Hard Emotion Test Evaluation Sets for Language Models

Tiberiu Sosea **Cornelia Caragea**
Computer Science
University of Illinois Chicago
tsosea2@uic.edu cornelia@uic.edu

Abstract

Language models perform well on emotion datasets but it remains unclear whether these models indeed understand emotions expressed in text or simply exploit supperficial lexical cues (e.g., emotion words). In this paper, we present two novel test evaluation sets sourced from two existing datasets that allow us to evaluate whether language models make real inferential decisions for emotion detection or not. Our human-annotated test sets are created by iteratively rephrasing input texts to gradually remove explicit emotion cues (while preserving the semantic similarity and the emotions) until a strong baseline BERT model yields incorrect predictions. Using our new test sets, we carry out a comprehensive analysis into the capabilities of small and large language models to predict emotions. Our analysis reveals that all models struggle to correctly predict emotions when emotion lexical cues become scarcer and scarcer, but large language models perform better than small pre-trained language models and push the performance by 14% over the 5% BERT baseline. We make our evaluation test sets and code publicly available.[1]

1 Introduction

Emotions are an integral element of the human nature, often affecting our everyday life and activities. Detecting emotions expressed in language has many applications, including the detection of harmful behavior on social media (Mohammad, 2012; Wang et al., 2012; Mohammad and Kiritchenko, 2015; Volkova and Bachrach, 2016; Abdul-Mageed and Ungar, 2017; Demszky et al., 2020), understanding the impact of public policy making on the society (e.g., in large-scale crises and pandemics) (Sosea et al., 2022; Desai et al., 2020; Beck et al., 2021; Kabir and Madria, 2021; Adikari et al., 2021; Choudrie et al., 2021; Scarpina, 2020; Calbi et al., 2021; Halse et al., 2016), designing empathetic conversational agents (Buechel et al., 2018; Hosseini and Caragea, 2021a,b), or enabling emotional intelligence skills to computers (Picard, 1997). Therefore, an increasing number of datasets for emotion detection have been made available in recent years (Singh et al., 2024; Sabour et al., 2024; Zhan et al., 2022; Desai et al., 2020; Sosea and Caragea, 2020; Demszky et al., 2020; Abdul-Mageed and Ungar, 2017; Volkova and Bachrach, 2016; Mohammad and Kiritchenko, 2015), among which GoEmotions (Demszky et al., 2020) and CancerEmo (Sosea and Caragea, 2020) are two large human annotated datasets with multiple annotations per sample for quality assurance. GoEmotions (Demszky et al., 2020) is created from Reddit comments and enables fine-grained emotion classification (with 27 emotions in total), whereas CancerEmo (Sosea and Caragea, 2020) is created from a health forum with each sample classified into one of the Plutchik-8 emotions (Plutchik, 1980).

Although these datasets are instrumental in the progress of emotion detection, oftentimes emotions are expressed in a very explicit way in these datasets. For example, in the sentence *I've never been this sad in my life!* extracted from GoEmotions, the words "this sad" are highly indicative lexical cues of the emotion "sadness". Deep learning models are known to exploit idiosyncrasies from the data allowing them to imitate desired behavior such as pattern matching or negation (Papernot et al., 2017; Gururangan et al., 2018). In fact, Sosea and Caragea (2020) found that lexical cues such as emotion words (Mohammad and Turney, 2013) (e.g., the words "this sad" in the example above) are strong signals for pre-trained language models like BERT (Devlin et al., 2019). Thus, it remains unclear whether language models indeed understand emotions expressed in text or rely blindly on surface-level lexical cues and lack a deep understanding of the expressed emotions. To

[1] https://github.com/tsosea2/HardEmotionDatasets

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 7945–7959

Label	Text	Confidence
Sadness	Wife left me and I am just broken. relapsed tonight from 7 years clean and just not seeing a point anymore	0.78
R1	After being clean for 7 years, my wife leaving has caused me to relapse	0.47
R2	After being clean for 7 years, my wife abandoning me has led to be to stop being clean	0.42
R3	After being clean for 7 years, my wife not wanting to be with me has led to be to stop being clean	0.48
R4	After being clean for 7 years, my wife not wanting to be with me has led me to pick up my old habits	0.23⋆
Gratitude	Easy money, thank you [NAME]	0.99
R1	Easy money, shout out to you [NAME]	0.21⋆

Table 1: Examples in our model-annotator feedback loop annotation process. Given an initial text labeled correctly by the model, our annotators carry out several rounds of rephrasings (R) until a rephrased example manages to yield an incorrect prediction by the model. We show the model confidence on each rephrased text and indicate by ⋆ a successful rephrasing (or in other words, an incorrect prediction), i.e., when the confidence of the gold label is not the highest confidence.

this end, we present two novel and challenging test evaluation sets sourced from the existing datasets, GoEmotions (Demszky et al., 2020) and CancerEmo (Sosea and Caragea, 2020), that allow us to evaluate the capabilities of language models at understanding emotions—whether language models can make subtle inferential decisions for emotion detection when lexical cues or emotion words are less frequent in the data.

To construct these challenging test evaluation sets, we utilize a model-annotator feedback loop in which human annotators are instructed to iteratively rephrase examples from the test sets of GoEmotions and CancerEmo that are correctly predicted by a BERT model (Devlin et al., 2019) aiming to make the BERT model return an incorrect prediction on these examples. In rephrasing a text, we ask our annotators to remove potential spurious correlations and "de-explicitize" emotion words to the entent possible, i.e., use more implicit expressions of emotions and less emotion words (while preserving the original emotions and the overall semantic meaning of the text). Table 1 shows two test examples and their rephrasings produced by our humman annotators. The first example *Wife left me and I am just broken. relapsed tonight from 7 years clean and just not seeing a point anymore* is correctly predicted by BERT with the sadness emotion with 0.78 confidence. We observe that the 4^{th} rephrasing attempt eliminates lexical cues such as *left*, *broken*, and *relapse*, and yields an incorrect prediction (making the example more challenging for BERT). Similarly, in *Easy money, thank you...*, the rephrasing of *thank you*, a clear lexical cue indicative of gratitude leads to an incorrect prediction by BERT.

Using our new challenging test sets, we establish strong baselines and evaluate increasingly powerful language models—small pre-trained language models designed to address the limitations of BERT, i.e., RoBERTa (Liu et al., 2019b), XLNet (Yang et al., 2019), and eMLM BERT (Sosea and Caragea, 2021), and large language models, OPT-IML (Iyer et al., 2022) and ChatGPT to evaluate their capabilities on examples where a vanilla BERT performs poorly. We observe that all models struggle to correctly predict emotions when emotion lexical cues become scarcer and scarcer, and even powerful large language models incur significant performance degradations and are unable to obtain good performance. Notably, OPT-IML (LoRA fine-tuned) obtains 0.65 F1 on original GoEmotions test set but scores only 0.15 F1 on our rephrased GoEmotions test set, which is a significant performance gap demonstrating the difficulty of our dataset. Still, large language models perform better than fine-tuned small language models and push the performance by 14% over the 5% BERT baseline.

Our contributions are as follows: **1)** We introduce two new challenging emotion detection test sets: GoEmotions.v2 and CancerEmo.v2, using a model-annotator feedback loop; **2)** We carry out extensive experiments to analyze the capabilities of various language models on our challenging test sets; **3)** We show that using a small amount of rephrased examples during training or in the prompt for LLMs significantly boosts the capabilities of both small and large language models.

2 Related Work

Emotion Detection Emotion detection has been studied extensively (Singh et al., 2024; Hosseini and Caragea, 2023a,b; Sosea et al., 2023; Maratos et al., 2023; Hosseini and Caragea, 2022; Cambria et al., 2017; Poria et al., 2018; Cambria et al., 2020; Stappen et al., 2021; Cambria et al., 2013) with applications in music (Strapparava et al., 2012), social networks (Mohammad, 2012; Islam et al., 2019), online news (Bao et al., 2009), health communities (Sosea and Caragea, 2021; Khanpour and Caragea, 2018; Khanpour et al., 2018; Biyani et al., 2014a,b), and literature (Liu et al., 2019a). All these domains can be examined with the help of large curated datasets.

Using these datasets, many methods have been developed for emotion detection. In the past, most approaches used feature-based methods, which usually leveraged hand-crafted lexicons, such as EmoLex (Mohammad and Turney, 2013) or the Valence Arousal Lexicon (Mohammad, 2018). However, due to the recent advancements in deep learning as well as large pre-trained language models, all state-of-the-art approaches (Chen et al., 2023; Shah et al., 2023; Suresh and Ong, 2021; Sosea and Caragea, 2021; Desai et al., 2020; Sosea and Caragea, 2020; Demszky et al., 2020) employ language model-based (Devlin et al., 2019) classifiers. In this work, we propose two challenging emotion test evaluation datasets aimed at exploiting the weaknesses of language models that can serve as a test bed to facilitate improvements in language models' capabilities.

Language Models Pre-trained language models have been used in a *pre-train then fine-tune* manner where a language model is pretrained on a large unlabeled corpus then adapted to a target task by fine-tuning (Devlin et al., 2019). However, it was recently observed that scaling models to 100B+ parameters leads to capabilities of few-shot learning (Brown et al., 2020) by way of in-context learning. Therefore, these models have excelled at effectively leveraging very few examples to solve any NLP task (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; Wang et al., 2023) using prompting. Prompting is achieved by altering the input-output space of a model depending on the task at hand to effectively leverage the knowledge of the model. Various methods were proposed to improve the prompting mechanism ranging from the structure and quality of the prompt template (Shin et al., 2020; Gao et al., 2020; Schick and Schütze, 2020; Jiang et al., 2020) to chain-of-thought prompting (Wei et al., 2022) or optimizing the few-shot example ordering (Lu et al., 2021). To overcome the difficulties of prompt engineering, instruction tuning has been proposed as a method to improve the perfomance by fine-tuning LLMs on a wide variety of tasks, ranging from chat and summarization to text classification, sentiment analysis and entity extraction. Popular instruction-tuned LLMs such as the open-source LLaMa-based OPT-IML (Iyer et al., 2022) and ChatGPT have started to receive attention in emotion detection as well (Tu et al., 2023; Singh et al., 2023; Kang and Cho, 2024; Zhang et al., 2023; Lei et al., 2023). In this paper, we use these models to benchmark their capabilities on our challenging emotion datasets and analyze how well they perform in zero or few-shot setups.

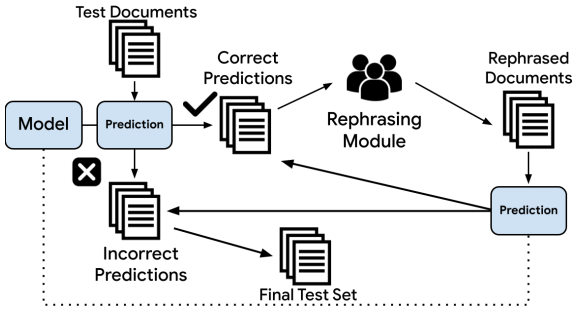

Figure 1: Overview of our annotation process.

3 Emotion Test Evaluation Datasets Construction

We now present the construction of our new test evaluation datasets sourced from GoEmotions (Demszky et al., 2020) and CancerEmo (Sosea and Caragea, 2020), designed to probe the capabilities of language models in understanding emotions. We show an overview of our annotation process in Figure 1 and describe it in more details below.

Setup Let $T = \{t_1, t_2, ..., t_n\}$ be the test set of an emotion dataset D (i.e., GoEmotions or CancerEmo in our case). In constructing our new test sets, we utilize a model-annotator feedback loop, with the BERT language model (Devlin et al., 2019) as the backbone model, where human annotators iteratively rephrase a test example to make it more difficult for BERT to return a correct prediction. We chose BERT as our baseline language model since BERT was established as a strong baseline in the original papers that introduced the datasets GoEmotions (Demszky et al., 2020) and CancerEmo (Sosea and Caragea, 2020), and, at the same time, it allows us to evaluate increasingly powerful language models—from small pre-trained language models that are trained in a more robust way compared with BERT to Large Language Models that often reach human performance (Chiang and Lee, 2023; Duong and Solomon, 2024).

Formally, we first train a BERT model M on the training set of D, and then use M to generate predictions on the test set T of D. Let $T^{correct} \subset T$ be the subset of examples from T where M makes **correct** predictions and let $T^{incorrect} = T \backslash T^{correct}$ be the subset of examples from T incorrectly predicted by M. We are mainly interested in rephrasing ex-

amples from $T^{correct}$ since the rest of the examples incorrectly predicted by M are already challenging for M, and hence, we do not consider them for human rephrasing. Our annotation goal is to generate a rephrased version t^R of $t \in T^{correct}$ such that: **1.** The semantic meaning of t^R does not diverge from t (i.e., they remain semantically similar), and **2.** t^R expresses the same emotion as t. In rephrasing a text, we asked our annotators to remove spurious correlations and "de-explicitize" emotion words to the entent possible, i.e., use more implicit expressions of emotions and less emotion words.

Annotation Process We hired five undergraduate students with expertize in natural language processing and linguistics and asked them to iteratively rephrase test examples that were correctly predicted by BERT model M. Our annotators are shown an example t and its gold label e. The annotation interface provides a textbox in which the annotators can type a candidate rephrasing t^c. After the rephrasing is typed in, we run inference on t^c using model M. If M incorrectly predicts t^c (i.e., the model prediction is no longer e), we consider $t^R = t^c$ as our successful rephrasing for t and continue with the rephrasing of the next test example in $T^{correct}$. If M continues to predict t^c in class e, we indicate that the rephrasing was unsuccessful and iterate the process (i.e., ask for additional rephrasings). We also show the confidence of M in class e on the rephrased text as a feedback signal to inform the annotators how well the generated text manages to change the confidence of the model. If our annotators are unable to create a successful rephrasing in *at most four trials*, they continue with the next test example. At any point, the annotators have the possibility to skip a particular example.

Each example $t \in T^{correct}$ is annotated by one of the five annotators. We provide more details on the annotator qualification and training in Appendix A. In total, the process took 3 months to complete.

Quality Assessment To assess the quality of our annotations we sampled 175 examples from each dataset and verified using an external set of annotators that the original and rephrased texts are semantically similar and they convey the same emotion. We computed the inter-annotator agreement between the external annotators and our student annotators and obtained a Krippendorff alpha of 0.75, indicating strong agreement.

	CancerEmo	GoEmotions
Total	2,254	3,067
Correct	1,603	1,745
Incorrect	651	1,322

Table 2: Total number of examples in the two test datasets and the number of examples correctly and incorrectly predicted by the baseline BERT model.

Model-in-the-Loop (BERT) Performance Table 2 shows the number of test examples correctly and incorrectly predicted by BERT on each dataset, GoEmotions and CancerEmo. On GoEmotions, BERT obtains an F1 score of 0.59 and correctly classifies 1,745 examples from the test set, while on CancerEmo, BERT obtains an F1 score of 0.73 and correctly classifies 1,603 examples. These examples constitute our $T^{correct}$ sets which we aimed to rephrase with our student annotators.

Training Subsets Additionally, we also used human annotators to rephrase 500 examples from the training set of each dataset to explore various ways of using the rephrasings during training to improve the model robustness. We call our datasets GoEmotions.v2 and CancerEmo.v2.

4 Datasets Characteristics

The annotation process of our test evaluation datasets, GoEmotions.v2 and CancerEmo.v2 produced 6,104 total rephrasings, yielding an average of 1.82 rephrasings per sample. Our test sets differentiate themselves from existing datasets in that they are created to be challenging by nature and can be viewed as a tool to understand the weaknesses of language models when emotions are expressed in more implicit ways. We also contrast our datasets with their backtranslated versions since backtranslation (Tiedemann and Scherrer, 2017) is well-known to introduce diversity in input text while maintaining similar semantic meaning. Backtranslation rephrases a text by translating it to a foreign language and back to English. We aim to understand how our manual rephrasings compare to rephrasings produced by this type of automatic approaches. To this end, we propose to analyze two backtranslation methods for obtaining rephrasings for $T^{correct}$. We leverage the OPUS-MT model (Tiedemann, 2012) to translate our text to German and back to English using top-50 sampling decoding and **1)** A softmax temperature of 1 (BT-1) and **2)** A softmax temperature of 10 (BT-10). Note that a softmax temperature of 10

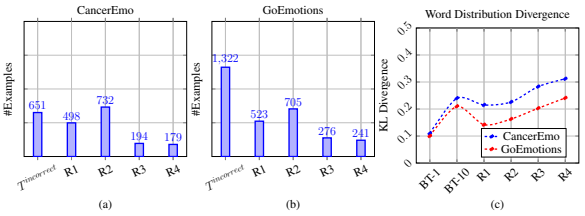

Figure 2: The number of examples at various rephrasings and on the $T^{incorrect}$ set for CancerEmo (a) and GoEmotions (b), respectively. (c) shows the average KL divergence between the word distribution of original examples and their rephrasings for both datasets.

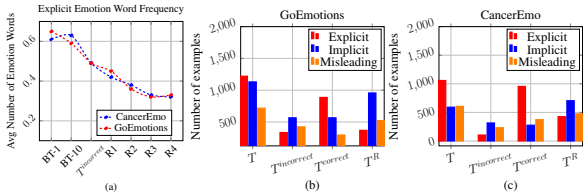

Figure 3: (a) Average amount of explicit emotion words for both datasets. (b) and (c) Distribution of explicit, implicit, and misleading examples in GoEmotions and CancerEmo, respectively.

introduces more variability in model predictions. We present characteristics of our datasets below.

Number of Rephrasings We show in Figures 2(a) and 2(b) the number of examples originally incorrectly predicted by BERT $T^{incorrect}$ and the number of examples from $T^{correct}$ that require from 1 to 4 rephrasings on CancerEmo and GoEmotions, respectively. We observe that 523 examples need a single rephrasing to reach an incorrect prediction by the model for GoEmotions and 498 for CancerEmo. Additionally, in total there are 93 examples in GoEmotions and 61 in CancerEmo where the annotators were unable to produce a successful rephrasing in four or less attempts.

Vocabulary Divergence We show in Figure 2(c) the differences in the word distribution of original examples in $T^{correct}$, their rephrasings T^R by our annotators, and the backtranslated versions of $T^{correct}$. Concretely, the values in the figure measure the KL divergence between original samples ($T^{correct}$) and their n^{th} rephrasing and the KL divergence between original samples and backtranslated versions of $T^{correct}$. We observe that our annotators tend to introduce more diversity the more attempts are needed to make an incorrect prediction. Notably, the KL divergence between the distribution of the 4^{th} rephrasings and the original samples (R4) is twice as large as that between the 1^{st} and the original (R1). Interestingly, we observe that BT-1 has a similar word distribution to $T^{correct}$, indicating that backtranslation tends to leverage the same words in their outputs. We see that raising the temperature of the softmax introduces significantly more diversity: the KL divergence increases from 0.1 in BT-1 to 0.24 in BT-10 on CancerEmo and from 0.1 to 0.21 on GoEmotions.

Explicit Emotion Word Frequency Let an explicit emotion word be any emotion word in the input text that is associated with the same emotion as the emotion label of the text. For example, in *I am happy* annotated with the emotion joy, *happy* is an explicit emotion word since it is associated with the emotion joy. We also denote a text that contains such words by **explicit** text. To obtain the labels of individual words we use the EmoLex (Mohammad and Turney, 2013) emotion-word association lexicon, which associates words with one of the Plutchik-8 basic emotions (Plutchik, 1980). We show in Figure 3(a) how the average amount of explicit emotion words changes across the rephrasings of $T^{correct}$ from R1 to R4. We observe a steady decrease in frequency on both datasets, indicating that the challenging examples created by our annotators contain fewer explicit emotion words with more rephrasing attempts. Interestingly, the average number of explicit emotion words in the $T^{incorrect}$ set of originally challenging examples is 0.59 on CancerEmo and 0.62 on GoEmotions, is similar to that of R1 and is significantly higher than that of R4 where the ratio of explicit words is 0.32 on CancerEmo and 0.33 on GoEmotions. Additionally, we can see that both BT-1 and BT-10 retain a large fraction of explicit emotion words after backtranslating the text.

Lexical Cues Besides the **explicit** text category defined above, we introduce here two additional categories: implicit and misleading. We define **implicit** text as any text where emotion lexical cues are absent. For example, in *Not sure what will happen to me when I get home*, no lexical information hints towards the *fear* emotion, hence this example is implicit. We define **misleading** text such as *amazing, I love when the rules change just before the deadline* any text that contains emotion-specific cues which are indicative of an emotion but that emotion is different than the gold annotated label. Here, *amazing* and *love* hint towards joy, however the conveyed emotion is *anger*.

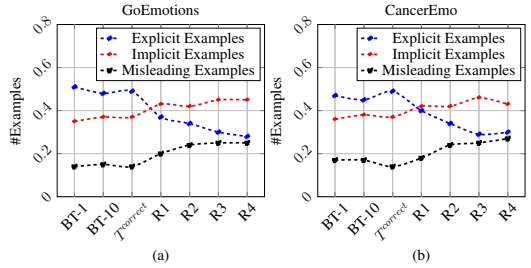

Figure 4: Ratio of explicit, implicit and misleading examples on variations of $T^{correct}$ on GoEmotions (a) and on CancerEmo (b).

We show in Figures 3(b) and 3(c) the frequency of explicit, implicit, and misleading examples in various subsets of the test data: T, $T^{correct}$, $T^{incorrect}$, and T^{R} for GoEmotions and CancerEmo, respectively. Interestingly, we see a clear pattern. The original correctly predicted examples $T^{correct}$ contain mostly explicit examples, which validates our assumption that these types of examples are easy to classify. Specifically, these easy examples can be classified by simply learning spurious correlations between the data and the labels (Wang et al., 2021). Notably, the $T^{correct}$ set contains 886 explicit examples in GoEmotions and 954 in CancerEmo, twice as many as in the set of incorrect predictions $T^{incorrect}$. $T^{incorrect}$ also predominantely contains implicit or misleading examples. Comparing $T^{correct}$ with T^{R} also offers valuable information: although the two subsets effectively contain the same examples in terms of semantic and emotional meaning, the rephrased texts contain 2.5 times as many implicit examples, 2 times as many misleading examples, and a much smaller amount of explicit examples. These results can therefore be viewed as strong evidence that de-explicitizing a piece of text, i.e., removing surface lexical cues that can artificially hint towards the emotion label or inserting misleading cues can pose significant challenges to language models.

Types of Examples at Each Rephrasing We show in Figures 4(a) and 4(b) the ratio of explicit, implicit, and misleading examples on $T^{correct}$, backtranslations of $T^{correct}$, and at different rephrasing attempts of $T^{correct}$. This result shows that the more rephrasings an example requires, the more likely it is that the rephrasing is an implicit or misleading example. Moreover, at the 4^{th} rephrasing, the ratio of misleading examples becomes close to that of explicit examples, denoting that the use of misleading lexical cues is the most effective way to make the model return an incorrect prediction. We also show in Appendix B the negation statistics and analysis in our rephrasings.

5 Language Models

In this section, we describe the language models that we use in our experiments to evaluate the difficulty of our new test datasets. We detail the experimental setup of our methods in Appendix C.

5.1 Small Language Models (SLMs)

Fine-tuning on the original training set We fine-tune the following small pre-trained language models on the original training set of each dataset, GoEmotions and CancerEmo: (a) **BERT** (Devlin et al., 2019) - pre-trained using the masked language modeling (MLM) objective; (b) **RoBERTa** (Liu et al., 2019b) - pre-trained in a more robust way than BERT, i.e., using dynamic masked language modeling where different words are masked in different epochs during training and is pre-trained on significantly more data than BERT; (c) **XLNet** (Yang et al., 2019) - pre-trained with a "Permutation Language Modeling" objective instead of MLM; and (d) **eMLM BERT** (Sosea and Caragea, 2021) - pre-trained with an emotion masked language modeling objective that assigns higher masking probabilities to emotion-relevant words during pre-training.

Fine-tuning on the original training set + rephrased challenging training examples For each dataset, GoEmotions and CancerEmo, in addition to the test evaluation datasets, we also created challenging rephrased examples for 500 original training examples. Thus, we fine-tuned BERT, RoBERTa, XLNet, and eMLM BERT on the combination of original training examples and the 500 rephrased challenging examples. However, instead of using a simple combination of original + challenging examples with a standard cross-entropy loss, we use a cross-entropy loss on the original examples and a KL-divergence loss on the rephrased examples as follows: Given a batch $B = \{(t_1, e_1), ..., (t_{|B|}, e_{|B|})\}$ of original examples, the cross entropy loss between the model predictions and the gold label is: $L_{CE} = \sum_{i=0}^{|B|} CE(M(t_i), e_i)$. For the subset of 500 training examples with rephrasings, we minimize the KL divergence between the outputs of the model on different rephrasings of an example. For each example t_i, we denote by $T^{R}(t_i) = \{t_1^{R}(t_i), t_2^{R}(t_i), ...\}$

Model \ Dataset	Original	BT-10	1 Rephrasing (1R)	2 Rephrasings (2R)	3 Rephrasings (3R)	4 Rephrasings (4R)
			GoEmotions.v2			
BERT	$0.59_{\pm.032}$	$0.55_{\pm.025}$	$0.49_{\pm.053}$	$0.23_{\pm.055}$	$0.11_{\pm.042}$	$0.05_{\pm.029}$
XLNet	$0.58_{\pm.033}$	$0.56_{\pm.026}$	$0.46_{\pm.061}$	$0.18_{\pm.043}$	$0.08_{\pm.027}$	$0.05_{\pm.022}$
RoBERTa	$0.58_{\pm.035}$	$0.55_{\pm.031}$	$0.46_{\pm.046}$	$0.22_{\pm.048}$	$0.08_{\pm.029}$	$0.06_{\pm.019}$
eMLM	$0.61_{\pm.029}$	$0.58_{\pm.039}$	$0.50_{\pm.038}$	$0.25_{\pm.036}$	$0.14_{\pm.028}$	$0.10_{\pm.029}$
BERT + KL	$0.61_{\pm.031}$	$0.60_{\pm.029}$	$0.52_{\pm.047}$	$0.26_{\pm.045}$	$0.16_{\pm.043}$	$0.08_{\pm.022}$
XLNet + KL	$0.57_{\pm.035}$	$0.55_{\pm.037}$	$0.47_{\pm.039}$	$0.19_{\pm.041}$	$0.10_{\pm.031}$	$0.08_{\pm.027}$
RoBERTa + KL	$0.59_{\pm.028}$	$0.55_{\pm.029}$	$0.48_{\pm.035}$	$0.23_{\pm.041}$	$0.13_{\pm.039}$	$0.11_{\pm.027}$
eMLM + KL	$0.64_{\pm.041}$	$0.61_{\pm.037}$	$0.52_{\pm.058}$	$0.27_{\pm.055}$	$0.17_{\pm.042}$	$0.13_{\pm.027}$
OPT-IML (ZS)	$0.61_{\pm.000}$	$0.59_{\pm.000}$	$0.51_{\pm.000}$	$0.35_{\pm.000}$	$0.21_{\pm.000}$	$0.14_{\pm.000}$
OPT-IML (FS)	$0.61_{\pm.000}$	$0.60_{\pm.000}$	$0.51_{\pm.000}$	$0.34_{\pm.000}$	$0.20_{\pm.000}$	$0.15_{\pm.000}$
OPT-IML (LoRA)	$\mathbf{0.65_{\pm.051}}$	$\mathbf{0.62_{\pm.048}}$	$0.51_{\pm.062}$	$\mathbf{0.43_{\pm.075}}$	$0.20_{\pm.053}$	$0.15_{\pm.032}$
OPT-IML (FS-R)	$0.62_{\pm.000}$	$0.61_{\pm.000}$	$0.52_{\pm.000}$	$0.41_{\pm.000}$	$\mathbf{0.23_{\pm.000}}$	$\mathbf{0.19_{\pm.000}}$
ChatGPT (ZS)	$0.63_{\pm.000}$	$0.59_{\pm.000}$	$\mathbf{0.55_{\pm.000}}$	$0.39_{\pm.000}$	$0.22_{\pm.000}$	$0.17_{\pm.000}$
ChatGPT (FS)	$0.61_{\pm.000}$	$0.59_{\pm.000}$	$0.54_{\pm.000}$	$0.40_{\pm.000}$	$0.20_{\pm.000}$	$0.18_{\pm.000}$
			CancerEmo.v2			
BERT	$0.73_{\pm.024}$	$0.71_{\pm.029}$	$0.62_{\pm.043}$	$0.45_{\pm.045}$	$0.15_{\pm.051}$	$0.09_{\pm.022}$
XLNet	$0.72_{\pm.027}$	$0.70_{\pm.031}$	$0.63_{\pm.052}$	$0.48_{\pm.048}$	$0.14_{\pm.045}$	$0.09_{\pm.043}$
RoBERTa	$0.75_{\pm.038}$	$0.71_{\pm.042}$	$0.62_{\pm.053}$	$0.46_{\pm.055}$	$0.16_{\pm.047}$	$0.11_{\pm.041}$
eMLM	$0.75_{\pm.028}$	$0.71_{\pm.033}$	$0.64_{\pm.035}$	$0.49_{\pm.056}$	$0.19_{\pm.059}$	$0.12_{\pm.042}$
BERT + KL	$0.74_{\pm.028}$	$0.72_{\pm.034}$	$0.63_{\pm.058}$	$0.49_{\pm.061}$	$0.17_{\pm.063}$	$0.15_{\pm.024}$
XLNet + KL	$0.72_{\pm.024}$	$0.71_{\pm.036}$	$0.64_{\pm.048}$	$0.48_{\pm.055}$	$0.16_{\pm.49}$	$0.15_{\pm.041}$
RoBERTa + KL	$0.75_{\pm.022}$	$0.72_{\pm.025}$	$0.63_{\pm.067}$	$0.48_{\pm.055}$	$0.19_{\pm.038}$	$0.15_{\pm.033}$
eMLM + KL	$0.77_{\pm.041}$	$0.74_{\pm.042}$	$\mathbf{0.66_{\pm.055}}$	$0.50_{\pm.058}$	$0.22_{\pm.047}$	$0.16_{\pm.044}$
OPT-IML (ZS)	$0.71_{\pm.000}$	$0.70_{\pm.000}$	$0.65_{\pm.000}$	$0.48_{\pm.000}$	$0.15_{\pm.000}$	$0.15_{\pm.000}$
OPT-IML (FS)	$0.75_{\pm.000}$	$0.74_{\pm.000}$	$0.65_{\pm.000}$	$0.45_{\pm.000}$	$0.21_{\pm.000}$	$0.17_{\pm.000}$
OPT-IML (LoRA)	$\mathbf{0.80_{\pm.041}}$	$\mathbf{0.76_{\pm.051}}$	$0.65_{\pm.066}$	$\mathbf{0.52_{\pm.059}}$	$0.23_{\pm.048}$	$0.14_{\pm.028}$
OPT-IML (FS-R)	$0.77_{\pm.000}$	$\mathbf{0.76_{\pm.000}}$	$0.64_{\pm.000}$	$0.49_{\pm.000}$	$\mathbf{0.25_{\pm.000}}$	$\mathbf{0.20_{\pm.000}}$
ChatGPT (ZS)	$0.73_{\pm.000}$	$0.73_{\pm.000}$	$0.61_{\pm.000}$	$0.48_{\pm.000}$	$0.22_{\pm.000}$	$0.16_{\pm.000}$
ChatGPT (FS)	$0.75_{\pm.000}$	$0.74_{\pm.000}$	$0.63_{\pm.000}$	$0.46_{\pm.000}$	$0.23_{\pm.000}$	$0.17_{\pm.000}$

Table 3: F1 score on the GoEmotions.v2 and CancerEmo.v2 stress tests. Blocks in order from top to bottom: **(1)** Small language models. **(2)** Small language models trained using the original training set + annotated rephrasings. **(3)** Open-source Large Language Model OPT-IML with its variants (with prompting and fine-tuning). **(4)** Closed-source Large Language Model ChatGPT (with prompting).

all the rephrasing attempts of the annotators. The size of this set for different input examples is variable and depends on how hard it is to produce a successful rephrasing to make the model return an incorrect prediction for that particular example. Given a batch B' of rephrased training examples, the pairwise KL divergence between the output of the model on every available rephrasing is:

$$L_{KL} = \sum_{i=0}^{|B'|} \sum_{j=0}^{|T^R(t_i)|} \sum_{k=j}^{|T^R(t_i)|} KL(M(t_j^R(t_i)), M(t_k^R(t_i))) \quad (1)$$

The final loss is the sum of the two independent losses: $L = L_{CE} + L_{KL}$.

5.2 Large Language Models (LLMs)

We benchmark an open-source large language model OPT-IML (Iyer et al., 2022) as well as a state-of-the-art closed-source model, ChatGPT.

Open-source LLM OPT-IML is instruction-tuned on 2,000 NLP tasks and achieves significant performance improvements over prior work in numerous evaluation benchmarks (Iyer et al., 2022). We explore two variants zero-shot (ZS) and few-shot (FS) using as many as 10 few-shot examples in the prompt (from the original training set) for all models. We detail our prompts for zero-shot and few-shot in Appendix D. Additionally, we train our OPT-IML using LoRA (Hu et al., 2021) on the entire original training set of each dataset. Finally, we integrate rephrasings of training examples into our few-shot prompts to boost the robustness and performance of our model. Specifically, for each example in the few-shot prompt, we indicate in the prompt that its rephrasings have the same meaning and hence express the same emotion. We detail our updated prompt in Appendix D and denote this modified few-shot approach by OPT-IML (FS-R).

Closed-source LLM We conduct a comparative evaluation using the closed-source ChatGPT model, where we explored both zero-shot (ZS) and few-shot (FS) using as many as 10 few-shot examples in the prompt (from the original training set) for all models. The prompts are the same as for OPT-IML.

6 Results and Observations

We show the results of the language models (SLMs and LLMs) in Table 3 in three settings: (a) on the original test sets of each dataset, GoEmotions and CancerEmo, (b) on the backtranslated test sets of each dataset, and (c) on our challenging test sets with various number of rephrasings. Specifically, the column denoted by *1 Rephrasing* (1R) indicates that the evaluation was performed on the union between the first rephrasings generated by our annotators on $T^{correct}$ (irrespective if the rephrasing is successful or not) and $T^{incorrect}$. Similarly, *2 Rephrasings* (2R) represents the test set composed of the union between the successful rephrasings of R1, the second rephrasings generated by our annotators (irrespective if the rephrasing is successful or not) and $T^{incorrect}$, and so on for 3R and 4R.

6.1 SLM Results

We show in the first block of Table 3 for each dataset the results using the small language models. We observe that BERT obtains an F1 of 59% and 73% on GoEmotions and CancerEmo but a low F1 score of 5% and 9%, respectively, across the fourth rephrasing (4R) of GoEmotions and CancerEmo, indicating that the rephrased examples pose significant challenges for the BERT model. We note that evaluating on examples that required more rephrasings leads to lower overall performance. For example, BERT obtains an F1 of 0.49 on the 1R test set and only 0.11 F1 on the 3R test set. RoBERTa and eMLM both ourperform BERT on the two datasets on the 4R test set. Critically, among the small language models, eMLM outperforms all approaches and obtains the best results. With an F1 score of 10% on the 4R test set, eMLM outperforms RoBERTa by 4% on GoEmotions. Moreover, eMLM obtains good results on the 2R and 3R test sets as well, improving upon BERT by an average of 3%. These results show the challenging nature of our test sets as well as that pretraining using the eMLM objective, which takes into consideration emotion information produces a better approach for emotion detection, which is more robust in the face of challenging examples. We also observe from the results on the backtranslated BT-10 test dataset that the performance drops only slightly compared to the original test set, indicating that backtranslation tends to leverage the same words in their translated outputs (as in the original input text).

Training with cross-entropy and KL divergence The second block in Table 3 shows the results of our language models trained using the original training set + 500 challenging training examples as described in Section 5.1. The results show that leveraging the rephrased training examples is extremely effective in improving the performance on both datasets. eMLM+KL outperforms eMLM by 3% on the 4R test datasets of GoEmotions and by 4% on CancerEmo. Interestingly, leveraging the KL-divergence loss term improves the performance on the original dataset as well. Specifically, eMLM+KL pushes the performance over eMLM by 3% on GoEmotions and by 2% on CancerEmo on the original test sets.

6.2 LLM Results

The LLM results are shown in the third and fourth blocks of Table 3 of each dataset. We see that even though LLMs such as ChatGPT have access to only zero or a few examples from the original training set, ChatGPT consistently outperforms the traditional language models on both datasets. Notably, on the 4R GoEmotions test dataset, ChatGPT (ZS) outperforms eMLM + KL by 5%. Additionally, ChatGPT (ZS) improves the performance over both OPT-IML (ZS) and OPT-IML (FS).

We observe that training OPT-IML using LoRA yields the best performance on the original test set, outperforming other approaches considerably. On GoEmotions original test set, OPT-IML (LoRA) outperforms BERT by 6% F1 and by 7% F1 on CancerEmo. However, we observe that the performance of OPT-IML (LoRA) is similar to few-shot (i.e., OPT-IML (FS)) on both GoEmotions and CancerEmo 4R test set. These results show that the proposed test evaluation datasets pose significant challenges to LLM generalization as well since the improvements on the original test set do not translate to the 4R test set. On the other hand, OPT-IML (FS-R) outperforms OPT-IML (FS) in most setups on both GoEmotions and CancerEmo, indicating that including challenging examples in the few-shot prompt boosts the performance of the model.

7 Conclusion

In this paper, we introduced GoEmotions.v2 and CancerEmo.v2, two novel test evaluation datasets which contain challenging examples in the context of emotion detection. We carried out a comprehensive analysis into the performance of deep learning methods including large language models to understand their weaknesses in emotion detection and found that these models frequently rely blindly on surface-level lexical cues and lack understanding of emotions. Our test sets can be viewed as a novel contribution not only for evaluating the capabilities of LLMs at understanding emotions, but also combatting data contamination (Golchin and Surdeanu, 2023; Shi et al., 2024; Roberts et al., 2023), since they have significantly different distributions than examples that appear in the current training sets of large language models. We make our annotated datasets available on GitHub and we hope that our work can spur research in this area.

Acknowledgements

This research is funded in part by the US National Science Foundation (NSF). Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

Limitations

In this paper, we proposed novel test evaluation datasets that probe the capabilities of small pretrained language models and large language models to evaluate their understanding of the emotions expressed in text. Using a human-and-model-in-the-loop annotation process, we created two datasets GoEmotions.v2 and CancerEmo.v2 having as a starting point the existing datasets, GoEmotions and CancerEmo. However, both these datasets are English datasets and cover only two limited domains: Reddit and Online Health Communities. In the future, we plan to study non-English languages, as well as other domains such as online news.

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.

S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, pages 699–704.

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. 2014a. 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.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014b. Identifying emotional and informational support in online health communities. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 827–836. ACL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. 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.

Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. 2017. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble

application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 105–114, New York, NY, USA. Association for Computing Machinery.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21.

Chih Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jyoti Choudrie, Shruti Patil, Ketan Kotecha, Nikhil Matta, and Ilias Pappas. 2021. Applying and understanding an advanced, novel deep learning approach: A covid 19, text based, emotions analysis study. *Information Systems Frontiers*, pages 1–35.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. *arXiv preprint arXiv:2004.14299*.

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.

Dat Duong and Benjamin D Solomon. 2024. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, 32(4):466–468.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Shane Errol Halse, Andrea H. Tapia, Anna Cinzia Squicciarini, and Cornelia Caragea. 2016. An emotional step towards automated trust detection in crisis social media. In *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016*. ISCRAM Association.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3713–3724. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021b. It takes two to empathize: One to seek and one to provide. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13018–13026. AAAI Press.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9266–9278. 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. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12881–12889. AAAI Press.

Mahshid Hosseini and Cornelia Caragea. 2023b. Semi-supervised domain adaptation for emotion-related tasks. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5402–5410. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Jumayel Islam, Robert E Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. 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 1355–1365.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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.

Yujin Kang and Yoon-Sik Cho. 2024. Improving contrastive learning in emotion recognition in conversation via data augmentation and decoupled neutral emotion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2194–2208, St. Julian's, Malta. Association for Computational Linguistics.

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. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 8099–8100. AAAI Press.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019a. Dens: A dataset for multi-class emotion analysis. *arXiv preprint arXiv:1910.11769*.

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 pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

George Maratos, Tiberiu Sosea, and Cornelia Caragea. 2023. Label smoothing for emotion detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16282–16283. AAAI Press.

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

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

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

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

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5986–6004. Association for Computational Linguistics.

Federica Scarpina. 2020. Detection and recognition of fearful facial expressions during the coronavirus disease (covid-19) pandemic in an italian sample: An online experiment. *Frontiers in Psychology*, 11:2252.

Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. 2023. Retrofitting light-weight language models for emotions using supervised contrastive learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3640–3654, Singapore. Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2023. Language models (mostly) do not consider emotion triggers when predicting emotion. *CoRR*, abs/2311.09602.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 603–614. Association for Computational Linguistics.

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, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6938–6947. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9550–9569. Association for Computational Linguistics.

Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5706–5707.

Carlo Strapparava, Rada Mihalcea, and Alberto Battocchi. 2012. A parallel corpus of music and lyrics annotated with emotions. In *LREC*, pages 2343–2346. Citeseer.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An empirical study on multiple knowledge from ChatGPT for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12160–12173, Singapore. Association for Computational Linguistics.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1567–1578, Berlin, Germany. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 587–592, Washington, DC, USA. IEEE Computer Society.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9436–9453. Association for Computational Linguistics.

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

A Annotator Qualification and Training

Prior to the annotation process, we carried out a $3-$ stage qualification process to train our annotators to produce difficult rephrasings. In each stage, we sampled 50 examples from $T^{correct}$ and asked our annotators to produce rephrasings for these examples. These rephrasings are then cross-checked by expert annotators that evaluate their quality. After cross-checking, we carried out a discussion with the annotators on various identified errors, provided feedback, and interactively produced adequate rephrasings. This training process proved extremely effective. While the 1^{st} stage produced an average annotator accuracy of 86%, the last stage indicated an accuracy of 96%.

B Negation statistics in rephrasings

A common way to rephrase an input example is to add a negation that does not change the meaning of the example. For instance, the text *I feel good* can be rephrased into *I don't feel bad at all*, maintaining the original meaning and conveyed emotion. To explore such cases in our annotation, we show in Figure 5 the percentage of input examples that contain negations. First, we observe that backtranslations and the $T^{correct}$ set of each dataset do not contain a significant number of negations. Additionally, we note that the number of negations rises with the first rephrasing, however, with three or four rephrasings, this number becomes constant. This may indicate that examples that are more challenging to rephrase require more sophisticated changes than simple negations (e.g, de-explicitization, creating a misleading example).

C Experimental Setup

We carry out all our experiments on a cluster of 4 Nvidia A5000 GPUs. We use the HuggingFace Transformers (Wolf et al., 2020) library for our model implementations and we will make the code for our methods and data available. For the closed source model, we use the ChatGPT API. We report the performance for emotion detection in terms of

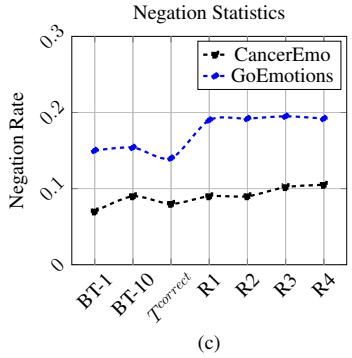

Figure 5: Negation Statistics.

macro F1 score. We run our approaches that require training and can produce variable results from run to run five times and report average values. These are traditional language models (i.e., BERT-like models) and the Lora-trained OPT-IML.

D Zero-shot and Few-shot Prompts

We show the design of our zero-shot (ZS), few-shot (FS), and few-shot adversarial (FS-ADV) prompts in Table 4.

t

Zero-shot Prompt (ZS)

Given the following text: (Instruction)
After being clean for 7 years, my wife not wanting to be with me has led me to pick up my old habits (Input Example)
Classify the text into one of the following categories depending on the emotion expressed by the text: (Instruction)
1. Anger 2. Joy 3. Sadness 4. Fear 5. Disgust 6. Anticipation 7. Trust 8. Surprise (Classes/Emotions)
model_completion

Few-shot Prompt (FS)

Given the text above: (Instruction)
Classify the text into one of the following categories depending on the emotion expressed by the text: (Instruction)
1. Anger 2. Joy 3. Sadness 4. Fear 5. Disgust 6. Anticipation 7. Trust 8. Surprise (Classes/Emotions)
I just cant stand seeing her like this. (Few-shot example #1 text)
3. Sadness (Few-shot example #1 label)
It is so awesome to hear news like yours! (Few-shot example #2 text)
8. Surprise (Few-shot example #2 label)
...
Guess I am more scared cause this has been very speedy. (Few-shot example #N text)
4. Fear (Few-shot example #N label)
After being clean for 7 years, my wife not wanting to be with me has led me to pick up my old habits (Input Example)
model_completion

Few-shot Adversarial Prompts (FS-ADV)

Given the text above: (Instruction)
Classify the text into one of the following categories depending on the emotion expressed by the text: (Instruction)
1. Anger 2. Joy 3. Sadness 4. Fear 5. Disgust 6. Anticipation 7. Trust 8. Surprise (Classes/Emotions)
A lot of amazing things happened after you left, however, the fact that we missed the mark by 2 meters really disappointed us. (Few-shot example #1 text)
The text expresses 3. Sadness (Few-shot example #1 label) because it has the same meaning as: (Instruction)
While good things happened after you left, unfortunately, we missed the mark which left us extremely disappointed. (Rephrasing of Few-shot Example #1)
which clearly expresses 3. Sadness (Few-shot example #1 label)
Was in a very happy mood for a while, but then exams started getting closer and closer. (Few-shot example #2 text)
The text expresses 4. Fear (Few-shot example #2 label) because it has the same meaning as: (Instruction)
Used to be happy but the exams coming up are terrifying. (Rephrasing of Few-shot Example #2)
which clearly expresses 4. Fear (Few-shot example #2 label)
...
After being clean for 7 years, my wife not wanting to be with me has led me to pick up my old habits (Input Example)
The text expresses:
model_completion

Table 4: Prompt designs for our LLM models.