Unsupervised Extractive Summarization of Emotion Triggers

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

Abstract

Understanding what leads to emotions during large-scale crises is important as it can provide groundings for expressed emotions and subsequently improve the understanding of ongoing disasters. Recent approaches (Zhan et al., 2022) trained supervised models to both detect emotions and explain emotion triggers (events and appraisals) via abstractive summarization. However, obtaining timely and qualitative abstractive summaries is expensive and extremely time-consuming, requiring highly-trained expert annotators. In time-sensitive, high-stake contexts, this can block necessary responses. We instead pursue *unsupervised* systems that extract triggers from text. First, we introduce COVIDET-EXT, augmenting (Zhan et al., 2022)'s abstractive dataset (in the context of the COVID-19 crisis) with extractive triggers. Second, we develop new unsupervised learning models that can jointly detect emotions and summarize their triggers. Our best approach, entitled Emotion-Aware Pagerank, incorporates emotion information from external sources combined with a language understanding module, and outperforms strong baselines. We release our data and code at https://github.com/tsosea2/CovidET-EXT.

1 Introduction

Language plays a central role in social, clinical, and cognitive psychology (Pennebaker et al., 2003), and social media presents a gold mine for such analysis: people turn to social media to share experiences around challenges in their personal lives and seek diagnosis, treatment, and emotional support for their conditions (Choudhury and De, 2014; Gjurković and Šnajder, 2018). During crises, such as natural disasters or global pandemics, large-scale analysis of language on social media — both *how people feel* and *what's going on in their lives to lead to these feelings* — can have a profound impact on improving mental health solutions as well as helping policymakers take better-informed decisions during a crisis.

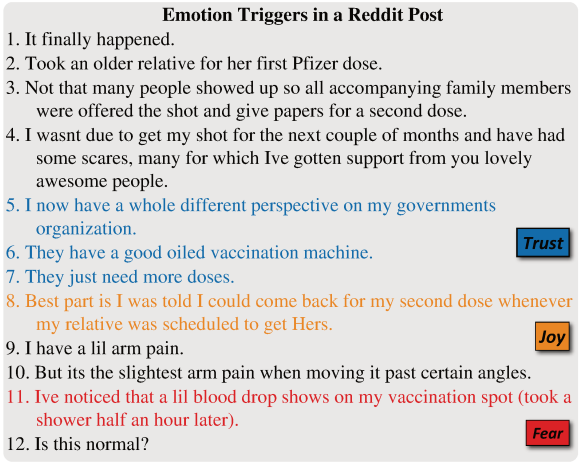

Figure 1: An example post from COVIDET-EXT annotated with emotion triggers. The highlighted sentences represent triggers of the tagged emotions.

Recent work (Zhan et al., 2022) taps into this broad challenge by jointly detecting emotions and generating a natural language description about what triggers them (triggers include both objective events and subjective appraisals of those events (Ellsworth and Scherer, 2003; Moors et al., 2013)). Trigger explanation is formulated as a supervised, abstractive summarization task that is emotion-specific. Unlike generic summarization however, due to the high cognitive load to provide judgments *for each emotion*, obtaining human-written summaries for this task is time-consuming and requires significant annotator training. This results in small, domain-specific datasets that are difficult to scale — especially in the face of new crisis events where the timing of such analysis is often pivotal.

This work instead takes a fully *unsupervised* approach such that we do not rely on any labeled data, thus becoming agnostic to distributional shifts in domain or types of crisis, and robust for time-

[*]Tiberiu Sosea and Hongli Zhan contributed equally.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9550–9569
July 9-14, 2023

critical events. We posit that emotion triggers can be summarized effectively in an extractive manner where unsupervised methods are well-suited; we thus tackle the challenge of *simultaneous* emotion prediction and trigger extraction.

For this new task, we first introduce COVIDET-EXT, augmenting Zhan et al. (2022)'s COVIDET with manually annotated extractive summaries corresponding to each of their abstractive summaries. The result is a dataset of $1,883$ Reddit posts about the COVID-19 pandemic, manually annotated with 7 fine-grained emotions (from COVIDET) and their corresponding **extractive** triggers (Figure 1). For every emotion present in a post, our annotators highlight sentences that summarize the emotion triggers, resulting in $6,741$ extractive summaries in total. Qualitative analyses of the dataset indicate good agreement among the annotators, and follow-up human validations of the annotations also reveal high correctness. COVIDET-EXT provides an ideal test bed to facilitate the development of extractive (supervised or unsupervised) techniques for the tasks of emotion detection and trigger summarization in crisis contexts.

We propose Emotion-Aware PageRank (EAP), a novel, fully unsupervised, graph-based approach for extractive emotion trigger summarization from text. The core of our method is to decompose the traditional PageRank (Page et al., 1999) ranking algorithm into multiple biased PageRanks (Haveliwala, 2003), one for each emotion. To bias our model towards various emotions, our approach harnesses lexical information from emotion lexicons (Mohammad and Turney, 2013; Mohammad, 2018). Critically, unlike previous graph-based unsupervised approaches (Mihalcea and Tarau, 2004; Liu et al., 2010; Gollapalli and Caragea, 2014; Florescu and Caragea, 2017; Patel and Caragea, 2021; Singh et al., 2019), which represent the text as a bag-of-words or word embeddings, EAP incorporates a language understanding module leveraging large language models to ensure that the summaries for an emotion are coherent in the context of that emotion. Results on our COVIDET-EXT indicate the effectiveness of our EAP, which significantly pushes the Rouge-L score of our summaries by an average of 2.7% over strong baselines.

Our contributions are as follows: **1)** We introduce COVIDET-EXT, a manually annotated benchmark dataset for the task of emotion detection and trigger summarization. **2)** We propose Emotion-Aware PageRank, a variation of PageRank that combines a language understanding module and external emotion knowledge to generate emotion-specific extractive summaries. **3)** We carry out a comprehensive set of experiments using numerous baselines to evaluate the performance on COVIDET-EXT and show that our proposed EAP significantly outperforms strong baselines.

2 Background and Related Work

Emotion Tasks. Most of the prior work on emotions on social media focuses solely on detecting emotions or emotional support from text (Wang et al., 2012; Biyani et al., 2014; Abdul-Mageed and Ungar, 2017; Khanpour et al., 2018; Khanpour and Caragea, 2018; Demszky et al., 2020; Desai et al., 2020; Sosea and Caragea, 2020; Adikari et al., 2021; Calbi et al., 2021; Kabir and Madria, 2021; Beck et al., 2021; Mohammed Abdulla et al., 2019; Sosea and Caragea, 2021; Hosseini and Caragea, 2021a,b; Saakyan et al., 2021; Ils et al., 2021; Sosea et al., 2022; Sosea and Caragea, 2022a,b). Our task is directly related to emotion cause extraction (Gao et al., 2015; Gui et al., 2016; Gao et al., 2017) which focused on identifying phrase-level causes from Chinese news or micro-blogs, which are distinct from the spontaneous writing on social media. In our context, similar to the work of Zhan et al. (2022), what *triggers* an emotion includes both what happened and how the writer appraised the situation. A major difference of our work from Zhan et al. (2022) is that we consider extractive summaries instead of abstractive and take a fully unsupervised perspective, eliminating the reliance on labeled data. For a comprehensive overview of COVIDET introduced by Zhan et al. (2022), refer to Appendix §A.

Unsupervised Extractive Summarization. Extractive summarization aims to condense a piece of text by identifying and extracting a small number of important sentences (Allahyari et al., 2017; Liu and Lapata, 2019; El-Kassas et al., 2021) that preserve the text's original meaning. The most popular approaches in unsupervised extractive summarization leverage graph-based approaches to compute a sentence's salience for inclusion in a summary (Mihalcea and Tarau, 2004; Zheng and Lapata, 2019). These methods represent sentences in a document as nodes in an undirected graph whose edges are weighted using sentence similarity. The sentences in the graph are scored and ranked using node cen-

trality, computed recursively using PageRank (Page et al., 1999). In contrast, our EAP considers words instead of sentences as nodes in the graph and employs multiple separate biased PageRanks (Haveliwala, 2003) to compute an emotion-specific score for each word, which is combined with a sentence-similarity module to produce one sentence score per emotion, indicating the salience of the sentences under each emotion.

3 Dataset Construction

Since there is no annotated data for extractive emotion triggers summarization in crisis contexts, we first bridge this gap by extending COVIDET, Zhan et al. (2022)'s abstractive-only dataset with extractive trigger summaries. Doing so **(a)** creates benchmark data for extractive systems; **(b)** allows in-depth analyses to understand how and when emotion triggers are expressed on social media. This will also create a parallel abstractive-extractive dataset for future research. We name our new dataset COVIDET-EXT (COVIDET {extractive, extension}).

Annotating Emotion Triggers. Given a post from COVIDET annotated with an emotion e, we ask annotators to highlight sentences in the post that best describe the trigger for e. An overview of our annotation scheme can be viewed in Appendix §B. We recruit both undergraduate students (in a Linguistics department) as well as pre-qualified crowd workers (from the Amazon Mechanical Turk) for this task.[1] Each post is annotated by two annotators. We monitor the annotation quality and work with the annotators during the full process. Similar to COVIDET, the test set is annotated by undergraduate students.

Benchmark Dataset. We follow the benchmark setup in Zhan et al. (2022) with $1,200$ examples for training, 285 examples for validation, and 398 examples for testing. If two annotators highlight different sentences as triggers for the same emotion, we consider both sets of sentences as the gold summaries and evaluate them using multi-reference ROUGE. We anonymize COVIDET-EXT. Note that since we explore *unsupervised* methods, the training set is *not* used in our summarization models. Nevertheless, we emphasize that while the focus of this work is the unsupervised setup, we hope that COVIDET-EXT can spur further research into both supervised and unsupervised methods, hence we maintain the splits in Zhan et al. (2022). For completeness, we carry out experiments in a fully supervised setup in Appendix §F.

[1]These crowd workers have an ongoing working relationship with our group and have prior experience in related complex tasks, and we make sure they are paid at least $10/hr.

	ANC	AGR	FER	SDN	JOY	TRS	DSG	Avg
Emotion	0.64	0.84	0.84	0.84	0.92	0.60	0.80	0.79
Trigger	0.56	0.64	0.76	0.76	0.80	0.56	0.72	0.69

Table 1: Human validation results on COVIDET-EXT.

Overlapping Status	55.5% of all summaries
Fleiss' Kappa	0.89 across 7 emotions
self-BLEU-2	0.429 (baseline: 0.151)
self-BLEU-3	0.419 (baseline: 0.139)
self-ROUGE-L	0.504 (baseline: 0.229)

Table 2: Inter-annotator statistics of COVIDET-EXT.

Human Validation. We validate the annotated extractive summaries of emotion triggers in COVIDET-EXT through inspections from third-party validators on the Amazon Mechanical Turk crowdsourcing platform. A subset of our training data including 300 randomly selected examples which contain annotations of extractive summaries of emotion triggers are validated. Given an annotated extractive trigger summary, we first ask the validators whether the summary leans towards the annotated emotion. It yes, we ask the validator to further point out if the *trigger* — rather than the *emotion* itself — is present in the summary. The percentage of examples that validators confirm for the two steps is shown in Table 1. Overall, the human validation results showcase moderately high correctness in the annotations of COVIDET-EXT, considering the subjective nature of our task.[2]

Inter-Annotator Agreement. We measure the inter-annotator agreement between two extractive trigger summaries for the same emotion in a post, as shown in Table 2. Results show that, within the examples where we find emotion overlaps, 29.9% of the extractive summaries of triggers for the same emotion share completely identical annotations from both annotators, and 25.6% have partial sentence-level overlaps. In total, we find overlaps

[2]The same sentence can be interpreted to be triggers for different emotions. For example, the sentence "I miss my room and I dont have many clothes or my meds here, but hes hitting these mics every fucking night and Im scared of contracting it" expresses *anger*, *sadness*, and *fear* simultaneously under the same context.

in 55.5% of the summaries, and the experts who were responsible for the test set (65.8%) have more overlapping summaries than the crowd workers who were responsible for the training and validation sets (52.3%). Furthermore, the average Fleiss' kappa (Fleiss, 1971; Randolph, 2005) is 0.89 across all the emotions in COVIDET-EXT. This suggests substantial agreement among our annotators.

In addition, we also employ automatic metrics including self-BLEU (with smoothing methods 1) and self-ROUGE to capture the overlap between annotators' summaries. To establish a baseline, we report these metrics between the annotators' work and a randomly selected sentence from the original post. We repeat this process five times. Results reveal that both the self-BLEU and self-ROUGE of our annotations significantly outperform that of the random baseline (as shown in Table 2). We also observed higher values of these measures for student annotators compared with crowd workers. (c.f. Appendix §D). These results indicate strong accordance among our annotators.

Dataset Statistics. Here we elaborate on the overview of COVIDET-EXT. On average, there are 1.35 sentences (std.dev = 0.79) consisting of 32.54 tokens (std.dev = 20.68) per extractive summary of emotion trigger in COVIDET-EXT. As shown in Figure 2, when broken down into unique trigger *sentences*, *fear* has the most trigger sentences in the dataset, closely followed by *anticipation*. On the other hand, *trust* has the lowest number of trigger sentences. This can be attributed to the calamitous nature of the domain of our dataset. Besides, unlike generic news summarization (Fabbri et al., 2021), the emotion-trigger extractive summarization task is *not* lead-based. This is manifested through our scrutiny of the position of emotion trigger sentences in the original posts (Figure 6 and Figure 7, Appendix §E), where a large number of triggers cluster in the later parts of the post.

Additional analyses of COVIDET-EXT can be found in Appendix §E.

Emotion Explicitness. To examine the explicitness of emotions in the extractive summaries of emotion triggers, we apply EmoLex (Mohammad and Turney, 2013), an English lexicon for the Plutchik-8 primary emotions. Specifically, for the extractive summaries of triggers to a certain emotion e, we measure the average ratio of e's words in EmoLex being present in the sentence-level lemmatized summaries. The results are presented in Figure 2. Interestingly, we notice that *sadness* is the most explicit emotion in the annotated extractive summaries of triggers in our dataset, while *anger* is the most implicit one.

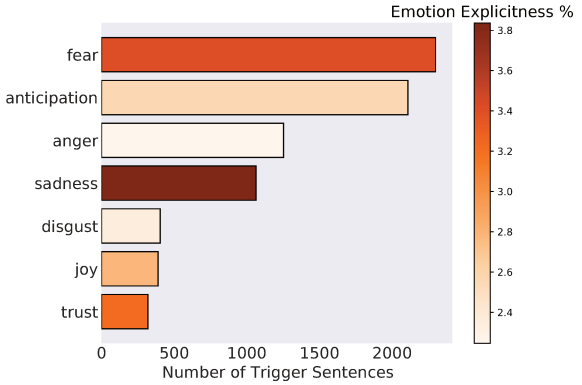

Figure 2: The sentence-level distribution of triggers in the original posts in COVIDET-EXT. The colorbar shows the explicitness of the emotion in the triggers.

4 Unsupervised Extractive Summarization

In this section we introduce Emotion-Aware Pagerank (EAP), our fully unsupervised, graph-based, emotion trigger extractive summarization method that incorporates information from emotion lexicons to calculate a biased PageRank score of each sentence in a post. EAP then fuses this score with an additional similarity-based sentence-level score that ensures the summary for a specific emotion e does not diverge in meaning from other summaries of the same emotion e. We show an overview of our model architecture in Figure 3.

Task Formulation. Let P be a Reddit post. P is composed of an ordered sequence of n sentences: $P = \{s_1, s_2, ..., s_n\}$. Generic extractive summarization aims to output an ordered set of sentences S with $S \subset P$ that captures the essence of post P. In our emotion trigger summarization, however, we aim to generate multiple extractive summaries conditioned on the expressed emotions. To this end, we are interested in a set of summaries $S^{emo} = \{S_{e_1}, S_{e_2}, ..., S_{e_m}\}$ where m is the total number of emotions present in P and S_{e_i} is the summary of the triggers that lead to the expression of emotion e_i with $S_{e_i} \subset P$. Note that P usually conveys a subset of emotions, in which case the summaries for the emotions that are not present in text are empty.

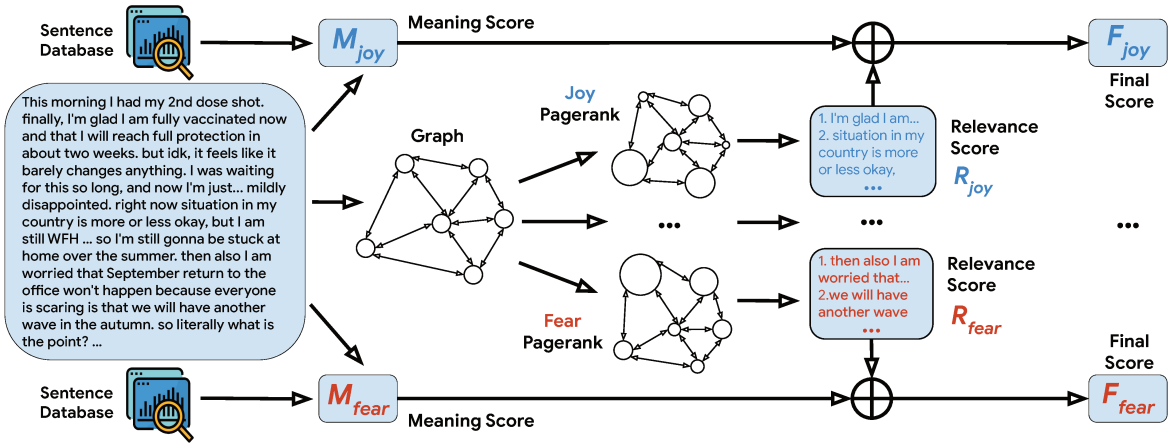

Figure 3: Diagram of our Emotion-Aware PageRank. EAP builds a word graph from a post, then runs separate biased PageRanks, one for each emotion, to score every candidate sentence under each emotion. The score is combined with an emotion-aware language understanding module to produce final rankings for each sentence under each emotion.

Graph Construction. We build an undirected graph $G = (V, E)$, where V is vocabulary set of words. To build V we employ various processing and filtering techniques. First, we only select nouns, adjectives, verbs, adverbs and pronouns and remove any punctuation. Next, we stem all the selected words to collapse them in a common base form. Finally, we remove infrequent words which appear less than 20 times in the entire training set. The remaining words form the vocabulary V. A pair of words $(w_i, w_j) \in E$ defines an edge between w_i and w_j and the operator $\beta(w_i, w_j)$ denotes the weight of edge (w_i, w_j). We compute the weight of an edge in our graph using word co-occurences in windows of text. Given a window size of ws, we say that two words w_i and w_j co-occur together if the number of words between them in text is less than ws. We build a co-occurence matrix C of size $|V| \times |V|$ from the documents in our training set where C_{ij} is the number of times words w_i and w_j co-occur together. Using C we simply define the weight of an edge as:

$$\beta(w_i, w_j) = \frac{2 \times C_{ij}}{\sum_{k=0}^{|V|}(C_{ik} + C_{jk})} \quad (1)$$

Intuitively, the more frequently two words co-occur together, the higher the weight of the edge between them becomes.

Emotion Decomposition. In PageRank, the importance or relevance $\mathcal{R}(w_i)$ of an arbitrary word w_i is computed in an iterative fashion using the following formula:

$$\mathcal{R}(w_i) = \lambda \sum_{k=1}^{|V|} \beta(w_k, w_i)\mathcal{R}(w_k) + (1 - \lambda)\frac{1}{|V|} \quad (2)$$

where $|.|$ is the set size operator and λ is the damping factor, a fixed value from 0 to 1 which measures the probability of performing a random jump to any other vertex in the graph. The idea of PageRank is that a vertex or word is important if other important vertices point to it. The constant term $\frac{1}{|V|}$ is called a random jump probability and can be viewed as a node *preference* value, which in this case assigns equal weights to all the words in the graph, indicating no preference.

In this current formulation, the PageRank model calculates the weights of words irrespective of the expressed emotion. We claim that for our purpose words should bear different importance scores in different emotion contexts. For example, the word *agony* should have a higher importance in the context of *sadness* or *fear* than in the context of *joy*.

To this end, we propose to decompose the text into multiple components, one for each emotion, where the relevance of a word differs from component to component. Biased PageRank (Haveliwala, 2003) is a variation of PageRank where the second term in Equation 2 is set to be non-uniform, which can influence the algorithm to prefer particular words over others. We propose to run a separate biased PageRank for each emotion and leverage a custom importance function $i_e(w_i)$ that yields high values for words that are correlated with an emotion e and low values otherwise. For-

mally, the relevance computation for the PageRank corresponding to emotion e becomes:

$$\mathcal{R}_e(w_i) = \lambda \sum_{k=1}^{|V|} \beta(w_k, w_i)\mathcal{R}_e(w_k) + (1-\lambda)\frac{i_e(w_i)}{N} \quad (3)$$

where N is a normalization factor such that $\sum_{w \in V} \frac{i_e(w)}{N} = 1$. Since the model prefers those vertices with higher random jump probabilies, using an accurate importance function $i_e(w_i)$ for emotion e can lead to accurate relevance scores in the context of e. We define this function using the NRC emotion intensity (Mohammad, 2018) lexicon. EmoIntensity associates words with their expressed emotions and also indicates the degree of correlation between a word and a particular emotion using real values from 0 to 1. For example, *outraged* has an intensity for anger of 0.964 while *irritation* has an intensity of 0.438. In our context, assigning importance values using intensity is appropriate since a sentence containing high intensity words for an emotion e is more likely to be relevant in the context of e compared to a sentence containing lower intensity words. Denoting the set of words in EmoIntensity correlated with emotion e by $\mathcal{I}_e$, all words $w \in \mathcal{I}_e$ also come with intensity value annotations denoted by $int_e(w)$. Therefore, we define the importance function as:

$$i_e(w) = \begin{cases} int_e(w) & if \quad w \in \mathcal{I}_e \\ c & if \quad w \in V \setminus \mathcal{I}_e \end{cases} \quad (4)$$

where c is a constant that we find using the validation set. Since our summaries are at the sentence level, we simply score a sentence s_i as the average relevance of its words:

$$R_e(s_i) = \frac{\sum_{w_j \in s_i} R_e(w_j)}{|s_i|} \quad (5)$$

Encoding the meaning. A major drawback of prior graph-based approaches is that they exclusively represent the input as a bag-of-words, ignoring the structure of text. We propose to solve this drawback by introducing a language model-based component to encode the meaning of a sentence. Our component is based on the assumption that a sentence s that is highly relevant for an emotion e should be similar in meaning to other sentences s_i relevant to e. We capture this property by scoring each sentence based on its similarity with other important (i.e., in the context of e) sentences. We leverage the popular Sentence-BERT (Reimers and Gurevych, 2019) model, which produces meaningful sentence embeddings that can be used in operations such as cosine similarity. Given a sentence s_i, let $\mathbf{s_i}$ be its embedding and $sim(\mathbf{s_i}, \mathbf{s_j})$ be the cosine similarity between the embeddings of sentences s_i and s_j. Denoting by $\mathcal{T}$ the set of sentences in the entire dataset, we score s_i in the context of emotion e as follows:

$$M_e(s_i) = \frac{\sum_{s \in \mathcal{T}} sim(\mathbf{s_i}, \mathbf{s}) * \mathcal{R}_e(s)}{|\mathcal{T}|} \quad (6)$$

Intuitively, $M_e(s_i)$ yields high values if s_i is similar in meaning to sentences relevant in the context of emotion e.

Constructing the Summaries. Given a post $P = \{s_1, s_2, ..., s_n\}$, we first combine the meaning and the relevance scores into a final, sentence level, per-emotion score, which we use to score every sentence s_i in P along all the emotions:

$$\mathcal{F}_e(s_i) = \mathcal{R}_e(s_i) * M_e(s_i) \quad (7)$$

We use this per-emotion score to rank the sentences in the post P. For an emotion e, we only select the sentences s_i where $\mathcal{F}_e(s_i) > t$ to be part of the final summary for e. t is a threshold value that we infer using our validation set. Note that given P, we compute the score $\mathcal{F}_e$ for every emotion e. In the case that none of the sentences in P exceed the threshold for a particular emotion, we consider that the emotion is not present in the post (i.e., we do not generate a summary).

5 Experiments and Results

In this section, we first introduce our emotion-agnostic and emotion-specific baselines. Next, we present our experimental setup and discuss the results obtained by EAP against the baselines.

Emotion-agnostic baselines. We explore two standard heuristic baselines, namely **1)** Extracting the first sentence in the post (1 sent) and **2)** Extracting the first three sentences in the post (3 sent). Next, we design three graph centrality measure-based methods: **3)** PacSum (Zheng and Lapata, 2019), **4)** PreSum (Liu and Lapata, 2019) and word-level **5)** TextRank (Mihalcea and Tarau, 2004). Note that these methods are emotion-oblivious and the generated summary will be identical for different emotions.

	ANGER		DISGUST		FEAR		JOY		SADNESS		TRUST		ANTICIPATION		AVG	
	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L
1-SENT	0.174	0.240	0.095	0.170	0.202	0.256	0.119	0.179	0.110	0.177	0.189	0.236	0.160	0.220	0.149	0.211
3-SENT	0.301	0.315	0.196	0.253	0.322	0.343	0.273	0.310	0.239	0.292	0.248	0.279	0.263	0.307	0.258	0.288
PACSUM	0.308	0.314	0.210	0.218	0.327	0.331	0.276	0.282	0.287	0.304	0.225	0.234	0.283	0.295	0.273	0.282
PRESUMM	0.306	0.312	0.219	0.221	0.332	0.335	0.268	0.274	0.295	0.317	0.222	0.227	0.284	0.291	0.275	0.282
TEXTRANK	0.296	0.301	0.236	0.235	0.319	0.326	0.272	0.276	0.286	0.306	0.225	0.231	0.218	0.221	0.264	0.270
EMOLEX	0.213	0.260	0.218	0.256	0.309	0.341	0.218	0.252	0.301	0.331	0.176	0.203	0.207	0.242	0.234	0.269
EMOINTENSITY	0.307	0.322	0.269	0.281	0.342	0.355	0.222	0.235	0.329	0.341	0.227	0.242	0.295	0.310	0.284	0.298
BERT-GOEMO	0.247	0.264	0.232	0.237	0.296	0.312	0.221	0.247	0.314	0.321	0.201	0.204	0.247	0.225	0.253	0.258
EAP	**0.324**†	**0.348**†	**0.285**†	**0.296**†	**0.364**†	**0.373**†	**0.285**†	**0.319**†	**0.348**†	**0.354**†	**0.258**†	**0.291**†	**0.319**†	**0.324**†	**0.309**†	**0.325**†

Table 3: Results of our models in terms of ROUGE-2 and ROUGE-L. We assert significance† using a bootstrap test where we resample our dataset 50 times with replacement (with a sample size of 500) and $p < 0.05$.

	ANG	DSG	FER	JOY	SDN	TRT	ANC	AVG
EMOLEX	0.561	0.572	0.568	0.613	0.563	0.581	0.593	0.578
EMOINTENSITY	0.581	0.583	0.557	0.632	0.573	0.589	0.585	0.584
GOEMOTIONS	0.516	0.532	0.562	0.576	0.531	0.556	0.574	0.537
EAP	**0.593**†	**0.595**†	**0.583**	**0.649**†	**0.581**†	**0.606**†	**0.612**†	**0.593**†

Table 4: Emotion detection results of our models in terms of Macro F-1. We assert significance† using a bootstrap test where we resample our dataset 50 times with replacement (with a sample size of 500) and $p < 0.05$.

Emotion-specific baselines. We first employ two lexical-based methods: **6)** EmoLex - we use the EmoLex (Mohammad and Turney, 2013) lexicon to identify lexical cues that indicate the expression of emotions. If a sentence contains a word that is associated with an emotion e, we consider the sentence to express e. The final summary for e contains all sentences expressing e. **7)** EmoIntensity - we leverage the NRC Affect Intensity Lexicon (Mohammad, 2018) to build a more fine-grained approach of identifying if a sentence expresses an emotion or not. For each sentence and emotion, we calculate the average emotion word intensity and compare it to a pre-defined threshold t. If the average intensity for e is higher than t we label the sentence with e. t is a tunable parameter that we select based on our validation set performance.

Finally, we leverage models trained on emotion detection datasets to build our emotion-specific summaries. For a post P, we use our model to make predictions on each sentence in P and build summaries by concatenating sentences that express the same emotions. We mainly experiment with a model trained on the **8)** GoEmotions (Demszky et al., 2020) dataset.

Experimental Setup. We carry out our experiments on an Nvidia A5000 GPU. We use the HuggingFace Transformers (Wolf et al., 2019) library for our Sentence-BERT implementation and we will make the code for our methods and data available for reasearch purposes. We report the performance in terms of Rouge-2 and Rouge-L (Lin, 2004) to evaluate the summarization performance. Additionally, we also calculate the performance in terms of F1 and show the results in Appendix I. We provide extensive details about the hyperparameters used in EAP and the baselines, such as our various thresholds and constants in Appendix §G.

Results. We show the results obtained in Table 3. First, we note that emotion-specific approaches outperform the emotion-oblivious methods considerably. Notably, EmoIntensity outperforms PacSum by an average of 1.1% in Rouge-2. Among the emotion-specific baselines, EmoIntensity, which uses the intensity of emotion words to extract relevant sentences for a particular emotion obtains good performance, outperforming the EmoLex method by 5.1% Rouge-2 on disgust and 3.3% on fear. This result emphasizes that having a degree of association between a word and an emotion (i.e., the intensity) is a stronger signal than the plain word-emotion association in our emotion-based extractive summarization context.

EAP consistently yields the highest results both in terms of Rouge-2 and Rouge-L compared to the other approaches. Concretely, we obtain an average improvement of 2.7% in Rouge-L and 2.5% in Rouge-2 score over our strongest EmoIntensity baseline. For example, on anger and joy we see improvements in Rouge-2 of 1.7% and 6.3% respectively. Moreover, our emotion-aware PageRank considerably outperforms TextRank (Mihalcea and Tarau, 2004) by as much as 5.5% Rouge-L and 4.5% Rouge-2 on average.

Emotion Detection. While EAP shows strong results in our emotion trigger summarization experiments, we want to evaluate our approach in a traditional emotion detection task. To this end,

	ANGER		DISGUST		FEAR		JOY		SADNESS		TRUST		ANTICIPATION		AVG	
	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L
EAP	**0.324**	**0.348**	**0.285**	**0.296**	**0.364**	**0.373**	**0.285**	**0.268**	**0.348**	**0.354**	**0.239**	**0.264**	**0.319**	**0.324**	**0.309**	**0.318**
-int	0.317	0.336	0.274	0.282	0.353	0.362	0.276	0.261	0.339	0.347	0.231	0.252	0.312	0.317	0.300	0.308
-sim	0.314	0.332	0.277	0.284	0.351	0.360	0.272	0.260	0.340	0.342	0.232	0.254	0.311	0.31	0.299	0.306
-int -sim	0.300	0.316	0.263	0.275	0.341	0.353	0.261	0.253	0.325	0.339	0.224	0.247	0.308	0.309	0.28	0.298

Table 5: Ablation study of our EAP.

Figure 4: Word-level Emotion-Aware PageRank scores and sentence-level meaning scores for the joy (Upper Box) and fear (Lower Box) emotions. The term relevance score is superscripted to each word (i.e., w^{score}), while the meaning score of sentences is superscripted at the end of the sentence (i.e., $[.]^{score}$). Gold summaries are highlighted.

we ask how well EAP can detect emotions at the post level. Given a post P, we label the post with emotion e if we identify any sentence $s \in P$ as a summary for e. If no sentence is selected to be included in the summary, we consider that EAP does not predict e.

We show the results obtained in Table 4, where we compare EAP to lexical methods (EmoLex and EmoIntensity) and a domain adaptation method, which trains a BERT (Devlin et al., 2019) model on the GoEmotions dataset (Demszky et al., 2020). We observe that EAP consistently outperforms prior work on all the emotions by an average of 0.9% in F1 score. Notably, we see 1.5% improvements in F1 on fear and 1.9% on anticipation.

Ablation Study. We perform a thorough ablation study to tease apart and analyze the components lead to the success of EAP. First, we analyze the influence of emotion intensity on the performance of the model. Here, we slightly modify the importance function from Equation 4 to a constant value. Instead of using the variable $int_e(w)$ we use a constant value c^e where $c^e > c$. Intuitively, we still bias the model towards a particular emotion e, however, every word associated with e weighs equal in this ablated version of EAP. We denote this modification of the algorithm by *-int*. Second, we remove the $meaning$ score M_e from our algorithm and use only the word-based relevance $\mathcal{R}_e$. This approach is denoted by *-sim*. We also analyze the behaviour of EAP when removing both components.

We show the results obtained in Table 5. Removing emotion intensity leads to a performance degradation of 1% in Rouge-L while the lack of our similarity module decreases the performance by 1.2% in Rouge-L. Removing both further decreases the performance by 2.9% in Rouge-2. These results emphasize that both similarity and intensity are core components of EAP and both consistently contribute to its success.

Anecdotal Evidence. To offer additional insights into our EAP, we provide anecdotal evidence in Figure 4, where we show a post expressing both joy and fear. We indicate for each word both its relevance for joy and for fear. Additionally, we show the meaning score for each sentence and emotion. Interestingly, we observe that the scores produced by our model are very relevant. For instance, *protection* has a very large value for joy of 0.531 and a very small value of 0.076 for *fear*. Along the same lines, *worried* has a relevance of 0.523 for *fear* and 0.074 for joy. The similarity scores are also accu-

rate. For example, *glad I am fully vaccinated* has a score for joy of 0.463, 9 times as large of the score of the same sentence for fear. We show additional analysis on the effect of the most relevant terms on EAP performance in Appendix §H.

6 Conclusion

We introduce COVIDET-EXT, a new benchmark dataset composed of $1,883$ Reddit posts annotated for the task emotion detection and extractive trigger summarization in the context of the COVID-19 pandemic. Our proposed Emotion-Aware Pagerank approach yields strong results on our datasets, consistently outperforming prior work in an unsupervised learning context. In the future, we plan to study abstractive trigger summarization from an unsupervised point of view to bridge the gap between the extractive and abstractive summarization performance.

Limitations

Since our EAP builds its graph representation from social media data, our method may carry inductive biases rooted in this type of data. Moreover, note that the scope of our study is limited to English social media posts and our approach does not consider inputs larger than 512 tokens. Therefore using our approach in long document summarization may be challenging. Finally, the general applicability of EAP in a different domain is highly dependent on the existence of high-quality lexicons for the domain in question, which may not be available.

Acknowledgements

This research was partially supported by National Science Foundation (NSF) grants IIS-1912887, IIS-2107487, ITE-2137846, IIS-2145479, IIS-2107524, IIS-2107487. We thank Jamie Pennebaker for useful discussions and comments. We also thank our reviewers for their insightful feedback and comments.

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.

Achini Adikari, Rashmika Nawaratne, Daswin De Silva, Sajani Ranasinghe, Oshadi Alahakoon, Damminda Alahakoon, et al. 2021. Emotions of covid-19: Content analysis of self-reported information using artificial intelligence. *Journal of Medical Internet Research*, 23(4):e27341.

Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys J. Kochut. 2017. Text summarization techniques: A brief survey. *CoRR*, abs/1707.02268.

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in German covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–13.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 827–836, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Marta Calbi, Nunzio Langiulli, Francesca Ferroni, Martina Montalti, Anna Kolesnikov, Vittorio Gallese, and Maria Alessandra Umiltà. 2021. The consequences of covid-19 on social interactions: an online study on face covering. *Scientific Reports*, 11(1):1–10.

R Sherlock Campbell and James W Pennebaker. 2003. The secret life of pronouns: Flexibility in writing style and physical health. *Psychological science*, 14(1):60–65.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological science*, 15(10):687–693.

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.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

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.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Corina Florescu and Cornelia Caragea. 2017. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada. 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.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Sujatha Das Gollapalli and Cornelia Caragea. 2014. Extracting keyphrases from research papers using citation networks. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1629–1635. AAAI Press.

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.

T.H. Haveliwala. 2003. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021b. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13018–13026.

Alexandra Ils, Dan Liu, Daniela Grunow, and Steffen Eger. 2021. Changes in European solidarity before and during COVID-19: Evidence from a large crowd- and expert-annotated Twitter dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1623–1637.

Md. Yasin Kabir and Sanjay Madria. 2021. Emocov: Machine learning for emotion detection, analysis and visualization using covid-19 tweets. *Online Social Networks and Media*, 23:100135.

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.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2018. Identifying emotional support in online health communities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

SDGV Akanksha Kumari and Shreya Singh. 2017. Parallelization of alphabeta pruning algorithm for enhancing the two player games. *Int. J. Advances Electronics Comput. Sci*, 4:74–81.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376, Cambridge, MA. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

G Mohammed Abdulla, Shreya Singh, and Sumit Borar. 2019. Shop your right size: A system for recommending sizes for fashion products. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 327–334, New York, NY, USA. Association for Computing Machinery.

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.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. In *The Web Conference*.

Krutarth Patel and Cornelia Caragea. 2021. Exploiting position and contextual word embeddings for keyphrase extraction from scientific papers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1585–1591, Online. Association for Computational Linguistics.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David Ian Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577. PMID: 12185209.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129.

Abraham Gerard Sebastian, Shreya Singh, P. B. T. Manikanta, T. S. Ashwin, and G. Ram Mohana Reddy. 2019. Multimodal group activity state detection for classroom response system using convolutional neural networks. In *Recent Findings in Intelligent Computing Techniques*, pages 245–251, Singapore. Springer Singapore.

Sarah Seraj, Kate G. Blackburn, and James W. Pennebaker. 2021. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences*, 118(7):e2017154118.

Rachel A Simmons, Dianne L Chambless, and Peter C Gordon. 2008. How do hostile and emotionally overinvolved relatives view relationships?: What relatives' pronoun use tells us. *Family Process*, 47(3):405–419.

Loveperteek Singh, Shreya Singh, Sagar Arora, and Sumit Borar. 2019. One embedding to do them all.

Shreya Singh, G Mohammed Abdulla, Sumit Borar, and Sagar Arora. 2018. Footwear size recommendation system.

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 and Cornelia Caragea. 2021. eMLM: A new pre-training objective for emotion related tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 286–293, Online. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2022a. EnsyNet: A dataset for encouragement and sympathy detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5444–5449, Marseille, France. European Language Resources Association.

Tiberiu Sosea and Cornelia Caragea. 2022b. Leveraging training dynamics and self-training for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4750–4762, Abu Dhabi, United Arab Emirates. 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.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

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, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

A CovidET[3]

Zhan et al. (2022) was the first to introduce the combined labeling of both emotions and (abstractive) summaries of their triggers on the domain of spontaneous speech (i.e., Reddit posts). They presented CovidET, a corpus of 1,883 Reddit posts manually annotated with 7 emotions (namely *anger*, *anticipation*, *joy*, *trust*, *fear*, *sadness*, and *disgust*) as well as abstractive summaries of the emotion triggers described in the post. The posts are curated from `r/COVID19_support`[4], a sub-Reddit for people seeking community support during COVID-19. To ensure the diversity of the data distribution, CovidET consists of Reddit posts from two different timelines (before and during the Omicron variant). The posts in CovidET are lengthy and emotionally rich, with an average of 156.4 tokens and 2.46 emotions per post. CovidET serves as an ideal dataset to spur further research on capturing triggers of emotions in long social media posts.

Nevertheless, the combined labeling of emotions and free-form abstractive summarization of their triggers is difficult and time-consuming as it requires annotators to comprehend the document in depth. This fails to meet the time-sensitivity requirement in the face of major crises like COVID-19. Our work instead proposes to generate an extractive summarization of emotion triggers and studies the task of emotion detection and trigger summarization from an unsupervised learning perspective, which is robust to domain variations and beneficial in boosting understanding in time-critical periods.

B Annotation Scheme of CovidET-EXT

The process of collecting annotations for CovidET-EXT is shown in Figure 5. Given a post and its annotations containing emotion e from CovidET, we ask annotators to highlight sentences in the post that best describe the trigger for emotion e. Rather than selecting text that expresses the emotion itself, we specifically instruct annotators to extract the events and how people make sense of the events that lead to the expression of the emotion. We use detailed examples provided by Zhan et al. (2022) to help our annotators better interpret the definition of emotion triggers.

[3] https://github.com/honglizhan/CovidET

[4] https://www.reddit.com/r/COVID19_support/

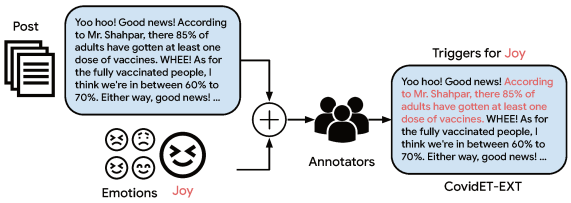

Figure 5: The process of collecting annotations for CovidET-EXT. The provided posts and annotated emotions are gathered from CovidET (Zhan et al., 2022).

C Crowd Workers

Both groups of annotators for CovidET-EXT come from the United States. The crowd workers are recruited from the Amazon Mechanical Turk crowdsourcing platform, with restrictions that their locale is the US and that they have completed 500+ HITs with an acceptance rate of at least 95%. The undergraduate students are hired from a university in the United States.

D Inter-annotator Agreement Among Undergraduate Students and Crowd Workers

As shown in Table 6, the inter-annotator performance of the undergraduate students consistently exceeds the crowd workers.

	Students	Crowd Workers
self-BLEU-2	0.466	0.418
self-BLEU-3	0.456	0.408
self-ROUGE-L	0.553	0.489

Table 6: Inter-annotator agreement among undergraduate students and crowd workers in CovidET-EXT.

E Additional Analyses of CovidET-EXT

Trigger Positions. We examine the position of the emotion trigger sentences in the original posts. The sentence-level distribution of the annotated triggers is reported in Figure 6. Results reveal that the trigger sentences spread evenly across the posts, with a large number of triggers clustering in the later parts of the post. This means that the emotion-trigger extractive summarization task is *not* lead-based, unlike generic news summarization (Fabbri et al., 2021; Sebastian et al., 2019). This is especially true for *anticipation*, as demonstrated in Figure 7.

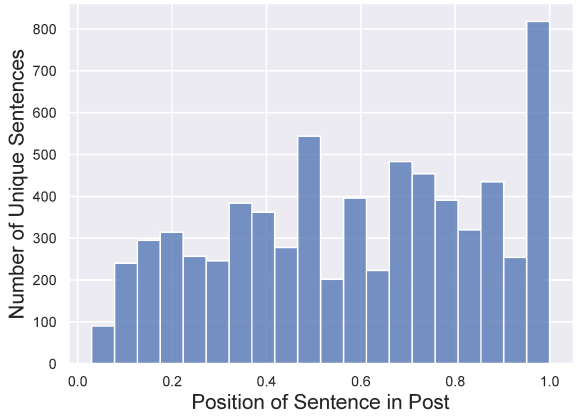

Figure 6: The sentence-level distribution of triggers in the original posts of COVIDET-EXT.

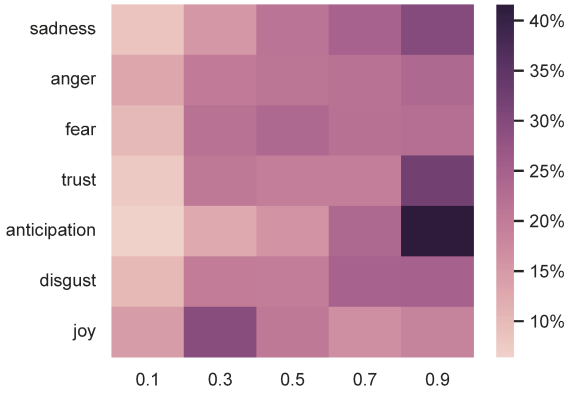

Figure 7: Heatmap of the distribution of triggers in the original posts of COVIDET-EXT. The X-axis stands for the position of trigger sentences in the original post, and the colorbar exhibits the percentage of trigger sentences under the emotion label.

Trigger Components. In addition to the explicitness of emotion triggers, we also examine the syntactic components of the extractive summaries of emotion triggers. Results are shown in Figure 8. We observe that nouns and verbs take up the majority of triggers, closely followed by the use of pronouns.

Pronoun Distributions. Psycho-linguistic studies reveal that the analysis of function words such as pronouns can disclose psychological effects of life experiences and social processes (Campbell and Pennebaker, 2003; Tausczik and Pennebaker, 2010; Pennebaker et al., 2014; Seraj et al., 2021; Singh et al., 2018). Specifically, overusing the first-person singular pronouns may imply a high level of self-involvement, whereas the increased use of other pronouns may signify improvement of social engagement (Cohn et al., 2004; Simmons et al., 2008; Kumari and Singh, 2017).

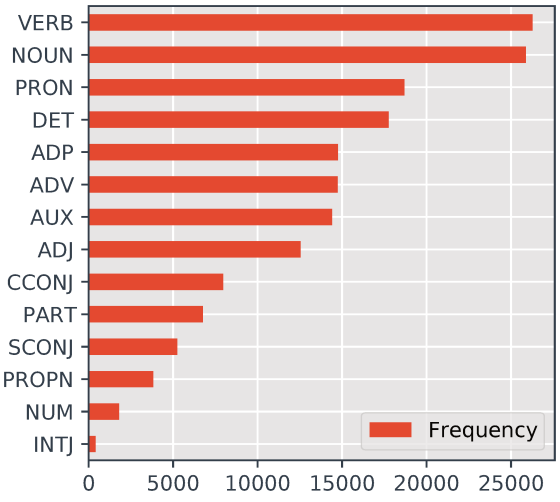

Figure 8: POS frequency distribution in COVIDET-EXT.

We evaluate the percentage of personal pronoun usage per annotated emotion trigger sentence. In particular, we discover an inverse correlation between first-person singular pronouns (e.g., *I*, *me*, *my*, *mine*, *myself*) and second-person pronouns (e.g., *you*, *your*, *yours*, *yourself*, *yourselves*). We provide the average percentage of the personal pronouns per emotion trigger in Figure 9. Further statistical tests reveal negative Pearson correlations between the percentage distribution of first-person singular pronouns and second-person pronouns in each emotion (with substantial significance in all 7 emotions; shown in Table 7). We note that when expressing negative emotions such as *sadness* and *fear*, authors used more first-person singular pronouns in triggers. On the other hand, authors used more second-person pronouns in expressing the triggers for positive emotions like *joy* and *trust*. The inverse correlation between first-person singular pronouns and second-person pronouns suggests more self-involvement in negative emotions and more social engagement in positive emotions in COVIDET-EXT.

Topical Variations. To better interpret the annotated emotion triggers, we train a multi-class bag-of-words logistic regression model to predict the emotion label of each annotated extractive emotion trigger sentence. The trained model's weights pertaining to each class of emotions are then extracted to locate the tokens that are most indicative of each emotion. The multi-class logistic regression model achieved a micro F1 score of 0.33 after

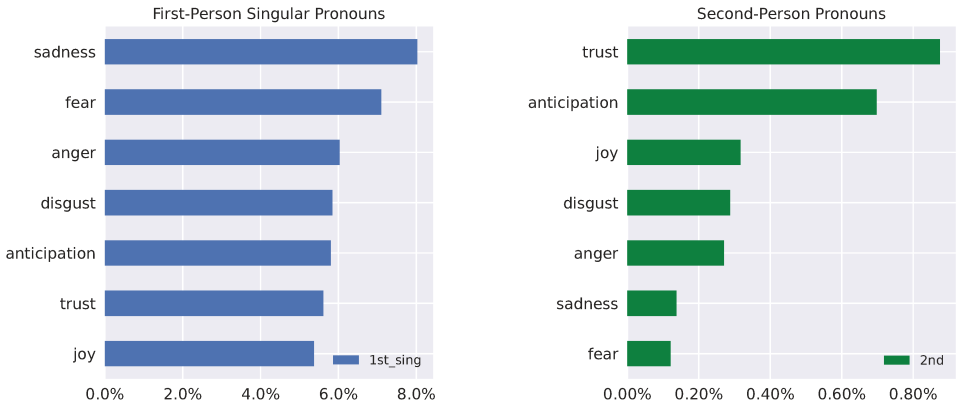

Figure 9: Average percentage of first-person singular and second-person pronouns in the annotated extractive summaries of emotion triggers of COVIDET-EXT.

	Pearson's r	p
anger	-0.1288	$4.77e^{-06}*$
fear	-0.0903	$1.45e^{-05}*$
anticipation	-0.1671	$1.13e^{-14}*$
joy	-0.1634	$1.22e^{-03}*$
sadness	-0.0945	$2.05e^{-03}*$
trust	-0.1873	$7.74e^{-04}*$
disgust	-0.1167	$1.90e^{-02}*$

Table 7: Pearson Correlation Coefficients between the percentage distribution of first-person singular pronouns and second-person pronouns among emotions in COVIDET-EXT. * indicates p value < 0.05.

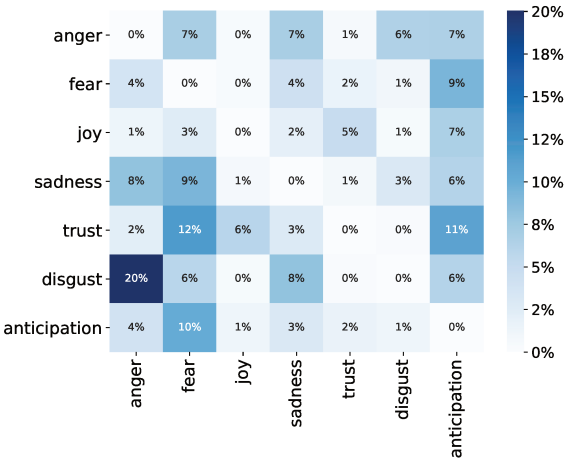

Figure 10: Overlapping trigger sentences for different emotions. Cell (i, j) represents the percentage of sentences that are triggers for emotion i that are also triggers for emotion j in COVIDET-EXT.

training and evaluating on our benchmark dataset. The most indicative tokens associated with each emotion are reported in Table 8.

Connections to COVIDET. To understand the ties between COVIDET-EXT and COVIDET, we measure the self-BERTScore between the extractive summaries of triggers from COVIDET-EXT and the abstraction summaries of triggers from COVIDET. Results reveal that the average BERTScore F1 is 0.872 between the extractive and abstractive summaries, indicating strong correlations between the two datasets.

Same Triggers for Different Emotions. The status of overlapping trigger sentences for different emotions is shown in Figure 10. Specifically, we measure the percentage of sentences that are triggers for an emotion i that are also triggers for emotion j in COVIDET-EXT.

F Supervised Extractive Summarization

Although our focus is exclusively on unsupervised approaches to eliminate the reliance on labeled data, we note that Covid-EXT can be a suitable benchmark for developing supervised methods as well. In this section, we compare two supervised methods against our unsupervised EAP. We experiment with two methods for emotion trigger extraction. **1)** First, we experiment with the BART-FT-JOINT (Zhan et al., 2022) model which is trained to jointly predict emotions and their summary. We train this model on the training set of Covid-EXT in a supervised manner. Second we employ a simple **2)** BERT (Devlin et al., 2019) classifier that is trained in a supervised manner to detect emotions at sentence level. We consider as positive examples the sentences that are included in the summary,

	ANGER		DISGUST		FEAR		JOY		SADNESS		TRUST		ANTICIPATION	
	Token	Weight	Token	Weight	Token	Weight	Token	Weight	Token	Weight	Token	Weight	Token	Weight
0	annoying	7.08	prodded	5.33	grandpa	7.11	grateful	6.84	unrelated	5.93	reacting	6.19	wreck	5.69
1	upset	6.42	guard	5.05	shd	6.92	thankfully	5.89	believing	5.93	okay	5.34	monger	5.53
2	angry	5.81	maskless	4.94	freaking	6.12	happy	5.43	sad	5.80	ineffective	5.12	harm	5.44
3	ridiculed	5.43	wiped	4.87	expose	6.09	fantastic	4.76	devastated	5.22	stock	5.06	statistic	5.43
4	milder	5.32	care	4.84	pass	6.07	glad	4.72	fault	5.09	cheer	5.00	question	5.28
5	ideal	5.25	nicely	4.75	afraid	5.69	provide	4.65	antivax	5.08	haven	5.00	waited	5.04
6	realized	5.20	experiencing	4.44	fear	5.60	ended	4.52	depression	5.00	accepted	4.55	infectious	5.04
7	strongly	5.20	beginning	4.41	pills	5.55	million	4.19	disappear	4.93	affirmed	4.26	questioning	4.97
8	wtf	5.18	coronavirus	4.40	scared	5.52	success	4.09	dead	4.81	psychiatrist	4.11	june	4.86
9	centered	5.08	dumbass	4.39	venting	5.49	effective	3.89	virtually	4.77	worried	4.04	wait	4.83

Table 8: The tokens with the most positive weights for each emotion in a multi-class bag-of-words logistic regression model trained to classify the emotion indicated by the trigger sentences.

	ANGER		DISGUST		FEAR		JOY		SADNESS		TRUST		ANTICIPATION		AVG	
	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L	R-2	R-L
BART-FT-JOINT	0.335	0.371	0.299	0.312	0.377	0.384	0.304	0.335	0.375	0.370	0.254	0.276	0.333	0.338	0.325	0.340
BERT	0.329	0.367	0.291	0.304	0.372	0.376	0.293	0.295	0.361	0.363	0.242	0.268	0.323	0.332	0.315	0.329
EAP	0.324	0.348	0.285	0.296	0.364	0.373	0.285	0.319	0.348	0.354	0.239	0.264	0.319	0.324	0.309	0.325

Table 9: Comparison between EAP and supervised approaches.

and negative examples the rest of the sentences. Note that we train 7 different models, one for each emotion.

We show the results obtained in Table 9. We observe that BART-FT-JOINT outperforms our EAP considerably by 1.5% in Rouge-L score. However, we see that the BERT-based approach is much closer to the performance of the unspervised EAP, outperforming it by less than 1% in Rouge-L and F1.

G Hyperparameters

In this section we detail the values of the hyperparameters used and the search space considered in the development of our EAP. First in terms of the constant c in Equation 4, we experiment with values in the range $0.1 \rightarrow 0.5$ but observed that 0.1 works well. We mentioned that the minimum frequency of a word necessary for selection in our vocabulary V is 20. We also experimented with other values ranging from 5 to 50. The threshold t from Equation 7 is emotion-specific and inferred using the validation set. We experiment with values between 0.2 and 0.7 and observed that 0.35 works well in general.

H Model Analysis

To offer additional insights into our approach, we show in Figure 11 an analysis on the effect of the top relevant terms on the performance of EAP. For each emotion, we experiment with completely dropping the top k most relevant terms (i.e., words) in the graph, with k ranging from 1 to 40 and report the average performance obtained. This analysis can be seen as a way to measure the reliance of EAP and the top relevant words. We observe that the performance drops considerably while dropping the first 28 terms and the starts to plateau.

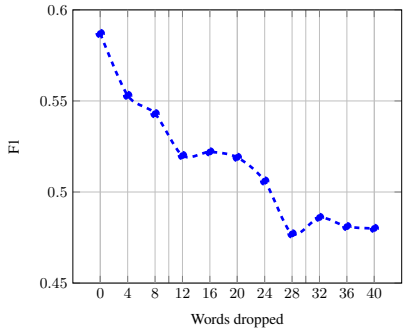

Figure 11: Average F-1 obtained when dropping the top k (with k from 0 to 40) highest relevance nodes in the graph.

I Extractive Summarization Results in terms of F1

In Table 10 we present the performance on extractive summarization in terms of F1. While Rouge captures the overlap between extracted summaries and human references at word level, F1 measures the number of extracted sentences from the post that are correctly part of the gold summary (human

references). Specifically, we compute F1 as if we dealt with a traditional classification problem. For every emotion, the sentences belonging to the trigger summaries are positive examples, and all the other sentences are negative examples. If our EAP model selects a sentence that does not appear in the trigger summary, we view it as a false positive. On the other hand, if our EAP model does not extract a sentence which belongs to the trigger summary, we count it as a false negative. We calculate F1 as the harmonic mean between precision and recall.

	ANGER	DISGUST	FEAR	JOY	SADNESS	TRUST	ANTICIPATION	AVG
1-SENT	0.14	0.07	0.159	0.113	0.097	0.197	0.235	0.144
3-SENT	0.306	0.182	0.300	0.275	0.241	0.270	0.268	0.263
PACSUM	0.297	0.179	0.296	0.280	0.246	0.271	0.276	0.263
PRESUMM	0.302	0.189	0.302	0.283	0.241	0.273	0.274	0.266
TEXTRANK	0.286	0.165	0.289	0.274	0.239	0.270	0.211	0.247
EMOLEX	0.238	0.248	0.320	0.238	0.298	0.200	0.218	0.253
EMOINTENSITY	0.298	0.221	0.347	0.293	0.325	0.274	0.272	0.284
BERT-GOEMO	0.264	0.215	0.308	0.216	0.312	0.201	0.253	0.269
EAP	$\mathbf{0.315}^{\dagger}$	$\mathbf{0.251}^{\dagger}$	$\mathbf{0.361}^{\dagger}$	$\mathbf{0.305}^{\dagger}$	$\mathbf{0.354}^{\dagger}$	$\mathbf{0.299}^{\dagger}$	$\mathbf{0.285}^{\dagger}$	$\mathbf{0.310}^{\dagger}$

Table 10: Results of our models in terms of F1. We assert significance† using a bootstrap test where we resample our dataset 50 times with replacement (with a sample size of 500) and $p < 0.05$.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☑ A2. Did you discuss any potential risks of your work?
Limitations Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 5

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5 + Appendix G

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 + Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5 experimental setup

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This data collection is reviewed and exempted by the IRB board of our institution

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C