Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts

Hongli Zhan[*1] **Tiberiu Sosea**[*2] **Cornelia Caragea**[2] **Junyi Jessy Li**[1]
[1]Department of Linguistics, The University of Texas at Austin
[2]Department of Computer Science, University of Illinois Chicago
{honglizhan,jessy}@utexas.edu {tsosea2,cornelia}@uic.edu

Abstract

Crises such as the COVID-19 pandemic continuously threaten our world and emotionally affect billions of people worldwide in distinct ways. Understanding the triggers leading to people's emotions is of crucial importance. Social media posts can be a good source of such analysis, yet these texts tend to be charged with multiple emotions, with triggers scattering across multiple sentences. This paper takes a novel angle, namely, *emotion detection and trigger summarization*, aiming to both detect perceived emotions in text, and summarize events and their appraisals that trigger each emotion. To support this goal, we introduce COVIDET (***Emotions and their Triggers during Covid-19***), a dataset of ~1,900 English Reddit posts related to COVID-19, which contains manual annotations of perceived emotions and abstractive summaries of their triggers described in the post. We develop strong baselines to jointly detect emotions and summarize emotion triggers. Our analyses show that COVIDET presents new challenges in emotion-specific summarization, as well as multi-emotion detection in long social media posts.

1 Introduction

Large-scale crises such as the COVID-19 pandemic continuously cause emotional turmoil worldwide. People are emotionally affected in different ways, e.g., online education has led to mental health issues among students (Akpınar et al., 2021) as well as parents (Cui et al., 2021); lock-down policies are protective for the vulnerable (Flaxman et al., 2020; Hsiang et al., 2020) while economically disastrous for many (Odii et al., 2021). Emotion analysis — both *detecting* emotion and understanding what *triggers* the emotion — brings invaluable insights both practically (e.g., for first responders, counselors, etc) and in scientific research (Arora et al., 2021; Uban et al., 2021).

[*]Hongli Zhan and Tiberiu Sosea contributed equally.

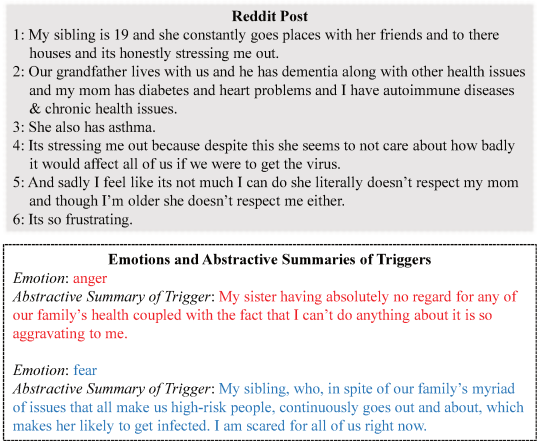

Figure 1: An example from COVIDET, with perceived emotion(s) identified and their trigger(s) summarized.

While emotion detection (typically formulated as a classification task among standard emotion taxonomies) is a well-established task (e.g., Mihalcea and Strapparava (2012); Wang et al. (2012); Abdul-Mageed and Ungar (2017); Khanpour and Caragea (2018); Demszky et al. (2020) and in crises contexts (Desai et al., 2020; Sosea et al., 2022)), fewer have studied ***what leads to these emotions*** in the scope of the text concerned in a data-driven manner. Xia and Ding (2019) adopt an extraction setup to identify emotion "causes" that is limited to the clause level, where only one (explicitly expressed) emotion and one cause are associated. This does not generalize to long, spontaneous social media posts that are emotionally charged. Illustrated in Figure 1, distinct emotions are triggered by different events across multiple sentences.

Additionally, how these events are subjectively evaluated, interpreted or *appraised*, e.g., "I can't do anything about it" in the first example of Figure 1, also contribute to the emotion (Smith and Ellsworth, 1985; Ellsworth and Scherer, 2003). The fact that different individuals may have distinct appraisals towards the same event (Moors et al., 2013) further highlights the challenging nature of

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9436 - 9453
December 7-11, 2022

understanding what triggers an emotion.

In this work we take a novel view, and formulate emotion-trigger detection as an *abstractive summarization* task that synthesizes a natural language description of the events and their appraisals that trigger a particular emotion. We frame our work as *emotion detection and trigger summarization* (Figure 1), which entails both detecting perceived emotions in text, and summarizing triggers for each emotion.

We present COVIDET (***E**motions and their **T**riggers during **Covid**-19*), a new dataset sourced from 1,883 English Reddit posts about the COVID-19 pandemic. Each post is annotated with 7 fine-grained emotion labels; for each emotion, annotators provided a concise, abstractive summary describing the triggers of the emotion. The triggers are further *validated* in a separate stage. COVIDET spans from June 2021 to January 2022, capturing various significant events as well as how they were emotionally appraised during the pandemic. Compared to prior emotion studies that consider only sentence-level texts (Sosea and Caragea, 2020; Demszky et al., 2020) or (short) tweets (Sosea et al., 2022; Abdul-Mageed and Ungar, 2017), COVIDET is challenging as it contains significantly longer texts. We showcase examples of COVIDET in Appendix §A.

Analyses of COVIDET reveal that negative emotions such as *fear* and *anger* are prevalent. These emotions co-occur most frequently with *anticipation*, which consistently rise after the Omicron subvariant became more dominant with *fear* dropping. Topic modeling over the trigger summaries points to irritations toward those who don't mask or get vaccinated, and positivity towards the vaccines.

Using COVIDET, we benchmark models for emotion detection and emotion-trigger summarization. We employ both separate emotion detection and trigger summarization models, as well as joint models that we designed to simultaneously detect emotions and generate trigger summaries. Our experiments showcase the distinct nature of our task, emphasizing that COVIDET is vital to training reliable emotion detection and trigger summarization approaches in a Covid-19 context. COVIDET bears various unique characteristics, ranging from its long sequences and invaluable context to the nature of the task itself. Therefore, general emotion detection or summarization models unsurprisingly lag behind in performance compared to our methods. Moreover, human evaluation of the generated trigger summaries tailored for emotion-trigger summarization indicates that our models are effective in capturing the underlying triggers of the post.

We release COVIDET and our code at https://github.com/honglizhan/CovidET.

2 Related Work

Summarization. Recent pre-trained models led to substantial progress in single document summarization. In the case of abstractive summarization, encoder-decoder transformer models are used to synthesize a concise description of the most salient concepts in the input (Lewis et al., 2020; Zhang et al., 2020). Significant efforts in summarization focus on news because of the availability of large datasets such as CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018); in the domain of social media, TL;DR sentences has been mined in Reddit to serve as summaries and train models (Völske et al., 2017; Kim et al., 2019). However, generic summaries tend not to be informative if users are concerned with specific emotions expressed.

In this sense our setup fits into settings where only a certain part of the content is of interest to the user. We could view our task as answering a query, "*Why does the writer feel [emotion]?*". However, such queries are more general than query-based summarization (Daumé III and Marcu, 2006; Otterbacher et al., 2009; Schilder and Kondadadi, 2008; Nema et al., 2017; Baumel et al., 2018; Laskar et al., 2020; Su et al., 2021; Zhong et al., 2021), where queries tend to be more document-specific. Perhaps a closer task is opinion summarization, or aspect-based summarization more generally. In opinion summarization, models need to summarize affect/opinions about a certain aspect of a service or product (Popescu and Etzioni, 2005; Angelidis and Lapata, 2018; Huy Tien et al., 2019; Suhara et al., 2020; Angelidis et al., 2021; Amplayo and Lapata, 2021); on the contrary, our setup entails identifying the emotions and summarizing the events and how they were made sense of with respect to each emotion. In aspect-based summarization, existing work has explored summarizing with respect to predesignated aspects of certain news (Frermann and Klementiev, 2019; Ahuja et al., 2022), and entities mentioned in text (Maddela et al., 2022).

Emotion Cause Extraction. Emotion Cause Extraction (ECE) is a task that aims to extract the

events triggering a particular emotion (Khunteta and Singh, 2021). ECE was first introduced by Lee et al. (2010), where they defined the task as extracting word-level causes to the given emotion in text. Chen et al. (2010) and Gui et al. (2016) expanded the task to clause-level cause detection; Xia and Ding (2019) aimed to removed the constraint that emotions must be human-annotated before conducting automatic cause extraction, and thus proposed Emotion-Cause Pair Extraction (ECPE) aiming to extract potential pairs of emotions and causes in a document. Most of the datasets are in Chinese, in either micro-blog or news domains (Gao et al., 2015; Gui et al., 2016; Gao et al., 2017).

In contrast, we study a more generalized notion of *triggers* of an emotion where readers are asked to actively appraise and interpret the emotions together with their stimuli in the document, rather than solely identifying the events behind each emotion. We use abstractive summarization to handle triggers, which can better synthesize interconnected complex events and abstract concepts, as well as making the output contextually independent.

3 Dataset Construction

We present COVIDET, a novel dataset from English Reddit posts that is manually annotated with emotions and summaries of their triggers. This section discusses the data creation process; in §4, we discuss inter-annotator agreement and our human verification process.

3.1 Selecting & Curating Reddit Posts

We gather posts from `r/COVID19_support`[1]. We select it as the source of our data because of its rich personal narration: rather than COVID-19 news snippets, this subreddit is targeted for people seeking any community support during the pandemic. We randomly sample posts before (from Jun 23, 2021 to Oct 1, 2021) and after (from Dec 1, 2021 to Jan 25, 2022) Omicron, a COVID-19 variant that emerged during December 2021.

We restrict posts to be between 50-400 tokens long (punctuation excluded); this allows us to have posts that are long enough, but still manageable for crowdsourcing tasks. Close scrutiny shows that the posts in COVIDET center around 100 tokens in length; the distribution of the length of the posts is given in Figure 2. The average length of posts

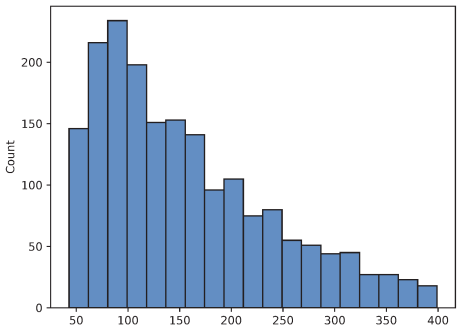

Figure 2: Distribution of the length of posts in COVIDET.

in COVIDET is 156.4 tokens (std.dev = 83.3). We mask web links with an `[url]` token and do not provide the metadata to our annotators. Note that 6 posts have length under 50 tokens: this is because we performed `[url]` masking *after* length filtering when collecting the source data. Details of the full preprocessing procedure are provided in Appendix §B.

3.2 Annotation Task

Instructions. Annotators are first asked to annotate Plutchik basic emotions (Plutchik, 2001) they perceive: *anger, anticipation, joy, trust, fear, sadness, and disgust.*[2] Multiple selection is allowed, and we also provide a *none of the above* option in case no emotion is perceived.

Once the annotators select an emotion, they are asked to summarize the trigger(s) to their perceived emotions, specifically an abstractive summary in their own words, in the author's voice. The summaries should contain trigger(s) to the emotion rather than just reflecting the emotion itself. We provide the detailed instructions to our annotation task in Appendix §C.

Annotators We recruit two different groups of annotators. The first group consists of trained turkers from Amazon Mechanical Turk. The workers' locale is the US, and they have completed 500+ HITs with an acceptance rate $\geq 95\%$. This group contributes to COVIDET's training and validation sets. The second group consists of 2 linguistic undergraduate students, who contributes to the test set. To ensure the quality of COVIDET, both groups of annotators are trained and qualified in a prequalification process. We also ask them to revise their work when needed during annotation.

[1] `https://www.reddit.com/r/COVID19_support/`

[2] After annotation, we found very little *surprise* in the data (59 in total), thus we leave out *surprise* for this work.

Pre-Annotation Training We trained the annotators before they annotate COVIDET. We set up a qualification task on the Amazon Mechanical Turk. The qualification task involves 3 posts, and annotators are required to complete the qualification task. Through manually examining the annotators' work on the qualification task and comparing the annotations to the *gold* annotations we develop, we filter high-quality annotators and give them the access to our annotation task. We also provide feedback to their annotations. The turkers are paid at least $10 per hour. To ensure this goal is reached, we keep track of their working time on the backstage and give out bonuses accordingly when needed.

Annotation Revisions During the process of the annotation on COVIDET, we regularly review the annotations and give feedback accordingly. When needed, we send the annotations back to the annotator along with the original post, and ask them to revise their work based on our suggestions. Note that the annotator is responsible for the revision of *their own* work only.

3.3 Benchmark Dataset

We annotated $1,485$ posts on the Amazon Mechanical Turk, each annotated by two independent workers. Since the neutral class is very infrequent, we remove it from our experiments. To facilitate our experiments, we split the examples into $1,200$ examples for training and 285 examples for validation. Our test set—which is annotated by linguistic undergraduates—contains 398 examples.

If at least one annotator labels a post with an emotion e, then we include emotion e as an emotion label. In cases where both annotators assign an emotion e to a post, we consider the trigger summaries as two separate training examples for the trigger summarization task. In cases where a post has two different trigger summaries in the validation or the test set, we consider them as multiple references when computing our metrics.

4 Agreement and Validation

To account for the quality of COVIDET, we measure the inter-annotator agreement in emotions (§4.1) and triggers (§4.2). The annotations are further validated through human inspection in §4.3. Results reveal that annotators tend to agree with each other in emotions whilst using varied vocabularies when summarizing the triggers.

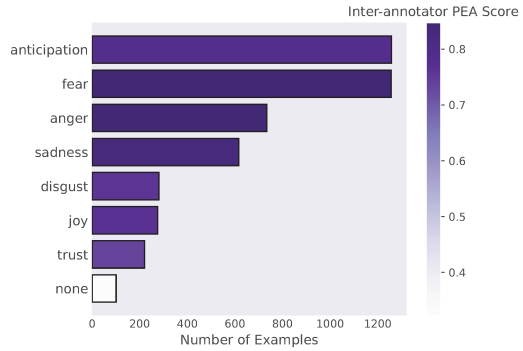

Figure 3: Emotion distribution of COVIDET, ranked by the number of examples. Colors indicate the inter-annotator agreement measured by the PEA score.

4.1 Agreement in Emotions

Percentage Overlap. For each example in COVIDET, we measure the number of emotions in which both annotators agree upon. Results show that in 81.4% of the examples in COVIDET, both annotators agree on at least 1 emotion label; in 26.6% of the examples, both annotators agree on at least 2 emotion labels.

PEA Score. To account for distances between emotions (e.g., disgust is further away from joy than from anger), we report the Plutchik Emotion Agreement (PEA) metric (Desai et al., 2020) for the inter-annotator agreement of emotions annotated in COVIDET. We first report the average PEA score among annotators weighted by their numbers of annotations, which is 0.8 for the training and validation sets combined, and 0.821 for the test set (0.804 for all three combined). These numbers indicate high agreement (Desai et al., 2020).

Figure 3 shows per-emotion PEA scores, along with the frequency of each emotion. All emotions have high agreement; the highest are among *fear* and *anger*, with the average PEA scores at around 0.85; the lowest is *trust*, with the average PEA score at around 0.74.

Finally, to calculate agreement between students and crowd workers, we randomly sample 208 examples from the training set and ask the linguistic undergraduate students to annotate them from scratch. We assign one student per example for validation. The average PEA score between crowd workers and linguistics students is 0.832, suggesting high agreement.

	AGR	DSG	FER	JOY	SDN	TRS	ANC	Avg
Emotion	0.96	0.92	0.96	1	0.96	0.88	0.92	0.94
Trigger	0.92	0.92	0.96	1	0.96	0.84	0.88	0.93

Table 1: Human validation results on the annotated emotions and abstractive summaries of triggers.

4.2 Similarity in Triggers

We further examine the similarity in the annotated summaries of triggers when two annotators both select the same emotion for one example, using ROUGE (Lin, 2004) for lexical overlap and BERTScore (Zhang et al., 2019) for semantic similarity. The average BERTScore (F1) between the two annotators is 0.883, indicating highly similar summaries. Yet the lexical overlap is low: the average ROUGE F scores between two annotators are: ROUGE-1: 0.255, ROUGE-2: 0.055, ROUGE-L: 0.190.

For those posts doubly annotated by linguistics students and crowd workers, the ROUGE values are similar for students vs. workers: BERTScore: 0.876; ROUGE-1: 0.246, ROUGE-2: 0.063, ROUGE-L: 0.188. [3]

4.3 Human Validation

In addition to the automatic evaluation metrics above, we also validate the emotion-trigger annotations in COVIDET through human inspections. We set up a human validation task on the Amazon Mechanical Turk, and recruit a new group of qualified workers. We randomly sample 300 examples from our training set for validation. The emotion annotations, as well as the abstractive summaries of triggers, are validated.

We describe the validation framework as follows. The validators are given an annotated trigger summary. We first validate whether the summary actually indicates the annotated emotion by asking a yes/no question. Next, if the validator confirms the presence of emotion in the summary, we then ask whether the summary indeed expresses the *trigger* and not the *emotion* by raising another yes/no question. We present the validation results based on the abstractive summaries in Table 1. The numbers indicate the proportion of examples on which validators confirm upon.

Overall, the human validation results indicate fairly high correctness in our annotations. It should be noted that annotators commonly adopt some sentence patterns that can be easily identified as emotion triggers. For example, in expressing the abstractive trigger for *anger*, an annotation in COVIDET is *I am angry that they would put me at risk of catching COVID and not tell me*, a sentence which is highly linguistically explicit of the emotion.

[3] Multi-reference ROUGE and BERTScore are applied in cases where all 3 annotators agree in the same emotion.

AGR	DSG	FER	JOY	SDN	TRS	ANC
covid	disgusted	covid	happy	sad	trust	covid
annoyed	covid	afraid	covid	covid	covid	expect
people	people	getting	vaccinated	feel	vaccine	looking
angry	feel	scared	pandemic	pandemic	information	know
don	don	going	vaccine	life	people	interested
vaccinated	pandemic	vaccine	getting	don	believe	people
pandemic	getting	worried	people	like	vaccines	symptoms
just	like	risk	feel	just	help	test
want	vaccine	concerned	better	people	pandemic	getting
life	just	health	good	friends	vaccinated	want
going	mask	vaccinated	able	time	credible	going
vaccine	going	symptoms	really	lost	protect	vaccinated
getting	vaccinated	fear	know	going	know	vaccine
family	want	don	news	really	end	positive
really	family	effects	vaccination	want	feel	guidance

Table 2: Results of topic modelling through LDA (Blei et al., 2003). The words are associated with the most prominent topic among the abstractive summaries of triggers of each emotion category in COVIDET.

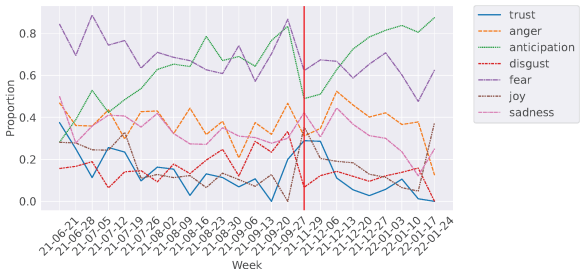

Figure 4: Emotion distribution in COVIDET over time (by week).

5 Data Analysis

Emotion Distribution. On average, there are 2.46 emotions ("none" excluded) per example in COVIDET. Figure 3 shows the general emotion distribution of COVIDET. *Fear* is the most common emotion in COVIDET, closely followed by *anticipation*. There is clearly a gap among the emotions, with positively valenced emotions such as *trust* and *joy* rarely present in COVIDET. This is predicted given the catastrophic nature of our domain.

We present the emotion co-occurrence heatmap in Figure 5. *Anticipation* co-occurs with *fear* and *anger* most frequently in COVIDET. Close scrutiny of the data reveals that the poster is either predicting negative events during COVID-19, or expecting advice on what do do under austere situations.

Emotion Trend. We present a temporal analysis of the emotion distribution by week in Figure 4, using a red vertical line to separate pre- and post-Omicron. Interestingly, we notice that the amount of *anticipation* consistently rises after the outbreak of the Omicron COVID-19 variant, whereas the expression of negative emotions including *anger* and *fear* becomes less prevalent, possibly due to the nature of the Omicron variant, which was less harmful compared to previous variants (Sigal, 2022). This result is also unsurprising in that people are getting weary and tired after two years of avoiding COVID-19.

Trigger Summary Abstractiveness. The average length of trigger summaries is: 130.9 characters / 26.9 words / 1.2 sentences. We measure the abstractiveness of the annotated abstractive summaries of triggers by computing the ROUGE score between the abstractive summaries and the post. we use ROUGE-n precision scores to calculate how abstractive the annotated abstractive summaries are compared to the post. Results are: ROUGE-1: 0.576, ROUGE-2: 0.149, ROUGE-L: 0.392. The results indicate that the trigger summaries are fairly abstractive with respect to the original posts in CovidET.

Topic Variation. To better understand the triggers of each emotion, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to extract the topics in the trigger summaries of each emotion. The triggers are lower-cased, and punctuation as well as stopwords are removed. We showcase the unigrams corresponding to the most prominent topics in Table 2. We observe a clear difference among the topics of triggers behind the emotions. For example, we notice words such as *don*, *vaccinated*, and *mask* in emotions like *anger* or *disgust*, suggesting that the posters are annoyed that people are not masking or vaccinated to prevent the spread of the pandemic. On the other hand, we see words such as *vaccine*, *believe*, and *credible* in *trust*, denoting that the posters believe in the capability of the vaccines to protect them from the virus.

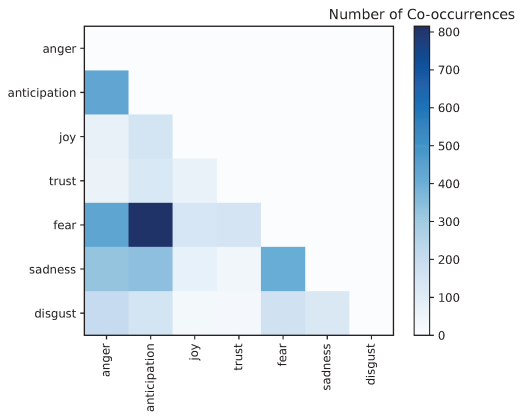

Figure 5: Emotion co-occurrences in CovidET.

6 Methods

We discuss our methods across three main dimensions: emotion detection, summarization, and joint emotion detection and trigger summarization.

6.1 Separate Models

Emotion Detection. To perform emotion detection, we experiment with **1)** EmoLex, a weak baseline based on the EmoLex lexicon (Mohammad and Turney, 2013), where words are associated with Plutchik basic emotions. For each post, we assign an emotion e if there exists a word from EmoLex associated with e. **2)** GoEmotions (Demszky et al., 2020), which involves training a BERT-large model (Devlin et al., 2019) on the GoEmotions dataset, which is composed of sentence-level examples from a general Reddit domain. **3)** HurricaneEMO (Desai et al., 2020), the same approach with the model trained on a Twitter disaster dataset. Finally, we use a **4)** BERT-large model fine-tuned on CovidET using the [CLS] token and an additional linear layer to classify the entire post.

Abstractive Summarization. We perform abstractive trigger summarization using two backbone models: **1)** Pegasus (Zhang et al., 2020) pretrained on Reddit TIFU (Kim et al., 2019) and **2)** BART (Lewis et al., 2020) pretrained on CNN/DailyMail (Hermann et al., 2015). For each model, we evaluate the summaries with and without fine-tuning on CovidET. We employ a separate summarization model for each emotion e, which we fine-tune using the abstractive summaries of triggers for e. We also experiment with two standard heuristic baselines: i.e., considering the first sentence in the post (1-SENT) or the first three sentences (3-SENT) as the trigger summary.

6.2 Joint Emotion Detection and Trigger Summarization

We propose a joint model based on BART that can be trained to simultaneously perform emotion detection and abstractive trigger summarization for a particular emotion e using a multitasking

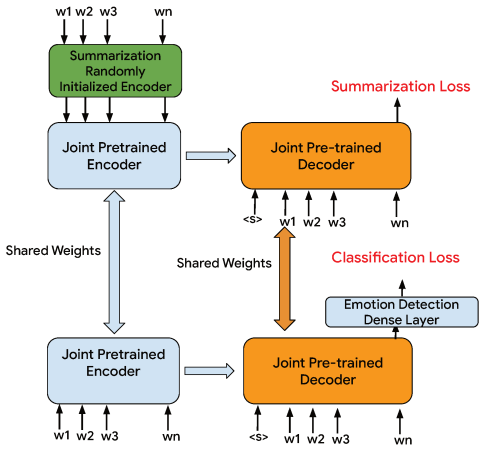

Figure 6: Architecture of our joint emotion detection and trigger summarization model.

framework. The model follows the architecture of BART (Lewis et al., 2020), where we add a single linear layer for emotion classification. We show the architecture of our model in Figure 6 and detail our training procedure as follows: Given an emotion e and a batch size of B, we first sample a positive set of $\frac{B}{2}$ examples: $X_p = \{(x_1, y_1, s_1), (x_2, y_2, s_2), ...(x_{\frac{B}{2}}, y_{\frac{B}{2}}, s_{\frac{B}{2}})\}$, where

$$y_i = e \quad | \quad i = 1...\frac{B}{2} \tag{1}$$

and s_i is an abstractive summary of the trigger for emotion y_i and post x_i. Next, we sample a set of negative examples for classification of the same size as follows: $X_n = \{(x_{\frac{B}{2}+1}, y_{\frac{B}{2}+1}), (x_{\frac{B}{2}+2}, y_{\frac{B}{2}+2}), ...(x_B, y_B)\}$, where:

$$y_i \neq e \quad | \quad i = \frac{B}{2}...B \tag{2}$$

Finally, we use a weighted combinations of the summarization and classification losses to train our model:

$$L = \lambda * \sum_{i=0}^{B} L_e(x_i, y_i) + (1 - \lambda) * \sum_{i=0}^{\frac{B}{2}} L_s(x_i, s_i) \tag{3}$$

where L_e and L_s are the regular classification and summarization losses.

7 Experiments and Results

7.1 Experimental Setup

We carry out all our experiments on an Nvidia A5000 GPU. We use the HuggingFace Transformers (Wolf et al., 2020) library for our model implementations and we will make the code for our methods and data available. We report the performance for emotion detection in terms of F1 and use automatic approaches such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) to evaluate the summarization performance. To enable a fair comparison with the joint model, for summarization, we only consider test examples where the joint model emotion predictions are correct to compute summarization metrics. We run our approaches five times with different model initializations and report average values. We provide extensive details about our hyperparameters, such as batch size or loss weighting λ in Appendix §D. Additionally, we carry out an extensive human evaluation of trigger summaries generated by our BART-FT-JOINT model and a general BART (Lewis et al., 2020) model trained on CNN/DailyMail.

	AGR	DSG	FER	JOY	SDN	TRS	ANC	AVG
EMOLEX	35.6	20.5	56.7	48.7	42.5	13.5	17.8	33.6
GOEMOTIONS	45.4	20.1	65.3	50.4	58.3	15.1	41.3	42.2
HURRICANEEMO	37.1	16.8	58.3	45.2	60.7	17.2	43.7	39.9
BERT-LARGE	68.1	20.2	86.8	54.2	69.5	20.3	64.5	54.8
BART-FT-JOINT	$69.5^{\dagger}$	20.6	$87.8^{\dagger}$	54.7	$71.3^{\dagger}$	20.8	$65.9^{\dagger}$	$55.8^{\dagger}$

Table 3: Results of our models in terms of F1 on emotion detection. We report the average performance of five independent runs. We use bootstrap statistical significance† testing over BERT-LARGE with $p < 0.05$ and 200 samples of size 10% of the test set.

7.2 Results

Emotion Detection. We show the F1s obtained using our models on emotion detection in Table 3. First, we observe that our lexicon-based EmoLex approach performs poorly compared to other methods. We also note that approaches trained outside our domain lag behind considerably compared to approaches trained on our data. Specifically, a BERT large model trained on our data outperforms the GoEmotions model by as much as 23% in F1 on anger and 28% in fear. We observe the same trend for models trained on hurricane disasters, which decrease the performance by 38% on fear and 9% on joy. This result indicates that models trained on natural disasters generalize poorly to Covid-19, further emphasizing the uniqueness of our dataset. We also note that our BART-FT-JOINT model, which is trained on our data to perform both detection and summarization obtains an average improvement of 1% over the BERT-large model.

Trigger Summarization. We show in Table 4 the results obtained in terms of ROUGE-L and

	ANGER		DISGUST		FEAR		JOY		SADNESS		TRUST		ANTICIPATION	
	R-L	BERTSc	R-L	BERTSc	R-L	BERTSc	R-L	BERTSc	R-L	BERTSc	R-L	BERTSc	R-L	BERTSc
1-SENT	0.121	0.575	0.112	0.545	0.122	0.528	0.103	0.518	0.115	0.506	0.118	0.537	0.119	0.507
3-SENT	0.142	0.598	0.129	0.562	0.153	0.535	0.154	0.537	0.134	0.517	0.152	0.548	0.142	0.527
PEGASUS	0.164	0.594	0.141	0.560	0.161	0.548	0.155	0.536	0.153	0.562	0.151	0.546	0.153	0.542
BART	0.161	0.587	0.138	0.558	0.164	0.529	0.149	0.551	0.157	0.559	0.158	0.571	0.164	0.558
PEGASUS-FT	0.185	0.681	0.155	0.713	0.199	0.739	0.158	0.683	0.173	0.705	0.164	0.663	0.193	0.736
BART-FT	0.190	0.705	0.159	0.695	0.206	0.748	0.165	0.699	0.177	0.718	0.162	0.653	0.198	0.749
BART-FT-JOINT	0.190	0.701	0.158	0.706	0.203	0.729	0.163	0.694	0.175	0.713	0.165	0.659	0.196	0.746

Table 4: Results of our models in terms of ROUGE-L and BERTScore on the trigger summarization subtask of emotion detection and trigger summarization. We report the average performance of five independent runs.

METRIC	*Coherence*	*Consistency*	*Fluency*	*Relevance*	*Extractiveness*
BART	4.947	5.000	4.974	2.158	4.970
BART-FT-JOINT	4.262	3.548	4.286	4.048	2.530

Table 5: Results of our trigger summary human evaluation procedure along four quality assessment dimensions.

BERTScore on the summarization task. First, we note that basic approaches such as 1-SENT or 3-SENT, which select the first sentences in a post as the trigger summaries, perform similarly to general summarization models like the Pegasus model trained on Reddit TIFU or the BART trained on CNN/DailyMail. This result highlights the distinct nature of our trigger summarization task, which bears very few similarities with a general summarization task. Fine-tuning these models on our data, however, brings substantial improvements. We see improvements as large as 18% in terms of BERTScore by fine-tuning a BART model on anger and 19% on anticipation. Our fine-tuned models also consistently outperform the baselines in ROUGE-L. For instance, our fine-tuned Pegasus obtains an improvement of 4.2% ROUGE-L on fear and 2% on sadness. We note that applying our joint model results in no loss of performance across all emotions.

We emphasize that in practice, generating trigger summaries and detecting emotions using a joint model has various advantages over single-task approaches, such as reduced memory footprint (i.e., by using a single model) and reduced inference time. Moreover, our approach improves the performance in emotion detection.

7.3 Human Evaluation of Model Summaries

We perform human evaluation and qualitative analysis of our model-generated trigger summaries to measure the overall quality and compare our BART-FT-JOINT model against a general BART summarization model.

Following Fabbri et al. (2021), we instruct two

Reddit Post

1: I have been realizing that I've been spiraling out of control lately on account of the Delta Variant reports, particularly the WHO message.
2: As of right now we have not been seeing many case increases here in the US.
3: The occasional rise, but so far it hasn't been huge.
4: That's why I feel I should take time off from looking at the updates of the Delta Variant.
5: The fact that we are still making progress with vaccination should remind myself that we are still on the path to beating the pandemic and that these restrictions are soon going to be gone.
6: THAT'S the motivation I should have.
7: Yes, we still have to be vigilant and yes, we DO have more to vaccinate, but things are just so much better now than before.
8: And for that, I personally think we should stay hopeful, not fearful.
9: This is probably the best I can give to others who feel similar worries.

Emotions and Abstractive Summaries of Triggers

Emotion: joy
Annotated Abstractive Summary of Trigger: The pandemic situation is objectively not as bad as I think it is right now. The pandemic is not going to be forever and we are continuing to vaccinate. We're doing a lot better compared to when the pandemic first started.
Model Generated Summary: I am happy that we are on the path to beating the Delta pandemic because that gives me some reason to be hopeful and hopeful that things will get better. We still have a long way to go but we have a lot more to be happy about than we did before.

Emotion: anticipation
Annotated Abstractive Summary of Trigger: I believe that as long as we continue to get people vaccinated the pandemic will be over soon.
Model Generated Summary: I expect that we are on the path to beating the pandemic and that these restrictions are soon going to be gone, so I expect that I should take a break from all of the Delta variant updates and focus on other things that matter, like the progress we are making with vaccination.

Figure 7: Human Evaluation Example.

expert annotators with linguistics expertise to grade with a score from 1 to 5 (where 1 is the lowest score) 21 trigger summaries generated by our joint model (three per emotion) along four dimensions: *Coherence*, *Consistency*, *Fluency*, and *Relevance*. Coherence refers to the collective quality of all the sentences in the summary and consistency measures the factual alignment between the summary and the summarized source. Next, we evaluate the quality of individual sentences from the post using fluency and measure how well the summary captures the emotion triggers through relevance. To offer a better understanding of these metrics, we detail them further in Appendix §E. Additionally, we also evaluate the summaries for the amount of *Extractiveness* (i.e., the amount of information copied from the orginial post).

We show the evaluation results in Table 5. The reported metrics are the average scores of the two individual annotators' scores. We measure the

agreement between the two annotators by computing the average score differences between their responses.

Evaluation of BART-FT-JOINT yields a small average difference of 0.690, indicating that the two annotators have good agreement on the assigned scores. The generated summaries have a good quality, with an average score of 4. We also note that the lowest score of 3.548 is obtained on consistency, indicating that the model can introduce non-factual details, and emphasize that our summarization model performs well identifying triggers, where it obtains a score of 4.048. To offer additional insights into the summaries generated by our joint model, we show an example in Figure 7. The post is annotated as joy and anticipation, and we provide both the gold and the model generated summaries. The summary for joy emotion is extremely effective capturing the trigger; i.e., the progress towards beating the Delta variant. However, we also note some model errors, such as the repetition of the word "hopeful". The annotators indicate that the model outputs tends to be two sentences long and the overall quality is good. Besides scoring the summaries, we also instruct annotators to spot such mistakes of the model in order to identify potential areas of improvement. We detail our findings in Appendix §F.

As mentioned, we also provide in Table 5 the Likert scoring of the generic summarization model by linguistic experts. Inspection of the data reveals that the generic summaries tend to be word-to-word extractive of the original post, leading to high scores in coherence, consistency, and fluency. However, the generic summaries perform badly in terms of relevance, suggesting that the models are not capturing the triggers of the emotions. This is also reflected in the low BERTScore performance for the generic models.

8 Conclusion

We propose a new task entitled *emotion detection and trigger summarization*, which aims to jointly detect perceived emotions in text and summarize the events as well as their appraisals that trigger each emotion. To address the task, we introduce COVIDET, a dataset of 1,883 English Reddit posts on COVID-19 annotated with emotions and abstractive summaries of their triggers. Experiments using our proposed joint model on the dataset reveal that COVIDET is vital resource for training models to capture emotions and their triggers in text. Our thorough evaluation of model-generated summaries emphasize that COVIDET is a challenging benchmark, and our error analysis indicates potential areas of improvements (e.g., improving the factuality of the summaries).

Limitations

This work presents a new dataset to address the task of detecting perceived emotions and summarizing their triggers in text. While picking the COVID-19 as our topic enables meaningful, real-world applications and allows us to access emotionally rich text, the emotion labels in COVIDET are highly unbalanced: negative emotions such as *fear* and *anger* are more prevalent. This makes it particularly challenging to train emotion detection and summarization models on emotions with few examples (e.g., *trust* and *joy*). Moreover, due to the lack of controllability and interpretability of end-to-end summarization models, we acknowledge the potential risks of generating biased or inappropriate trigger summaries for certain posts. In particular, our results revealed consistency and factuality issues that exist in modern abstractive summarization systems.

Acknowledgements

We thank Desmond Ong and the anonymous reviewers for their valuable comments and feedback, which helped improve our paper. We also thank our annotators for their dedication and hard work. We thank John Henry Cruz for his help with Reddit data. This research is partially supported by Good Systems,[4] a UT Austin Grand Challenge to develop responsible AI technologies, and NSF grants IIS-2145479, IIS-2107524, IIS-2107487, and BigData-1912887. We acknowledge the Texas Advanced Computing Center (TACC)[5] at UT Austin and AWS for many of the results within this paper.

References

Muhammad Abdul-Mageed and Lyle Ungar. 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. Association for Computational Linguistics.

[4] https://goodsystems.utexas.edu/

[5] https://www.tacc.utexas.edu

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Ezgin Akpınar et al. 2021. The effect of online learning on tertiary level students mental health during the covid-19 lockdown. *The European Journal of Social & Behavioural Sciences*.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Anshika Arora, Pinaki Chakraborty, M. P. S. Bhatia, and Prabhat Mittal. 2021. Role of Emotion in Excessive Use of Twitter During COVID-19 Imposed Lockdown in India. *Journal of Technology in Behavioral Science*, 6(2):370–377.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *CoRR*, abs/1801.07704.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China.

Shu Cui, Chao Zhang, Shijiang Wang, Xingong Zhang, Lei Wang, Ling Zhang, Qiuyu Yuan, Cui Huang, Fangshuo Cheng, Kai Zhang, and Xiaoqin Zhou. 2021. Experiences and attitudes of elementary school students and their parents toward online learning in China during the COVID-19 pandemic: Questionnaire study. *Journal of Medical Internet Research*, 23(5):e24496.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

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

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion.* Oxford University Press.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell,

Samir Bhatt, and Imperial College COVID-19 Response Team. 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517–4528.

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

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

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*.

Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, Peiley Lau, Jaecheol Lee, Esther Rolf, Jeanette Tseng, and Tiffany Wu. 2020. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*, 584(7820):262–267.

Nguyen Huy Tien, Le Tung Thanh, and Nguyen Minh Le. 2019. Opinions summarization: Aspect similarity recognition relaxes the constraint of predefined aspects. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 487–496, Varna, Bulgaria. INCOMA Ltd.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

Arunima Khunteta and Pardeep Singh. 2021. Emotion cause extraction - a review of various methods and corpora. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 314–319.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. 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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Enamul Hoque, and Xiangji Huang. 2020. Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Canadian Conference on AI*.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause events: Corpus construction and analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. EntSUM: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 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. Association for Computational Linguistics.

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

Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion Review*, 5(2):119–124.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Aloysius Odii, Millicent Odinakachukwu Ngwu, Mmesoma Clara Aniakor, Ifunanya Crystal Owelle, Maureen Chimaoge Aniagboso, and Onyinye Winifred Uzuanwu. 2021. Effect of covid-19 lockdown on poor urban households in nigeria : Where do we go from here? *Ianna Journal of Interdisciplinary Studies*, 2(1):75–85.

Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev. 2009. Biased lexrank: Passage retrieval using random walks with question-based priors. *Information Processing & Management*, 45(1):42–54.

Robert Plutchik. 1958. Section of psychology: Outlines of a new theory of emotion*. *Transactions of the New York Academy of Sciences*, 20(5 Series II):394–403.

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

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Frank Schilder and Ravikumar Kondadadi. 2008. FastSum: Fast and accurate query-based multi-document summarization. In *Proceedings of ACL-08: HLT, Short Papers*, pages 205–208, Columbus, Ohio. Association for Computational Linguistics.

Alex Sigal. 2022. Milder disease with Omicron: is it the virus or the pre-existing immunity? *Nature Reviews Immunology*, 22(2):69–71.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Tiberiu Sosea and Cornelia Caragea. 2020. CancerEmo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904, Online. Association for Computational Linguistics.

Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea, and Junyi Jessy Li. 2022. Emotion analysis and detection during COVID-19. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6938–6947, Marseille, France. European Language Resources Association.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3124–3131, Online. Association for Computational Linguistics.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and A. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

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

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Dataset Examples

An example of COVIDET is shown in Figure 8. This example includes annotations from both annotators. Annotations for different emotions are in distinct colors.

B Data Curation Details

Here we detail the preprocessing procedure over the source data. We preprocess the source data using regular expressions. As the first step, we tokenize posts into individual words. Specifically, we apply the following regular expressions in combination with the NLTK word_tokenize package to tokenize posts into words:

```
re.sub("\s+","␣", post)
re.sub(r'(?<=[.,!?:])(?=[^\s])', r'␣', post)
re.sub(r'\s([?.!,:"](?:\s|$))', r'\1', post)
nltk.tokenize.word_tokenize(post)
```

Then we exclude punctuation from the tokenized posts and filter the posts that are 50-400 tokens long. Finally, we mask web links by substituting them into `[url]` tokens using the following regular expressions:

```
pandas.Series.str.replace(r'http\S+', '[url]').
    str.strip()
pandas.Series.str.replace(r'''(?i)\b((?:https
    ?://|www\d{0,3}[.]|[a-z0-9.\-]+[.][a-z
    ]{2,4}/)(?:[^\s()<>]+|\(([^\s()<>]+|(\([^\s
    ()<>]+\)))*\))+(?:\(([^\s()<>]+|(\([^\s()
    <>]+\)))*\)|[^\s`!()\[\]{};:'".,<>?«»“”‘’]))
    ''', '[url]').str.strip()
```

C Annotation Instructions

Comprehensive instructions are provided to the annotators, as demonstrated in Figure 9. Note that the instruction page pops up as a modal before every annotation, so as to remind the annotators of the task framework. We also ask the annotators to pay special attention to a few principles as follows. For the emotion annotations, we ask annotators to follow the emotion guidelines on the Six Seconds website[6] and interpret *anticipation* as (good or bad) *expectancy* (Plutchik, 1958). For the trigger annotations, we instruct annotators to annotate summaries containing *triggers* that lead to the emotion instead of sentences expressing the *emotion* itself.

The layout of our annotation task is shown in Figure 10.

D Hyperparameters

In this section, we detail the hyperparameter search space and the final hyperparameters used by our joint BART-FT-JOINT model, which were chosen based on the best validation performance. Specifically, we show the values for the learning rate, batch size and multitasking loss weighting term λ in Table 6. In terms of search space, we tried batches in the range $4 \rightarrow 64$ and learning rates in the range $1e-5 \rightarrow 9e-5$ with a step of $1e-5$. We also search a suitable λ in the range $0.1 \rightarrow 0.9$. We decode our summaries using beam search decoding and a beam size of 4. Training BART-FT-JOINT model on our A5000 GPU takes ~1 hour to complete for each emotion.

E Human Evaluation Instructions

We provide the detailed instructions for human evaluation in Figure 11.

F Human Evaluation Summary Errors

We instructed our expert human evaluators to find potential areas of improvement of our BART-FT-JOINT summarization model by identifying frequent errors the model makes. In this section, we analyze our findings and present a few examples in Table 7. Specifically, the annotators pointed out four main model errors: **1)** Non-factual relative clauses; **2)** Model summary includes information in the summary that is not discussed in the post; **3)** At least a few sentences in the model summary are formatted to make the text difficult to read; and **4)** The overall model summary is not well-structured.

[6] https://www.6seconds.org/2020/08/11/plutchik-wheel-emotions/

Reddit Post

Aaron Astor has made an interesting discovery on the Delta Variant, according to his Twitter. According to graphs in Scotland, the variant may hit hard and fast, but it ultimately doesn't do as bad a damage as other variants. In fact, Scotland's cases peaked at June 30 after having a big spike. But now the cases have since crashed. Big time. More importantly, the hospitalization's peak, two weeks after, topped out at 1/4 that of Alpha. Again, most of that has to do with how vaccinated Scotland was. More interestingly, the most vaccinated regions didn't have such a major impact with it and barely had any huge numbers. The unvaccinated ones, on the other hand, did. What does all this mean? One, it means that perhaps the Delta Variant wave won't be as long or as massively damaging as some people are fearing, and two, the vaccine helps. Again, the more we vaccinate, the faster we'll be out of this. But having said that, the Delta Variant's wave thankfully may not be as lengthy. That's attributed to how much vaccinations we have made. The more people we do this to, the better. I hope I am not giving any false hopes, but this post DID have me intrigued.

Annotator 1

- Anticipation

→ *Abstractive Summary of Trigger*: I hope that as vaccination rates rise the pandemic will improve.

- Joy

→ *Abstractive Summary of Trigger*: It's comforting to know that we might not have to worry as much about Delta, especially in highly vaccinated areas.

- Trust

→ *Abstractive Summary of Trigger*: I trust Aaron Astor as a reliable source for COVID.

Annotator 2

- Anticipation

→ *Abstractive Summary of Trigger*: With how weak Delta has been in comparison to the other variants, the wave is probably not going to wreak havoc or last long.

- Joy

→ *Abstractive Summary of Trigger*: I'm happy that Delta doesn't seem to be as bad as other variants, that Delta cases are falling, and that the peak isn't as bad as the first variants. I'm also happy to see that vaccination is working.

- Trust

→ *Abstractive Summary of Trigger*: I believe Aaron Astor is credible. Vaccination is working and I trust it to shorten the pandemic.

Figure 8: Example of COVIDET.

	ANGER	DISGUST	FEAR	JOY	SADNESS	TRUST	ANTICIPATION
Batch Size	32	32	32	8	32	8	16
Learning Rate	$2e-5$	$4e-5$	$5e-5$	$3e-5$	$3e-5$	$5e-5$	$3e-5$
Loss weight λ	0.1	0.1	0.2	0.1	0.1	0.2	0.1

Table 6: Hyperparameters of our BART-FT-JOINT model.

POST	SUMMARY
Non-factual relative clauses	
I am visiting family this week. All of my family members who are eligible for the vaccine, including myself. have gotten vaccinated. The only people who arent are my niece and nephew, because they are only 4 and 1 years old. I wanted to see an old friend from high school, but I found out that this friend did not get vaccinated. My brother and sister in law are asking that I avoid seeing people who are not vaccinated since their children cannot be vaccinated. Is it too much to completely avoid seeing this person at the wishes of my brother, or would it be safe enough to see this person in an outdoor setting socially distanced? I know this is a stupid question, but Im curious what you all would do.	I'm looking for any advice about whether it's safe to see a friend who isn't vaccinated **when their children can't be vaccinated**, or whether I should follow the wishes of my brother and sister-in-law, who want me to avoid this person. .
My country is in a third wave. Its never ending. Im sure this post will get removed because Im too depressed/depressing for Reddit. Im broke because of the pandemic. Im struggling to work because my anxiety gives me physical symptoms. My friends dont talk to me anymore, or when I do talk to them they just tell me to see a therapist even though I already do (and can barely afford it). Im thinking about stopping eating so I can afford therapy. Theres no point to any of this. Every death that is about to happen here could have been prevented, and no one cares and they call me crazy. Im tired of the endless hurt.	I expect that this post will be removed from Reddit **because I'm too depressed to post** it because I expect that no one will want to read it and everyone will think that I'm crazy for thinking that I should stop eating so that I can afford to see a therapist. .
Model Summary includes information in the summary not discussed in the post	
israel imposed their mask mandate, despite being one of the most vaccinated countries. i feel like this will never end and I don't need stupid replies like "hang in there" or "it will be okay." and don't remove this post because it "causes anxiety." I'm not. I'm simply worried that we'll never get back to normal.	I'm afraid that we're never going to be able to get back to normal after COVID, because Israel has a mandate and Israel is one of the most vaccinated countries in the world, and **that's putting a strain on our health systems** and on our mental health.
I am fully vaccinated with the Moderna shot, and have been getting back to my regular life. I live in a state (MA) with extremely high vaccination rates, but nonetheless Im concerned about the Delta variant. Ive been hearing stories of breakthroughs of the variant in fully vaccinated people, so I am concerned. At the same time, Im extremely tired of this. I feel selfish for saying this, Ive done all I can do to protect myself and others against COVID, and I want to continue to live my life after over a year of taking precautions. Anyone else feel the same?	I am disgusted to see myself being selfish because I have done all I can to prevent COVID and I want to live my life as if I had not had any COVID side effects at all since I have taken every precaution possible to prevent the COVID variant.
Some sentences in the model summary are formatted to make the text difficult to read	
So the Vaccine team in Iceland is taking a summer holiday for a month that extends over the time when I was suppose to get my second Astra Zeneca shot. They offered me to get it sooner but I heard it will decrease it's effectiveness by allot. Should I get the shot 7 weeks after my first shot or should I wait until they come back and get it at least 15 weeks after after my first shot. Iceland has stopped all restrictions so i am a bit nervous.	I trust that the vaccine will help protect me from catching COVID and **I'll get it as soon as I can get it**. I trust that COVID **will do what it's supposed to do and do it it will do its best** to protect me and my family from COVID.
Is anyone else experiencing bad post-outing anxiety? Ive been trying to push myself out of my comfort zone (and sometimes I even get excited to) and so I get ahead of myself and leave my house. No bars or clubs, but I did attend an outdoor gathering thats weighing heavy on my mind. While Im out, Ive surprisingly found Im quite bored when I leave my house, but the real problem comes the day afterwards. I sit and think was that too soon? Can I re-enter my bubble now that people have seen me? Am I a hypocrite? And these questions flow through my brain in a never ending sequence. Feeling that Ive been perceived by others and I cant take it back feels unbearable, yet I took the decision to leave my house so I then encounter feelings of embarrassment, guilt and shame. Not to mention the obvious fear of the delta variant, and overall uncertainty over cdc recommendations. (I wish someone could spell out a good plan for reintegration besides - rip off the mask and live!) Everyone I know has returned to life normally and Im here in a weird limbo. I know a lot of people in this struggle with just leaving the house, but is anyone else out there struggling with how they feel once they do?	I find it hard to leave the house and it surprises me when **I find out that I'm not feeling the same way when I do**. I feel embarrassed and ashamed that I've been seen out in public and that I can't go back and change what I've done.
The overall model summary was not well-structured.	
I've recently heard of stories that people who are getting mildly sick after being vaccinated are still coming down with long covid. People on the covid long hauler subreddit are saying that we're going to have a huge problem on our hands because no vaccine protects against long covid....This isn't going to end, isn't it? Right as I think I've tasted freedom, I find out information that makes me want to stay inside forever like a hermit...	I feel sad and hopeless because I think I've tasted freedom and then I find out more information that makes me want to stay inside like a hermit. I wish I could just be free from this virus for a while but it doesn't look like that will be possible.
This makes me really just not want to go out and about again.... I've been on this sub for a while and posted a lot. More or less this pandemic has crushed my mental health and with having some health issues makes me really hesitant to do anything. I was finally getting my life back a little and this Delta variant makes me want to go back to old habits and just stay home and see no one... I really am at a loss of what to do and am feeling super overwhelmed.	I'm at a loss for what to do and don't know what I can do to get back on track with my health issues, so I just want to go back to my old ways and stay home and see no one. I was finally getting my life back before the pandemic hit.

Table 7: Example of common model errors identified by the expert evaluators.

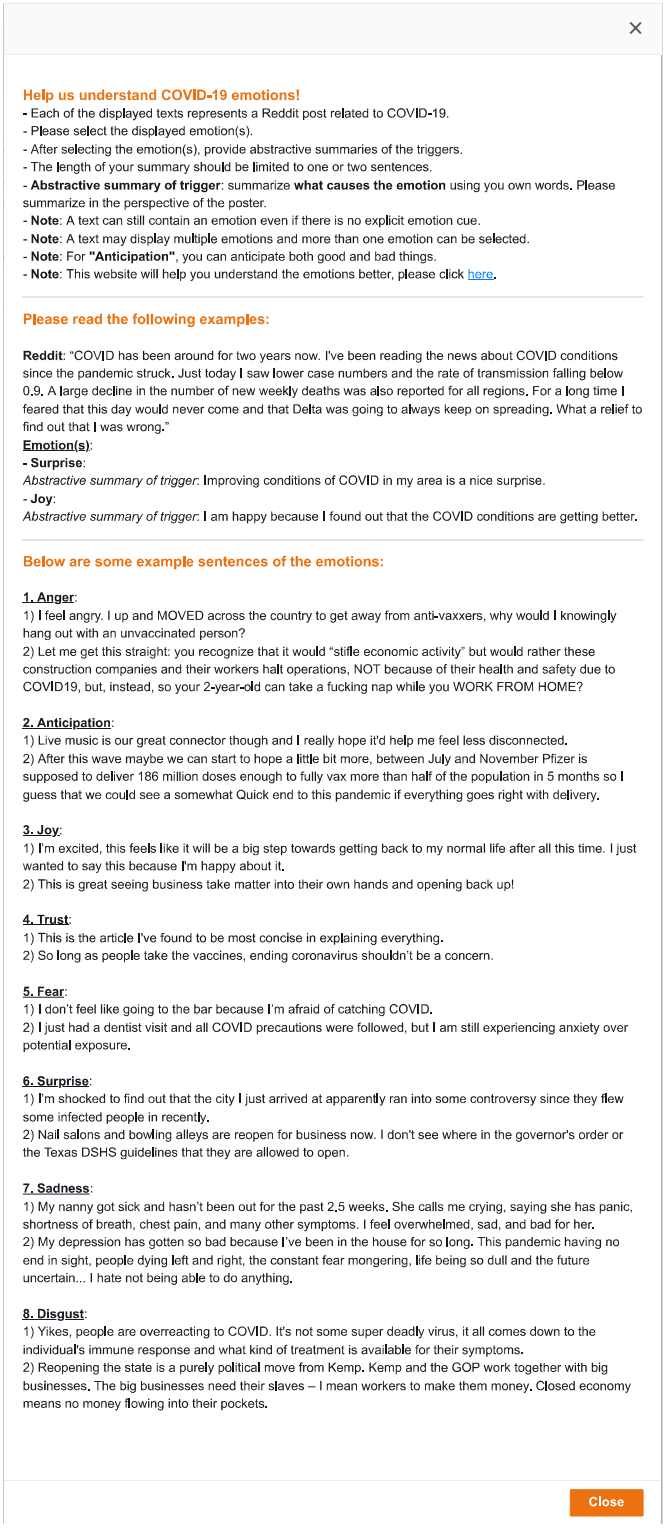

Figure 9: Annotation instructions (always shown before annotating).

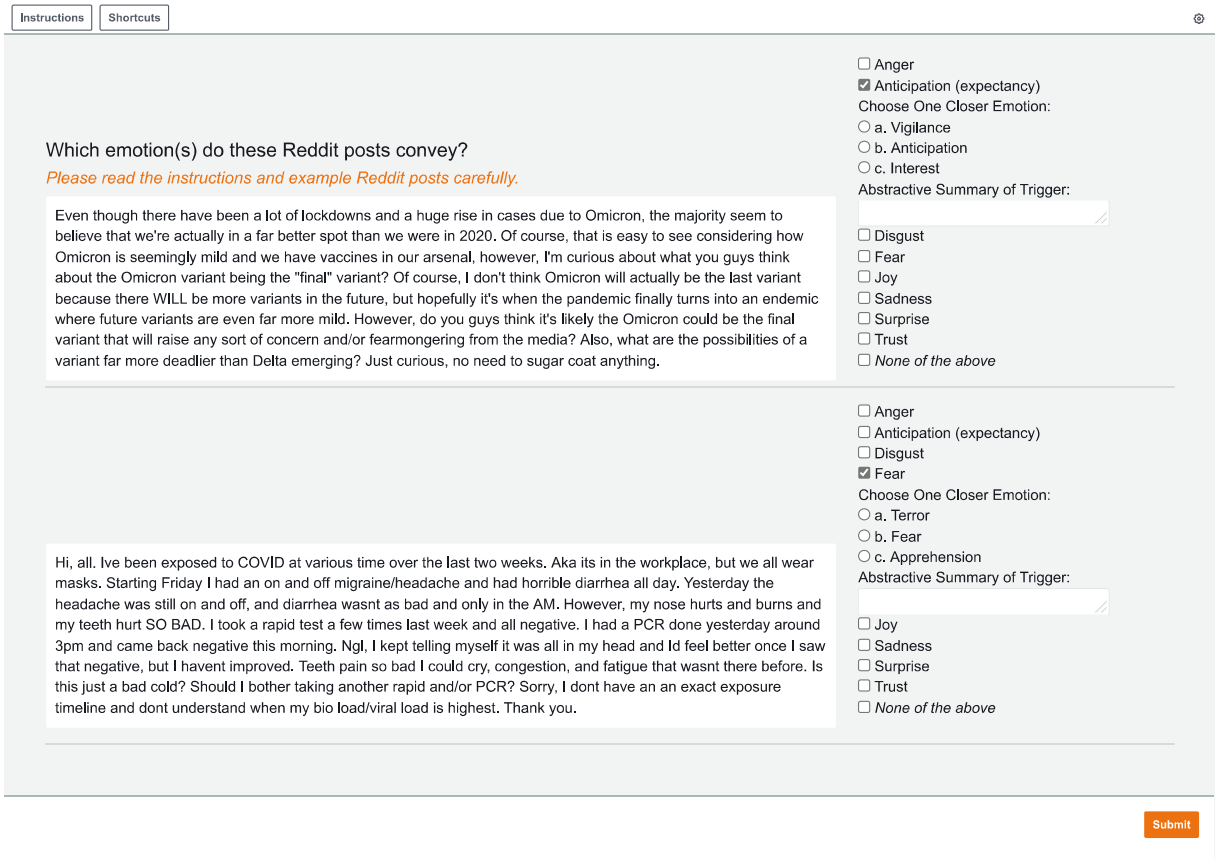

Figure 10: The annotation task layout of an example hit on the Amazon Mechanical Turk.

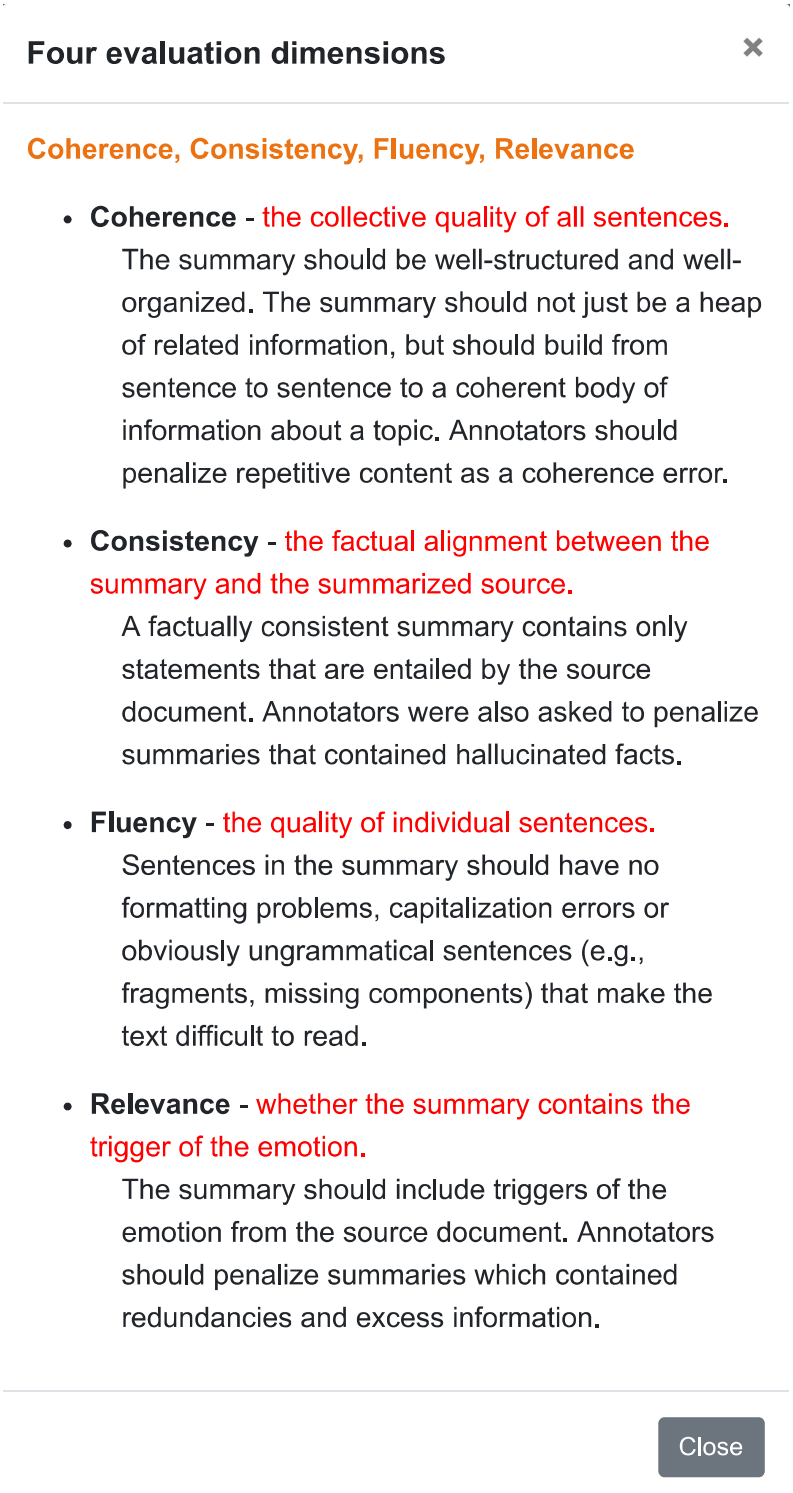

Figure 11: Human Evaluation Instructions.