In most cases, *Cues* are single tokens (*e. g.*, “infuriates”, “slams”), *Causes* have the largest proportion of annotations that span more than seven tokens on average (65% out of all annotations in this category).

For the role of *Experiencer*, we see the lowest number of annotations (19%), which is a very different result to the one presented by Kim and Klinger (2018), where the role *Experiencer* was the most annotated. We hypothesize that this is the effect of the domain we annotated; it is more likely to encounter explicit experiencers in literature (as literary characters) than in news headlines. As we can see, the *Cue* and the *Cause* relations dominate the dataset (27% each), followed by *Target* (25%) relations.

Table 9 also shows how many times each emotion triggered a certain relation. In this sense, *Negative Surprise* and *Positive Surprise* has triggered the most *Experiencer*, and *Cause* and *Target* relations, which due to the prevalence of the annotations for this emotion in the dataset.

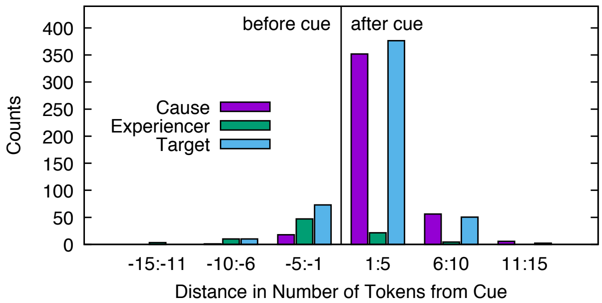

Figure 2: Distances between emotion cues and the other relations: cause, experiencer, and target.

Further, Figure 2, shows the distances of the different roles from the cue. The causes and targets are predominantly realized right of the cue, while the experiencer occurs more often left of the cue.

4.3. Emotions across News Sources

Table 10 shows the top three media sources for each emotion that has been annotated to be the dominating one and the respective sources for the reader’s emotion.

Unsurprisingly for the positive emotions, *Joy*, *Love*, *Positive Surprise*, and *Pride* there is one common source, namely Positive.News. For strong negative emotions such as *Anger* and *Disgust* the top three across the different emotions vary. Though the annotated data for each of the sources is comparably limited, there are a set of interesting findings. Infowars, which the Media Bias Chart categorizes as most right wing and least reliable is found in the list of most frequently being associated with *Fear* in the reader. Breitbart is found to be associated with *Negative Surprise* in the reader. However, both these sources are not in the list of the text-level emotion annotation. Surprisingly, BBC and LA Times are in the list of the most associated with fear on the text-level, despite of both sources being relatively neutral and moderately factual. Further, it is noteworthy that Reuters, ABC News, as being categorized as maximally reliable, are not in the top emotion list at all.

This analysis regarding emotions and media sources is also interesting the other way round, namely to check which

Emotion	Dominant Emotion	Reader Emotions
Anger	The Blaze, The Daily Wire, BuzzFeed	The Gateway Pundit, The Daily Mail, Talking Points Memo
Annoyance	Vice, NewsBusters, AlterNet	Vice, The Week, Business Insider
Disgust	BuzzFeed, The Hill, NewsBusters	Mother Jones, The Blaze, Daily Caller
Fear	The Daily Mail, Los Angeles Times, BBC	Palmer Report, CNN, InfoWars
Guilt	Fox News, The Daily Mail, Vice	The Washington Times, Reason, National Review
Joy	Time, Positive.News, BBC	Positive.News, ThinkProgress, AlterNet
Love	Positive.News, The New Yorker, BBC	Positive.News, AlterNet, Twitchy
Pessimism	MotherJones, Intercept, Financial Times	The Guardian, Truthout, The Washinghton Post
Neg. Surprise	The Daily Mail, MarketWatch, Vice	The Daily Mail, BBC, Breitbart
Optimism	Bussines Insider, The Week, The Fiscal Times	MarketWatch, Positive.News, The New Republic
Pos. Surprise	Positive.News, BBC, MarketWatch	Positive.News, The Washington Post, MotherJones
Pride	Positive.News, The Guardian, The New Yorker	Daily Kos, NBC, The Guardian
Sadness	The Daily Mail, CNN, Daily Caller	The Daily Mail, CNN, The Washington Post
Shame	The Daily Mail, The Guardian, The Daily Wire	Mother Jones, National Review, Fox News
Trust	The Daily Signal, Fox News, Mother Jones	Economist, The Los Angeles Times, The Hill

Table 10: Top three media sources in relation to the main emotion in the text and the reader's emotion.

emotions are dominating which source. From all sources we have in our corpus, nearly all of them have their headlines predominantly annotated with surprise, either negative or positive. That could be expected, given that news headlines often communicate something that has not been known. Exceptions are *Buzzfeed* and *The Hill*, which are dominated by disgust, *CNN*, *Fox News*, *Washington Post*, *The Advocate*, all dominated by *Sadness*, and *Economist*, *Financial Times*, *MotherJones*, all dominated either by *Positive* or *Negative Anticipation*. Only *Time* has most headlines annotated as *Joy*.

Note that this analysis does not say a lot about what the media sources publish – it might also reflect on our sampling strategy and point out what is discussed in social media or which headlines contain emotion words from a dictionary.

5. Baseline

As an estimate for the difficulty of the task, we provide baseline results. We focus on the segmentation tasks as these form the main novel contribution of our data set. Therefore, we formulate the task as sequence labeling of emotion cues, mentions of experiencers, targets, and causes with a bidirectional long short-term memory networks with a CRF layer (biLSTM-CRF) that uses ELMo embeddings (Peters et al., 2018) as input and an IOB alphabet as output.

The results are shown in Table 11. We observe that the results for the detection of experiencers performs best, with .48F_1, followed by the detection of causes with .37F_1. The recognition of causes and targets is more challenging, with .14F_1 and .09F_1. Given that these elements consist of longer spans, this is not too surprising. These results are in line with the findings by Kim and Klinger (2018), who report an acceptable result of .3F_1 for experiencers and a low .06F_1 for targets. They were not able achieve any correct segmentation prediction for causes, in contrast to our experiment.

Category	P	R	F_1
Experiencer	0.44	0.53	0.48
Cue	0.39	0.35	0.37
Cause	0.19	0.11	0.14
Target	0.10	0.08	0.09

Table 11: Results for the baseline experiments.

6. Conclusion and Future Work

We introduce *GoodNewsEveryone*, a corpus of 5,000 headlines annotated for emotion categories, semantic roles, and reader perspective. Such a dataset enables answering instance-based questions, such as, "who is experiencing what emotion and why?" or more general questions, like "what are typical causes of joy in media?". To annotate the headlines, we employ a two-phase procedure and use crowdsourcing. To obtain a gold dataset, we aggregate the annotations through automatic heuristics.

As the evaluation of the inter-annotator agreement and the baseline model results show, the task of annotating structures encompassing emotions with the corresponding roles is a difficult one. We also note that developing such a resource via crowdsourcing has its limitations, due to the subjective nature of emotions, it is very challenging to come up with an annotation methodology that would ensure less dissenting annotations for the domain of headlines.

We release the raw dataset including all annotations by all annotators, the aggregated gold dataset, and the questionnaires. The released dataset will be useful for social science scholars, since it contains valuable information about the interactions of emotions in news headlines, and gives exciting insights into the language of emotion expression in media. Finally, we would like to note that this dataset is also useful to test structured prediction models in general.

7. Acknowledgements

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) and partially funded by the German Research Council (DFG), projects SEAT (Structured Multi-Domain Emotion Analysis from Text, KL 2869/1-1). We thank Enrica Troiano and Jeremy Barnes for fruitful discussions.

8. Bibliographical References

Abdul-Mageed, M. and Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July. Association for Computational Linguistics.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In Václav Matoušek et al., editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Bostan, L.-A.-M. and Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2017a). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (2017b). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April. Association for Computational Linguistics.

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

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Chen, Y., Lee, S. Y. M., Li, S., and Huang, C.-R. (2010). Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China. Association for Computational Linguistics.

Chen, Y., Hou, W., Cheng, X., and Li, S. (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, October-November. Association for Computational Linguistics.

Cheng, X., Chen, Y., Cheng, B., Li, S., and Zhou, G. (2017). An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–19, November.

Ding, Z., He, H., Zhang, M., and Xia, R. (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*, pages 6343–6350. AAAI, Association for the Advancement of Artificial Intelligence.

Dodds, P. S. and Danforth, C. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, 11(4):441–456, August.

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

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Gao, K., Xu, H., and Wang, J. (2015). A rule-based approach to emotion cause detection for chinese microblogs. *Expert Systems with Applications*, 42(9):4517–4528.

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

Gildea, D. and Jurafsky, D. (2000). Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong, October. Association for Computational Linguistics.

Gui, L., Yuan, L., Xu, R., Liu, B., Lu, Q., and Zhou, Y. (2014). Emotion cause detection with linguistic construction in chinese weibo text. In *Natural Language Processing and Chinese Computing*, pages 457–464. Springer.

Gui, L., Wu, D., Xu, R., Lu, Q., and Zhou, Y. (2016). Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November. Association for Computational Linguistics.

Gui, L., Hu, J., He, Y., Xu, R., Lu, Q., and Du, J. (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, September. Association for Computational Linguistics.

Haider, T., Eger, S., Kim, E., Klinger, R., and Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France, May. European Language Resources Association (ELRA).

Hellrich, J., Buechel, S., and Hahn, U. (2019). Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Minneapolis, USA, June. Association for Computational Linguistics.

Herman, M. (2019). Djokovic happy to carry on cruising. https://www.reuters.com/article/us-tennis-frenchopen-djokovic/djokovic-happy-to-carry-on-cruising-idUSKCN1T41QK.

Kenton, L. (2019). A couple infuriated officials by landing their helicopter in the middle of a nature reserve. https://www.dailymail.co.uk/news/article-6858233/Couple-infuriate-officials-landing-helicopter-middle-California-nature-reserve.html.

Kim, E. and Klinger, R. (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, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

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

Lee, S. Y. M., Chen, Y., and Huang, C.-R. (2010). A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA, June. Association for Computational Linguistics.

Lee, Y. M. S. (2010). *A linguistic approach to emotion detection and classification*. Ph.D. thesis, The Hong Kong Polytechnic University.

Li, W. and Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749, March.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Liew, J. S. Y., Turtle, H. R., and Liddy, E. D. (2016). EmoTweet-28: A fine-grained emotion corpus for sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Liu, H., Li, S., Zhou, G., Huang, C.-R., and Li, P. (2013). Joint modeling of news reader's and comment writer's emotions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 511–515, Sofia, Bulgaria, August. Association for Computational Linguistics.

Liu, V., Banea, C., and Mihalcea, R. (2017). Grounded emotions. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 477–483, San Antonio, Texas, October.

Ma, Y., Peng, H., and Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5876–5883, New Orleand, Louisiana USA, February. Association for the Advancement of Artificial Intelligence.

Mihalcea, R. and Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea, July. Association for Computational Linguistics.

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

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. and Kiritchenko, S. (2018). Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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

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

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Mohammad, S. (2012). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Otero, V. (2018). Media Bias Chart. https://www.adfontesmedia.com/wp-content/uploads/2018/08/Media-Bias-Chart_4.0_8_28_2018-min.jpg.

Otero, V. (2019). Ad Fontes Media's First Multi-Analyst Content Analysis Ratings Project White Paper. https://adfontesmedia-demo.ehspook.com/wp-content/uploads/2019/08/Multi-Analyst-Ratings-Project-White-Paper-Aug-2019.pdf.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Plutchik, R. (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.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multiparty dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July. Association for Computational Linguistics.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (2016). Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15. Association for Computational Linguistics.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., and Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Schuff, H., Barnes, J., Mohme, J., Padó, S., and Klinger, R. (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. Workshop at Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Stieglitz, S. and Dang-Xuan, L. (2013). Emotions and information diffusion in social media – sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4):217–248.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tafreshi, S. and Diab, M. (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)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Tang, Y.-j. and Chen, H.-H. (2012). Mining sentiment words from microblogs for predicting writer-reader emotion transition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1226–1229, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Troiano, E., Padó, S., and Klinger, R. (2019). Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, July. Association for Computational Linguistics.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 616–626, Austin, Texas, November. Association for Computational Linguistics.

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

Xia, R., Zhang, M., and Ding, Z. (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, August. International Joint Conferences on Artificial Intelligence.

Xu, R., Hu, J., Lu, Q., Wu, D., and Gui, L. (2017). An ensemble approach for emotion cause detection with event extraction and multi-kernel svms. *Tsinghua Science and Technology*, 22(6):646–659.

Xu, B., Lin, H., Lin, Y., Diao, Y., Yang, L., and Xu, K. (2019). Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.

Xue, W. and Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July. Association for Computational Linguistics.

Yang, C., Lin, K. H.-Y., and Chen, H.-H. (2009). Writer meets reader: Emotion analysis of social media from both the writer's and reader's perspectives. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, volume 1 of *WI-IAT'09*, pages 287–290, USA, September. IEEE Computer Society.