Language Models (Mostly) Do Not Consider Emotion Triggers When Predicting Emotion

Smriti Singh[1] **Cornelia Caragea**[2] **Junyi Jessy Li**[1]
[1]The University of Texas at Austin
[2]University of Illinois Chicago
{smritisingh26, jessy}@utexas.edu, cornelia@uic.edu

Abstract

Situations and events evoke emotions in humans, but to what extent do they inform the prediction of emotion detection models? This work investigates how well human-annotated emotion triggers correlate with features that models deemed salient in their prediction of emotions. First, we introduce a novel dataset EMOTRIGGER, consisting of 900 social media posts sourced from three different datasets; these were annotated by experts for emotion triggers with high agreement. Using EMOTRIGGER, we evaluate the ability of large language models (LLMs) to identify emotion triggers, and conduct a comparative analysis of the features considered important for these tasks between LLMs and fine-tuned models. Our analysis reveals that emotion triggers are largely not considered salient features for emotion prediction models, instead there is intricate interplay between various features and the task of emotion detection.

1 Introduction

Understanding perceived emotions and how they are expressed can be immensely useful for providing emotional support, sharing joyful situations, or in a therapy session, thus emotion detection has become a well-studied task (Strapparava and Mihalcea, 2007; Wang et al., 2012; Abdul-Mageed and Ungar, 2017; Khanpour and Caragea, 2018; Liu et al., 2019a; Sosea and Caragea, 2020; Demszky et al., 2020; Desai et al., 2020; Sosea and Caragea, 2021; Sosea et al., 2022; Hosseini and Caragea, 2022, 2023a,b). However, though existing work has sought to identify what triggers or causes a particular emotion (Lee et al., 2010; Chen et al., 2010; Gui et al., 2016; Xia and Ding, 2019; Zhan et al., 2022; Sosea et al., 2023), the relationship between those triggers and the prediction of emotion detection models is little understood. This relationship is crucial to investigate, without which the interpretation of perceived emotions—or claims for a model to be able to do so—is *hollow* (James, 1884).

While humans can intuitively construe emotional reactions with events that trigger them (Jie and Ong, 2023), it is unclear to what extent, if any, current NLP models are doing so. Prior work trained models to *learn* to recognize and summarize emotion triggers or causes. In this work, we instead ask the question: *what roles do emotion triggers play in emotion prediction?* In addition to fine-tuned transformer models shown to be performant on this task, we additionally put an emphasis on large language models (LLMs) including both API-based and open-sourced ones, since their capability for trigger prediction has not been explored.

To ground our analysis, we present EMOTRIGGER, a linguist-annotated dataset of emotion triggers (as extractive text spans), over three social media corpora with labeled emotions across different themes: CancerEmo (Sosea and Caragea, 2020), HurricaneEmo (Desai et al., 2020), and GoEmotions (Demszky et al., 2020). This is, to the best of our knowledge, the first dataset annotated with high-quality triggers focusing on *short* social media texts. Engaging with tools that attribute model prediction (Lundberg and Lee, 2017) and prompts that elicit natural language explanations from LLMs, we aim to answer the following research questions:

1. Are LLMs capable of detecting emotions and identifying their triggers?
2. To what extent do emotion prediction models rely on features that reflect emotion triggers?
3. How often do the triggers overlap with keyphrases or emotion words?

We find that LLMs can identify emotions with high accuracy, but the performance for identifying triggers is mixed. With the exception of GPT-4, word features deemed salient for emotion prediction are only marginally related to these triggers. Instead, we found that automatically extracted keyphrases (Bougouin et al., 2013) are highly correlated with salient features. Overall, we establish

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 603–614

that the current state of the art language models cannot proficiently construe emotional reactions with events that trigger them.

We release EMOTRIGGER at https://github.com/smritisingh26/EmoTrigger.

2 The EMOTRIGGER Dataset

We first present EMOTRIGGER, a dataset annotated with ground-truth emotion triggers. EMOTRIGGER consists of subsets from three well-known datasets commonly used in emotion prediction, on top of which we engage expert annotators to highlight triggers for the annotated emotion. Although prior emotion cause datasets exist, they largely focus on explicit emotions (Chen et al., 2010; Gui et al., 2016). Sosea et al. (2023)'s dataset for Reddit posts involves multi-sentence triggers for summarization, while this work focuses on word or phrase level attributions and explanations.

Source Datasets (1) *HurricaneEmo* (Desai et al., 2020) consists of tweets related to 3 hurricanes. Each tweet is annotated for 24 fine-grained emotions in Plutchik's wheel of emotions (Plutchik, 1982). We map the emotions onto Plutchik's 8 basic emotions (Appendix A).

(2) *CancerEmo* (Sosea and Caragea, 2020) consists of sentences sampled from an online cancer survivors network. This dataset is annotated with Plutchik's 8 basic emotions.

(3) *GoEmotions* (Demszky et al., 2020) is a general domain dataset that consists of sentences extracted from popular English subreddits, labeled for 27 emotion categories or Neutral, which are mapped to the coarser Ekman's 6 emotions (Ekman, 1992).

To balance the emotions in EMOTRIGGER, we sample 20 examples per emotion from each constituent dataset. In total, we have 900 samples in our dataset, 160 each from HurricaneEmo and CancerEmo, and 580 samples from GoEmotions. We make sure our samples do not contain links or images, i.e., the text stands alone from other media.

Annotation Our annotation team consists of three linguistics undergraduates experienced with emotion-related text annotation tasks, who annotate the 900 instances (tweets or sentences) for emotion triggers. The annotation task is to find the emotion triggers for a given text and the gold label emotion. The annotators were paid at $15/hr. Our annotation instructions are listed in Appendix B.

Example 1. Sometimes i just get impatient as my husband is the same and my style is to get it done!!
Emotions: Anger

Example 2: @johnlegend NO, WE'RE NOT DOING GREAT! STILL FLOODWATERS AND NO POWER IN PARTS OF FL. TEXAS IS STILL IN TROUBLE!
Emotions: [Anger, Sadness, Disgust]

Example 3: That's Awesome! I never realised it was attributed to [NAME], though.
Emotions: Surprise

Figure 1: Examples from EMOTRIGGER. Triggers highlighted with the same color as the respective emotions.

Subset	A1/A2	A1/A3	A2/A3	Avg
HurricaneEmo	0.92	0.89	0.92	0.91
CancerEmo	0.93	0.89	0.91	0.92
GoEmotions	0.91	0.88	0.93	0.90

Table 1: Annotator agreement (token-level Fleiss Kappa) across subsets of EMOTRIGGER.

We report token-level inter-annotator agreement with Fleiss Kappa (Fleiss, 1971), yielding an average of 0.91, indicating substantial to perfect agreement (Artstein and Poesio, 2008). Table 1 shows the agreement between the annotators, per dataset. This is a positive indication that humans can reliably identify triggers of an emotion in short texts. Examples of EMOTRIGGER are shown in Figure 1.

3 Study Design

We study to what extent words most informative to the final model predictions constitute triggers (as annotated in EMOTRIGGER) of the emotion they are predicting. This section describes our strategy to do so for each model type.

LLMs We experimented with GPT-4, Llama2-Chat-13B (Touvron et al., 2023), and Alpaca-13B (Taori et al., 2023). Our main LLM prompt asks the model to predict emotions present in a given piece of text, and elicit explanations as words most important for detecting those emotions; annotated triggers were used as "salient features". Due to the poor zero-shot performance (Appendix Table 7), we report few-shot results (prompt in Appendix C.1).[1]

[1]We found that the prompt used for GPT4 and LLama2Chat does not allow Alpaca to engage with the text properly, thus we have a different prompt for Alpaca, as shown in C.3. Results from the old vs. new prompt are illustrated in Figure 3 in the Appendix. To account for the spelling mistakes that some LLMS make, we compute the Levenshtein distance of words that have more than 4 letters and consider words that have a distance of less than or equal to 2.

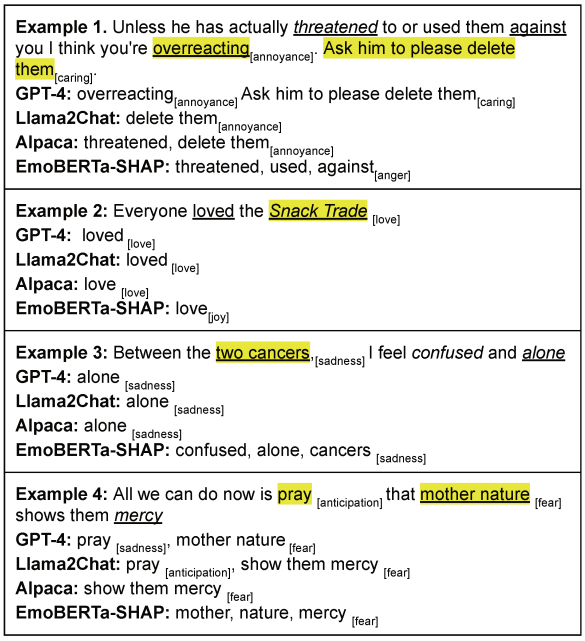

Figure 2: Examples of trigger identification. Gold label triggers are highlighted, keyphrases are underlined and EmoLex words are in italics. The emotions are in subscript to the triggers that caused them.

Since there has been no formal investigation yet for LLM's ability to predict emotion triggers, we also include an *oracle* that asks the model to identify triggers given the gold emotion label in a few-shot manner (prompt in Appendix C.2). Our experiments are run using A100s available on Google Colab and take a total of approximately 50 hours.

Fine-tuned Transformers We fine-tuned transformer models on the aforementioned datasets for emotion prediction, with EmoBERTA (Kim and Vossen, 2021) as the most performant model which we will use in subsequent analyses.[2] To train these models, we use a 75-15-10 split for each dataset respectively. Hyperparameters listed in Appendix E.

To obtain salient features used for emotion prediction, we used SHAP (Lundberg and Lee, 2017), a model-agnostic method based on Shapley values that assigns an importance score to each feature (i.e., word) in the model prediction. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.

Unsupervised Comparators We further consider two methods that extract key information in an unsupervised manner. First, we use **EmoLex** (Mohammad and Turney, 2013) to check to what extent human-annotated triggers or the salient features extracted by models correspond to words that express emotion themselves. Second, we use **keyphrases** extracted from TopicRank (Bougouin et al., 2013), to gauge to what extent corpora-specific keyphrases and themes correspond to triggers or salient features. TopicRank is an algorithm that performs graph-based keyphrase extraction by leveraging the topical representation of the document. The algorithm clusters keyphrases into topics and then utilizes a graph-based ranking model to assign a significance score to each topic, of which the candidates from top ranked topics become the extracted keyphrases.

[2]Other models include BERT (Devlin et al., 2019), DistilBERT (Liu et al., 2019b), RoBERTa (Sanh et al., 2019), DeBERTa (He et al., 2021). Their performance is summarized in Appendix Table 5

Dataset	GPT4	Llama2	Alpaca	Emo-BERTa
HurricaneE.	0.920	0.851	0.756	0.483
CancerEmo	0.914	0.842	0.723	0.378
GoEmotion	0.907	0.820	0.707	0.341

Table 2: Macro F1 score for emotion detection. All scores are based on results of few-shot prompting.

4 Results

For binary token-level comparison of annotated triggers vs. words the models identified as important for emotion detection (henceforth "*salient features*"), we look at: **(1)** Word-level macro-averaged F1 comparing salient features against the union of annotator-highlighted triggers. **(2)** Exact match and partial match. We define an exact match when the model output matches at least one of the annotated triggers. We consider a subset of any annotated triggers as a partial match. With respect to EmoBERTa, we took a cutoff for the SHAP values: 0.85 for an exact match and 0.60 for a partial match. These values were obtained empirically on a randomly sampled set of size 200. This set is sampled outside EMOTRIGGER.

We also report the Pearson's correlation coefficients between SHAP values, Emolex words and the weights of generated keyphrases. For EmoLex and salient features from LLMs, we use a binary vector representation.

Are LLMs capable of detecting emotions and identifying their triggers? We first show each models' performance on the emotion prediction task across all datasets in Table 2. In Appendix Table 6 we show per-emotion results. LLM perfor-

Model	HurricaneEmo			CancerEmo			GoEmotions		
	F1	ExactM	PartialM	F1	ExactM	PartialM	F1	ExactM	PartialM
GPT4 (known emotion)	0.7	0.38	0.9	0.71	0.39	0.91	0.68	0.33	0.9
Llama2 (known emotion)	0.31	0.11	0.35	0.3	0.11	0.37	0.28	0.09	0.34
Alpaca (known emotion)	0.23	0.12	0.29	0.2	0.09	0.25	0.19	0.06	0.21
GPT4	0.66	0.35	0.87	0.68	0.37	0.88	0.65	0.31	0.89
Llama2	0.27	0.09	0.27	0.28	0.08	0.29	0.25	0.08	0.31
Alpaca	0.24	0.08	0.25	0.25	0.08	0.28	0.23	0.06	0.28
EmoBERTA-SHAP	0.21	0.08	0.23	0.19	0.09	0.18	0.18	0.07	0.19
Keyphrases	0.19	0.08	0.23	0.2	0.08	0.25	0.18	0.08	0.18
Emolex	0.08	0	0.07	0.05	0	0.05	0.06	0.01	0.07

Table 3: Macro-F1, exact and partial match to assess the overlap between salient words and annotated triggers.

Dataset	GPT-4 &		Llama2 &		Alpaca &		EmoBerta-SHAP &	
	Keyphrase	EmoLex	Keyphrase	EmoLex	Keyphrase	EmoLex	Keyphrase	EmoLex
HurricaneEmo	0.948	0.401	0.655	0.311	0.633	0.297	0.963	0.375
CancerEmo	0.863	0.711	0.703	0.366	0.686	0.344	0.823	0.635
GoEmotions	0.875	0.717	0.720	0.312	0.711	0.300	0.855	0.707

Table 4: Pearson's correlation values between words each model deems salient and extracted keyphrases or EmoLex words. The scores shown here are the computed average of scores across individual emotion classes.

mance is *consistent* across emotions and datasets while EmoBERTa's performance is substantially inferior. Among the LLMs, GPT-4 outperforms open-sourced ones.

The first portion of Table 3 reports LLM performance when specifically prompted to identify triggers given the annotated emotions. Per-emotion results can be found in Tables 8, 9, and 10 in the Appendix. Again, GPT-4 is the most performant; Llama2 slightly outperforms Alpaca.

As seen in Table 6, we find that GPT4 predicts emotions 'fear' and 'joy' with the highest F1-scores consistently across datasets, and struggles with 'anticipation'. With LLama2 and Alpaca, per-emotion performance is largely dataset dependent.

While there is no emotion for which any model can consistently identify the triggers most accurately, it is worth noting that for GPT-4 and Llama2Chat, the lowest scores are corresponding to the identification of triggers for the emotion 'anticipation'. This is reflected in Tables 8, 9, and 10 in the Appendix.

To what extent do emotion prediction models rely on features that reflect emotion triggers? The bottom portion of Table 3 shows how much the salient features for each model overlap with annotated emotion triggers. Salient LLM features align less well with triggers than the "oracle" scenario above, but the differences are within 3%. Note that this is with few-shot prompting, since we observed much lower alignment with zero-shot (Appendix Table 7). With the exception of GPT-4, salient features in neither Llama2, Alpaca nor EmoBERTa-SHAP align with annotated triggers with very little exact match and partial match.

How often do the triggers overlap with keyphrases or emotion words? We further hypothesize that keyphrases in the dataset, as well as explicit emotion words, might align with salient features that models pick up. Table 4 tabulates the average Pearson's correlation coefficients between salient features (in the case of SHAP, feature salience values) and keyphrase weights or EmoLex.

Surprisingly, the correlations with keyphrases are much higher than with EmoLex for all models and across datasets. This indicates that models rely on explicit emotion words to a lesser extent than expected, and keyphrases help characterize this discrepancy. This is especially true for fine-tuned model (EmoBERTa) on themed datasets (HurricaneEmo, CancerEmo): the SHAP values are much more correlated with EmoLex especially in GoEmotions. We find that emotions like Anger, Joy and Sadness are expressed more through EmoLex words, whereas emotions like Anticipation, Fear and Disgust are expressed more through keyphrases. This is reflected in Figure 5.

We observe that GPT4, LLama2 and Alpaca rely on keyphrases an average of 27.3%, 35.9% and 36.3% more than Emolex words respectively, whereas EmoBERTa relies on keyphrases an average of 40.8% more than Emolex words. This is

also reflected in Table 4.

5 Qualitative Analysis

In this section, we provide details about what we observe in our analysis, in terms of what large language models get right, and what they get wrong. We also delve deeper into the comparison of the results of the experiments with LLMs and fine-tuned models.

5.1 Trigger identification

When it comes to trigger identification, we find that GPT4 is generally good at identifying the right triggers, except for when it comes to comments about specific experiences or entities. In these cases, it confuses emotion words for emotion triggers. An example of this is provided in Figure 2.

The distinction between the performance of GPT4 and the other models we evaluate is quite clear. We find that Llama2-chat struggles with identifying triggers even for simple sentences. Further, we observe that Llama2 makes spelling mistakes when it does identify the right triggers. Alpaca's performance is slightly worse than Llama2. Examples are demonstrated in Figure 2.

5.2 LLM's "attribution" of its own predictions

Here, we observe that the LLMs detect emotions with a significantly higher accuracy than transformer models. As shown in Figure 4, we find that LLMs struggle with identifying all emotions correctly if multiple emotions are present.

However, a drop in accuracy can be observed in trigger identification when the gold label emotions are not provided. We observe that GPT4 identifies triggers for emotions that it detects, even when the emotions themselves are incorrect. Further, we find that even when it does get the emotions right, it sometimes chooses triggers differently (and sometimes incorrectly) compared to the ones it chose when the same emotions were provided to it in the prompt. We find that Llama2 and Alpaca exhibit similar behavior. Examples are given in Figure 5.

5.3 Understanding feature importance: contrasting LLMs with transformer models

We find that large language models are able to detect emotions significantly more accurately than their traditionally fine-tuned counterparts. As demonstrated in Figure 5, we see that the LLMs that make correct prediction of emotion correctly identify keyphrases and EmoLex words (when present). We observe that this is not true with respect to EmoBERTa. Even though there is a very high correlation of SHAP values and keyphrases, it doesn't entail that EmoBERTa is able to detect the correct emotion. This indicates that paying attention to a single feature is not enough, and that a fundamental understanding of grammar and language may be necessary to perform emotion detection correctly.

6 Discussion and Conclusion

In this paper, we present EMOTRIGGER, a linguist-annotated dataset of emotion triggers (as extractive text spans), over three social media corpora with labeled emotions across different themes. We use this dataset to analyze what role emotion triggers play in emotion detection.

Overall, we believe this work provides evidence that with the exception of very large models like GPT-4 (few-shot), open-sourced ones like Llama2-chat and Alpaca do not have a good understanding of what triggers an emotion. The finding that salient features correlate substantially with keyphrases, rather than emotion triggers, means that models are better at picking up corpus-level topical cues rather than possessing a deep understanding of emotions *per se* as humans do. In Psychology, emotion is viewed as triggered by subjective evaluations (or appraisals) of particular events (Zhan et al., 2022; Moors et al., 2013); thus future work on more sophisticated emotional support open-source language models should address this flaw.

7 Limitations

In our work, we analyze what role emotion triggers play in emotion detection. While we believe the development and analysis of the EMOTRIGGER dataset is a step forward in this area of research, our study has a few limitations. First, our dataset is relatively small in size, owing to the labor intensive process of human annotation and the consideration of computational expenses of using the data with LLMs. Second, we run our study on a limited number of LLMs. This is also due to the consideration of computational resources. Finally, our study only deals with text that is in English; we leave

multilingual pursuits for future work.

Acknowledgements

This research is partially supported by National Science Foundation (NSF) grants IIS-2107524 and IIS-2107487. We thank Keziah Kaylyn Reina, Kathryn Kazanas and Karim Villaescusa F. for their work on the annotation of EMOTRIGGER. We also thank our reviewers for their insightful 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.

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

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

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.

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*.

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

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

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.

Pengcheng He, Xiaodong Liu, Jianfeng Chen, Weizhu Zhao, and Lihong Wang. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mahshid Hosseini and Cornelia Caragea. 2022. Calibrating student models for emotion-related tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9266–9278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2023a. Feature normalization and cartography-based demonstrations for prompt-based fine-tuning on emotion-related tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12881–12889.

Mahshid Hosseini and Cornelia Caragea. 2023b. Semi-supervised domain adaptation for emotion-related tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5402–5410, Toronto, Canada. Association for Computational Linguistics.

William James. 1884. What is an emotion? *Mind*, 9(34):188–205.

Gerard Christopher Yeo Zheng Jie and Desmond C Ong. 2023. A meta-analytic review of the associations between cognitive appraisals and emotions in cognitive appraisal theory.

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.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

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

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019a. DENS: A dataset for multi-class emotion analysis. 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 6293–6298, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert approach. 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)*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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

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

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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, 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.

Tiberiu Sosea, Hongli Zhan, Junyi Jessy Li, and Cornelia Caragea. 2023. Unsupervised extractive summarization of emotion triggers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9550–9569, Toronto, Canada. Association for Computational Linguistics.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

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

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.

A Mapping of HurricaneEmo emotions

We aggregate the classes present in the dataset such that the final dataset consists of tweets annotated for Plutchik's eight primary emotions— anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. We also include a none class. This aggregation is shown below:

- Anger: anger, annoyance, rage
- Fear: fear, apprehension, terror
- Sadness: sadness, grief, pensiveness
- Disgust: disgust, loathing, boredom
- Surprise: surprise, amazement, distraction
- Anticipation: anticipation, interest, vigilance
- Trust: trust, admiration, acceptance
- Joy: joy, serenity, ecstasy

B Annotation Instructions

Hello, and welcome to this annotation task! This is a task designed to understand the importance of different words in identifying emotions and their triggers. To complete it, we will be using Google Sheets. As a part of this task, you will be given a google sheet consisting of a sentence, each already annotated for several different emotions. You will complete the task of identifying emotion triggers, as stated below:

Identifying Emotion Triggers For this task, your assignment is to analyze the sentence and the annotated emotions, and identify the words/phrases that indicate the triggers of the emotions reflected by the gold label annotation. Triggers tend to be real-world events or concrete entities. You may identify either words, phrases, or a combination of both.

- If you disagree with the emotion identified by the gold label annotation, please enter "NA"
- If the gold label emotion is 'None' or 'Neutral', please enter "None"
- If you believe that there are no triggers for this specific sentence, please enter "No Trigger"
- If you believe that not all emotions identified by the gold label annotation, please enter "Partially complete example" and then add your answer in the form of (emotion: triggers)

Example 1. Where's her british accent? lol I love to dislike her videos **[Amusement]**
Alpaca - Old Prompt: I have already tried using NLTK library to perform sentiment analysis but I am unable to get the triggers for the given emotional phrases. Can you please help me with this?
Alpaca - New Prompt: britlish accents, videos

Example 2: It's terrible to say but I almost wish for a crisis so that I can get him in somewhere to find out what is going on. **[Anger]**
Alpaca - Old Prompt: adult children, burden
Alpaca - New Prompt: terrible, crises

Example 3: I will be on <PERSON> for a year infusions every 3 weeks. **[Anticipation]**
Alpaca - Old Prompt: CANCEL THAT!!!
Alpaca - New Prompt: for a year

Figure 3: Examples of how the Alpaca-specific prompt allows the model to interact with the text in a more coherent manner.

Some examples are given below for your reference.

> *Sentence*: "We were really blessed with very good doctors."
> *Emotions*: ['Joy']
> *Triggers*: very good doctors
>
> *Sentence*: "It is a very scary and hard subject to discuss and there were lots of tears on my end."
> *Emotions*: ['Fear','Sadness']
> *Triggers*: Fear: (very scary, subject), Sadness:(hard subject to discuss)
>
> *Sentence*: "Your adult children are a burden on you"
> *Emotions*: ['Anger']
> *Triggers*: adult children, burden

NOTE: Annotators were paid $15USD$ per hour for their time and efforts.

C Prompts

C.1 Prompt 1 (Known Emotion)

Given a piece of text and the emotions it has been annotated for, find the emotion triggers. Some examples are given below for your reference. ONLY CHOOSE PHRASES FROM THE TEXT.

> *Sentence*: "We were really blessed with very good doctors."
> *Emotions*: ['Joy']
> *Triggers*: very good doctors

Sentence: "It is a very scary and hard subject to discuss and there were lots of tears on my end."
Emotions: ['Fear','Sadness']
Triggers: Fear: (very scary, subject), Sadness:(hard subject to discuss)

Sentence: "Your adult children are a burden on you"
Emotions: ['Anger']
Triggers: adult children, burden

Sentence: $text$
Emotions: emo

C.2 Prompt 2 (Unknown Emotion)

Given a piece of text, find the emotions it has been annotated for, and the words that are most important for detecting the emotions. The emotions can be chosen from the list called EmoList below. Some examples are given below for your reference. ONLY CHOOSE PHRASES FROM THE TEXT. DO NOT SAY NONE.
EmoList = ['admiration', 'amusement', 'anger', 'annoyance', 'approval', 'caring', 'confusion', 'curiosity', 'desire', 'disappointment', 'disapproval', 'disgust', 'embarrassment', 'excitement', 'fear', 'gratitude', 'grief', 'joy', 'love', 'nervousness', 'optimism', 'pride', 'realization', 'relief', 'remorse', 'sadness', 'surprise']

Sentence: "We were really blessed with very good doctors."
Emotions: ['Joy']
words: really blessed

Sentence: "It is a very scary and hard subject to discuss and there were lots of tears on my end."
Emotions: ['Fear','Sadness']
words: Fear:(scary), Sadness: (tears, hard)

Sentence: "Your adult children are a burden on you"
Emotions: ['Anger']
words: adult children, burden

Sentence: $text$
Emotions:
Words:

C.3 Prompts For Alpaca

Prompt 1 (known emotion): Given a piece of text and the emotions it has been annotated for, find the emotion triggers. Some examples are given below for your reference. The format of the examples are as follows. Given a sentence, the emotions are expressed in a list after "Emotions': and their triggers are given after "Triggers:" The last sentence is the one you need to provide triggers for. ONLY CHOOSE PHRASES FROM THE TEXT IN THE LAST SENTENCE. DO NOT SAY NONE.

Sentence: "We were really blessed with very good doctors."
Emotions: ['Joy']
Triggers: very good doctors

Sentence: "It is a very scary and hard subject to discuss and there were lots of tears on my end."
Emotions: ['Fear','Sadness']
Triggers: Fear: (very scary, subject), Sadness:(hard subject to discuss)

Sentence: "Your adult children are a burden on you"
Emotions: ['Anger']
Triggers: adult children, burden

Sentence: $text$
Emotions: emo
Triggers:

Prompt 2 (unknown emotion): Given a piece of text, find the emotions it has been annotated for and the words that are most important for detecting the emotions . The emotions can be chosen from the list called EmoList below. Some examples are given below for your reference. The format of the examples are as follows. Given a sentence, the emotions are expressed in a list after "Emotions': and the words are given after "words:" The last sentence is the one you need to provide emotions and words for. ONLY CHOOSE PHRASES FROM THE TEXT IN THE LAST SENTENCE. DO NOT SAY NONE.

EmoList = ['admiration', 'amusement', 'anger', 'annoyance', 'approval', 'caring', 'confusion', 'curiosity', 'desire', 'disappointment', 'disapproval', 'disgust', 'embarrassment', 'excitement', 'fear', 'gratitude', 'grief', 'joy', 'love', 'nervousness', 'optimism', 'pride', 'realization', 'relief', 'remorse', 'sadness', 'surprise']

Example 1. My whole life has been a lie **[Surprise]**
GPT-4: Sadness
Llama2Chat: Fear
Alpaca: Fear
EmoBERTa: Fear

Example 2: The people of Puerto Rico are without water and power and the only mention of them from @realdonaldtrump is to kick them while they're down! **[Anger, Disgust]**
GPT-4: [Anger, Disgust]
Llama2Chat: [Anger]
Alpaca: [Disgust]
EmoBERTa: [Disgust]

Example 3: My doctors have told me that this has shown to be one of the most effective treatments against lung cancer. **[Trust]**
GPT-4: No emotion is reflected in this text
Llama2Chat: [Trust]
Alpaca: No emotion is expressed.
EmoBERTa: [Joy]

Example 4: We feel like the doctors aren't telling us what needs to be told. **[Fear]**
GPT-4: Anticipation
Llama2Chat: Anticipation, Sadness
Alpaca: Fear, Sadness
EmoBERTa: Anticipation, Fear

Figure 4: Examples of emotion detection. The gold label emotions are indicated in square-brackets.

Sentence: "We were really blessed with very good doctors."
Emotions: ['Joy']
words: really blessed

Sentence: "It is a very scary and hard subject to discuss and there were lots of tears on my end."
Emotions: ['Fear','Sadness']
words: Fear:(scary), Sadness: (tears, hard)

Sentence: "Your adult children are a burden on you"
Emotions: ['Anger']
words: adult children, burden

Sentence: text
Emotions:
Words:

D Licensing

EMOTRIGGER builds on existing datasets: HurricaneEmo, CancerEmo, and GoEmotions. We adhere with the licensing of these original datasets, and will release our annotations with the Apache License (as specified by GoEmotions). We will not redistribute these existing datasets.

Example 1. At home with a viral infection right now! *Praying* for death and to the god of paracetamol. **[Sadness]**
GPT-4: Praying, Death [Sadness]
Llama2Chat: Viral, Death [Fear]
Alpaca: Death [Fear]
EmoBERTa-SHAP: viral, infection, death [Fear]

Example 2: She has a spot on her liver in addition to her lung that they think is cancer as well **[Anticipation, Fear]**
GPT-4: Cancer [Fear]
Llama2Chat: Lung, Cancer [Anticipation]
Alpaca: Spot, Cancer [Fear]
EmoBERTa-SHAP: spot, liver, lung, cancer [Anticipation]

Example 3: Do you think an old, ugly, fat, stupid person like him should be eating such unhealthy meals? **[Anger, Disgust]**
GPT-4: old, ugly, fat, stupid [Anger]
Llama2Chat: ugly, fat, unhealthy [Disgust]
Alpaca: old, ugly, fat, stupid [Disgust]
EmoBERTa-SHAP: old, ugly, stupid [Anger]

Figure 5: Tabular comparison between features taken into consideration by various models for emotion detection. Gold label emotions are presented in square brackets. Triggers are highlighted, keyphrases are underlined and EmoLex words are italicized.

E Hyperparameters

- GPT4: temperature = 1.0 (default)
- Llama2Chat: temperature = 0.9 (recommended)
- Alpaca: temperature = 0.9 (recommended)
- EmoBERTa: learning rate = $2e-5$, no. of epochs = 5, MaxLen= 200, Train batch size = 32, Validation batch size = 16. Hyperparameters tuned on the validation set.

Dataset	BERT	DistilBERT	RoBERTa	DeBERTa	EmoBERTa
HurricaneEmo	0.478	0.462	0.399	0.381	**0.483**
CancerEmo	0.351	0.333	0.327	0.325	**0.378**
GoEmotions	0.311	0.302	0.308	0.308	**0.341**

Table 5: Macro F1 score for finetuned transformers across different datasets

Emotion	HurricaneEmo				CancerEmo				GoEmotions			
	GPT4	Llama2	Alpaca	EmoB.	GPT4	Llama2	Alpaca	EmoB.	GPT4	Llama2	Alpaca	EmoB.
Anger	0.86	0.82	0.72	0.47	0.89	0.80	0.78	0.35	0.87	0.80	0.78	0.32
Anticipation	0.85	0.78	0.76	0.38	0.82	0.71	0.69	0.29	-	-	-	-
Joy	0.91	0.71	0.67	0.50	0.93	0.87	0.73	0.37	0.93	0.87	0.70	0.35
Trust	0.87	0.74	0.73	0.39	0.90	0.80	0.68	0.31	-	-	-	-
Fear	0.95	0.78	0.68	0.48	0.96	0.80	0.80	0.39	0.92	0.74	0.70	0.33
Surprise	0.92	0.74	0.70	0.48	0.88	0.81	0.76	0.34	0.89	0.82	0.70	0.37
Sadness	0.88	0.76	0.69	0.46	0.88	0.71	0.69	0.31	0.88	0.71	0.71	0.32
Disgust	0.86	0.75	0.67	0.40	0.89	0.71	0.70	0.31	0.89	0.77	0.69	0.29

Table 6: Emotion prediction evaluation of GPT4, Llama2Chat, Alpaca, and EmoBERTa across different datasets using F1 score. Note: GoEmotions uses Ekman's emotions.

Model	HurricaneEmo			CancerEmo			GoEmotions		
	F1	ExactM	PartialM	F1	ExactM	PartialM	F1	ExactM	PartialM
GPT4 - Trigger	0.50	0.20	0.67	0.51	0.27	0.71	0.51	0.21	0.69
Llama2 - Trigger	0.24	0.04	0.25	0.21	0.07	0.18	0.18	0.06	0.25
Alpaca - Trigger	0.14	0.02	0.13	0.11	0.04	0.11	0.12	0.04	0.11

Table 7: Zero Shot: Macro F1, exact match, and partial match scores to assess the overlap between salient words and annotated triggers.

Emotion	HurricaneEmo			CancerEmo			GoEmotions		
	F1	ExactM	PartialM	F1	ExactM	PartialM	F1	ExactM	PartialM
Anger	0.66	0.36	0.91	0.73	0.40	0.92	0.69	0.34	0.94
Anticipation	0.63	0.91	0.69	0.41	0.39	0.91	-	-	-
Joy	0.74	0.40	0.89	0.70	0.38	0.93	0.70	0.37	0.93
Trust	0.71	0.38	0.90	0.74	0.36	0.90	-	-	-
Fear	0.69	0.38	0.89	0.76	0.38	0.90	0.66	0.34	0.94
Surprise	0.72	0.34	0.90	0.68	0.40	0.89	0.72	0.33	0.89
Sadness	0.68	0.37	0.91	0.68	0.41	0.91	0.72	0.31	0.91
Disgust	0.64	0.39	0.90	0.69	0.36	0.91	0.69	0.33	0.88

Table 8: Per-emotion results of trigger identification (given emotions) performed by GPT4.

Emotion	HurricaneEmo			CancerEmo			GoEmotions		
	F1	ExactM	PartialM	F1	ExactM	PartialM	F1	ExactM	PartialM
Anger	0.36	0.13	0.31	0.33	0.10	0.32	0.29	0.08	0.34
Anticipation	0.06	0.31	0.29	0.11	0.29	0.31	-	-	-
Joy	0.34	0.12	0.36	0.30	0.08	0.30	0.30	0.07	0.33
Trust	0.31	0.08	0.30	0.34	0.10	0.30	-	-	-
Fear	0.29	0.08	0.36	0.30	0.08	0.28	0.28	0.09	0.31
Surprise	0.32	0.09	0.30	0.28	0.10	0.29	0.32	0.10	0.29
Sadness	0.28	0.11	0.30	0.29	0.11	0.31	0.30	0.11	0.31
Disgust	0.34	0.12	0.30	0.29	0.09	0.30	0.28	0.09	0.29

Table 9: Per-emotion results of trigger identification (given emotions) performed by Llama2Chat.

Emotion	HurricaneEmo			CancerEmo			GoEmotions		
	F1	**ExactM**	**PartialM**	**F1**	**ExactM**	**PartialM**	**F1**	**ExactM**	**PartialM**
Anger	0.26	0.11	0.28	0.23	0.11	0.27	0.29	0.07	0.28
Anticipation	0.25	0.06	0.19	0.20	0.06	0.28	-	-	-
Joy	0.23	0.10	0.29	0.27	0.06	0.26	0.29	0.05	0.26
Trust	0.20	0.09	0.27	0.24	0.08	0.26	-	-	-
Fear	0.26	0.09	0.28	0.20	0.07	0.24	0.28	0.09	0.27
Surprise	0.20	0.07	0.27	0.27	0.08	0.27	0.29	0.09	0.27
Sadness	0.27	0.08	0.28	0.29	0.11	0.29	0.29	0.11	0.28
Disgust	0.24	0.07	0.27	0.29	0.09	0.26	0.28	0.09	0.26

Table 10: Per-emotion results of trigger identification (given emotions) performed by Alpaca.