Understanding Negative Emotional Reactions and Their Triggering Events: Interpretable Multitask Models

Thesis Proposal

Elsbeth Turcan
Columbia University
Department of Computer Science
April 2, 2021

Abstract

Distress is more prominently visible in the modern world than ever before, particularly in the online sphere. While emotion recognition from text is a well-studied problem in natural language processing, with many diverse datasets and interesting sub-problems, comparatively little research has gone into studying the specifics of negative emotional reactions, where we assert there is greatest potential to identify and help humans in need. Further, we believe that human-understandable models, especially those crafted with an eye to psychological theory, are particularly vital for sensitive human applications such as emotion detection which deal with the intents and feelings of potentially unknown writers. In this thesis, we focus on developing computational models that further our understanding of negative emotional reactions and can potentially find those in need so that they may receive help. We focus on (1) being able to detect such emotional reactions as well as (2) detecting their causes, and we approach these problems from the perspective that our models should be interpretable and understandable.

With respect to problem (1), we first present our completed work on detecting distress, a type of negative stress, including a novel social media dataset for the problem and a suite of *emotion-infused* models which use multi-task learning with emotion detection in order to predict distress using human-intuitive information without any loss in performance. We then propose similar work with grief, where we particularly note that grief has a well-defined trigger event (loss) that should be important to understanding the mood itself. We propose the collection of another new dataset for grief detection and propose models in the same vein as our emotion-infused models to make use of emotional information, including knowledge of loss.

Finally, with respect to problem (2), we leverage the stimuli that trigger emotions to better understand the emotions themselves. We present our completed work on using common-sense knowledge and multi-task learning for the general task of emotion prediction and their associated causes, including novel architectures where one of our tasks (emotion classification or emotion cause span tagging, where the cause can be any text span) is performed and used to inform the prediction of the other task. We then present our final proposed work, in which we focus in again on negative emotional reactions and propose to model emotion stimuli specifically as hierarchical events and predict the change in their associated emotional reaction over time and for different writers. It is our hope that the work in this thesis inspires emotion and semantics researchers to look to established psychological theory in developing their models and to make interpretability a forefront concern of their research.

Contents

1 Introduction

Distress is more visible in the modern world than ever before. Surveys of stress in the United States continue to report rising numbers of adults and teenagers experiencing significant negative emotions and moods like stress and anger[1] in the wake of tragic and stressful events. Natural disasters such as hurricanes in the South Atlantic and wildfires in the western United States and Australia; disease events like the COVID-19 pandemic; numerous mass shootings, particularly in the United States; and the events sparking social movements such as #MeToo and #BlackLivesMatter affect many people directly, as well as countless others indirectly as they are repeatedly publicized by news sources and social media.

These events have long-lasting effects on those who experience them, from physical and psychological consequences of COVID-19 (Sood, 2020), to loss of home and family to natural disasters and violence, to emotional trauma including studied disorders such as Post-Traumatic Stress Disorder (PTSD) (McFarlane, 2010). Research has also suggested that continued and repeated exposure to these stressful events, even for those who have not experienced them personally, has an effect of "secondhand trauma", contributing to ambient feelings of distress (Wayne, 2016).

Fortunately, we have not only increased opportunity to observe this distress, but also better tools and data sources to examine and alleviate it. Social media and the Internet, where users constantly express their worries, seek help and community from others, and bond over their shared experiences, allow us to examine how people express, communicate, and cope with negative emotional reactions to stressful events. In this proposal, we present completed and proposed work around the central theme of identifying and understanding negative emotional reactions and their causes. While there has been much work on detecting emotions from text, there has been comparatively little work specifically on understanding these negative emotions across a range of circumstances. Being able to identify people in distress has many potentially lifesaving applications, from school counseling to disaster relief, and being able to identify the *causes* of these negative emotional reactions is even more helpful, with benefits to clinical professionals and policymakers, among others.

In building computational models, we emphasize the importance of understandable models for these sensitive tasks. The consequences of deploying some solution which purports to help those in distress but fails to identify people in need are serious; falsely identifying someone as distressed may also incur some negative stigma, especially if done in a systematically biased way. Furthermore, black-box models are notorious for developing bias and picking up on spurious correlations

[1] `https://www.apa.org/news/press/releases/stress/index`; `https://news.gallup.com/poll/249098/americans-stress-worry-anger-intensified-2018.aspx`

from their limited datasets (Gururangan et al., 2018; Clark et al., 2019), leading to potentially disastrous results. We believe for these reasons it is vital to include humans in the loop for sensitive applications such as ours; we envision tools that can help professionals to help others, but not replace their decision-making. To this end, we propose to develop more human-understandable models that make decisions by focusing on information that humans find intuitively useful or relevant. We propose that this will make their mistakes more obvious to a human operator and lessen the potential for harm. In this work, we endeavor to incorporate domain knowledge such as psychological models of emotion and narratives of grief into our computational models in order to fuse expert knowledge and the power of big data while exposing our models' prediction processes more directly, for informed decision-making.

1.1 Contributions

We propose two major classes of contributions in this document: detecting negative emotional reactions (section 2) and detecting their causes (section 3). We present completed work on detecting distress (subsection 2.2) and propose to extend that work to detect grief, which is extreme distress particularly in response to loss (subsection 2.3). Then, we present completed work on jointly detecting emotions and their causes (subsection 3.2) and propose further work on linking negative emotional reactions to large-scale events that affect many people (subsection 3.3). Across all of this work, we focus on developing human-understandable models that use intuitive information to make their predictions in this sensitive space, and we employ multitask learning to achieve this.

The fine-grained research contributions we have completed and propose in this document are as follows:

1. New datasets for a) psychological distress detection and b) grief detection in reaction to events;
2. A suite of models for both distress and grief detection, including those that make use of multitask information to incorporate emotional knowledge;
3. Several new methods of analyzing interpretability for multitask models;
4. New models for emotion-cause extraction, including those that take into account commonsense reasoning; and
5. New models for attaching emotional reactions to events, including some that work in a streaming setting and some that draw on established psychological theory.

2 Detecting Negative Emotional Reactions

The first step to understanding negative emotional reactions is learning to detect their presence. This is a challenging problem, since general distress can be a highly personal and subjective experience, and there is little labeled data to support machine learning models. Our first section details completed and proposed work on detecting negative emotional reactions from text, including distress and grief. We introduce new datasets for these problems and develop models that use multi-task learning to predict negative emotion in an interpretable way, by modeling extra tasks like general emotion detection that contribute to negative emotional reactions.

2.1 Related Work

Emotion detection is a mature subfield of natural language processing (Canales and Martínez-Barco, 2014; Seyeditabari et al., 2018; Poria et al., 2020). Existing work typically classifies input text into some number of emotion categories based on the emotions the author is thought to express. Emotion label schemes usually come from psychological research (Ekman, 1992; Plutchik, 2001; Russell, 1980; Mehrabian and Russell, 1974), schemes which are typically not collected with textual evidence in mind. In this work, we focus on detecting the emotions and moods that are known to co-occur with distress, especially those that may help explain it (e.g., grief may be identified because certain emotions are detected along with certain behaviors). Emotion intensity prediction is a related problem in this area (Mohammad and Bravo-Marquez, 2017a,b; Mohammad et al., 2018; Yu et al., 2015; Wang et al., 2016; Buechel and Hahn, 2017; Zhu et al., 2019; Navas Alejo et al., 2020), where not only the emotion category but also a numerical estimate of the emotion's intensity is predicted.

Some researchers have previously examined moods, which are considered less intense but longer-lasting than emotions (Luomala and Laaksonen, 2000). These research problems are often framed as classification problems similar to emotion (Balog et al., 2006; Mishne and Rijke, 2006; Chen et al., 2010; Nguyen, 2010; Choudhury et al., 2012; Alam et al., 2016). We find that the literature does not always make a strong distinction between moods and emotions, and thus some research defines moods that we would consider emotions by our definitions.

The first psychological mood we examine is stress, an emotional feeling of strain or pressure resulting from an individual's uncertainty about their capabilities to meet the demands of a changing situation (Selye, 1956; Lazarus and Folkman, 1984). Stress detection is well-studied in the psychological sciences (Giannakakis et al., 2019), where researchers broadly study many types of stress. We specifically study distress, which is stress with negative valence (some events, like a job

interview or wedding, may be stressful but also positive). Researchers in the psychological disciplines often look for biological markers of stress, such as heart rate or cortisol levels; a number of markers are significantly correlated with stress but are inconvenient for stress detection on a large scale, as they require additional hardware or data. Computational research has also examined distress detection in non-text modalities (Zuo et al., 2012; Kumar et al., 2020; Lefter et al., 2016). Work that does examine distress in text often makes use of additional modalities that are not always available or diagnostic tools which require interaction with the posters (Winata et al., 2018; Lin et al., 2017; Guntuku et al., 2018).

The next negative emotional reaction we study is grief, which sometimes appears in the computational literature as part of an emotion detection problem (Abdul-Mageed and Ungar, 2017; Demszky et al., 2020) because it is included in Plutchik's wheel of emotion as the most intense form of sadness (Plutchik, 2001). However, grief is uniquely situated as an emotion with a highly specific and identifiable trigger event: grief is a reaction to loss, including but not limited to the death of a loved one. Some researchers have studied loss in a computational form, such as Blevins et al. (2016); Chang et al. (2018). We also find some computational research on grief in the psychological literature (Laricchiuta et al., 2018), with a small sample size but expert-annotated data.

In this work, we make use of the multi-task learning paradigm (Li et al., 2019; Caruana, 1993; Duong et al., 2015), where one set of model parameters is trained to perform multiple tasks. Multi-task learning has been successfully applied to many domains across NLP (Sun et al., 2019; Kiperwasser and Ballesteros, 2018; Liu et al., 2019b; Xu et al., 2018), improving the performance of some target task by training on related but distinct tasks.

Models for sensitive applications, especially those related to emotions, crucially must be able to explain their predictions and thus, interpretability has recently become a popularly desirable goal across a range of tasks. Many existing frameworks use post-hoc methods to explain otherwise non-explainable models (Ribeiro et al., 2016; Kumar and Talukdar, 2020; Lei et al., 2016), while others bake interpretability directly into their model design (Chen et al., 2018a; Hase et al., 2019). Many other frameworks blur the line between the two types by probing the internals of seemingly uninterpetable models (Thorne et al., 2019; Lundberg and Lee, 2017; Liu et al., 2019a; Serrano and Smith, 2019; Bach et al., 2015; Chen et al., 2020). Some work has specifically examined multi-task neural networks, suggesting they may have enhanced potential for interpetability compared to black-box models (Wang et al., 2020; Chaplot et al., 2019; Tang et al., 2020).

Domain	Subreddit Name	Total Posts	Avg Tokens/Post	Labeled Segments
abuse	r/domesticviolence	1,529	365	388
	r/survivorsofabuse	1,372	444	315
	Total	**2,901**	**402**	**703**
anxiety	r/anxiety	58,130	193	650
	r/stress	1,078	107	78
	Total	**59,208**	**191**	**728**
financial	r/almosthomeless	547	261	99
	r/assistance	9,243	209	355
	r/food_pantry	343	187	43
	r/homeless	2,384	143	220
	Total	**12,517**	**198**	**717**
PTSD	r/ptsd	4,910	265	711
social	r/relationships	107,908	578	694
All		**187,444**	**420**	**3,553**

Table 1: We include ten total subreddits from five domains in Dreaddit. We endeavor to label a comparable amount of data from each domain for training and testing.

2.2 Work to Date: Distress

In this section, we describe our completed work on distress detection from text, including our collection of a new social media dataset, Dreaddit (subsubsection 2.2.1), our models for distress prediction (subsubsection 2.2.2), our results (subsubsection 2.2.3), and a study of our models' interpretability (subsubsection 2.2.4).

2.2.1 Data

2.2.1.1 Datasets

We collect our own dataset, **Dreaddit**, for distress detection and leverage existing data (GoEmotions) and data from our collaborators (Vent) for emotion detection.

Dreaddit. Reddit is a social media website where users post in topic-specific communities (*subreddits*), and other users comment and vote on these posts. The lengthy nature of these posts makes Reddit an ideal source of data for studying the nuances of phenomena like distress. To collect expressions of distress, we select categories of subreddits where members are likely to discuss stressful topics:

- **Interpersonal conflict**: abuse and social domains. Posters in the abuse subreddits are largely survivors of an abusive relationship or situation sharing stories and support, while posters in the social subreddit post about any difficulty in a relationship and seek advice for how to handle the situation.

- **Mental illness**: anxiety and Post-Traumatic Stress Disorder (PTSD) domains. Posters in these subreddits seek advice about coping with mental illness and its symptoms, share support and successes, seek diagnoses, and so on.
- **Financial need**: financial domain. Posters in the financial subreddits generally seek financial or material help from other posters.

We include ten subreddits in these five domains, as detailed in Table 1, and our analysis focuses on the domain level. The dataset contains mostly first-person narrative accounts of personal experiences and requests for assistance or advice. Our data displays a range of topics, language, and agreement levels among annotators.

We annotate a subset of the data for distress using Amazon Mechanical Turk. We partition the posts into contiguous five-sentence chunks for labeling; we wish to annotate segments of the posts because we are ultimately interested in what parts of the post depict distress, but we find through manual inspection that some amount of context is important. However, it would be difficult for annotators to read and annotate entire posts, as they are quite long.

We set up an annotation task in which English-speaking Mechanical Turk Workers are asked to label five randomly selected text segments (of five sentences each) with binary distress labels. We specifically ask Workers to decide whether the author is expressing both stress and a negative attitude about it, not whether the situation itself seems stressful.

We submit 4,000 segments, sampled equally from each domain and uniformly within domains, to Mechanical Turk to be annotated by at least five Workers each. After quality control, we are left with 3,553 labeled data points from 2,929 different posts. We take the annotators' majority vote as the label for each segment. The resulting dataset is nearly balanced, with 52.3% of the data labeled stressful. Our agreement on all labeled data is $\kappa = 0.47$, using Fleiss's Kappa (Fleiss, 1971). We observe that annotators achieved perfect agreement on 39% of the data, and for another 32% the majority was 3/5 or less.[2] This suggests that our data displays significant variation in how distress is expressed.

Emotion Datasets. Because our collected Dreaddit 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 includes 58,009 Reddit comments labeled by crowd workers with one or more of 27 emotions (or Neutral), for its large size and genre similarity to Dreaddit. We will 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 into the Ekman 6 basic emotions (anger, disgust, fear, joy, sadness, surprise, neutral) (Ekman, 1992) (GoEmotions$_{Ekman/E}$), and one into

[2]It is possible for the majority to be less than 3/5 when more than 5 annotations were solicited.

Domain	"Negemo" %	"Negemo" Coverage	"Social" %	"Anxiety" Coverage
Abuse	2.96%	39%	12.03%	58%
Anxiety	3.42%	37%	6.76%	62%
Financial	1.54%	31%	8.06%	42%
PTSD	3.29%	42%	7.95%	61%
Social	2.36%	38%	13.21%	59%
All	2.71%	62%	9.62%	81%

Table 2: Results from our domain analysis using LIWC word lists. Each term in quotations refers to a specific word list curated by LIWC; percentage refers to the percent of words in the domain that are included in that word list, and coverage refers to the percent of words in that word list which appear in the domain.

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 where users post vents of any length, tag them as they like, and other users react to them or post comments. The benefit of Vent is that posters self-identify some emotion they are feeling from a large list of pre-made emotions. We select Vent data that has been labeled with fear or sadness, which we hypothesize to be related to distress, 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 genre labels – we subsample the existing "all" dataset to select only data points 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.

2.2.1.2 Dreaddit Data Analysis

While all of Dreaddit has the same genre and personal narrative style, we find distinctions among domains and between stressful and non-stressful data that can influence model design for this dataset. Many of the posters express distress, but we expect that their different functions and stressors lead to differences in how they do so among subreddits. We find that data analysis is an important part of understanding and interpreting our models because models can learn anything and everything from their data. We present completed analysis of Dreaddit, and we propose to complete similar analyses of our future proposed data later on.

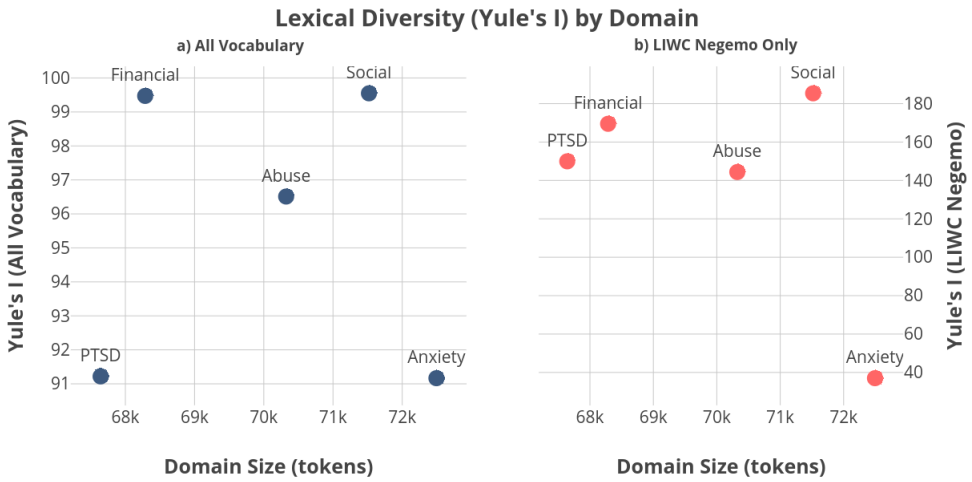


Figure 1: Yule's I measure (on the y-axes) is plotted against domain size (on the x-axes) and each domain is plotted as a point. Graph a) measures the lexical diversity of all words in the vocabulary, while b) includes only words from LIWC's negative emotion word list.

By domain. We examine the vocabulary patterns of each domain using our training data only. First, we use the word categories from the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a lexicon-based tool that gives scores for psychologically relevant categories such as sadness or cognitive processes, as a proxy for topic prevalence and expression variety. We calculate both the percentage of tokens per domain which are included in a specific LIWC word list, and the percentage of words in a specific LIWC word list that appear in each domain ("coverage" of the domain).

Results of the analysis are highlighted in Table 2. We first note that variety of expression depends on domain and topic. We also see clear topic shifts among domains: for example, we observe varying coverage of the anxiety word list.

We also examine the lexical diversity of each domain by calculating Yule's I measure (Yule, 1944). Figure 1 shows the lexical diversity of our data, both for all words in the vocabulary and for only words in LIWC's "negemo" word list. Yule's I measure reflects the repetitiveness of the data (as opposed to the broader coverage measured by our LIWC analysis). We notice exceptionally low lexical diversity for the mental illness domains, which we believe is due to the structured, clinical language surrounding mental illnesses. When we restrict our analysis to negative emotion words, this pattern persists only for anxiety.

By label. We perform similar analyses on data stratified according to its label. We confirm some common results in the mental health literature, including that

Label	1st-Person %	"Posemo" %	"Negemo" %	"Anxiety" Cov.	"Social" %
Stress	9.81%	1.77%	3.54%	78%	8.35%
Non-Stress	6.53%	2.78%	1.75%	67%	11.15%

Table 3: Results from our label analysis using LIWC word lists, with the same definitions as in Table 2. First-person pronouns use the LIWC "I" word list.

Measure	Stress	Non-Stress
% Conjunctions	0.88%	0.74%
Tokens/Segment	100.80	93.39
Clauses/Sentence	4.86	4.33
F-K Grade	5.31	5.60
ARI	4.39	5.01

Table 4: Measures of syntactic complexity by label.

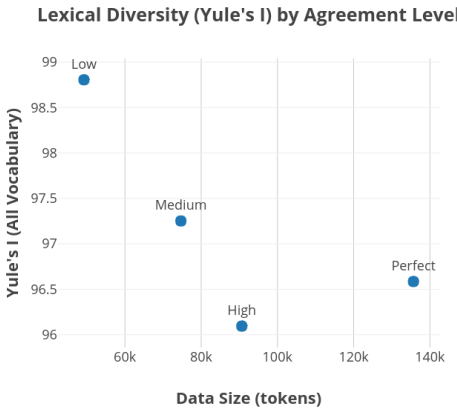


Figure 2: Yule's I measure (on the y-axis) is plotted against domain size (on the x-axis) for each level of annotator agreement. Perfect means all annotators agreed; High, 4/5 or more; Medium, 3/5 or more; and Low, everything else.

distressed data uses more first-person pronouns (perhaps reflecting increased self-focus) and that non-distressed data uses more social words (perhaps reflecting a better social support network).

Additionally, we calculate measures of syntactic complexity, including the percentage of tokens that are conjunctions, average number of tokens per labeled segment, average number of clauses per sentence, Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Automated Readability Index (Senter and Smith, 1967). These scores are comparable for all splits of our data; however, as shown in Table 4, we do see non-significant but persistent differences between distressed and non-

distressed data, with distressed data being generally longer and more complex but also rated simpler by readability indices.

By agreement. Finally, we examine the differences among annotator agreement levels. We find an inverse relationship between the lexical variety and the proportion of annotators who agree, as shown in Figure 2. Yule's I measure controls for corpus size, so we believe that this trend reflects a difference in the type of data that encourages high or low agreement.

2.2.2 Methods

We frame the distress detection problem as a binary classification problem and present several models for this task. We perform preliminary experiments when we collect the data, which yield a best non-neural classifier comparable to the neural state of the art. We then present several multi-task models that make use of emotional information to predict distress in an interpetable way with no loss of performance.

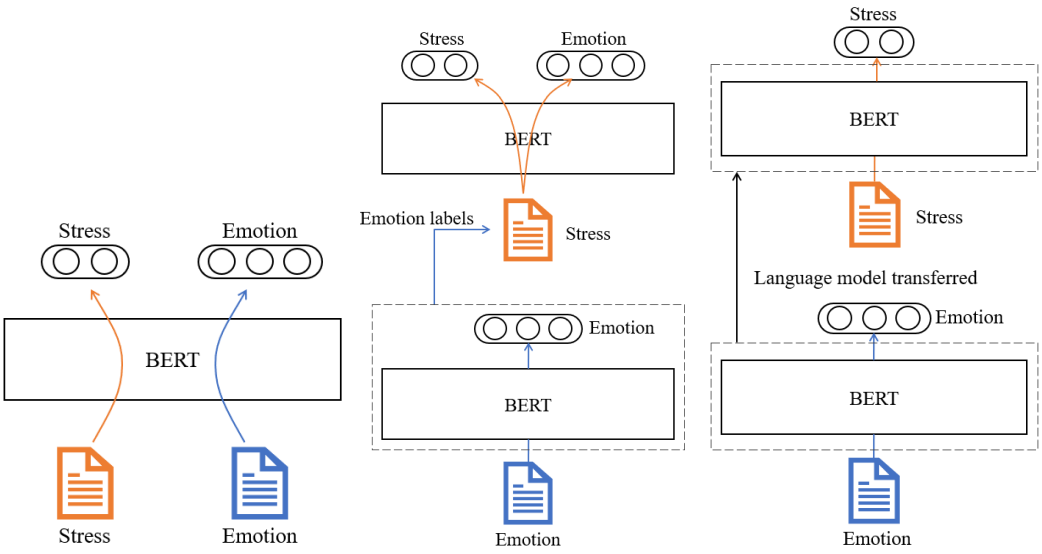
Linear Classifiers. We experiment with a group of linear classifiers and a variety of word representations and discrete features to explore the data. These models include Support Vector Machines (SVMs), logistic regression, Naïve Bayes, Perceptron, and decision trees. We experiment with a variety of feature sets, including bag-of-words representations, pre-trained word embeddings (Mikolov et al., 2013; Devlin et al., 2019), lexical features derived from LIWC and the Dictionary of Affect in Language (Whissel, 2009), syntactic features based on sentence complexity, and social features related to social network engagement. We experiment with various subsets of these features, as well as with limiting the training data by annotator agreement, to discover which ones are most beneficial.

BERT Baseline. In our preliminary work, we also apply BERT directly to our task, fine-tuning the pretrained BERT-base[3] on our classification task for three epochs (as performed in Devlin et al. (2019) when applying BERT to any task).

Emotion-Infused Models. With the preliminary work completed, we also experiment with three types of *emotion-infused* distress detection models that incorporate emotion information: alternating multi-task, multi-task, and fine-tuning.

Alternating Multi-Task Models. Our first set of multi-task models, which we refer to as Multi^{Alt}, are multi-task in the sense that two distinct classification layers (one for distress and one for emotion) share the same representation. These models are alternating in that we train them with two different datasets with two different sets of labels–i.e., we train the distress 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$^{Alt}_{GE_A}$, Multi$^{Alt}_{GE_E}$, etc. The Multi^{Alt} models can be seen in

[3] `https://github.com/huggingface/transformers`



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

Figure 3: The emotion-informed architectures we use in our experiments.

Figure 3a. 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$_{A,E,S}$).

Multi-task Models. We also experiment with a classical 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 Dreaddit is labeled only with distress, 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 distress. The Multi model can be seen in Figure 3b. Since distress detection is our main task in this work, we focus on this task where we have gold labels for distress, but note that it will be interesting in future work to experiment with other task settings, such as whether distress detection can improve emotion classification. In these models, the training losses of the distress task and the emotion task are summed together for each batch with a tunable weight parameter, i.e., $\mathfrak{L} = \lambda\mathfrak{L}_{\text{distress}} + (1-\lambda)\mathfrak{L}_{\text{emotion}}$.

Fine-Tuning Models. We also 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 distress 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 pa-

Model	P	R	F
Majority baseline	51.61	100.0	68.08
LogReg w/ domain Word2Vec + features*	74.33	83.20	79.80
BERT-base*	75.18	86.99	80.65

Table 5: Precision (P), recall (R), and F1-score (F) for our baseline models. "Features" is our best-performing feature set. * indicates significance over the majority baseline (approximate randomization test, $p < 0.01$).

rameters and continue to fine-tune on Dreaddit. We denote these models as, e.g., Fine-Tune$_{GE_A \rightarrow Dreaddit}$ for a model that was first trained on GoEmotions$_{all}$ and then on Dreaddit. These fine-tuning models can be seen in Figure 3c.

2.2.3 Results

We present the results of our baselines in Table 5. Our best linear model is a logistic regressor with Word2Vec embeddings trained on our unlabeled corpus, high-correlation features (≥ 0.4 absolute Pearson's r with the training data), and high-agreement data (at least 4/5 annotators agreed); this model achieves an F-score of 79.8 on our test set, a significant improvement over the majority baseline by the approximate randomization test ($p < 0.01$). The high-correlation features used by this model are LIWC's clout, tone, and "I" pronoun features. We find that this logistic regression classifier achieves comparable performance to BERT-base (approximate randomization test, $p > 0.5$) with the added benefits of increased interpretability and less intensive training.

We also perform an error analysis of these two nontrivial baselines. Although the dataset is nearly balanced, both BERT-base and our best logistic regression model greatly overclassify distress, and they broadly overlap but do differ in their predictions (disagreeing with one another on approximately 100 instances).

We note that the examples misclassified by both models are often, though not always, ones with low annotator agreement (with the average percent agreement for misclassified examples being 0.55 for BERT and 0.61 for logistic regression). Both models seem to have trouble with less explicit expressions of distress, framing negative experiences in a positive or retrospective way, and stories where another person aside from the poster is the focus; these types of errors are difficult to capture with the features we used (primarily lexical).

Emotion-Infused Models. In our work on emotion-infused models, we present our results as an average of 3 distinct runs with distinct random seeds. Because of this, we re-implement our BERT-base model for comparison.

Model	Binary F1	Accuracy
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

(a) Results of our multitask models.

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
$\text{FT}_{GE_{FSJ} \to Dr}$	$\mathbf{80.25} \pm 0.24$	$\mathbf{80.98} \pm 0.20$

(b) Results of our fine-tuning and pre-training models.

Table 6: Our emotion-infused models' results. The best result under each metric is bolded. GE is GoEmotions, Dr is Dreaddit, and FT is Fine-Tune.

We report the results of our multi-task models in Table 6a[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, $\text{Multi}^{Alt}_{GoEmo_E}$ performs best. The 28-way classification of GoEmotions_A naturally leads to lower numerical performance than the tasks with fewer classes, and we expect that GoEmotions_S may group too many distinct emotions together under the same emotion labels. We also note that the $\text{Multi}^{Alt}_{Vent}$ model performs equally well, which indicates that genre mismatch is not an issue for this problem, or that Vent and Reddit have sufficient genre similarity. Somewhat surprisingly, $\text{Multi}^{Alt}_{GoEmo_{FSJ}}$ 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 6b. 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 and not Vent. Here, we observe that our best model, $\text{Fine-Tune}_{GoEmo_{FSJ} \to Dreaddit}$, scores at least one standard deviation above BERT. We see higher increases in performance for the simpler classification problems GoEmotions_S and GoEmotions_{FSJ} and worsened performance forGoEmotions_A, suggesting that in the sequential paradigm, more complex problems are not able to interact appropriately with the main task and instead interfere.

Overall, the inclusion of emotion information results in modest improvements, even though not statistically significant, as compared to BERT. However, our true

[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 distress + pred. emotion)	0.3396	0.2554	0.0565	0.3207
GoEmotions (gold emotion + pred. distress)	0.1274	0.2668	0.2786	0.4115

Table 7: Correlations of the gold labels for each dataset with labels predicted by the other classifier in a Multi^{Alt} model. GoEmotions$_{FSJ}$ uses the correlation ratio η, while the other GoEmotions variants use the coefficient of determination R^2.

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

goal is to analyze the explainability of our models, to which we turn next.

2.2.4 Interpretability

We perform several analyses to probe our emotion-infused 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 distress 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.

We perform an analysis of our Multi^{Alt} models to see what information they learn about emotion[5]. We take the development sets of both Dreaddit and GoEmotions and predict labels for 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 7. In this case, the GoEmotions$_{FSJ}$ variant is a single-label 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 8.

We observe that our multi-task models learn a moderate correlation between

[5]We did perform an equivalent analysis on the Multi models, which show similar trends.

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%

Table 9: A comparison of how often several of our models rely on words from several LIWC categories to make their decisions on the dev set. These numbers represent the percentage of available LIWC words each model selected in the top 10 LIME explanations. Dr is Dreaddit, GE is GoEmotions, and FT is Fine-Tune.

the distress labels and the emotion labels; they learn that negative emotions like fear and sadness are linked to distress and neutral or positive emotions are linked to non-distress. These emotion predictions can help explain the distress 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 distress classifier detects distress. From a machine learning perspective, this correlation also suggests the potential for using emotion data as distantly-labeled distress data to supplement the small extant distress datasets.

We also investigate the information each model is using to make its decisions. In this section, we again use LIWC to categorize the information our different models use to predict distress.

We first analyze our models' most important unigrams using LIME. LIME accepts an input from our development set, perturbs it in the bag-of-unigrams space, and runs our classifiers on each perturbation to calculate the importance of various unigrams. We acquire the 10 unigrams with the highest magnitude output by LIME for each data point 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 labeling decisions for distress. An abbreviated list of results showing our best models from each category, is shown in Table 9. We observe consistent effects suggesting that, in comparison to plain BERT, our emotion-enhanced models 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 for Multi$^{Alt}_{GoEmotions_E}$, which makes 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, $\text{Multi}_{GoEmotions_{FSJ}}$ uses Certainty twice as often as BERT. These cognitive words seem to target the mental aspects of distress. Rumination and a focus on absoluteness are known signs of anxiety disorders, an extreme form of chronic distress (Nolen-Hoeksema et al., 2008; Miranda and Mennin, 2007).

In conclusion, we find our emotion-infused models focusing more on emotional information to make predictions of distress than our baseline BERT models without any loss of performance. These results are promising for the development of models that focus on information that humans consider intuitive.

2.3 Proposed Work: Grief and Related Emotions

With our work on distress detection completed, we propose to turn next to another type of negative emotional reaction: grief. Grief is a reaction to loss (such as a death, the loss of a job, the end of a relationship, etc.), and where distress tends to be short-lived, grief is typically expected to last for a long time. We propose to apply similar methods to grief as we used for distress, with a particular focus on whether knowledge of loss can improve our models.

2.3.1 Data

For our work on grief detection, we propose to collect a new dataset with a platform we call *ieso*: Integrating Emotional Stories Online. This platform will recruit participants from Columbia University and the New York City area to write journal-like posts about their emotional experiences and support their peers anonymously.

On ieso, users will create posts whenever they wish about whatever topic they wish, although our recruitment focuses particularly on negative emotional reactions. Users will write text describing their feelings, as well as rate several emotions on a scale from 1-10. In particular, we will include "stressed" (relevant for comparison with our previous work) and "lonely" (which we hypothesize will be relevant to grief), as well as all of the Ekman basic emotions, for comparison with existing resources. Thus, the emotion set of interest will be *stressed*, *lonely*, *guilty*, *sad*, *angry*, *afraid*, *surprised*, *disgusted*, and *happy*. Posters will additionally have the opportunity to describe an event that caused their emotions in writing, identify broadly when this event happened, and tag their posts as being relevant to some large-scale topic like the COVID-19 pandemic. We wish to highlight that the authors are expected to provide their own emotion labels for their own text; we expect this to help mitigate bias introduced when annotators (who may have different experiences and perhaps demographics from the writers) annotate the text of other, unknown writers for subjective states like emotion.

Finally, expert annotators from the Columbia University School of Social Work will label this data for several dimensions as part of an existing collaboration. In particular, these annotators will label whether or not the person who wrote the post is experiencing grief.

We expect this platform to yield a small dataset of 1,000-2,000 posts. Thus, once again we will also look to external datasets for possibly useful training data. We will again examine the utility of GoEmotions (which has an infrequently used grief label) and Vent (which includes some emotions which may be expected to co-occur with grief, like loneliness). We also expect that the aggression-loss data collected by Blevins et al. (2016) may be useful, as loss is the triggering event for grief.

We propose to complete an analysis of this data similar to our completed analysis of Dreaddit (paragraph 2.2.1.2), examining what types of words are important in determining emotion and grief (using tools such as LIWC) and what concepts we can see expressed in the data.

2.3.2 Problem Description

With this dataset, we propose to explore two problems for which we will have labeled data. First, our primary target task will be **grief detection**, which we will frame as a binary classification task in accordance with the annotation; the input to this task will be the poster's description of their feelings concatenated with their description of the triggering event, if any. Second, we will also perform **emotion intensity prediction**, a multi-regression problem where we will output emotion scores from 0 to 1 based on the nine emotions which the posters of the ieso data will self-label. In such a way, we hope to extend our techniques developed for multi-task distress prediction to other types of negative emotional reactions. We will evaluate both problems on a held-out test subset of the ieso data.

2.3.3 Proposed Models

We propose to develop a series of models to perform binary grief prediction and normalized emotion intensity regression. Our baselines will naturally include existing state-of-the-art models for both of these tasks individually, (as we are not aware of any existing models for grief detection, we will apply popular multi-purpose models, such as fine-tuning BERT (Devlin et al., 2019)) and our proposed models will attempt to combine them, following our work on distress prediction.

A feature of this dataset will be that both tasks will be labeled for the same data, so we will develop a multi-task neural model which predicts both tasks at once, as in our Multi models for distress prediction. We will additionally develop Multi^{Alt} models for each task, where we will experiment with additional tasks on other labeled data such as emotion classification (drawing on existing datasets which

may include grief as one of many emotions (Abdul-Mageed and Ungar, 2017), and our Vent data which includes emotion labels such as *lonely* and *resentful*), emoji prediction, event detection (since the emotions elicited on ieso may be caused by particular events), and loss prediction (using the loss-aggression dataset of Blevins et al. (2016)). We believe that implicit emotion is an important extension of our work; therefore, we also propose three possible tasks that may incorporate this information: common-sense reasoning (Sap et al., 2018), implicit emotion itself (Mohammad and Bravo-Marquez, 2017b), and connotation prediction (Allaway and McKeown, 2021)). We will also experiment with combining both models to create a Multi-Enhanced model, which predicts grief and emotion intensity on the main ieso dataset alternately with additional tasks on other labeled data. Finally, we also propose to experiment with a Holistic model that performs the additional tasks, such as emotion classification, separately, and then uses its predictions as input to the grief detection task, which we propose would result in a more clearly interpretable model.

In order to select the emotions for our emotion classification tasks, we will draw on psychological literature about grief (Bonanno, 2009; Arizmendi and O'Connor, 2015; Brinkmann and Kofod, 2018); this literature suggests that useful emotion labels may include guilt, anger, sadness, surprise, and joy (Institute of Medicine, 1984, Chap. 3; Zisook and Shear, 2009; Shear et al., 2011) which we include in our label set for ieso. However, in order to validate the use of these emotions in our online textual setting, we will also examine our ieso data to find correlations between user-reported emotion scores and professionally-assigned grief labels. Since most existing emotion classification datasets are labeled by external annotators and not the author of the data, we will also conduct an annotation study to label some of our ieso data with binary emotion labels from an outside perspective and examine correlations in that setting. This will also give us an interesting side problem to evaluate, namely, the extent to which self-reported emotion labels correlate with annotator labels. We also anticipate our collaborators from the School of Social Work being able to provide some analysis of the emotions felt by our participants based on interviews with a number of them.

2.3.4 Interpretability

We propose to examine these models for "humanlike" interpretability first in similar ways to our work on distress classification. We will examine whether the additional tasks (emotion classification, loss classification, etc.) correlate at all with the main tasks (grief classification and emotion intensity regression) and whether the main tasks correlate with each other. We will also experiment with rationale words, similar to our LIME analysis for distress prediction, using some similar contemporary methods for model interpertation (e.g., Chen et al. (2020); etc.).

Because grief is an emotional reaction to a loss, our intuition is that some understanding of these two components (loss and emotion) should be vital for the ability to classify grief. We expect that a model that predicts grief in an interpretable way should also be able to predict these aspects accurately. Therefore, we also propose that the performance on additional loss and emotion tasks will be a form of interpretability, especially for our Holistic model.

3 Emotions in Reaction to Events

Building on our proposed work in grief prediction, where grief is a specific emotion triggered by a specific loss event, we now take the view that all emotions are triggered by some event or stimulus (Schachter and Singer, 1962) and explicitly model these stimuli. In this section, we review our completed work on identifying the cause of an emotion as a span in a sentence and detail proposed work on modeling the cause of a negative emotional reaction as an event to be detected.

3.1 Related Work

Given our perspective above, emotion has a *cause* which triggers it, and this can be expressed as some semantic role related to the detected emotion. Comparatively few researchers have looked at the semantic roles related to emotion. The largest corpus of emotion-cause work to date is a Chinese dataset compiled by Gui et al. (2016). This dataset characterizes the emotion and cause detection problems as two clause-level selection problems, where individual clauses are identified as containing the emotion or the cause. A large amount of work has been done using this dataset and problem formulation (Chen et al., 2018b; Xia and Ding, 2019; Xia et al., 2019; Fan et al., 2020; Wei et al., 2020; Ding et al., 2020).

Some previous work has also addressed the cause detection problem from a span-tagging perspective, the problem formulation we use for this work, often using small, hand-crafted data (Russo et al., 2011; Ghazi et al., 2015). Some datasets of this type have begun to include roles and entities aside from the emotion itself, such as the experiencer, cause, or target (Kim and Klinger, 2018). In this area, we particularly highlight the work of Bostan et al. (2019), who developed a dataset of 5,000 news headlines annotated with various semantic roles including cause, and tested these tasks with a BiLSTM-CRF model.

There is a small body of computational work which explicitly calls out events as a contributor to emotions, rather than trusting that generalized representations will encode such information. Oberländer et al. (2020) particularly explore the use of annotated semantic roles for emotion detection (i.e., whether roles like the stimulus are useful when included and when explicitly marked) and find them helpful to

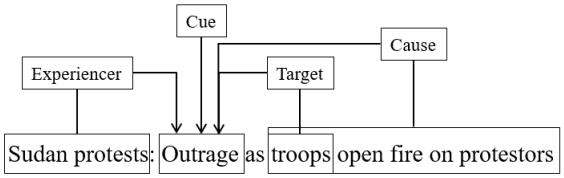


Figure 4: An example of the semantic roles annotated by Bostan et al. (2019)

understanding what emotion is being expressed. Other work has used event words as features for classification (Li et al., 2016a); a shared task has also focused on using stimuli descriptions to predict implicit emotions (Klinger et al., 2018). Some work also uses this relationship between events and emotions as a lens through which to view problems like commonsense reasoning (Rashkin et al., 2018). Finally, some prior work has looked at how emotions change over time and whether this is correlated with particular events (Mishne and Rijke, 2006; Balog et al., 2006), although this work typically analyzes reported or predicted emotion trends for spikes in some emotion correlating with some event rather than attempting to directly predict a causal relationship.

We also highlight work on affective events, which attaches some emotional appraisal to events, rather than necessarily finding an emotion and then finding its triggering event (Riloff et al., 2013; Ding and Riloff, 2016; Zhuang et al., 2020); we believe this is a complimentary perspective on the problem of emotion cause detection. Events in this work are expressed as event triples (subject, verb, object) extracted via dependency parsing; they are assigned sentiment polarity scores through methods such as semi-supervised learning based on how they affect their participants. Some work has also looked into understanding the reasons for these polarities using information such as psychological theory (Ding et al., 2018; Ding and Riloff, 2018; Ding and Feng, 2020). In terms of topic prediction, the most similar work to ours includes the MuSe-Topic subchallenge of the MuSe-2020 Challenge (Stappen et al., 2020), which asks participants to predict bucketed emotion intensity and topic identity for multimodal data.

3.2 Work to Date: Emotion and Cause

3.2.1 Data

For our completed experiments, we use the GoodNewsEveryone corpus (Bostan et al., 2019), which contains 5,000 news headlines labeled with emotions and semantic roles such as the target, experiencer, and cause of the emotion, as shown in Figure 4. We focus on the emotion detection and cause tagging tasks.

In our completed experiments, we limit ourselves to the data points for which a cause span was annotated (4,798). We also note that this dataset uses a 15-way emotion classification scheme, an extended set including the eight basic Plutchik emotions as well as additional emotions like *shame* and *optimism*. While a more fine-grained label set is useful for capturing subtle nuances of emotion, many external resources focus on a smaller set of emotions. Therefore, for our work, we choose to limit ourselves to the six Ekman emotions (*anger*, *fear*, *disgust*, *joy*, *surprise*, and *sadness*). We also choose to keep *positive surprise* and *negative surprise* separated, to avoid severely unbalancing the label distribution. The remaining data, which we use for our experiments, numbers 2,503 data points.

3.2.2 Models

For the GoodNewsEveryone data, we approach the cause detection problem as a sequence tagging problem using the IOB scheme (Ramshaw and Marcus, 1995): $\mathcal{C} = \{\text{I-cause}, \text{O}, \text{B-cause}\}$. We tackle the emotion detection task as a seven-way classification task with $\mathcal{E} = \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{negative surprise}, \text{positive surprise}\}$.

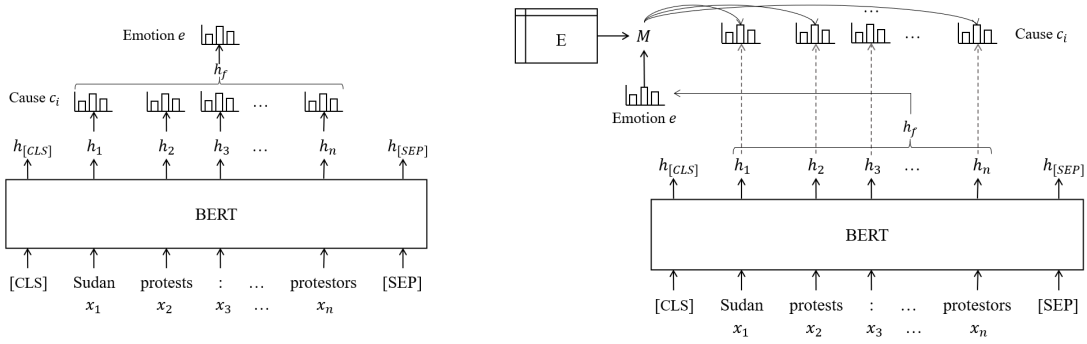
3.2.2.1 Single-Task Models

As a baseline, we train single-task models for each of emotion classification and cause span tagging. We use a pre-trained BERT language model (Devlin et al., 2019), which we fine-tune on our data, as the basis of this model. For a sequence of n WordPiece tokens, our input to the BERT model is a sequence of $n+2$ tokens, $X = [[\text{CLS}], x_1, x_2, ...x_n, [\text{SEP}]]$, where [CLS] and [SEP] are BERT's begin and end tokens. Passing X through BERT yields a sequence of vector hidden states $H = [h_{[CLS]}, h_1, h_2, ..., h_n, h_{[SEP]}]$ with dimension d_{BERT}. For emotion classification, we pool these hidden states and allow hyperparameter tuning to select the best type: selecting the [CLS] token ($h_f = h_{[CLS]}$), mean pooling ($h_f = \frac{\sum_{i=1}^{n} h_i}{n}$), max pooling ($h_{f,j} = \max h_{i,j}$), or attention as formulated by Bahdanau et al. (2015):

$$h_f = \sum_{i=1}^{n} \alpha_i h_i \tag{1}$$

where $\alpha_i = \frac{\exp(W_a h_i + b_a)}{\sum_{j=1}^{n} \exp(W_a h_j + b_a)}$ for trainable weights $W_a \in \mathbb{R}^{1 \times d_{BERT}}$ and $b_a \in \mathbb{R}^1$. Then, the final distribution of emotion scores is calculated by a single dense layer and a softmax:

$$e = \text{softmax}(W_e h_f + b_e) \tag{2}$$



(a) The Multi$_{C\to E}$ model.

(b) The Multi$_{E\to C}$ model.

Figure 5: The architectures of our multitask models.

with $e \in \mathbb{R}^{|\mathcal{E}|}$ and for trainable parameters $W_e \in \mathbb{R}^{|\mathcal{E}| \times d_{BERT}}$ and $b_e \in \mathbb{R}^{|\mathcal{E}|}$. For cause tagging, a tag probability distribution is calculated on each hidden state:

$$c_i = \text{softmax}(W_c h_i + b_c) \tag{3}$$

3.2.2.2 Multi-Task Models

We present three multi-task models to test our hypothesis that the emotion detection and cause tagging tasks can inform one another. For all multi-task models, we use the same base architecture (BERT) as the single models. Additionally, for these models, we combine the losses of both tasks and weight them with a tunable weight parameter: $\lambda \text{NLL}_{\text{emo}} + (1 - \lambda)\text{NLL}_{\text{cause}}$.

Multi. The first model, Multi, is the classical multi-task learning framework with hard parameter sharing, where both tasks share BERT layers. Two dense layers for emotion classification and cause tagging operate at the same time from the same BERT layers, and we train both of the tasks simultaneously. That is, we calculate emotion scores e and cause tag scores c from the same hidden states H.

We further develop two additional multi-task models with the intuition that we can design more explicit and concrete task dependencies than simple parameter sharing in the representation layer.

Multi$_{C\to E}$. We assume that if a text span is given as the cause of an emotion, it should be possible to classify that emotion while looking only at the words of the cause span. Therefore, we propose the Multi$_{C\to E}$ model, the architecture of which is illustrated in Figure 5a. This model begins with the single-task cause detection model, $BERT_C$, which produces a probability distribution $P(y_i|x_i)$ over IOB tags for each token x_i, where $P(y_i|x_i) = c_i$ from Equation 3. Then, for each token, we calculate the probability that it is part of the cause as $P(\text{Cause}|x_i) = P(B|x_i) + P(I|x_i) = 1 - P(O|x_i)$. We feed the resulting probabilities through a

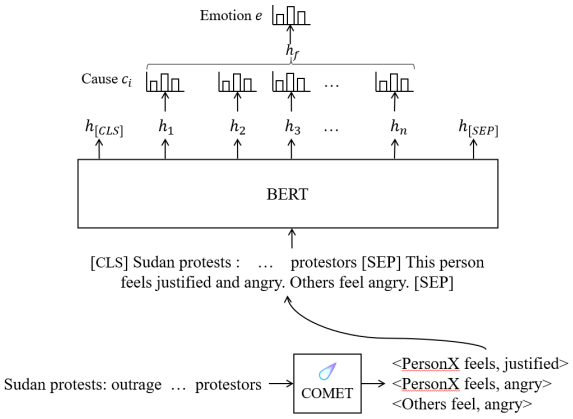


Figure 6: The architecture of our Multi$^{COMET}_{C\rightarrow E}$ model.

softmax and use them as an attention distribution over the input tokens in order to pool the hidden representations and perform emotion classification: attention is computed as in Equation 1, with $\alpha_i = \frac{\exp P(\text{Cause}|x_i)}{\sum_{j=1}^{n} \exp P(\text{Cause}|x_i)}$, and emotion classification as in Equation 2. For the Multi$_{C\rightarrow E}$ model, we apply teacher forcing at training time, and the gold cause spans are used to create the attention weights before emotion classification (which means that $P(\text{Cause}|x_i) \in \{0, 1\}$). At inference time, the model uses the predicted cause span instead.

Multi$_{E\rightarrow C}$. Next, we hypothesize that knowledge of the predicted emotion should help us identify salient cause words. The Multi$_{E\rightarrow C}$ model first performs emotion classification, which results in a probability distribution over predicted emotion labels, as in the BERT$_E$ model and Equation 2. We additionally keep an emotion embedding matrix E, where $E[i]$ is a learnable representation of the i-th emotion label (see Figure 5b) with dimension d_e. We use the predicted label probabilities e to calculate a weighted sum of the emotion embeddings, i.e., $M = \sum_i e_i \cdot E[i]$. We then concatenate M to the hidden representation of each token and perform cause tagging with a final dense layer, i.e., $c_i = \text{softmax}(W_{c'}[h_i; M] + b_{c'})$, where ; is the concatenation operator and $W_{c'} \in \mathbb{R}^{|\mathcal{C}|\times(d_{BERT}+d_e)}$ and $b_{c'} \in \mathbb{R}^{|\mathcal{C}|}$ are trainable parameters. In the Multi$_{C\rightarrow E}$ model, we again do teacher forcing and use the gold emotion labels before computing M (i.e., e is a one-hot vector). At inference time, the model uses the predicted emotion distribution instead.

3.2.2.3 Adapted Knowledge Models

Recent work has shown that fine-tuning pre-trained language models such as GPT-2 on *knowledge graph tuples* such as ConceptNet (Li et al., 2016b) or ATOMIC (Sap et al., 2018) allows these models to express their implicit knowledge directly

(Bosselut et al., 2019). These adapted *knowledge models* (e.g., COMET (Bosselut et al., 2019)) can produce commonsense knowledge on-demand for any entity, relation or event. Considering that common-sense knowledge plays an important role in understanding implicitly expressed emotions and the reasons for those emotions, we explore the use of common-sense knowledge for our tasks, in particular the use of COMET adaptively pre-trained on the ATOMIC event-centric knowledge base. ATOMIC's event relations include "xReact" and "oReact", which describe the feelings of certain entities after the input event occurs. For example, using the headline "Sudan protests: Outrage as troops open fire on protestors", COMET-ATOMIC outputs that PersonX feels justified, PersonX feels angry, Others feel angry, and so on (Figure 6). To use this knowledge model, we modify our approach by reframing our single-sequence classification task as a sequence-pair classification task (for which BERT can be used directly). We feed our input headlines into pre-trained COMET-ATOMIC, collect the top two outputs for xReact and oReact using beam search decoding, and then feed them into BERT alongside the input headlines, as a second sequence using the SEP token. That is, our input to BERT is now $X = [[\text{CLS}], x_1, x_2, ..., x_n, [\text{SEP}], z_1, z_2, ..., z_m, [\text{SEP}]]$, where z_i are the m WordPiece tokens of our COMET output and are preprocessed in the same way as x_i. We hypothesize that, since pre-trained BERT is trained with a next sentence prediction objective, expressing the COMET outputs as a grammatical sentence will help BERT make better use of them, so we formulate this second sequence as complete sentences (e.g., "This person feels... Others feel...") (Figure 6).

This approach allows us incorporate information from COMET into all our single- and multi-task BERT-based models; the example shown in Figure 6 is our $\text{Multi}_{C \rightarrow E}$ model. We refer to the COMET variants of these models as: BERT^{COMET} (single-task models) and Multi^{COMET}, $\text{Multi}^{COMET}_{C \rightarrow E}$, $\text{Multi}^{COMET}_{E \rightarrow C}$ for the three multitask models.

3.2.3 Results

Because the subset of the data we use is relatively small, we present results using the average performance of five models with five fixed random seeds (e.g., reported emotion F1 is the average of the emotion F1 scores for each of our five runs). For our joint models, since our novel models revolve around using one task as input for the other, we separately tune versions of each model, one based on each of the single-task metrics, yielding, for example, one Multi model optimized for predicting emotion and one optimized for predicting cause. We present the results of our models in Table 10. We see that the overall best model for each task is a multi-task adapted knowledge model, with $\text{Multi}^{COMET}_{C \rightarrow E}$ performing best for emotion (which is a statistically significant improvement over BERT by the paired t-test, $p < 0.05$)

	Emotion Macro F1	Emotion Accuracy	Cause Span F1
BERT	37.25 ± 1.30	38.50 ± 0.84	37.49 ± 1.94
BERT^{COMET}	37.74 ± 0.84	38.50 ± 1.14	39.27 ± 1.85
Multi	36.91 ± 1.48	38.34 ± 1.94	38.35 ± 3.89
Multi$_{C\rightarrow E}$	37.74 ± 2.12	38.74 ± 2.07	39.08 ± 3.73
Multi$_{E\rightarrow C}$	38.26 ± 3.28	39.69 ± 3.41	38.83 ± 1.60
Multi^{COMET}	37.06 ± 2.04	39.05 ± 0.98	**39.50** ± 2.25
Multi$^{COMET}_{C\rightarrow E}$	**39.26*** ± 1.13	**40.79** ± 2.17	38.68 ± 1.36
Multi$^{COMET}_{E\rightarrow C}$	37.44 ± 1.37	38.58 ± 1.44	36.27 ± 1.31

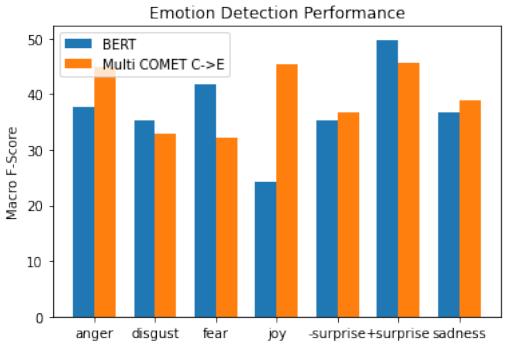
Table 10: The results of our models, averaged over five runs with the same five distinct random seeds. The model with the highest mean performance under each metric is bolded. Results marked with a * are statistically significant above the single-task BERT baseline by the paired t-test ($p < 0.05$).

and Multi^{COMET} performing best for cause. These results seem to support our two hypotheses: 1) emotion recognition and emotion cause detection can inform each other and 2) common-sense knowledge is helpful to infer the emotion and the cause for that emotion expressed in text. Additionally, we note that BERT^{COMET} performs better than BERT for cause, as does Multi, but for emotion, both COMET and multi-task learning seem to be needed for consistent improvement over BERT.

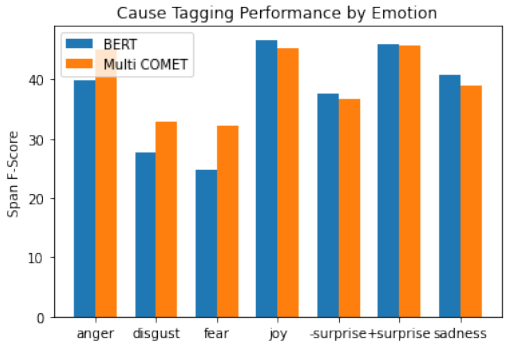
Finally, we also present per-emotion results for our best model for each task (Multi$^{COMET}_{C\rightarrow E}$ for emotion and Multi^{COMET} for cause) against the single-task BERT baselines in Figure 7a and Figure 7b; these per-emotion scores are again the average performance of models trained with each of our five random seeds. We see that each task improves on a different set of emotions: for emotion classification Multi$^{COMET}_{C\rightarrow E}$ consistently improves over BERT by a significant margin on joy and to a lesser extent on anger and sadness. Meanwhile, for cause tagging, Multi^{COMET} improves over BERT on anger, disgust, and fear, while yielding very similar performance on the rest of the emotions. We note that these include emotions we found to be helpful for stress detection (fear, sadness, anger, joy).

3.2.4 Analysis and Discussion

Upon inspection of the GoodNewsEveryone data, we note that annotators' emotion labels can vary significantly depending on what emotional event they choose to annotate. For example, our development set includes the input *Simona Stuns Serena at Wimbledon: Game, Set and "Best Match" for Halep*. Its gold adjudicated emotion label is *negative surprise*, but annotators also included emotion labels such as *joy*, *negative surprise*, *positive surprise*, *pride*, and *shame*, which can be understood as



(a) Performance of the BERT and Multi$^{COMET}_{C \rightarrow E}$ models on emotion classification.



(b) Performance of the BERT and Multi^{COMET} models on cause tagging, broken down by emotion.

Figure 7: The performance of our best models, broken down by emotion class.

various emotions felt by the different entities and even reader.

Inspired by this variation in the data, we also examine our models' performance at selecting emotions that were suggested by some human annotator but not selected as the gold labels. We refer the reader to the appendix for details on this analysis; our conclusions are that language models have general knowledge about emotion already, but common-sense knowledge helps pare down the space of plausible outputs to those that are most commonly selected by human annotators. Further, we see that also predicting the cause of an emotion causes the model to narrow down the space of possible emotion labels to only the most prominent.

In conclusion, we find that the Multi$^{COMET}_{C \rightarrow E}$ model achieves the best performance on emotion detection, an important target for the rest of our work, by improving positive emotion detection and helping the model to narrow in on the gold emotion labels overall. Therefore, we intend to use similar models in our remaining work, incorporating common-sense knowledge and a pipeline structure where some tasks directly inform other tasks.

3.3 Proposed Work: Emotions and Large-Scale Events

Moving forward from characterizing the cause of an emotion as a simple span of text, in this section we view causes as meaningful constructions in their own right and look particularly through the lens of negative emotional reactions. We aim to characterize the events that trigger emotions (especially negative ones) and connect them together into larger concepts and narratives. That is, we now propose to leverage the temporal and topic aspects of our collected ieso dataset to predict how emotions change over time in reaction to large-scale events. Furthermore, in order

to join our work in subsection 2.2 and subsection 2.3 to our work on the causes of emotions, we will leverage the grief and distress labels given to our ieso data to characterize the experiences and triggers of especially these emotions and moods over time.

3.3.1 Problem Description

Given that our ieso dataset will include self-reported emotion scores as well as self-reported triggering events, we propose to predict both of these things (the emotions and the event that triggered them) in tandem, much as in our completed emotion-cause work. In this work, we will express the emotion problem the same as for our grief prediction task, i.e., as a normalized intensity regression problem, and we will frame the event prediction problem as a two-part hierarchical prediction task in which we first predict the topic (such as COVID, #BlackLivesMatter, etc.) as a classification problem and then predict the exact triggering event as a span tagging problem.

For example, suppose we received the following fictitious data from ieso for our dataset in subsection 2.3:

Emotions	I feel so scared and so angry...I can't believe this is happening.
Event	My grandmother has been coughing and having trouble breathing. She just tested positive this morning and I don't know what's going to happen to her.

We might expect the gold self-reported emotion labels to include high scores for fear, anger, and sadness, and naturally the topic would likely be COVID-19. In this setting, we would expect our models to output the correct emotion intensity scores and topic label, as well as a "tested positive" event, which would be output as a generated span of text. We propose to predict emotion scores, topic, and triggering event for each post individually as well as concurrently. Additionally, we believe that this problem includes two important aspects: its temporal nature, and its shared topics. Therefore, we also propose to build models which operate in a time-series setting, as well as models that explicitly model shared topics and use them to predict emotions and triggering events.

Finally, we note that our dataset is not formulated to include labels for the fine-grained triggering events, only the broad topics and emotions. We will collect a small seed set of annotations and attempt to use an off-the-shelf event classifier to match these annotations and create event labels for the entire dataset. The human annotations will be used for evaluation on the test set.

3.3.2 Proposed Models

For this task, we propose a number of models in which we predict each of our three tasks (emotion intensity, topic classification, event description) in a multi-task way (i.e., in tandem) and in sequence (i.e., we first predict one task and then use our prediction to predict the other tasks, as in our sequential models for emotion-cause detection). We will explore a number of important questions when developing these models:

1. What is the best way to express our hierarchical topic-event pairs? We will experiment with our proposed three-way classification problem and with generating event strings from continuous representations. We will additionally experiment with handling new events, including whether some form of zero-shot learning or topic modeling can be useful.
2. What architecture best predicts the emotions we consider? We will experiment with multi-task and pipelined models of the type described throughout this proposal. We will experiment with secondary tasks that may provide useful information, such as whether existing work on affective events may be informative.
3. What temporal and event-based information can help us better predict emotions? We will also experiment with (1) structuring our predictions in time and (2) creating a streaming setting in which our models explicitly maintain some global knowledge about all seen users' emotions about some events to make future judgments about the same and similar events. We will particularly investigate whether models with a temporal aspect can more accurately predict emotions and events later in the timeline compared to models without this temporal aspect.
4. With multiple posts over time from the same users, can we make predictions about individual users' narratives? Given that the ieso data will be labeled for grief, we will investigate whether we can also predict the trajectory of grief over time–that is, whether it will resolve or become complicated during the timespan of our dataset (Weaver, 2010). We will also be able to examine whether our data follows psychological theories of grief, such as the well-known five stages (Kübler-Ross, 1969).
5. How can models for this problem be made interpretable? We will experiment with our previous methods of interpretability (e.g., investigating whether the models learn that some events are likely overall to have certain types of reactions, connecting narratives of grief to work from the psychological field, etc.) and strive to develop new ones for this new problem setting.

We propose to investigate these questions and select a number of directions which offer the most promising results and potential for study to be included in our final thesis.

4 Limitations

We note that large-scale machine learning, especially with respect to sensitive issues such as the emotional states of human beings, has inherent potential for misuse. Current emotion models are not infallible (and neither will the models we develop be); it is our belief that they should be used always with a human in the loop to think critically about the decisions the models are making. In this proposal, we aim to develop models that make this task easier for humans, not models that replace humans: by exposing their decision-making processes and relying on information that is intuitively understandable to humans, we hope that our models will be more understandable to their users.

The datasets we present in this work are all small by neural-network standards. We therefore also note the possibility of overfitting our training data and inducing dataset bias, and we will take steps to investigate this and mitigate it where we can. Much of our data is also annotated after the fact by outside annotators with emotion labels (Dreaddit, GoEmotions, etc.), and we are very interested to characterize the differences between this data and data that is labeled with emotions by its writers (such as ieso).

In this work, we limit ourselves to predicting emotional states purely from text; we do not include other modalities of information such as video, audio, or bio-signals (e.g., heart rate, breathing rate, etc.) which may be useful. There are additionally many more possible negative emotional reactions than we specifically list and study here, including clinical diagnoses such as depression. We select some psychological theory that is accessible and may feasibly be explored through text data, but emotion is a wide and rich field of psychological study with many more competing theories than we use to inform our models and analysis here.

5 Timeline to Completion

We propose the timeline in the table below for the completion of our proposed work.

We expect that interpretable grief detection will be finished by mid-fall 2021, leaving a year to perform the exploration and many experiments in subsection 3.3. For this problem, we will focus first on modeling emotions and large-scale events, including annotating data for fine-grained events, and then move forward to

Proposed Work Section	Estimated Completion Date
Grief Detection	Spring 2021
Grief Detection Analysis	Fall 2021
Emotions & Large-Scale Events Modeling	Fall 2021 & Spring 2022
Emotions & Large-Scale Events Interpretability	Summer 2022
Emotions Over Time Modeling	Fall 2022
Dissertation Writing	Fall 2022 & Spring 2023
Dissertation Defense	Spring 2023

creating interpretable models and finally to creating temporal models of negative emotions and stitching them together into narratives. Finally, we allot six months for the writing and defense of the dissertation.

6 Conclusion

In this proposal, we have proposed two large central problems which will contribute to our understanding of negative emotional reactions: detecting negative emotional reactions and detecting their causes. We show our completed work on detecting distress and show how we will extend that work to another negative emotional reaction, grief; this work relies on multi-task learning to incorporate psychological theory and yield more knowledgeable, interpetable models for emotion detection. In the process, we additionally present new datasets for distress and grief prediction and dedicate time to analyzing the contents of these datasets so as to understand what our models may learn from them. Additionally, we evaluate these models for interpetability, showing that our novel distress detection models use emotional information to make decisions about the distress problem; this is one of the most important threads of this work, and we plan to evaluate all of our models for interpetability.

Because emotions are not present in isolation, but are reactions to something else, we expand our understanding of negative emotions by focusing on the events that cause them. We present our completed work on using common-sense reasoning and multi-task learning to improve emotion-cause detection, where causes are realized as spans of text, and propose work characterizing the causes of emotions as events. We hope that our experiments in predicting events and their emotional reactions will inspire detailed representations of emotions and their semantic roles, as well as of events and their effects.

References

Muhammad Abdul-Mageed and Lyle Ungar. 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, July 2017. Association for Computational Linguistics.

Firoj Alam, Fabio Celli, Evgeny A. Stepanov, Arindam Ghosh, and Giuseppe Riccardi. The social mood of news: Self-reported annotations to design automatic mood detection systems. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 143–152, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Emily Allaway and Kathleen McKeown. A unified feature representation for lexical connotations, 2021.

Brian J. Arizmendi and Mary-Frances O'Connor. What is "normal" in grief? *Australian Critical Care: Official Journal of the Confederation of Australian Critical Care Nurses*, 28(2), 2015.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), Jul 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Krisztian Balog, Gilad Mishne, and Maarten de Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In *Demonstrations*, 2006.

Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

George A. Bonanno. *The other side of sadness: What the new science of bereavement tells us about life after loss*. Basic Books, 2009.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *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 4762–4779. Association for Computational Linguistics, 2019.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *CoRR*, abs/1912.03184, 2019.

Svend Brinkmann and Ester Holte Kofod. Grief as an extended emotion. *Culture & Psychology*, 24(2):160–173, 2018.

Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics.

Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador, October 2014. Association for Computational Linguistics.

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

Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathleen McKeown. Detecting gang-involved escalation on social media using context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. Embodied multimodal multitask learning. *CoRR*, abs/1902.01385, 2019.

Mei-Yu Chen, Hsin-Ni Lin, Chang-An Shih, Yen-Ching Hsu, Pei-Yu Hsu, and Shu-Kai Hsieh. Classifying mood in plurks. In *Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010)*, pages 172–183, Nantou, Taiwan, September 2010. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *CoRR*, abs/1806.10574, 2018.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online, July 2020. Association for Computational Linguistics.

Munmun De Choudhury, S. Counts, and M. Gamon. Not all moods are created equal! exploring human emotional states in social media. In *ICWSM*, 2012.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *CoRR*, abs/1909.03683, 2019.

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

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *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, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

Haibo Ding and Zhe Feng. Learning to classify events from human needs category descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4698–4704, Online, November 2020. Association for Computational Linguistics.

Haibo Ding and Ellen Riloff. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2935–2942. AAAI Press, 2016.

Haibo Ding and Ellen Riloff. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Haibo Ding, Tianyu Jiang, and Ellen Riloff. Why is an event affective? classifying affective events based on human needs. In *AAAI Workshops*, pages 8–15, 2018.

Zixiang Ding, Rui Xia, and Jianfei Yu. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170, Online, July 2020. Association for Computational Linguistics.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China, July 2015. Association for Computational Linguistics.

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

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717, Online, July 2020. Association for Computational Linguistics.

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

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *CICLing*, 2015.

Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, pages 1–1, 2019.

Nina Grant, Mark Hamer, and Andrew Steptoe. Social Isolation and Stress-related Cardiovascular, Lipid, and Cortisol Responses. *Annals of Behavioral Medicine*, 37(1):29–37, 02 2009.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November 2016. Association for Computational Linguistics.

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324, 2018.

Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. *CoRR*, abs/1906.10651, 2019.

US Institute of Medicine. *Bereavement: Reactions, Consequences, and Care*. National Academies Press, 1984.

Elisabeth Kübler-Ross. *On Death and Dying*. Collier Books/Macmillan Publishing Co., 1969.

Evgeny Kim and Roman Klinger. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

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

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, July 2020. Association for Computational Linguistics.

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. Stressnet: Detecting stress in thermal videos, 2020.

Daniela Laricchiuta, Francesca Greco, Barbara Cordella, Debora Cutuli, Eleonora Picerni, Francesca Assogna, Carlo Lai, Gianfranco Spalletta, and Laura Petrosini. "the grief that doesn't speak": Text mining and brain structure. 06 2018.

Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

Iulia Lefter, Gertjan J. Burghouts, and Leon Rothkrantz. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing*, 7(2):162–175, 2016.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Rationalizing neural predictions. *CoRR*, abs/1606.04155, 2016.

Minglei Li, Da Wang, Qin Lu, and Yunfei Long. Event based emotion classification for news articles. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 153–162, Seoul, South Korea, October 2016.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany, August 2016. Association for Computational Linguistics.

Jianquan Li, Xiaokang Liu, Wenpeng Yin, Min Yang, and Liqun Ma. An empirical evaluation of multi-task learning in deep neural networks for natural language processing. *CoRR*, abs/1908.07820, 2019.

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

Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 5570–5581, Florence, Italy, July 2019. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *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, 2019.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

Harri T. Luomala and Martti Laaksonen. Contributions from mood research. *Psychology & Marketing*, 17(3):195, 03 2000.

Alexander C. McFarlane. The long-term costs of traumatic stress: intertwined physical and psychological consequences. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, 9(9):3–10, 2010.

Albert Mehrabian and James A. Russell. *An approach to environmental psychology*. The Mit Press, 1974.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

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

Gilad Mishne and Maarten de Rijke. Capturing global mood levels using blog posts. In *Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, January 2006.

Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

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

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. Cross-lingual emotion intensity prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

Thin Nguyen. Mood patterns and affective lexicon access in weblogs. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, page 43–48, USA, 2010. Association for Computational Linguistics.

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

Laura Oberländer, Kevin Reich, and Roman Klinger. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions?, 2020.

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

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

Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing (TAFFC)*, 2020.

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040, 1995.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. Event2Mind: Commonsense inference on events, intents, and reactions. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 463–473, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *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, 2016.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. EMOCause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon, June 2011. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146, 2018.

Stanley Schachter and Jerome E. Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69(5):379—-399, Sep 1962.

Hans Selye. *The Stress of Life*. McGraw Hill, 1956.

R.J. Senter and E.A. Smith. Automated readability index. November 1967.

Sofia Serrano and Noah A. Smith. Is attention interpretable? *CoRR*, abs/1906.03731, 2019.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. *CoRR*, abs/1806.00674, 2018.

Katherine M Shear, Naomi Simon, Melanie Wall, Sidney Zisook, Robert Neimeyer, Naihua Duan, Charles Reynolds, Barry Lebowitz, Sharon Sung, Angela Ghesquiere, Bonnie Gorscak, Paula Clayton, Masaya Ito, Satomi Nakajima, Takako Konishi, Nadine Melhem, Kathleen Meert, Miriam Schiff, Mary-Frances O'Connor, Michael First, Jitender Sareen, James Bolton, Natalia Skritskaya, Anthony D. Mancini, and Aparna Keshaviah. Complicated grief and related bereavement issues for dsm-5. *Depression and Anxiety*, 28:103–117, 2011.

Sadhika Sood. Psychological effects of the coronavirus disease-2019 pandemic. *Research& Humanities in Medical Education*, 7:23–26, Apr. 2020.

Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, MuSe'20, page 35–44, New York, NY, USA, 2020. Association for Computing Machinery.

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

Zheng Tang, Gus Hahn-Powell, and Mihai Surdeanu. Exploring interpretability in event extraction: Multitask learning of a neural event classifier and an explanation decoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–175, Online, July 2020. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Generating token-level explanations for natural language inference. 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 963–969, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230, Berlin, Germany, August 2016. Association for Computational Linguistics.

Shenhao Wang, Qingyi Wang, and Jinhua Zhao. Multitask learning deep neural networks to combine revealed and stated preference data. *Journal of Choice Modelling*, 37:100236, 2020.

Teddy Wayne. The trauma of violent news on the internet. *New York Times*, Sep 2016.

Janalee Weaver. Narratives from grief counseling: Client perspectives on effective interventions and strategies for recovery. Master's thesis, 2010.

Penghui Wei, Jiahao Zhao, and Wenji Mao. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online, July 2020. Association for Computational Linguistics.

Cynthia Whissel. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2):509–521, oct 2009.

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

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. *CoRR*, abs/1906.01267, 2019.

Rui Xia, Mengran Zhang, and Zixiang Ding. RTHN: A rnn-transformer hierarchical network for emotion cause extraction. *CoRR*, abs/1906.01236, 2019.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 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, 2018.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-jie Zhang. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 788–793, Beijing, China, July 2015. Association for Computational Linguistics.

George Udny Yule. *The statistical study of literary vocabulary*. Cambridge Univ. Pr., 1944.

Suyang Zhu, Shoushan Li, and Guodong Zhou. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480, Florence, Italy, July 2019. Association for Computational Linguistics.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5608–5617, Online, November 2020. Association for Computational Linguistics.

Sidney Zisook and Katherine Shear. Grief and bereavement: what psychiatrists need to know. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 8(2):67–74, 2009.

Xin Zuo, Tian Li, and Pascale Fung. 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, May 2012. European Language Resources Association (ELRA).

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

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

Table 12: A comparison of how often several of our models rely on words identified as salient for distress or non-distress to make their decisions on the dev set. These numbers represent the percentage of available relative salience words each model selected in the top 10 LIME explanations. Dreaddit is Dr, GoEmotions is GE, and FT is Fine-Tune.

Appendices

Appendix A Extended Analysis for Distress Detection

During our distress experiments, we also investigate the Dreaddit data itself for highly significant words using the measure of relative salience proposed by Mohammad (2012), $RelativeSalience(w|T_1, T_2) = \frac{f_1}{N_1} - \frac{f_2}{N_2}$, which compares the relative frequency of a token w in two corpora (T_1, T_2). We compute this measure for all words in the Dreaddit training data, taking our two corpora to be the subsets labeled distress and non-distress. We take the top 200 unigrams for each label (distress as opposed to non-distress and vice-versa) and provide some examples in Table 11. We perform a manual analysis of the produced words: 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 distress (which echoes our prior finding that stressful data is typically longer with more clauses), while non-distress includes words expressing future-thinking like *if*, *will*, and *would*. We also naturally find negative words for distress and posi-

tive words for non-distress, as well as a dichotomy of isolation and helplessness for distress vs. support and community for non-distress which is supported by psychological literature (Grant et al., 2009).

We finally look at the intersection between relative salience and LIME explanations in the same manner as for LIWC categories, counting how many LIME explanations are highly salient words for distress or non-distress; abbreviated results are shown in Table 12. We see that our emotion-infused models learn to rely more often on words indicative of non-distress, the minority class, than of distress, the majority class.

Appendix B Ieso Interface

A mockup of the ieso interface is shown in Figure 8.

Appendix C Extended Analysis for Emotion-Cause Detection

BERT	**Multitask**	**BERT**^{COMET}	**Multitask**^{COMET}
Mexico reels from shooting attack in El Paso ***fear***			
negative surprise	negative surprise	fear	fear
Insane video shows Viking Sky cruise ship thrown into chaos at sea ***fear***			
negative surprise	fear	negative surprise	fear
Durant could return for Game 3 ***positive surprise***			
for game	could return for game		
Dan Fagan: Triple shooting near New Orleans School yet another sign of city's crime problem ***negative surprise***			
school yet another sign of city's crime	: triple shooting near new orleans school yet another sign of city's crime		

Table 13: Example outputs from our systems. For each example, the gold cause is highlighted in yellow and the gold emotion is given under the text; the first two examples give our models' emotion outputs; the latter two, their causes. Joined cells show that multiple models produced the same output. To make this table easier to read, "Multitask" here may refer to Multi, Multi$_{E\to C}$, or Multi$_{C\to E}$ (most multitask models gave similar outputs).

Figure 8: The proposed ieso posting interface.

Metric	BERT	BERT^{COMET}	Multi	Multi$_{E\rightarrow C}$	Multi$_{C\rightarrow E}$	Multi^{COM}	Multi$^{COMET}_{E\rightarrow C}$	Multi$^{COMET}_{C\rightarrow E}$
Acc. (Gold)	38.50	38.50	38.34	39.68	38.74	39.05	38.58	40.79
Acc. (¬Gold)	23.48	23.24	22.37	21.11	22.85	21.26	22.45	20.08

Table 14: Comparison of gold accuracy and non-gold (¬gold) accuracy for our emotion classification models.

We present some example outputs from our emotion-cause models from subsection 3.2 in Table 13, and we present the full results of our label analysis for our models in Table 14. In the latter table, Gold accuracy is the conventionally accepted form where a model's prediction is correct if and only if it matches the gold emotion label. ¬Gold, on the other hand, is accuracy where a prediction is correct if it was suggested by any human annotator but was not selected as the gold.