Emotion trigger summarization when augmented with synthetic data

Anuva Banwasi
ab5084@columbia.edu

Applying 4 late days

Abstract

Emotion detection and classification are well-established tasks in natural language processing. One area that is becoming increasingly important and has been less researched is emotion summarization. In essence, the objective is to detect and summarize both the emotion and its triggering factors. This is significant for enhancing overall comprehension of a conversation or text. The authors of the 2022 paper "Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts" presented the COVIDET dataset (Emotions and their Triggers during Covid-19) sourced from over 1,000 Reddit posts where each post was manually annotated with an emotion as well as an abstractive summary describing the triggers of the emotion. However, a limitation of the COVIDET dataset is the uneven representation of certain emotions, such as disgust, joy, and trust. The goals of this paper are two-fold: (1) can we use LLMs like GPT-4/Chat GPT to generate synthetic data for the emotion disgust and (2) can we fine-tune BART and T-5 models on this data along with the original COVIDET dataset for the task of emotion trigger summarization? We experiment with prompt-engineering strategies including zero-shot and few-shot learning to generate the synthetic data. We then test these strategies against each other and analyze the resulting performance of the BART and T-5 models. We compare the model performance with and without the additional LLM-generated dataset for each variation of the synthetic data.

1 Introduction

Emotion is a key part of our daily conversations as well as in the text and content that we produce online. Especially during large-scale crises like the COVID-19 pandemic, people are heavily emotionally affected. To better understand people's emotions during crisis, we need a better understanding of what causes emotion. Within the field of emotion recognition, detection and classification are well-established tasks. However, a largely unresolved question is understanding the cause of an emotion. What causes someone to feel fear or anger? Given a conversation or online post, what leads to the speaker/user to feel a certain way? This is the area of emotion trigger summarization, the task of understanding what causes emotion in texts or conversations. The goal is to not only detect the emotion in a conversation/text but to also summarize the trigger leading to that emotion. This is extremely important because understanding what causes an emotions can help to improve overall understanding of a conversation or piece of text. Xia and Ding (2019) utilize an extraction framework to identify emotion "causes" within the confines of the clause level. However, their setup involves the association of one explicitly expressed emotion within a single cause. In their 2022 paper "Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts", Zhan et al. (2022) argued that the Xia and Ding study does not generalize to extended, impromptu social media posts characterized by high emotional charge/intensity. To fill this gap in the literature, they presented the COVIDET dataset (Emotions and their Triggers during Covid-19); this dataset was sourced from over 1,000 Reddit posts where each post was manually annotated with an emotion as well as an abstractive summary describing the triggers of the emotion. However, one of the main limitations of this dataset is that there were several emotions such as disgust, joy, and trust that were not as well-represented. For example, the COVIDET dataset had about 1200 examples of posts for the emotion anticipation vs. less than 400 examples of posts for the emotion disgust. The goal of this paper is to use LLMs such as GPT-4 to generate data like COVIDET for the underrepresented emotion disgust and then use the LLM-generated dataset along with the original COVIDET dataset to fine-tune

the BART and T-5 models for the task of emotion trigger summarization.

We consider this augmentation of the COVIDET dataset with synthetic data for an underrepresented emotion, such as disgust, to be a valuable and novel contribution in the realm of emotion trigger summarization. This endeavor not only enables us to assess the efficacy of LLMs in generating social-media-like data but also allows for a comparative analysis with the original dataset. The exploration of whether this augmented dataset enhances the performance of emotion trigger summarization, particularly for emotions like disgust, is a crucial step in understanding the potential impact of synthetic data on improving model robustness and generalization. Looking ahead, the methodology developed for this augmentation could serve as a blueprint for addressing imbalances in other datasets by leveraging synthetic data generated from LLMs, thereby contributing to more comprehensive and representative training sets for various NLP tasks.

2 Related Work

Emotion detection. Previous research on emotions in social media predominantly centers around the detection of emotions. Wang et al. (2012) explored sentiment analysis and emotion detection in microblogs, employing machine learning techniques to discern emotions expressed in short, user-generated content such as tweets on Twitter. Biyani et al. (2014) delved into the challenges of detecting sentiment and emotions in online health forums. They modeled the task as a binary classification problem with the goal being to identify emotional support and informational support in the user messages on these forums. Abdul-Mageed and Ungar (2017) focused on enabling better emotion detection using a larger dataset. They collected a large dataset of labeled tweets over 24 emotion types and then employed recurrent neural networks to perform emotion detection. More recently, semi-supervised learning has emerged as a key technique for enhancing deep learning models when there is less training data. Sosea and Caragea (2022) introduced a novel self-training approach that utilizes training dynamics to evaluate the adequacy of the pseudo-labels generated by the teacher.

Summarization. Language summarization research has witnessed significant strides, driven by advancements in pre-trained models. Lewis et al. (2019) highlights the progress in single-document summarization achieved through abstractive approaches employing encoder-decoder transformer models. Their work showcases the effectiveness of these models in synthesizing concise and informative summaries of textual content, contributing to the broader landscape of language summarization. Language summarization has been further driven by 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, Völske et al. (2017) and Kim et al. (2019) utilized TL;DR sentences from Reddit as summaries to train their models. However, a key limitation of these papers is that these language summaries do not always adequately summarize emotion expressed in the discussion or news content. To address this gap, Zhan et al. (2022) developed the COVIDET dataset with manual emotion trigger summaries for over 1,000 Reddit posts in the COVID-19 support forum. Their research goes beyond language summarization and into emotion summarization and emotion cause extraction.

Emotion cause extraction. Emotion cause extraction has emerged as a significant research area within natural language processing, aiming to uncover the factors that trigger emotions in textual content. Chen et al. (2010) initially introduced ECE, defining it as the extraction of word-level causes associated with a given emotion in text. Gao et al. (2015) expanded the scope of emotion cause extraction by introducing clause-level cause detection tasks. Their research aimed to identify causes associated with specific emotions within clauses, contributing to a deeper understanding of the nuanced relationship between emotions and their triggers. Gui et al. (2016) further advanced this domain by exploring the task of emotion cause detection at the clause level. Xia and Ding (2019) contributed to this field by adopting an extraction setup focused on identifying emotion "causes" at the clause level. Their work provided valuable insights into associating explicitly expressed emotions with their respective causes within individual clauses. However, the study acknowledged limitations in generalizing this approach to longer, emotionally charged social media posts, highlighting the need for advancements in handling the complexities of emotional expression across various textual contexts

3 Data

For this research, we will be using the COVIDET dataset from the Zhang et al. paper as well as synthetic data generated from GPT-4. The COVIDET dataset is sourced from 1,8883 English Reddit posts on the topic of the COVID-19 pandemic. The posts are each manually annotated with seven emotion labels (anticipation, fear, anger, sadness, disgust, joy, trust). Furthermore, for each emotion, the annotators also produce an abstractive summary on the triggers of that emotion. They sampled their posts from the r/COVID19_support channel. Specifically, they restrict posts that are 50-400 tokens long, with the most common length being around 100 tokens. COVIDET differentiates itself from other emotion studies in that it is comprised of longer texts as opposed to short sentence-level tweets. Another advantage of COVIDET is that they validated their posts and the manual trigger descriptions. However, one limitation of the COVIDET data is that it is unbalanced in such a way that the emotions joy, trust, and disgust are much less represented than the emotions anger, anticipation, fear, and sadness. When analyzing the model performance for emotion trigger summarization, we see that BART performs most accurately (highest ROUGE-L) for the more well-represented emotions like anger and fear but less accurately for the less-represented emotions like disgust and trust. One hypothesis is that the lack of data (posts) for these emotions adversely impacts the model performance.

To address this issue, we generate data with an LLM (GPT-4) for the rare emotion disgust. We chose to focus on disgust out of the other under-represented emotions like joy and trust because the BART model performed the worst on this emotion. Moreover, we believe that the emotion disgust should be more common in COVID-19 posts because many people feel this way when they have the disease. We prompt GPT-4 to generate posts about COVID-19 that are 100-200 tokens long where the main emotion in the post is disgust. For each generated post, we also ask the model (GPT-4) to produce an abstractive summary of the trigger for the emotion. We experiment with different prompt-engineering strategies to generate the synthetic data. These are discussed more in the Experiments section. For each prompt-engineering strategy, we generate around 300 posts and their associated annotation trigger summary. We also do a comparison of the generated data and the COVIDET data by manually evaluating on the features of consistency, fluency, coherence, and relevance. Any generated posts that have the wrong emotion (aka not disgust) are flagged and removed.

4 Methods

The second task of this project after data generation is emotion trigger summarization. The goal is to investigate the impact on performance when fine-tuning models on LLM generated data in addition to COVIDET data. We perform trigger summarization using two main models: 1) BART from Lewis et al. (2019) and 2) T-5 from Raffel et al. (2023). BART (Bidirectional and Auto-Regressive Transformers) is a sequence-to-sequence model developed by Facebook AI Research. It combines bidirectional pre-training and auto-regressive fine-tuning, employing a Transformer architecture to generate coherent and contextually rich text bidirectionally. BART particularly excels in abstractive text summarization, demonstrating state-of-the-art performance by learning to reconstruct corrupted versions of input sequences during pre-training. T5 (Text-To-Text Transfer Transformer) is a pre-trained language model developed by Google Research that frames all NLP tasks as a text-to-text problem. It employs a transformer architecture and is pre-trained on a diverse range of tasks, making it highly adaptable for various natural language processing applications. T5 has achieved strong performance across tasks like translation, summarization, question answering, and more, showcasing its versatility and effectiveness in understanding and generating human-like text.

For this midterm report, we have so far utilized BART trained on the COVIDET and LLM-generated datasets. We also trained and test the T-5 model on the COVIDET data. Our baseline model is the existing BART-FT from the Zhang et al. paper where they fine-tuned BART on only the COVIDET dataset. We compare the results of this baseline model with our BART model fine-tuned on both the COVIDET dataset and the LLM-generated dataset as well as the T5 model fine-tuned on both datasets. In order to evaluate the models and compare their performance, we use two main metrics: ROUGE-L and BERTScore. ROUGE-L is an evaluation metric commonly used for assessing the quality of summaries by measuring the longest common subsequence (LCS) between the generated summary and the reference summary. BERTScore is a

metric that leverages contextual embeddings from BERT to evaluate the semantic similarity between the generated and reference summaries, providing a more nuanced and context-aware assessment of the quality of generated text.

5 Experiments

We have so far used a variety of prompt-engineering techniques and fine-tuned the BART model on the resulting LLM-generated data along with the original COVIDET dataset. The first prompt-engineering technique utilized was the zero-shot learning strategy. Here, we used the GPT-4 API and provided a prompt asking it to generate a Reddit post about COVID-19 with the main emotion represented in the post being disgust. A sample prompt is "You are a person from 2020-2021 posting in the Reddit r/COVID19_support channel. This is a subreddit dedicated to people seeking community support during the pandemic. Generate a Reddit post that is 100-400 tokens in length where the main emotion expressed in your post is disgust." After generating the post, we then prompt GPT-4 to produce an abstractive summary of the trigger. A sample prompt for the trigger summary generation is as follows: "Based on the Reddit post you just generated, summarize what caused the emotion of disgust in 1-2 sentences." We repeated this to generate around 300 Reddit posts and their associated abstractive trigger summaries. We then fine-tuned the BART model on the joint COVIDET and the LLM-generated data.

The second prompt-engineering strategy utilized was few-shot learning. In the COVIDET paper, the authors show in the Appendix how they provided their manual annotators with examples of Reddit posts with the specific emotions represented as well as the abstractive trigger summaries. In a similar manner of providing examples to manual annotators, we can provide samples to the GPT-4 model via few-shot learning. The motivation behind few-shot learning, as described in "Meta-Dataset: A Benchmark and Analysis for Multi-Task and Meta-Learning in the Wild" (Triantafillou et al., 2020), is to enable models to generalize and perform well on new tasks with minimal training examples. Few-shot learning aims to address scenarios where obtaining large labeled datasets for each specific task may be impractical or costly. In our case, providing labeled examples to an LLM like GPT-4 can help it understand the kind of content we are asking it to generate. This is one such instance of what we would include in the prompt to GPT-4: "Consider the following example Reddit post followed by the emotion expressed in the post and a summary of trigger causing that emotion. Generate your own Reddit post that is 100-200 tokens where the main emotion is disgust. Also generate the emotion and a summary of the trigger that caused the disgust in the post." We generated 300 such Reddit posts and then fine-tuned the BART model on this data alongside the COVIDET data.

6 Results

The first analysis we conducted was to evaluate the LLM-generated data and compare it to the original COVIDET dataset. For this, we had manual annotators (myself and another student) each evaluate around 50 of the LLM- generated posts along with their emotion trigger summaries. For comparison, we also provided annotators with 10 Reddit posts and trigger summaries from the COVIDET dataset for the emotion disgust. They were asked to compare the LLM-generated data and the COVIDET data and judge the consistency, fluency, coherence, and relevance to the topic of disgust around COVID-19. Using the same system from Fabbri et al. (2021) and Zhang et al. (2022), we had the annotators grade the summaries and posts for each feature on a score of 1-5. We then took the average of their responses. ZS refers to the posts and summaries generated by zero-shot strategy and FS refers to the posts and summaries generated by the few-shot strategy. The results were as follows:

	Coherence	Consistency	Fluency	Relevance
ZS Posts	4.632	4.975	4.984	2.434
ZS Summ	4.751	4.882	4.985	2.510
FS Summ	4.843	4.968	4.989	2.612
FS Summ	4.794	4.973	4.947	2.745

Table 1: Coherence, Fluency, Relevance, etc. scores for GPT-4 generated content (posts and summaries).

As can be seen in Table 1, the consistency and fluency were marked to be quite high. The most common problem was relevance with the annotators finding that GPT-4 sometimes generated content that included irrelevant information. This was more common for posts that were generated with the zero-shot strategy. The relevance was higher in the posts generated using the few-shot learning strategy. This shows that providing GPT-4

Model	Anger		Disgust		Fear		Joy		Sadness	
	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
3-SENT	0.142	0.598	0.129	0.562	0.153	0.535	0.154	0.537	0.134	0.517
BART	0.161	0.587	0.138	0.558	0.164	0.529	0.149	0.551	0.157	0.559
BART-COVIDET	0.190	0.705	0.159	0.695	0.206	0.748	0.165	0.699	0.177	0.718
BART-FT-Zero	0.184	0.715	0.166	0.711	0.212	0.694	0.126	0.712	0.143	0.629
BART-FT-Few	0.191	0.718	0.170	0.714	0.209	0.713	0.154	0.699	0.184	0.615

Model	Trust		Anticipation	
	R-L	BERTSc	R-L	BERTSc
1-SENT	0.118	0.537	0.119	0.507
3-SENT	0.152	0.548	0.142	0.527
BART	0.158	0.571	0.164	0.558
BART-COVIDET	0.162	0.653	0.198	0.749
BART-FT-Zero	0.164	0.621	0.187	0.727
BART-FT-Few	0.160	0.609	0.192	0.764

Table 2: Results of models in terms of ROUGE-L and BERT-Score for trigger summarization.

with few-shot examples can help the LLM generate more relevant/on-topic content.

Next, we present the results for the BART model trained on the COVIDET and LLM-generated data for emotion trigger summarization. We calculated the ROUGE-L and BERTScore for the summarization task on each emotion. The baseline model scores are represented in the row labeled BART-COVIDET. These are the original scores from the Zhang et al. paper where they trained on the COVIDET dataset alone. We also include baselines from basic approaches like 1-SENT and 3-SENT that just consider the first sentence or first three sentences for the trigger summary. Our own results are included in the last two rows of the table: BART-FT-Zero, from BART trained on COVIDET and LLM-generated data using the zero-shot learning strategy and BART-FT-Few, from BART trained on COVIDET and LLM-generated data using the few-shot learning strategy. Table 2 at the top of the page displays the results.

The ROUGE-L and BERTScores are relatively comparable for the emotions like anger, disgust, fear, sadness, trust, and anticipation. Our focus for this study is on the comparison of the scores for disgust. We see that between BART-COVIDET and BART-FT-Zero, the ROUGE-L score for disgust increased by around 4% and the BERTSc increased by 2%. The BART-FT-Few-shot fine-tuned model showed even larger improvement. Between BART-COVIDET and BART-FT-Few, we see an increase in 7% for the ROUGE-L and an increase in 2.7%. This is promising and demonstrates that there could be potential benefits from including LLM-generated data to improve the accuracy for trigger summarization on less represented emotions like disgust, joy, and trust. We posit that the incorporation of synthetic data enhanced the model performance. The infusion of additional instances of posts and summaries featuring disgust facilitated a more robust training regimen for the BART model. Consequently, during testing, the model exhibited improved proficiency in recognizing and summarizing this particular emotion. This observation underscores the utility of synthetic data in addressing the limitations of imbalanced datasets, offering a practical solution to bolster the model's comprehension and handling of emotions, ultimately contributing to more accurate and nuanced results in emotion trigger summarization tasks.

7 Error Analysis

The most common issue that we faced during the data generation phase using LLMs was that it would output a post that was not relevant. This was especially a problem in the zero-shot learning strategy where we did not provide it with sample posts. For example, an initial post generated by GPT-4 was the following: "Hey r/COVID19_support, feeling the need to share a bizarre yet annoying experi-

ence. So, I was doing my laundry, trying to maintain some sense of normalcy in these weird times, right? Well, it turns out I've lost yet another sock. Like, seriously, where do they go? The frustration of having mismatched socks is just...disgusting. I mean, who needs this added stress right now? Anyone else dealing with the absurdity of sock disappearances during quarantine? Let's commiserate. #SockDrama #LaundryWoes." Although the emotion of disgust is present in the post, it does not match the content of most of the other posts on the channel that are about frustrations around COVID-19. The issue was that the prompt was not specific enough to focus on the topic of the channel. To address this problem, we added more details in the prompt asking the LLM to generate content about distress and disgust around the COVID-19 situation, disgust around symptoms, etc. We chose 40 LLM-generated posts and had 2 annotators mark they were relevant to the topic at hand. After improving the prompt to include more detail about disgust around COVID-19, the percentage for relevant posts increased from 64% to 87%.

8 Future Work

We have effectively created a pipeline for creating LLM-generated posts. We have fine-tuned the BART and T-5 models on data that includes COVIDET and this synthetic data. We have also worked with a couple different prompt-engineering techniques like zero-shot and few-shot learning. We tested and evaluated the performance of the fine-tuned model against the original baselines from the Zhang et al. paper. The next steps that we are looking at is if we can better evaluate the LLM-generated content. Right now, we are having manual annotators review the generated posts and evaluate them against the COVIDET data for fluency, consistency, etc. We want to see if we can somehow make this process more automatic (ex: are their models that can be used to numerically compare the COVIDET and LLM-generated data). We especially need to make sure that posts with irrelevant content are not included so that it does not dilute the training data. We will be looking into ways to automatically flag LLM-generated content that does not fit within the confines of the topic of disgust around COVID-19. We also want to see if it is possible to improve the accuracy on the other emotions besides disgust. We see that with the BART model, there are instances where the score for disgust improves but the score for other emotions decreases. A future next step would be to analyze whether it is possible to train on even larger amounts of generated data and see if this helps increase the accuracy.

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.

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.

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

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.

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.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Tiberiu Sosea and Cornelia Caragea. 2022. 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.

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 Amit Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts.

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.