Emotion-Infused Models for Explainable Psychological Stress Detection

Elsbeth Turcan[1] and **Smaranda Muresan**[1,2] and **Kathleen McKeown**[1]
Department of Computer Science, Columbia University[1]
Data Science Institute, Columbia University[2]
{eturcan, smara, kathy}@cs.columbia.edu

Abstract

The problem of detecting psychological stress in online posts, and more broadly, of detecting people in distress or in need of help, is a sensitive application for which the ability to interpret models is vital. Here, we present work exploring the use of a semantically related task, emotion detection, for equally competent but more explainable and human-like psychological stress detection as compared to a black-box model. In particular, we explore the use of multi-task learning as well as emotion-based language model fine-tuning. With our emotion-infused models, we see comparable results to state-of-the-art BERT. Our analysis of the words used for prediction show that our emotion-infused models mirror psychological components of stress.

1 Introduction

As crises have begun to multiply worldwide, including the COVID-19 pandemic and the resulting economic downturn, psychological stress has risen dramatically[1]. The problem of detecting psychological stress, and more broadly, of detecting people in distress and in need of help, is a sensitive application; therefore, the ability to interpret the results, in order to understand why, is vital. The consequences of blindly trusting a black-box model and mislabeling users' stress levels could be serious in a deployed application such as a therapeutic chatbot, where some users may not receive the immediate help they need. Furthermore, models that make decisions based on psychology theory about factors that impact stress will be easier for humans to understand, and their mistakes will be more obvious. Researchers have recently begun to study psychological stress, but in this work, we propose a new focus on examining the information our models use to make decisions and finding ways to incorporate psychological factors, like emotion, into them.

To approach the problem of stress detection, which has much less labeled data than many popular classification tasks, we first note that stress has been shown to interact with emotion (Lazarus, 2006; Thoern et al., 2016; Levenson, 2019), a task that has far more publicly available labeled data. For example, individuals who are stressed are likely to express emotions such as fear, sadness, or anger and unlikely to express emotions such as happiness.

Traditional multi-task learning would normally be helpful in this situation, but there are no currently available datasets labeled with both stress and emotion. Even if there were, it would be beneficial to incorporate external information without re-labeling new datasets for each new combination of useful tasks. Here, we present work exploring how to use semantically related tasks–here, emotion detection–to create *emotion-infused* models capable of equally competent, but explainable, psychological stress detection as compared to a black-box model. In particular, we explore the use of multi-task learning as well as emotion-infused language model fine-tuning, two existing frameworks which we examine through the lens of interpetability. Our code for this work is available at github.com/eturcan/emotion-infused.

Our contributions in this work are as follows: (i) consideration of factors suggested by psychological theory in deep learning methods for predicting stress, with a focus on emotion; (ii) an exploration of three different approaches to emotion-infused models, with experimental results showing comparable results to the state-of-the-art in all cases; and (iii) a framework for interpreting our models to show the impact of emotion and other factors in our models.

2 Related Work

Researchers who use natural language approaches for stress detection often rely on external resources such as diagnostic questionnaires (e.g., Guntuku

[1] https://www.apa.org/news/press/releases/stress/2020/report-october

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2895–2909
June 6–11, 2021. ©2021 Association for Computational Linguistics

et al. (2018)) or techniques like pattern matching (patterns such as "I am stressed", e.g., Winata et al. (2018); Lin et al. (2017)) to assign labels. Much of the work that has been done on psychological stress detection focuses either on establishing baseline models with little advancement in computational modeling, or on using external information about the text (e.g., author, time of posting, number of replies), which is usually, but not always available and may differ in meaning or importance across platforms and domains.

There has also been a substantial amount of work on detecting related mental health concerns such as anxiety (e.g., Shen and Rudzicz (2017); Gruda and Hasan (2019); Jiang et al. (2020)), but these are distinct from the generalized experience of stress.

The most similar work to ours is Turcan and McKeown (2019), our prior work publishing a dataset of psychological stress collected from the social media website Reddit and labeled by crowd workers, and presenting baselines with several basic non-neural and BERT-based models on this data. We use this dataset in our current work; however, we focus on exploring interpretable frameworks for this sensitive task and connecting the stress detection task concretely with emotion detection.

The models we propose in this work rely on two types of enhancements to the neural representation learned by models like BERT: multi-task learning and pre-training or fine-tuning. Multi-task learning is an increasingly popular framework in which some parameters in a model are shared between or used to inform multiple different tasks. Hard parameter sharing (Caruana, 1993), the variant we employ, uses some set of parameters as a shared base representation and then allows each task to have some private parameters on top and perform their own separate predictions. Multi-task learning has been successfully applied to many domains across NLP (Sun et al., 2019; Kiperwasser and Ballesteros, 2018; Liu et al., 2019); we are especially interested in instances where it has improved semantic and emotion-related tasks, such as Xu et al. (2018), who perform emotion detection with a suite of secondary semantic tasks including personality classification.

Pre-training and fine-tuning are another type of transfer learning where multiple tasks are trained in sequence rather than at the same time. Pre-trained language models are perhaps the most widely used example, where a large neural language model can be fine-tuned for many different tasks (Devlin et al., 2019). Additionally, continuing to pre-train the language model itself on language from the target domain has been shown to improve performance (Howard and Ruder, 2018; Chakrabarty et al., 2019; Gururangan et al., 2020) (also note Chronopoulou et al. (2019), who perform this task at the same time as the target task, in a form of multi-task learning). This methodology has been successfully extended to other domains, in which a model is first fine-tuned on some large, broadly useful task and then further fine-tuned for a smaller target task (e.g., Felbo et al. (2017), who first fine-tuned on emoji detection and then fine-tuned on target semantic tasks including emotion and sentiment detection).

It should be noted that the psychological stress is much better studied in settings where researchers have access to some physiological signals (e.g., Zuo et al. (2012); Allen et al. (2014); Al-Shargie et al. (2016); Kumar et al. (2020); Jaiswal et al. (2020)). This work is not as relevant to our task, since we have only text data available when detecting stress from online posts.

Dataset	Size
Dreaddit	3,553
GoEmotions$_{A,E,S}$	58K
GoEmotions$_{FSJ}$	4,136
Vent	1.6M

Table 1: The datasets we use in this work and their relative sizes (in terms of total number of data points).

3 Data

A comparison of all the datasets we use in this work can be seen in Table 1. The primary dataset we use for this work is Dreaddit (Turcan and McKeown, 2019), a dataset of 3,553 segments of Reddit posts from various support communities where the authors believe posters are likely to express stress. The stress detection problem as expressed in this dataset is a binary classification problem, with crowdsourced annotations aggregated as the majority vote from five annotators for each data point. We note that this paper frames the stress classification problem in terms of the author and the time–i.e., a post is labeled stressful only if the poster themselves is currently expressing stress.

Because this dataset is small for training a deep learning model, we also experiment with larger datasets to provide auxiliary information. We select the GoEmotions dataset (Demszky et al., 2020), which consists of 58,009 Reddit comments labeled by crowd workers with one or more of 27 emotions (or Neutral), for its larger size and genre similarity to Dreaddit. In this paper, we refer to the dataset in this form as GoEmotions$_{all}$ or GoEmotions$_{A}$. The authors also published two relabelings of this dataset, achieved by agglomerative clustering: one where labels are clustered together into the Ekman 6 basic emotions (anger, disgust, fear, joy, sadness, surprise, neutral) (Ekman, 1992) (GoEmotions$_{Ekman/E}$), and one into simple polarity (positive, negative, ambiguous, neutral) (GoEmotions$_{sentiment/S}$). We run our experiments with each version of this dataset.

We also explore the use of another social media website, Vent. Vent is a platform more similar to Twitter or Tumblr than Reddit, where users post vents of any length and tag them as they like, and other users react to them or post comments. The benefit of Vent for this purpose is that posters self-identify some emotion they are feeling from a large list of pre-made emotions. The data we use is collected by Malko et al. (2021)[2]. We select Vent data that has been labeled with fear or sadness, which we hypothesize to be related to stress, as well as joy, for a contrast. We note that this dataset is strictly single-class, whereas GoEmotions may have more than one emotion label per data point. In all, there are 1.6M vents in our dataset, much larger than Dreaddit or GoEmotions; we randomly sample this data in a stratified manner to create a training, development, and test set with an 80/10/10 ratio. To examine the effects of domain similarity, we also select a subset of GoEmotions with the corresponding emotion labels – we subsample the existing "all" dataset to select only data points originally labeled with fear, joy, or sadness, for a final set of 4,136 data points (3,342 of which are the train set). We call this subset GoEmotions$_{FSJ}$, and we compare it against Vent to see whether genre similarity or data size is more important in this multitask setting.

4 Models

We experiment with three types of emotion-infused models; that is, we present three different ways to incorporate emotion information into our stress detection models, divided into multi-task learning and fine-tuning.

4.1 Alternating Multi-Task Models

Our first multi-task models, which we refer to as Multi^{Alt}, are simply two single-task models sharing the same base BERT representation layers. The models are alternating in that we train them with two datasets with two different sets of labels–i.e., we train the stress task with the Dreaddit data and the emotion task with the GoEmotions or Vent data. We refer to the variants with a subscript, i.e., Multi$^{Alt}_{GoEmotions_A}$ (i.e., GoEmotions with *all* emotions), Multi$^{Alt}_{GoEmotions_E}$ (i.e., the *Ekman* GoEmotions relabeling), Multi$^{Alt}_{Vent}$ (i.e., the Vent data), etc. The Multi^{Alt} models can be seen in Figure 1a. One loss step for these models consists of only one dataset and task, so they are trained with the negative log-likelihood (NLL) loss for single-label tasks (Dreaddit, Vent, GoEmotions$_{FSJ}$) and the binary cross-entropy (BCE) loss for multi-label tasks (GoEmotions$_{A,E,S}$).

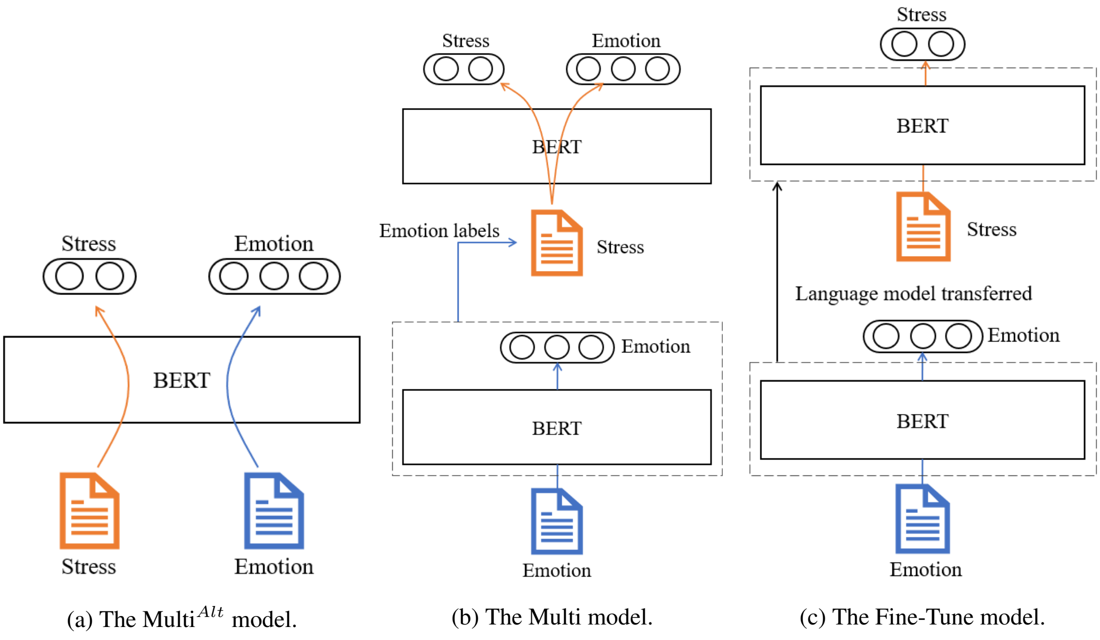
4.2 Classical Multi-Task Models

We also experiment with a multi-task learning setup where we perform the two tasks at the same time on the *same input data*. We call this architecture Multi. However, because the Dreaddit data is labeled only with stress, we first separately train BERT models on the various versions of GoEmotions and use them to predict emotion labels for Dreaddit. We then take these emotion labels to be "silver data" and train on them alongside stress. The Multi model can be seen in Figure 1b. Since stress detection is our main task in this work, we focus on this task where we have gold labels for stress, but note that it will be interesting in future work to experiment with other task settings, such as whether stress detection can improve emotion classification. In these models, the losses of the stress task and the emotion task are summed together for each batch using a tunable weight parameter, i.e., $\mathfrak{L} = \lambda \mathfrak{L}_{\text{stress}} + (1 - \lambda)\mathfrak{L}_{\text{emotion}}$.

4.3 Fine-Tuning Models

We experiment with models in which we first endow the BERT representation with knowledge of the emotion task by fine-tuning and then apply it to stress detection (as in Phang et al. (2018)). We perform a sequential version of the Multi^{Alt} models, in which we fine-tune a pre-trained BERT language model on another task, and then extract the language model parameters to initialize a BERT model that we continue to fine-tune

[2] Due to license and ethics policy restrictions, we currently do not make this data publicly available.



(a) The Multi^{Alt} model. (b) The Multi model. (c) The Fine-Tune model.

Figure 1: The emotion-informed architectures we use in our experiments.

on Dreaddit. We denote these models as, e.g., Fine-Tune$_{GoEmotions_A \rightarrow Dreaddit}$ for a model that was first trained on GoEmotions$_{all}$ and then on Dreaddit (for space, we will abbreviate Fine-Tune as FT). These fine-tuning models can be seen in Figure 1c. These models are trained with the NLL and BCE losses as in the Multi^{Alt} models.

5 Experimental Setup and Results

5.1 Baselines

We present a re-implementation of the same BERT-based fine-tuning model used in Turcan and McKeown (2019), where this model performed best on Dreaddit. We report this as an average of 3 runs with distinct random seeds, and our results are, on average, lower than the single model reported, but with high variance. Because of this, we assume that the previously reported performance is from the high end of this variance and use our average score as our baseline in this work. This model is a pre-trained BERT language model (released as `bert-base-uncased` by Wolf et al. (2019); we use this same pre-trained language model as the basis for all our models) followed by a dropout layer and a dense classification layer. We also report a recurrent neural network (RNN) model, which uses either a long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) or a gated recurrent unit (GRU) (Cho et al., 2014) in place of the transformer from BERT and is otherwise the same. These models are trained with the NLL and BCE losses as with the Multi^{Alt} models.

5.2 Training

We train all of our models with minibatch gradient descent using the Adam optimizer (Kingma and Ba, 2015) with a batch size of 16, given GPU space constraints. We perform gradient clipping to 1.0 to prevent exploding gradients. When training any model, we perform early stopping based on the F1 score on the Dreaddit development set and select the model parameters from the epoch that achieved the best development score for our final evaluated model.

5.3 Hyperparameter Tuning

We tune hyperparameters for all our models using Bayesian Optimization from the Python library `ax`[3]. All models train the initial learning rate of the Adam optimizer and the dropout probability before the final classification layer; the Multi models also tune the loss weight parameter λ, and we also note that the RNN model tunes additional parameters such as the type of RNN, hidden dimension, etc. For all models, we tune parameters based on the F1 score on the Dreaddit development set; we train an ensemble of three models with three different, fixed random seeds and average their performance for a given parameter setting. We report the mean and standard deviation of three models, with three

[3] `https://github.com/facebook/Ax`

Model	Binary F1	Accuracy
RNN	67.58 ± 1.22	68.86 ± 1.10
BERT	78.88 ± 1.09	79.11 ± 1.32
$\text{Multi}^{Alt}_{GE_A}$	79.02 ± 0.35	79.72 ± 0.69
$\text{Multi}^{Alt}_{GE_E}$	80.24 ± 1.39	$\mathbf{81.07} \pm 1.13$
$\text{Multi}^{Alt}_{GE_S}$	79.46 ± 1.05	79.86 ± 0.50
$\text{Multi}^{Alt}_{GE_{FSJ}}$	79.17 ± 0.61	78.69 ± 1.86
$\text{Multi}^{Alt}_{Vent}$	$\mathbf{80.34} \pm 1.39$	79.67 ± 2.03
Multi_{Dr_S}	78.97 ± 0.24	78.55 ± 0.07
$\text{Multi}_{Dr_{FSJ}}$	78.90 ± 0.59	78.55 ± 0.07

Table 2: Results of our multitask models. The best result under each metric is bolded. GE is GoEmotions.

Model	Binary F1	Accuracy
BERT	78.88 ± 1.09	79.11 ± 1.32
$\text{FT}_{GE_A \to Dr}$	76.40 ± 0.50	76.83 ± 0.40
$\text{FT}_{GE_E \to Dr}$	79.44 ± 0.29	79.53 ± 0.46
$\text{FT}_{GE_S \to Dr}$	79.75 ± 0.52	80.61 ± 0.40
$\text{FT}_{GE_{FSJ} \to Dr}$	$\mathbf{80.25} \pm 0.24$	$\mathbf{80.98} \pm 0.20$

Table 3: Results of our fine-tuning models. The best result under each metric is bolded. GE is GoEmotions, and Dr is Dreaddit.

different random seeds, trained with the best hyperparameters. More details about hyperparameter tuning can be found in the appendix.

5.4 Results

We report the results of our multi-task models in Table 2[4]. In general, our Multi^{Alt} models perform similarly, and outperform the Multi models; we assume this is due to the introduction of noise in labeling the silver emotion data. Of these models, $\text{Multi}^{Alt}_{Vent}$ performs best. With regards to GoEmotions, the 28-way classification of GoEmotions_A naturally leads to lower numerical performance than the tasks with smaller numbers of classes, and we expect that GoEmotions_S may group too many distinctly labeled emotions together under the same emotion labels; it seems GoEmotions_E is the happy medium for this model. We also note that the $\text{Multi}^{Alt}_{Vent}$ and $\text{Multi}^{Alt}_{GoEmotions_E}$ models perform equally well, which indicates that the genre mismatch is not an issue for this problem, or that Vent has a similar enough genre to Reddit that it does not affect the results. Somewhat surprisingly, $\text{Multi}^{Alt}_{GoEmotions_{FSJ}}$ does not do as well as $\text{Multi}^{Alt}_{Vent}$; however, the GoEmotions data is much smaller than Vent, especially when subsampled to select specific emotions.

We further report the results of our fine-tuning models in Table 3. Because we expect that genre similarity should play a larger role when the secondary task can offer no direct training signal during the primary task fine-tuning, we evaluate on GoEmotions here and not Vent. Here, we observe that our best model, $\text{Fine-Tune}_{GoEmotions_{FSJ} \to Dreaddit}$, scores at least one standard deviation above BERT. We see higher increases in performance for the simpler classification problems in GoEmotions_S and GoEmotions_{FSJ} and worsened performance for GoEmotions_A, suggesting that in the sequential paradigm, more complex tasks are not able to interact appropriately with the main task and instead interfere.

Dataset	Macro F1
$\textbf{GoEmotions}_A$	48.98
$\textbf{GoEmotions}_E$	62.16
$\textbf{GoEmotions}_S$	69.65
$\textbf{GoEmotions}_{FSJ}$	91.87

Table 4: Performance of our fine-tuning BERT models on the different GoEmotions labelings and datasets.

We also report the performance of the fine-tuning BERT models we trained on GoEmotions in order to label Dreaddit with emotion in Table 4; these results track well with the fine-tuning results reported by Demszky et al. (2020). Because these models are intermediates used for labeling, we report the F1 scores of the single model we actually used for labeling, although we tuned their parameters with an average of 3 different instances as with all other models. Many-way classification problems have much more opportunity for error and noise in an already-noisy process of labeling unlabeled data, so we use only the two best-performing GoEmotions models, which are those trained on the fewest-label datasets, GoEmotions_S and GoEmotions_{FSJ}, for our Multi models.

Overall, the inclusion of emotion information results in modest improvements, even though not statistically significant, as compared to BERT. However, our true goal in this work is to analyze the explainability of all of these models, to which we turn next.

[4] We did compute statistical significance by calculating the majority vote of each of the models' 3 runs and using the approximate randomization test, but no model is significantly different from BERT.

	GoEmo$_A$	GoEmo$_E$	GoEmo$_S$	GoEmo$_{FSJ}$*
Dreaddit (gold stress + pred. emotion)	0.3396	0.2554	0.0565	0.3207
GoEmotions (gold emotion + pred. stress)	0.1274	0.2668	0.2786	0.4115

Table 5: Correlations of the gold labels for each dataset with labels predicted by the other task's classifier in a Multi^{Alt} model. GoEmotions$_{FSJ}$ (abbreviated for space as GoEmo$_{FSJ}$) is starred because its emotion data is not multi-label and therefore the correlation ratio η is used instead of the coefficient of determination R^2 (which is used for the other, multilabel GoEmotions variants).

	GoEmotions$_S$				GoEmotions$_{FSJ}$		
	neutral	negative	ambiguous	positive	fear	sadness	joy
Dreaddit	-0.3960	0.6128	-0.0106	-0.2759	0.9697	0.7113	0.1386
GoEmotions	-0.1021	0.4866	0.0751	-0.3323	0.9545	0.8921	0.0235

Table 6: Per-class scores of emotion and stress for Dreaddit (with gold stress and predicted emotion) and GoEmotions (with gold emotion and predicted stress). For GoEmotions$_S$, these numbers are the Pearson correlation r of each individual emotion label with the stress labels; for GoEmotions$_{FSJ}$, these are the average stress label assigned to data points in each emotion category, where 0 is non-stress and 1 is stress.

6 Analysis

We perform three different analyses to probe our trained models and discover what information they learn to use. For our Multi^{Alt} models, we investigate the usefulness of the emotion prediction layers in explaining stress classifications, and for all models, we use Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) to show that our emotion-infused models rely on meaningfully different types of words than BERT in order to make their predictions.

6.1 Multi-task Knowledge

We perform an analysis of our Multi^{Alt} models to see what information they learn about emotion[5]. We take the development sets of each of the datasets (Dreaddit and GoEmotions) and predict their labels under the other task (i.e., emotion for Dreaddit and vice-versa). We report the correlation of these predicted labels with the gold labels in Table 5[6]. In this case, the GoEmotions$_{FSJ}$ variant is a single-label three-way classification problem, so we report the correlation ratio η (Fisher, 1938). The other GoEmotions variants are multi-label, so we report the coefficient of determination R^2 (Cohen et al., 2015). We further present breakdowns of the correlations per emotion category for the polarity and FSJ subsets of GoEmotions in Table 6 and include the All and Ekman sets as well as the Vent data in the appendix.

We observe that our multi-task models generally learn a moderate correlation between the stress labels and the emotion labels; they learn that negative emotions like fear and sadness are linked to stress and neutral or positive emotions are linked to non-stress, which makes intuitive sense. These emotion predictions can help explain the stress classifier's predictions; imagine, for example, showing a patient or clinician that the patient's social media shows a strong pattern of fear and anger as a more detailed explanation for places a stress classifier detects stress. From a machine learning perspective, this correlation also suggests the potential for using emotion data as distantly-labeled stress data to supplement the small extant stress datasets.

6.2 LIWC Analysis

We also investigate the types of information each model is using to make its decisions. In this section, we use the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a hand-crafted lexicon which collects words belonging to psychologically meaningful categories like positive emotion and cognitive processes, to categorize the information our different models use to predict stress.

We first analyze the unigrams our various models use to perform stress classification using LIME. LIME accepts an input from our development set, perturbs it in the bag-of-unigrams space, and runs one of our classifiers on each perturbation to calculate the importance of various unigrams; through

[5] We did perform an equivalent analysis on the Multi models, which shows similar trends, but as Multi^{Alt} shows better performance, we omit it for space.

[6] We also note the possibility that different combinations of emotions are relevant to stress; however, not enough of our data is labeled with multiple emotion labels (4% of Dreaddit's silver labels from GoEmotions$_S$, 9% of GoEmotions$_E$) to test this hypothesis in this work.

LIWC	**BERT**	**Multi**$^{Alt}_{GE_E}$	**Multi**$^{Alt}_{Vent}$	**Multi**$_{Dr_{FSJ}}$	**FT**$_{GE_{FSJ}\rightarrow Dr}$
Affective Processes	19%	22%	19%	16%	22%
Positive Emotion	8%	10%	9%	9%	12%
Anger	31%	40%	30%	25%	31%
Cognitive Processes	16%	17%	17%	17%	17%
Certainty	8%	13%	12%	16%	11%
Perceptual Processes	17%	15%	14%	14%	15%
Biological Processes	15%	19%	17%	16%	17%
Achievement	17%	19%	19%	13%	17%

Table 7: A comparison of how often several of our models rely on words from several LIWC categories to make their decisions, according to LIME. These numbers represent the percentage of available LIWC words each model selected in the top 10 LIME explanations for the entire development set. Dr is Dreaddit, and GE is GoEmotions.

this process, we acquire the 10 unigrams with the highest magnitude output by LIME for each development example and consider them "explanations". We thus have 2,760 individual unigram explanations for the entire development set to analyze.

We then use the word lists from LIWC 2015's 72 psychological categories to see what types of words each classifier tends to use to make decisions of stress vs. non-stress. An abbreviated list of results showing our best models from each category is shown in Table 7[7]. We observe small but consistent effects suggesting that, in comparison to the basic BERT model, our emotion-enhanced models broadly learn to use the following information:

Affective information. Most emotion-infused models except for Multi learn to use affective information, which includes both positive and negative emotion words, more often. We see the largest increase in anger, one of the emotions we had identified as relevant to stress, for Multi$^{Alt}_{GoEmotions_E}$, which makes intuitive sense because anger is one of the Ekman six basic emotions and thus, is explicitly predicted by this model.

Cognitive processes. All models show some increase in using words related to cognitive processes as compared to BERT; however, its subcategory Certainty, which includes words about absoluteness such as *never*, *obvious*, and *clearly*, shows larger changes. For example, Multi$_{Dreaddit_{FSJ}}$ uses Certainty twice as often as BERT. These cognitive words seem to target the mental aspects of stress. Rumination and a focus on absoluteness are known signs of anxiety disorders, an extreme form of chronic stress (Nolen-Hoeksema et al., 2008; Miranda and Mennin, 2007).

Additional differences. We observe other, smaller patterns among LIWC usage for these models. For example, the Multi^{Alt} models use the most achievement-oriented words (although most models show modest increases), suggesting that this information, which includes words about success and failure, is relevant to emotion and to stress. This makes sense, since failing to achieve (e.g., failing a class) can be a major stressor. We also see larger proportions of biological process words used by all emotion-infused models. We suggest this is because Dreaddit includes posts taken from Reddit communities about anxiety and PTSD, where posters are likely to describe their physical and mental symptoms while seeking help.

6.3 Salient Words

We then investigate the data itself for highly significant words using the measure of relative salience proposed by Mohammad (2012), $RelativeSalience(w|T_1, T_2) = \frac{f_1}{N_1} - \frac{f_2}{N_2}$. That is, it measures the importance of a token w in two different corpora T_1, T_2 by subtracting their two relative frequencies (where f_1, f_2 are the counts of token w in each corpus and N_1, N_2 are the total tokens in each corpus). We compute this measure for all words in the Dreaddit training data, taking our two corpora to be the subsets labeled stress and not-stress. We take the top 200 unigrams for each label (stress as opposed to non-stress and vice-versa) and provide some examples in Table 8 with the full list of words available in the appendix. We examine the words and divide them into related groups in order to understand what types of information should theoretically be most important to classifying the data. For example, we see that different sets of function words are actually among the most important for both classes, with words like conjunctions typically appearing more indicative of stress

[7] More detail on the full table is available in the appendix.

	Category	**Example Words**
Stress	Function Words	and, but, how, like, no, not, or, where, why
	Negative Sentiment	awful, bad, cry, fear, hate, stress, stupid
	Helplessness	alone, can't, nothing, nowhere, trying
Non-Stress	Function Words	a, for, if, some, the, was, who, will, would
	Positive Sentiment	amazing, best, good, great, hope, nice
	Support	email, helped, support, thank, together, we

Table 8: Some examples of words identified by relative salience on the Dreaddit training data as indicative of stress or non-stress. We group the words by hand into semantically meaningful categories for ease of understanding.

Label	**BERT**	**Multi**$^{Alt}_{GE_E}$	**Multi**$^{Alt}_{Vent}$	**Multi**$_{Dr_{FSJ}}$	**FT**$_{GE_{FSJ} \rightarrow Dr}$
Stress	33%	36%	32%	32%	33%
Non-Stress	15%	15%	19%	18%	17%

Table 9: A comparison of how often several of our models rely on words identified as salient for stress or non-stress to make their decisions, according to LIME. These numbers represent the percentage of available relative salience words each model selected in the top 10 LIME explanations. Dreaddit is Dr, and GoEmotions is GE.

(which echoes Turcan and McKeown (2019)'s finding that stressful data is typically longer with more clauses), while non-stress includes words expressing future-thinking like *if*, *will*, and *would*. We also naturally find negative words for stress and positive words for non-stress, as well as a dichotomy of isolation and helplessness for stress vs. support and community for non-stress which is supported by psychological literature (Grant et al., 2009).

We then look at the intersection between relative salience and LIME explanations, counting how many LIME explanations are highly salient words for stress or non-stress; abbreviated results are shown in Table 9 and the full table is available in the appendix. We see that our emotion-infused models learn to rely more often on words identified as indicative of non-stress, the minority class, instead of stress, the majority class.

6.4 Error Analysis

We note that the presented models do sometimes make some new errors when incorporating emotional information, and that while these methods successfully incorporate such information with no feature crafting, some further innovation may be needed in order to use this information optimally. For example, we reproduce an example from our development set, with profanity censored:

*And everyone was passive aggressive. The manager tried to peg down my salary multiple times like a f**king haggler at a market. Anyway, I decided to go get some antidepressants and the bottle fell out of my pocket, a coworker noticed and reported it to my boss. Who smiled and asked if there was anything I'd like to tell her. The passive aggressive s**t really got to me, and then I realized that I was being illegally paid.*

The annotators for Dreaddit label this post not stress, presumably because there is not enough context for how the poster feels about this story presently, and the poster conveys more anger than anything else. The LIME explanations for the BERT model, which labels this correctly, include some profanity, but largely focus on function words. However, all four of our Multi$^{Alt}_{GoEmotions}$ models misclassify this example as stressed and rely on words like *aggressive* (from *passive aggressive*) and the profanity to do so. Meanwhile, the emotion classifiers of our Multi$^{Alt}_{GoEmotions}$ models are misled by words like *smiled* and label this example *joy* or *positive*. This is a difficult example; without noticing that the event happened in the past, it is easy to assume the poster is presently stressed. We believe examples like this require some grounding– for example, an understanding of what *passive aggressive* means and some representation of the timeline involved, that language models simply cannot express in the traditional classification setup.

We also reproduce an anonymized example where our emotion-infused models improve upon BERT:

She comes crying to me and formulates a plan to break up. She talks to <name> about their issues and her will to leave him wilts. She stays with him. Rinse and repeat, except it gets worse over time. How can I break the cycle, or help her break the *cycle?*

BERT misclassifies this example, where the author is stressed about a friend's situation, as non-stressful, relying on words like *break* and *help*, while our $\text{Multi}^{Alt}_{GoEmotions}$ models successfully use the word *crying* to predict stress. We notice that *crying* or *worse* is the highest-ranked explanation for most of our emotion-infused models. These results are promising for the development of models that focus on information that humans consider intuitive.

7 Conclusion

In this work, we present a suite of emotion-enhanced models that incorporate emotional information in various ways to enhance the task of binary stress prediction. All three types of our models achieve comparable performance to a state-of-the-art fine-tuning BERT baseline, and, more importantly, we show that they result in more explainable models. We also introduce a new framework for model interpretation using LIME and show that our emotion-enhanced multi-task models offer a new dimension of interpetability by using the predictions of auxiliary tasks to explain the primary task. In our future work, we hope to expand these analyses to tasks in other domains and devise model architectures that can make more direct use of multi-task learning to make and explain their predictions.

8 Ethical Considerations

Our intended use of stress detection is to help those in distress. We envision systems such as therapeutic chatbots or assistants that can understand users' emotions and identify those in need so that a person can intervene. We would urge any user of stress detection technology to carefully control who may use the system.

Currently, the presented models may fail in two ways: they may either misclassify stress, or they may use the wrong information to make their predictions. Obviously, there is some potential harm to a person who is truly in need if a system based on this work fails to detect them, and it is possible that a person who is not truly in need may be irritated or offended if someone reaches out to them because of a mistake. In terms of explanations, we note that previous work has shown that focusing on incorrect rationales can unfairly target some groups of people (Zhong et al., 2019), although in this work we see that function words truly differ across the stressed and non-stressed populations and we do not observe any language that we know to be representative of minority groups in our explanations.

We emphasize our intention that emotional systems such as this be used responsibly, with a human in the loop–for example, a guidance counselor who can look at the predicted labels and offered explanations for their students' stress levels and decide whether or not they seem sensible.

We note that because most of our data was collected from Reddit, a website with a known overall demographic skew (towards young, white, American men[8]), our conclusions about what stress looks like and how to detect it cannot necessarily be applied to broader groups of people. We also note that we have no way to determine the demographic information of the specific posters in any of our datasets and whether they differ from the overall Reddit statistics. We hope that we, and other researchers, can find ways to consider the specific ways in which minority groups express stress as well.

Acknowledgements

We thank our reviewers, as well as the members of our Natural Language Processing research group at Columbia University, for their insightful and constructive comments.

References

Fares Al-Shargie, Masashi Kiguchi, Nasreen Badruddin, Sarat C. Dass, and Ahmad Fadzil Mohammad Hani. 2016. Mental stress assessment using simultaneous measurement of eeg and fnirs. *Biomedical Optics Express*, 7(10):3882–3898.

Andrew P. Allen, Paul J. Kennedy, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. 2014. Biological and psychological markers of stress in humans: Focus on the trier social stress test. *Neuroscience & Biobehavioral Reviews*, 38:94–124.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

[8] https://social.techjunkie.com/demographics-reddit

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. 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 2089–2095, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. 2015. *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd edition. Routledge.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. 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, ACL 2020, Online, July 5-10, 2020*, pages 4040–4054. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(5):550–553.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1615–1625. Association for Computational Linguistics.

Ronald A. Fisher. 1938. *Statistical methods for research workers.*

Nina Grant, Mark Hamer, and Andrew Steptoe. 2009. Social Isolation and Stress-related Cardiovascular, Lipid, and Cortisol Responses. *Annals of Behavioral Medicine*, 37(1):29–37.

Dritjon Gruda and Souleiman Hasan. 2019. Feeling anxious? perceiving anxiety in tweets using machine learning. *CoRR*, abs/1909.06959.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, and Lyle H. Ungar. 2018. Understanding and measuring psychological stress using social media. *CoRR*, abs/1811.07430.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2020. MuSE: a multimodal dataset of stressed emotion. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1499–1510, Marseille, France. European Language Resources Association.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Trans. Assoc. Comput. Linguistics*, 6:225–240.

Satish Kumar, A S M Iftekhar, Michael Goebel, Tom Bullock, Mary H. MacLean, Michael B. Miller, Tyler Santander, Barry Giesbrecht, Scott T. Grafton, and B. S. Manjunath. 2020. Stressnet: Detecting stress in thermal videos.

Richard S. Lazarus. 2006. *Stress and emotion: a new synthesis*, 1st edition. Springer Publishing Company.

Robert W. Levenson. 2019. Stress and illness: A role for specific emotions. *Psychosomatic Medicine*, 81(8):720–730.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(09):1820–1833.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4487–4496. Association for Computational Linguistics.

Anton Malko, Cecile Paris, Andreas Duenser, Mervi Kangas, Diego Mollá, Ross Sparks, and Stephen Wan. 2021. Expressing and reacting to emotions in a specialised online community. Technical report, CSIRO.

Regina Miranda and Douglas S. Mennin. 2007. Depression, generalized anxiety disorder, and certainty in pessimistic predictions about the future. *Cognitive Therapy and Research*, pages 71–82.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decis. Support Syst.*, 53(4):730–741.

Susan Nolen-Hoeksema, Blair E. Wisco, and Sonja Lyubomirsky. 2008. Rethinking rumination. *Perspectives on Psychological Science*, (5):400–424.

James Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR*, abs/1907.12412.

Hanna A. Thoern, Marcus Grueschow, Ulrike Ehlert, Christian C. Ruff, and Brigit Kleim. 2016. Attentional bias towards positive emotion predicts stress resilience. *PLoS ONE*, 11(3).

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis LOUHI@EMNLP 2019, Hong Kong, November 3, 2019*, pages 97–107. Association for Computational Linguistics.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based LSTM for psychological stress detection from spoken language using distant supervision. *CoRR*, abs/1805.12307.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 292–298. Association for Computational Linguistics.

Ruiqi Zhong, Steven Shao, and Kathleen R. McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *CoRR*, abs/1908.06870.

Xin Zuo, Tian Li, and Pascale Fung. 2012. A multilingual natural stress emotion database. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1174–1178, Istanbul, Turkey. European Language Resources Association (ELRA).

Name	**Type**	**Range**
learning rate	continuous	$[10^{-6}, 10^{-3}]$
P(dropout)	continuous	[0, 1]
λ	continuous	[0, 0.9]
embedding dim.	integer	[32, 256]
hidden dim.	integer	[32, 512]
n_{layers}	integer	[1, 3]
RNN	categorical	{LSTM, GRU}

Table 10: Hyperparameter ranges for our models. BERT-based models tuned the first two; the Multi models additionally tuned λ; the RNN additionally tuned the remainder.

A Reproducibility

We report the contents of the NAACL 2021 Reprodicibility Checklist that apply to our work.

A.1 Training and Tuning

Our Multi$^{Alt}_{Vent}$, Multi$_{Dreaddit_S}$, and Multi$_{Dreaddit_{FSJ}}$ models were trained on one Tesla V100 GPU with one CPU. All other models were trained on one Nvidia P100 GPU with one CPU.

Hyperparameter tuning was done the same way for every model, with Bayesian optimization as implemented by `ax`, with the F1 score on the Dreaddit development set as the criterion to optimize. Multi$_{Dreaddit_S}$ and Multi$_{Dreaddit_{FSJ}}$ were given 35 trials for time constraints; all other models were given 50 trials. All models were trained with a patience of 5 epochs and a tolerance of 0.0001 for dev set improvement, and allowed to run for a maximum of 20 epochs. All models tuned the initial learning rate and the dropout probability, with the Multi models also tuning the lambda weight parameter between their two task losses. Additionally, the RNN model was initialized from scratch and additionally tuned the embedding dimension, hidden dimension, number of layers, and type of RNN. Our parameter ranges are shown in Table 10.

The selected values of learning rate and dropout for all our models are shown in Table 11, rounded to two decimal places. We also note the remaining hyperparameters here: our RNN used a hidden dimension of 506, an embedding dimension of 137, and a 2-layer GRU. Additionally, Multi$_{Dr_{FSJ}}$ selected $\lambda = 0.90$ and Multi$_{Dr_{FSJ}}$, $\lambda = 0.67$.

All of our models are similar in architecture and therefore take similar runtimes and have a similar number of parameters. Running our entire hyperpa-

Model	Learning Rate	P(dropout)
RNN	1.40×10^{-4}	0.86
BERT	4.27×10^{-5}	0.13
Multi$^{Alt}_{GE_A}$	8.47×10^{-6}	0.40
Multi$^{Alt}_{GE_E}$	1.08×10^{-5}	0.00
Multi$^{Alt}_{GE_S}$	1.69×10^{-5}	0.00
Multi$^{Alt}_{GE_{FSJ}}$	8.98×10^{-6}	0.00
Multi$^{Alt}_{Vent}$	4.44×10^{-5}	0.00
Multi$_{Dr_S}$	1.14×10^{-5}	0.00
Multi$_{Dr_{FSJ}}$	1.79×10^{-5}	0.00
FT$_{GE_A\rightarrow Dr}$	7.30×10^{-5}	0.05
FT$_{GE_E\rightarrow Dr}$	1.35×10^{-5}	0.00
FT$_{GE_S\rightarrow Dr}$	1.95×10^{-5}	0.09
FT$_{GE_{FSJ}\rightarrow Dr}$	5.03×10^{-6}	0.03

Table 11: Our models' selected hyperparameters. So that the table fits in a column, GE is GoEmotions, and Dr is Dreaddit.

Model	**Binary F1**	**Accuracy**
RNN	72.58 ± 0.50	74.15 ± 1.46
BERT	81.79 ± 0.45	82.97 ± 0.30
Multi$^{Alt}_{GE_A}$	81.31 ± 0.81	82.97 ± 0.51
Multi$^{Alt}_{GE_E}$	80.30 ± 0.85	82.25 ± 0.59
Multi$^{Alt}_{GE_S}$	80.79 ± 1.31	82.00 ± 0.74
Multi$^{Alt}_{GE_{FSJ}}$	81.87 ± 2.21	82.61 ± 2.42
Multi$^{Alt}_{Vent}$	82.30 ± 1.16	82.49 ± 2.01
Multi$_{Dr_S}$	81.40 ± 1.54	82.49 ± 1.20
Multi$_{Dr_{FSJ}}$	82.58 ± 1.11	83.21 ± 1.46
FT$_{GE_A\rightarrow Dr}$	82.58 ± 1.53	82.13 ± 1.04
FT$_{GE_E\rightarrow Dr}$	82.58 ± 1.53	83.57 ± 1.71
FT$_{GE_S\rightarrow Dr}$	80.87 ± 1.15	82.49 ± 0.68
FT$_{GE_{FSJ}\rightarrow Dr}$	82.88 ± 0.92	84.54 ± 0.74

Table 12: Results of all our presented models on the Dreaddit development set. So that the table fits in a column, GE is GoEmotions, and Dr is Dreaddit.

rameter tuning setup described above takes about one day, and training one ensemble of three models takes about 25 minutes. BERT makes up the vast majority of our models' parameters, putting all of them at about 109B parameters (aside from the RNN, which has 7B).

Our performance for each model on the dev set is shown in Table 12.

A.2 Evaluation

We note the standard equations of our reported metrics. For our Dreaddit models, we report binary F1, i.e., the harmonic mean of precision and recall for the positive class (here, stress): $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP}+\text{FN})}$, where TP represents the number of examples that were correctly classified as stress (true positives), FP those incorrectly classified as stress (false positives), TN those correctly classified as non-stress (true negatives), and FN those incorrectly classified as non-stress (false negatives). We also report classification accuracy, which is the fraction of samples classified correctly, i.e., $\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$.

We note that Table 4 reports macro-averaged F1 in the multi-label and single-label cases. For both of these, we calculate the macro-average F1, which sums up TP, TN, etc. across all emotion labels and then calculates F1 score. For a multi-label input, we treat each label as a separate classification–e.g., if the model is incorrect with respect to one class but correctly identifies a second, the example counts towards incorrect examples for the first class and then again towards correct examples for the second.

A.3 Data

Our data is all English social media data. Dreaddit and GoEmotions are taken from Reddit, and Vent from the social media platform Vent. Dreaddit consists of 3,553 labeled segments of posts taken from 10 subreddits: r/domesticviolence, r/survivorsofabuce, r/anxiety, r/stress, r/almosthomeless, r/assistance, r/food_pantry, r/homeless, r/ptsd, and r/relationships. 52.3% of the data is labeled stress, with the remaining 47.4% labeled non-stress. We use the train-dev-test split of Turcan and McKeown (2019) into 2,562 train, 276 development, and 715 test examples. GoEmotions consists of 58,009 labeled Reddit comments taken from non-disclosed selection of subreddits Demszky et al. (2020). We refer the reader to the original publication's Figure 1 for details on the label distribution; GoEmotions uses 28 labels with a widely varying amount of data for each. We use the label groupings that the authors provide in order to evaluate on the Ekman labels and sentiment labels; these schemes group several of the 28 original labels together into smaller sets.

Our Vent data consists of 1.6 million Vents gathered in collaboration with the Vent platform. A much larger amount of data was collected, but we select the data with self-labeled emotion tags related to joy, sadness, and fear. These data were collected from 2013 to 2016; in their current form, we do not retain metadata about the posters. A group with whom we collaborate has collected this data, and due to licensing and ethics requirements, we are not able to release it publicly. We selected sadness, fear, and happiness based on intuition that they should be relevant to stress; we partitioned off a label-stratified 10% of the data for development and test each and the training set is 1.3 million examples. The label distribution of this dataset is 24.0% fear, 36.0% sadness, and 40.0% happiness.

We do not filter or remove any examples. The only preprocessing we perform is to apply the pretrained `bert-base-uncased` tokenizer from Wolf et al. (2019).

B Extended Analysis

We include the full tables of per-emotion correlations for $\text{Multi}^{Alt}_{GoEmotions_A}$ in Table 13 and $\text{Multi}^{Alt}_{GoEmotions_E}$ in Table 14. We note that the alternating multi-task models do not predict all of their possible emotions on Dreaddit, although all possible emotions do occur in the GoEmotions development set. We also report the correlation coefficients η and mean stress prediction for the $\text{Multi}^{Alt}_{Vent}$ model in Table 15.

The full table of LIWC/LIME explanation counts for every model is too large to fit comfortably on a page, so we make a spreadsheet available at `www.cs.columbia.edu/~eturcan/data/emotion_infused_explanations.csv`.

We include the top 200 relative salience unigrams (in order of relative salience) for stress and non-stress from the Dreaddit train set here. These are tokens as split by the Natural Language Toolkit (NLTK). We reproduce these exactly as they appear, and caution that they may include explicit or

	neutral	**anger**	**fear**	**annoyance**	**surprise**
gold stress + pred. emotion	-0.3761	–	–	–	–
gold emotion + pred. stress	-0.0728	0.2175	0.1066	0.1420	0.0188
	gratitude	**desire**	**optimism**	**admiration**	**confusion**
gold stress + pred. emotion	–	–	–	–	–
gold emotion + pred. stress	-0.0663	-0.0288	-0.0388	-0.1036	0.0471
	amusement	**approval**	**caring**	**embarrass.**	**realization**
gold stress + pred. emotion	–	–	–	–	–
gold emotion + pred. stress	-0.0195	-0.0788	-0.0095	0.0304	-0.0073
	disappoint.	**grief**	**sadness**	**curiosity**	**joy**
gold stress + pred. emotion	-0.1119	–	–	-0.2070	-0.0644
gold emotion + pred. stress	0.0465	0.0320	0.1075	0.0262	-0.0704
	love	**excitement**	**disapproval**	**remorse**	**disgust**
gold stress + pred. emotion	-0.2735	–	–	–	–
gold emotion + pred. stress	-0.0563	-0.0292	0.0342	0.0683	0.1142
	relief	**pride**	**nervousness**		
gold stress + pred. emotion	-0.2070	–	-0.1450		
gold emotion + pred. stress	0.0026	-0.0222	0.0318		

Table 13: Full emotion correlations for the Multi$^{Alt}_{GoEmotions_A}$ model.

	neutral	**anger**	**fear**	**surprise**	**joy**	**sadness**	**disgust**
gold stress + pred. emotion	-0.0419	–	–	–	-0.4986	0.0565	–
gold emotion + pred. stress	-0.0876	0.2936	0.1686	0.0276	-0.3286	0.2762	0.1548

Table 14: Full emotion correlations for the Multi$^{Alt}_{GoEmotions_E}$ model.

	fear	**sadness**	**joy**	**correlation coefficient** η
gold stress + pred. emotion	0.9697	0.7113	0.1386	0.3207
gold emtion + pred. stress	0.9545	0.8921	0.0235	0.4115

Table 15: Mean stress predictions by the Multi$^{Alt}_{Vent}$ model for given emotions in Dreaddit and GoEmotions, as well as the correlation coefficient η for those predictions.

	BERT	**Fine-Tune**$_{GE_A}$	**Fine-Tune**$_{GE_E}$	**Fine-Tune**$_{GE_S}$
stress	33.4%	31.9%	36.2%	33.8%
non-stress	15.1%	17.5%	15.5%	15.3%
	Fine-Tune$_{GE_{FSJ}}$	**Multi**$^{Alt}_{GE_A}$	**Multi**$^{Alt}_{GE_E}$	**Multi**$^{Alt}_{GE_S}$
stress	33.3%	33.1%	33.4%	31.9%
non-stress	17.3%	17.5%	16.5%	17.9%
	Multi$^{Alt}_{GE_{FSJ}}$	**Multi**$^{Alt}_{Vent}$	**Multi**$_{GE_S}$	**Multi**$_{GE_{FSJ}}$
stress	30.9%	31.5%	33.4%	32.0%
non-stress	17.6%	18.7%	17.4%	18.4%

Table 16: Full results for the relative salience analysis. So that the table fits on the page, GE is GoEmotions, and Dr is Dreaddit.

offensive language.

Stress. i, my, me, do, and, ’m, n’t, just, ’, feel, because, like, am, what, even, ?, he, but, anxiety, m, so, myself, this, know, ca, it, now, have, out, get, no, about, t, feeling, up, bad, how, ’ve, scared, not, him, over, going, all, tell, right, stop, want, anxious, past, to, fucking, need, hate, s, really, why, panic, where, happened, trying, still, when, days, makes, job, tired, or, shit, hard, getting, day, life, nothing, tl, dr, afraid, has, sorry, boyfriend, felt, crying, school, worse, don, go, attacks, sick, leave, deal, attack, anymore, being, work, im, having, constantly, thinking, almost, feels, been, worried, is, stress, which, family, due, fear, something, keep, everything, enough, every, back, worst, ..., point, home, sometimes, car, down, making, angry, literally, feelings, actually, cry, horrible, wo, think, anyone, end, move, .., help, terrified, fuck, head, then, pain, losing, situation, depression, depressed, ve, made, money, coming, mom, safe, else, everyday, gets, honestly, thing, unable, turn, whole, terrible, alone, room, heart, saying, wake, awful, sleep, against, mentally, come, absolutely, nightmares, stupid, remember, lot, without, does, abuse, lose, class, sad, stuck, hell, suffer, cant, severe, emotions, leaving, /, flashbacks, hospital, close, memories, off, night, nowhere, abused, knowing, issues, trigger, sexually

Non-Stress. ,, you, the, a, her, she, we, for, ., in, your, would, be,), !, will, (, *, :, <, that, are, who, >, as, was, url, more, if, years, -, first, were, their, thank, us, met, people, his, them, our, an, they, said, one, together, others, share, let, best, food, other, &, person, interested, please, study, each, here, asked, link, treatment, those, free, could, ”, take, great, same, support, good, “, [, some, make, months, may, older, finally, bit, research, online, experience, little, through, hope, #, $, many, helped, edit, decided, friend, see, took, few, homeless, wanted, nice, information, thanks, around, ”, questions, any, date, went, later, everyone, looking, guys, ask, than, relationship, ago, ’ll, sister, post, complete, ’d, dating, year, both, current, mental, ’s, send, 18, moved, amazing, community, provide, items, read, however, name, x200b (i.e., a zero-width space), world, willing, different, guy, 3, turned, area, visit, health, open, well, case, survivors, 10, hear, ’re, give, university, own,], hi, learn, couple, access, old, long, eventually, choose, agreed, began, love, reading, stories, loving, hey, experiences, include, preferences, forward, ;, write, sub, 1, posted, also, loved, page, email, start, away, sleeping, note, app, liked, helping, seemed, grateful, background, girl, talked, based, amazon, 2

We believe the numbers appearing in the non-stress list are indicative of financial posts where the authors indicate some amount of money has been raised or needs to be raised.

We include the full table of relative salience/LIME explanation counts in Table 16.