

Detecting and Explaining Emotional Reactions in Personal Narrative

Elsbeth Turcan

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

© 2024

Elsbeth Turcan

All Rights Reserved

Abstract

Detecting and Explaining Emotional Reactions in Personal Narrative

Elsbeth Turcan

It is no longer any secret that people worldwide are struggling with their mental health, in terms of diagnostic disorders as well as non-diagnostic measures like perceived stress. Barriers to receiving professional mental healthcare are significant, and even in locations where the availability of such care is increasing, our infrastructures are not equipped to find people the support they need. Meanwhile, in a highly-connected digital world, many people turn to outlets like social media to express themselves and their struggles and interact with like-minded others.

This setting—where human experts are overwhelmed and human patients are acutely in need—is one in which we believe artificial intelligence (AI) and natural language processing (NLP) systems have great potential to do good. At the same time, we must acknowledge the limitations of our models and strive to deploy them responsibly alongside human experts, such that their logic and mistakes are transparent. We argue that models that make and explain their predictions in ways guided by domain-specific research will be more understandable to humans, who can benefit from the models' statistical knowledge but use their own judgment to mitigate the models' mistakes.

In this thesis, we leverage domain expertise in the form of psychology research to develop models for two categories of emotional tasks: identifying emotional reactions in

text and explaining the causes of emotional reactions. The first half of the thesis covers our work on detecting emotional reactions, where we focus on a particular, understudied type of emotional reaction: psychological distress. We present our original dataset, Dreaddit, gathered for this problem from the social media website Reddit, as well as some baseline analysis and benchmarking that shows psychological distress detection is a challenging problem. Drawing on literature that connects particular emotions to the experience of distress, we then develop several multitask models that incorporate basic emotion detection, and quantitatively change the way our distress models make their predictions to make them more readily understandable.

Then, the second half of the thesis expands our scope to consider not only the emotional reaction being experienced, but also its cause. We treat this cause identification problem first as a span extraction problem in news headlines, where we employ multitask learning (jointly with basic emotion classification) and commonsense reasoning; and then as a free-form generation task in response to a long-form Reddit post, where we leverage the capabilities of large language models (LLMs) and their distilled student models. Here, as well, multitask learning with basic emotion detection is beneficial to cause identification in both settings.

Our contributions in this thesis are fourfold. First, we produce a dataset for psychological distress detection, as well as emotion-infused models that incorporate emotion detection for this task. Second, we present multitask and commonsense-infused models for joint emotion detection and emotion cause extraction, showing increased performance on both tasks. Third, we produce a dataset for the new problem of emotion-focused explanation, as well as characterization of the abilities of distilled generation models for this problem. Finally, we take an overarching approach to these problems inspired by psychology theory that incorporates expert knowledge into our models where possible, enhancing explainability and performance.

Table of Contents

Acknowledgments	xiv
Dedication	xvii
1 Introduction	1
1.1 Motivation and Goals	1
1.2 Organization and Contributions of the Thesis	5
I Detecting Emotional Reactions	7
2 Related Work for Detecting Emotional Reactions	12
2.1 Emotion in Psychology	12
2.2 Emotion in Natural Language Processing	14
2.2.1 <i>Emotion Detection</i>	15
2.2.2 <i>Related Problems</i>	17
2.3 Mental Health and Psychological Distress	18
2.3.1 <i>Psychology Research on Stress</i>	18
2.3.2 <i>NLP Research on Stress and Mental Health</i>	19
2.4 Multi-task Learning	21

3	Detecting Distress in Social Media	24
3.1	Dreaddit	26
3.1.1	<i>Data Collection</i>	26
3.1.2	<i>Data Annotation</i>	29
3.1.3	<i>Data Analysis</i>	32
3.2	Basic Modeling	36
3.2.1	<i>Data Features</i>	36
3.2.2	<i>Models</i>	37
3.3	Results	38
3.4	Discussion and Conclusions	41
4	Multitask Models for Distress Detection	45
4.1	Data	47
4.2	Methods	49
4.2.1	<i>Models</i>	49
4.2.2	<i>Baselines</i>	51
4.2.3	<i>Training</i>	52
4.3	Results	54
4.4	Analysis	56
4.4.1	<i>Multitask Knowledge</i>	57
4.4.2	<i>LIWC Analysis</i>	58
4.4.3	<i>Salient Words</i>	60

4.5	Discussion and Conclusions	63
5	Discussion for Detecting Emotional Reactions	66
5.0.1	<i>Contributions</i>	68
5.0.2	<i>Limitations and Future Work</i>	69
II	Understanding Emotions and Their Causes	71
6	Related Work for Understanding Emotion and Events	75
6.1	Emotion-Cause Detection	75
6.1.1	<i>Related Problems</i>	77
6.2	Emotional Summaries and Explanations	79
7	Emotion-Cause Pair Extraction	81
7.1	Data	83
7.2	Modeling	84
7.2.1	<i>Single-Task Models</i>	85
7.2.2	<i>Multi-Task Models</i>	87
7.2.3	<i>Adapted Knowledge Models</i>	89
7.2.4	<i>Experimental Setup</i>	91
7.3	Results	95
7.4	Analysis	97
7.4.1	<i>Example Outputs</i>	97
7.4.2	<i>Label Agreement</i>	98

7.5	Discussion and Conclusion	100
8	Generating Emotion-Focused Explanations	104
8.1	Problem Definition	106
	<i>8.1.1 Criteria for an Explanation</i>	<i>107</i>
8.2	Data	109
	<i>8.2.1 Human Annotations</i>	<i>114</i>
	<i>8.2.2 Data Generation by Prompting</i>	<i>116</i>
	<i>8.2.3 Test Set Generation</i>	<i>119</i>
8.3	Methods	120
	<i>8.3.1 Models</i>	<i>121</i>
	<i>8.3.2 Evaluation</i>	<i>124</i>
8.4	Results and Analysis	129
	<i>8.4.1 Automated Metrics</i>	<i>129</i>
	<i>8.4.2 Human Evaluation</i>	<i>132</i>
	<i>8.4.3 Qualitative Analysis</i>	<i>133</i>
8.5	Discussion and Conclusion	139
9	Discussion for Understanding Emotions and Events	142
9.1	Contributions	144
9.2	Limitations and Future Work	146
10	Conclusion	148

10.1 Contributions	149
10.2 Limitations and Future Work	151
References	156
Appendix A Emotion-Cause Joint Extraction Extended Examples . .	191
Appendix B Emotional Explanation Prompts	193
Appendix C Full Individual Model Results for Explanation Generation Models	194

List of Figures

3.1	Annotation Instructions. Instructions given to our human annotators on Amazon Mechanical Turk for the distress labeling task. Annotators were given definitions and examples and asked to label five datapoints at once.	30
3.2	Lexical Diversity by Domain. Yule’s I measure (on the y-axis) is plotted against domain size (on the x-axis) and each domain is plotted as a point on two graphics. a) measures the lexical diversity of all words in the vocabulary, while b) deletes all words that were not included in LIWC’s negative emotion word list.	33
3.3	Lexical Diversity by Agreement. 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.	35
4.1	Emotion-Infused Models. The emotion-informed architectures we use in our experiments.	50
7.1	Semantic Roles in GoodNewsEveryone. An example of the semantic roles annotated by Bostan et al. (2020)	83
7.2	GoodNewsEveryone Label Distribution. Distribution of adjudicated emotion labels in the GoodNewsEveryone train data, as a percentage of the data points. “Positive” and “Negative” are abbreviated as + and -.	84
7.3	Multi-Task Models. Our multi-task models for concurrent emotion detection and emotion cause tagging.	87

7.4	Adapted Knowledge Model Architecture. The architecture of our proposed Multi ^{COMET} _{C→E} model.	89
7.5	Emotion Detection Performance by Label. Performance of the BERT and Multi ^{COMET} _{C→E} models on emotion classification.	96
7.6	Emotion Cause Tagging Performance by Label Performance of the BERT and Multi ^{COMET} models on cause tagging, broken down by emotion.	96
8.1	Example Annotation. A sample datapoint and explanation from GEDdit. Please note that the source datapoint has been paraphrased (using ChatGPT) so as to limit searchability.	115
8.2	Human Evaluation Instructions. The annotation instructions for our human evaluation of our generated emotional explanations. Annotators could collapse the HTML element containing the instructions for readability, and a link to contact the primary author was provided at the bottom.	126
8.3	Human Evaluation Interface. The annotation interface for our human evaluation of our generated emotional explanations. Student social workers completed annotations for 21 items each using this interface on Amazon Mechanical Turk.	127
8.4	Explanation Automatic Evaluation Results. The results of our models on the test set, for the explanation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	129
8.5	Emotion Automatic Evaluation Results. The results of our models on the test set, for the emotion task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	130
8.6	Event Automatic Evaluation Results. The results of our models on the test set, for the explanation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	131

C.1	Explanation Automatic Evaluation Results (Full) . The results of all of our models on the test set, for the explanation generation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	194
C.2	Emotion Automatic Evaluation Results (Full) . The results of all of our models on the test set, for the emotion detection task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	195
C.3	Event Automatic Evaluation Results (Full) . The results of all of our models on the test set, for the event extraction task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is bolded (see subsection 8.3.2 for discussion).	195

List of Tables

3.1	Dreaddit Data Statistics. We include ten total subreddits from five domains in our dataset. Because some subreddits are more or less popular, the amount of data in each domain varies. We endeavor to label a comparable amount of data from each domain for training and testing.	28
3.2	Dreaddit Data Examples. Examples from our dataset with their domains, assigned labels, and number of annotators who agreed on the majority label (reproduced exactly as found, except that a link to the GoFundMe has been removed in the last example). Annotators labeled these five-sentence segments of larger posts.	29
3.3	Dreaddit Distress by Domain. Percentages of each Dreaddit domain that are labeled as “distress” (according to majority vote).	31
3.4	LIWC Analysis by Domain. Results from our 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.	33
3.5	LIWC Analysis by Label. Results from our analysis using LIWC word lists, with the same definitions as in Table 3.4. First-person pronouns (“1st-Person”) use the LIWC “I” word list.	34
3.6	Complexity by Label. Measures of syntactic complexity for stressful and non-stressful data.	34

3.7	Supervised Results. Precision (P), recall (R), and F1-score (F) for our supervised models. Our best model achieves 79.80 F1-score on our test set, comparable to the state-of-the-art pretrained BERT-base model. In this table, “features” always refers to our best-performing feature set (≥ 0.4 absolute Pearson’s r). Models marked with a * show a significant improvement over the majority baseline (approximate randomization test, $p < 0.01$).	39
3.8	Feature Sets and Data Sets. The results of our best classifier trained on different subsets of features and data. Features are grouped by type and by magnitude of their Pearson correlation with the train labels (no features had an absolute correlation greater than 0.5); data is separated by the proportion of annotators who agreed. Our best score (corresponding to our best non-neural model) is shown in bold.	40
3.9	Error Analysis Examples. Examples of test samples our models failed to classify correctly: “BERT” refers to the state-of-the-art BERT-base model, while “LogReg” is our best logistic regressor.	44
4.1	Distress and Emotion Data. The datasets we use in this work and their relative sizes (in terms of total number of data points).	47
4.2	Hyperparameter Search Ranges. Hyperparameter ranges for our models. BERT-based models tuned the first two; the Multi models additionally tuned λ ; the RNN additionally tuned the remainder.	53
4.3	Selected Hyperparameter Values. Our models’ selected hyperparameters for Dreddit.	53
4.4	Multitask Distress Detection Performance. Results of our multitask models. The best result under each metric is bold.	54
4.5	Fine-Tuning Distress Detection Performance. Results of our fine-tuning models. The best result under each metric is bold.	55
4.6	Fine-Tuning Emotion Detection Performance. Performance of our fine-tuning BERT models on the different GoEmotions labelings and datasets.	55

4.7	Learned Task Correlations. Correlations of the gold labels for each dataset with labels predicted by the other classifier in a Multi ^{Alt} model. GoEmotions _{FSJ} (abbreviated for space as GoEmo _{FSJ}) is starred because its emotion data is not multi-label and therefore the correlation ratio η is used instead of the coefficient of determination R^2 (which is used for the other, multilabel GoEmotions variants).	57
4.8	Learned Task Correlations Per Emotion Label. Per-class scores of emotion and distress for Dreddit (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.	57
4.9	Rationale Prevalence by LIWC Category. A comparison of how often several of our models rely on words from several LIWC categories to make their decisions, according to LIME. These numbers represent the percentage of available LIWC words each model selected in the top 10 LIME explanations for the entire dev set. Dr is Dreddit, and GE is GoEmotions.	59
4.10	Example Relative Salience Terms. Some examples of words identified by relative salience on the Dreddit training data as indicative of distress or not-distress. We group the words by hand into semantically meaningful categories for ease of understanding.	61
4.11	Rationale Prevalence by Relative Salience. A comparison of how often several of our models rely on words identified as salient for distress or non-distress to make their decisions, according to LIME. These numbers represent the percentage of available relative salience words each model selected in the top 10 LIME explanations. So that the table fits on a page, Dreddit is Dr, and GoEmotions is GE.	62
7.1	Emotion Detection and Cause Extraction Hyperparameter Search Ranges. Our hyperparameter search ranges for all models. . .	92
7.2	Single-Task Model Hyperparameters. The selected hyperparameter values for our single-task models.	92
7.3	Multi-Task Hyperparameters. The selected hyperparameter values for our multi-task BERT models.	93

7.4	Multi-Task Adapted Knowledge Model Hyperparameters. The selected hyperparameter values for our multi-task COMET models. . . .	94
7.5	Emotion Detection and Emotion Cause Tagging Performance. 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 ($p < 0.05$).	95
7.6	Example Outputs. 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 $E \rightarrow C$, or Multi $C \rightarrow E$ (details on selection and results for each individual model available in appendix; most multi-task models gave similar outputs). . . .	98
7.7	Model Accuracy on Alternative Emotion Labels. Comparison of gold accuracy and non-gold (\neg gold) accuracy for our emotion classification models.	99
8.1	Emotional Explanation Criteria Examples. Manually constructed positive and negative examples of factually consistent and informative explanations for a sample social media post. Note that the input is from our dataset but paraphrased in order to minimize searchability.	110
8.2	GEDdit Data Statistics. Data contained in our GEDdit dataset, broken down by subreddit. The average length of a post in each subreddit, in tokens (rounded to the nearest integer) according to Huggingface’s pretrained <code>google/flan-t5-large</code> tokenizer, is given as well.	111
8.3	Emotions in GEDdit. Scores are broken down by domain and further by subreddit. For the eight Plutchik basic emotions and two polarities (positive, negative), we present the average percentage of words in a single post that are marked by the NRC Emolex (Mohammad and Turney, 2013) for that emotion. Note that r/sad and r/relationships each make up their own domain.	112

8.4	VAD in GEDdit. Scores are broken down by domain and further by subreddit. We present the average magnitude of all posts in that category, where the magnitude of a single post is defined as the average magnitude of all VAD-bearing words in the post. Note that r/sad and r/relationships each make up their own domain.	113
8.5	External Tasks. External tasks and datasets used in the Flan-T5 _{Split} +dataset models.	123
8.6	Sample Human Annotation Results. The collected results of our human evaluation. Each cell shows the mean \pm standard deviation of the collected scores under each metric. The mean of the best model under each metric is bolded. No differences are significant under the approximate randomization test.	132
8.7	Emotion vs. All Multitask Model Example. A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.	134
8.8	Multitask Emotion vs. Combined Model Example. A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.	136
8.9	QA vs. All Multitask Model Example. A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.	137
A.1	All Outputs for Emotion-Cause Example 1. Full model outputs for our first provided example from Chapter 7.	191
A.2	All Outputs for Emotion-Cause Example 2. Full model outputs for our second provided example from Chapter 7.	191
A.3	All Outputs for Emotion-Cause Example 3. Full model outputs for our third provided example from Chapter 7.	192
A.4	All Outputs for Emotion-Cause Example 4. Full model outputs for our fourth provided example from Chapter 7.	192
B.1	Task Prompts. Individual instructions or prompts given to the Flan-T5 _{Split} +dataset models in Chapter 8.	193

Acknowledgements

First, and most importantly: to my advisor, Kathleen McKeown, who supervised nearly every chapter of this dissertation work¹, thank you for your unending patience, wisdom, and encouragement. Without your guidance, I would surely not have made it through even the first chapter. I often want to jump from one project to the other, and I am very good at things like getting sick or (completely hypothetically) twisting my ankle so badly I need complete bedrest one semester before my dissertation needs to be finished. I want to extend heartfelt gratitude to Dr. McKeown for always being flexible and expertly balancing the competing needs of keeping me on track and giving me space to develop my own skills.

To Julia Hirschberg and Smaranda Muresan, who have served both as my dissertation committee members and as frequent mentors and advisors at Columbia: thank you for your insight and your support. It is always important to consider multiple perspectives on a project, and I appreciate every time you have blocked off parts of your busy day to listen to me explain my research projects. Thank you for helping make Columbia's CS department a fun and educational place to be!

To Cecile Paris and Mona Diab, who have served on both my proposal committee and my dissertation defense committee: thank you truly for helping guide me on my journey. I am very proud to have fellow alumnae on my committee, both from Dr.

¹Aside from my work at Amazon in [Chapter 7](#).

McKeown's lab and from the George Washington University!

To my supervisors and collaborators at Amazon, including Yaser Al-Onaizan, Rishita Anubhai, Shuai Wang, and again Smaranda Muresan, who contributed to the work in [Chapter 7](#): thank you for your mentorship and encouragement, both academically and professionally.

To my undergraduate advisors and supervisors at the George Washington University, including Rahul Simha, Poorvi Vora, and Jennifer Pizzuro: thank you for encouraging me to pursue my dreams and inspiring me to begin this journey.

To my parents, and all of their four-legged children over the years: thank you, thank you, thank you for always supporting me and believing in me. I could never have taken even the first steps on the road to this dissertation without your encouragement and love.

To my friends and labmates at Columbia University who have joined our lab family before and after me, including Jessica Ouyang, Christopher Hidey, Christopher Kedzie, Noura Farra, Faisal Ladhak, Amith Ananthram, Melanie Subbiah, Nicholas Deas, Zachary Horvitz, and Yanda Chen: thank you for inspiring me, celebrating with me, and commiserating with me. To Fei-Tzin Lee and Emily Allaway: thank you for all of the above and for keeping me sane while doing my thesis work; I hope we still have many more lunches to come. To the undergraduate and graduate students who have worked with me on various projects at Columbia University: thank you so much for your hard work, creativity, and fresh perspectives. I am glad to have been part of all of your journeys, and I hope that each one of you can follow a path that brings you joy, satisfaction, and fulfillment.

To my friends, Kay, Emma, Rose, Grace, Allan, Marwa, Lex, and Monica: thank you always for your support and companionship through good times and bad, including listening to my technobabble about fields that are not your areas of expertise. I am

lucky to know you and hope we continue our journeys together for many years. And, of course, to my own four-legged child who cannot read these acknowledgements but ought to be thanked regardless: thank you to Pippin for his silent love, earnest companionship, and diligent coating of cat hair on every fabric and surface I own.

Dedication

To my parents, William and Elaine Turcan, who have always believed in me.

Chapter 1

Introduction

1.1 Motivation and Goals

Humans are social creatures. We love to talk to other people—we express our joys and sorrows and worries to our family and friends every day. In the modern age, we post them on social media, too, discussing everything from our dinner to our celebrity crushes to our passionate stances on government policy or social ethics. Our deeply connected digital world allows us to talk to and hear from people all over the planet and access a wealth of information at incredible speeds, and we benefit greatly from this. If we want to know how long to cook spaghetti squash, or how long it will take to drive to Washington, D.C., or how to solve the video game puzzle that’s been frustrating us for an hour—we can just look it up. If we need to talk to our family and friends for help with a problem, we can message them even on the other side of the world. We can listen to their worries, as well—and those of anyone who has a public social media account, in fact. The proliferation of worrisome news we can access on social media is so great that it can actually be harmful¹.

We have a lot to worry about, too. Ongoing military crises, health crises like the COVID-19 pandemic (and its associated financial recession), climate change, and myriad societal issues such as gun violence, police brutality, or poverty, among many others,

¹<https://www.apa.org/monitor/2022/11/strain-media-overload>

all contribute to collective distress the world over. Often, these kind of structural stressors can contribute to poor individual mental health—sometimes to the extent that it interferes with daily functioning and qualifies as a diagnostic mental disorder. Individual factors like genetics and social factors like family relationships contribute, too, resulting in recent estimates that as many as one-half of all adults worldwide will experience a diagnostic mental disorder in their lifetime (McGrath et al., 2023), with one in five of adults struggling with depression² and one in eight adults living with such a disorder at any given time (World Health Organization, 2022). Poor mental health can negatively impact individuals’ careers, education, finances, living conditions, relationships, physical health, and even mortality.

Worldwide, barriers to receiving professional mental healthcare are significant, including high cost, limited availability (due in part to a lack of resources including government funding), and a pervasive stigma about mental health, resulting in a massive number of people who cannot access needed health services (World Health Organization, 2022). In countries like the United States where the availability of professional help is growing, its effectiveness is still hampered as we lack the resources to give each patient the quality, personalized therapeutic treatment they need.³ It is important to understand that existing treatments, including medication and psychotherapy, are both an art and a science; it may take a long time for a patient and their doctor to understand why their particular symptoms occur and what combination of treatments can effectively manage them.

It is in this context we situate our work and in this area we believe it has the most potential to do good for our society. Indeed, a vast body of prior work has identified mental health as a fruitful area for artificial intelligence (AI) and natural language

²<https://wellcome.org/reports/wellcome-global-monitor-mental-health/2020>

³<https://time.com/6308096/therapy-mental-health-worse-us/>

processing (NLP) (Le Glaz et al., 2021; Espejo et al., 2023; Haque and Rubya, 2023).⁴ We as a research community can imagine many situations in which the power of AI could lead to real, visible benefits in mental health: dedicated computer therapists who can help patients on-demand right from their phones; digital assistants and wearable devices that can detect the manifestation of negative emotions in real time and help patients make sense of them in the moment; tools that can assist healthcare professionals in triage and treatment by summarizing ongoing cases and pointing out subtle markers of various conditions; and many, many more. We can envision a world where our technologies enable us to identify large-scale societal stressors and comprehensively map out their effects on the population, guiding government policy and the efforts of healthcare organizations in the directions that will create the most social good.

However, we must also acknowledge the difference between this idealized world and the world we actually live in today. While AI has made great strides in healthcare, it still has a long way to go before reality meets ideal (Tornero-Costa et al., 2023). Although their performance on a wide range of tasks is impressive, modern large AI models are more often than not opaque “black boxes” whose training data is mysterious and whose mistakes are presented with just as much confidence as their quality outputs. In NLP, modern large language models (LLMs) are far from infallible, producing outputs that contradict their inputs or “hallucinate” new information at an alarming rate (Zhang et al., 2023c). These qualities are unacceptable in high-stakes, sensitive applications such as mental health—imagine the consequences if a model misjudges a high-risk patient as low-risk, or suggests an inappropriate treatment plan, and no human oversight catches the mistakes. Of course, these are just some of the many ethical considerations that should be taken into account as part of any AI tool for mental health. In this thesis, we advocate strongly for the development of AI tools that can assist human experts, not

⁴<https://www.forbes.com/sites/bernardmarr/2023/07/06/ai-in-mental-health-opportunities-and-challenges-in-developing-intelligent-digital-therapies>

replace them, and whose predictions are founded in established theory and explained as accurately as possible. With tools like these, humans can benefit from the tools' statistical knowledge but ultimately make their own careful judgments to mitigate the tools' mistakes.

In this thesis, we will present our work towards developing technologies that can help improve mental health while remaining open and interpretable to humans. Our work focuses on the first two foundational steps in this process: identifying our emotions, and understanding our emotions. In order to make meaning out of our emotions, we must first understand what those emotions are and where they come from—identifying homesickness after a move, grief after the loss of a loved one, or pride after a promotion helps us deal appropriately with our feelings in the moment and gives us a foundation to build on in the future. This thesis will be split into two parts, each of which tackles one of these foundational steps: first we will focus on detecting emotional reactions, where we will develop classification models for a specific type of emotional reaction (psychological distress); and then we will proceed to understanding emotions and their causes, where we will focus on detecting the cause of an emotional reaction in order to gain a fuller understanding of the reaction.

Existing state-of-the-art models in NLP for problems like these are typically not interpretable; they calculate their outputs with a massive stack of statistical parameters that are opaque to humans and hard for us to predict. While progress has been made in causing these models to explain themselves (that is, to use the same uninterpretable parameters to produce a sensible explanation for the output that is also consistent with the model's behavior), we argue in this thesis that changes to the architecture, training, and prediction are key for helping humans understand a model's outputs. Models that make predictions and explain them in a way that humans intuitively understand (e.g., being able to explain in some way related to existing psychology theory) will both be

easier to trust and expose their mistakes more obviously to a human expert, at least if one is familiar with the theory being referenced. In this work, we will particularly draw on multitask learning, where a model is trained to perform multiple tasks with some shared parameters, as a source of explainability. For example, a model that predicts the presence of distress can also be trained to predict the presence of certain emotions known to be related, such as fear or anger, and use these emotions as an explanation for the distress prediction. It is our hope that the work we present will inspire others to continue developing new models that are grounded in expert knowledge and explain their predictions as clearly as possible.

1.2 Organization and Contributions of the Thesis

This thesis is split into two sections. The first ([Part I](#)), which spans [Chapter I](#) through [Chapter 5](#), covers our work on detecting emotional reactions. After introducing this problem ([Chapter I](#)) and briefly providing background on relevant prior work ([Chapter 2](#)), we proceed to lay out our work in this area. Specifically, [Chapter 3](#) introduces our Reddit dataset, Dreddit, for the distress detection problem, including benchmark results against contemporary models as well as analysis of the types of text and emotional states in the dataset. Then, [Chapter 4](#) covers our multitask emotion-infused models for explainable prediction on this dataset. We conclude the first part of the thesis with a discussion of our findings ([Chapter 5](#)).

Then, the second part ([Part II](#)), from [Chapter II](#) to [Chapter 9](#), focuses on the triggers of those emotional reactions and covers our work understanding emotional states in a more structured way. Again we begin with an introduction ([Chapter II](#)) and a survey of related work ([Chapter 6](#)). [Chapter 7](#) tackles the joint problems of emotion detection and emotion cause extraction in a multitask way and informed by commonsense

reasoning, showing improved performance on both tasks on the existing GoEmotions dataset (Demszky et al., 2020). Chapter 8 aims to generate freeform explanations of emotional reactions and the events that trigger them using the capabilities of large language models prompted with examples from student social workers; we present the formulation of this new task, a new dataset collected from Reddit, and a characterization of the capabilities of distilled models (from large language models) in generating these explanations. We conclude the second part of the thesis with a discussion (Chapter 9), and we proceed to conclude the entire thesis at a higher level with a discussion of themes and recommendations for future work in the field (Chapter 10).

The contributions of this thesis, therefore, are as follows:

1. A dataset, Dreddit, for psychological distress detection (Chapter 3), and emotion-infused models that incorporate emotion detection with existing emotion datasets (Chapter 4);
2. Multitask and commonsense-infused models for joint emotion detection and emotion cause extraction in news headlines, showing performance gains (Chapter 7);
3. A dataset, GEDdit, for the new problem of emotion-focused explanation, and a characterization of the abilities of distilled generation models (based on large language models) for this problem (Chapter 8); and
4. An overarching approach that is inspired by psychology theory and incorporates expert knowledge where possible, enhancing explainability and performance.

Part I

Detecting Emotional Reactions

Introduction to Detecting Emotional Reactions

The first step in the process of realizing our goals for reliable, understandable mental health technology is developing models that can accurately understand the emotional state of the user. Before we can fully understand our experiences and how they shape us, we need to be able to correctly identify what we are feeling. As we will discuss in [Chapter 2](#), identifying emotional and affective states from text is a well-studied problem in NLP; researchers typically tackle the problem of *emotion detection* and either classify a document as expressing one or more emotions from a predefined set, or rate the intensity of one or more emotions in the document. These label sets are often borrowed from the *basic emotion* sets of psychology research, which are sets of emotional states thought to be universal across cultures based on various evidence.⁵

The work we will present in this part of the thesis is slightly different from emotion detection. We do not focus on basic emotion sets; rather, we will focus on a specific type of emotional state, distress (not necessarily classifiable as an *emotion*), in [Chapter 3](#) and [Chapter 4](#); when we move into [Part II](#) of the thesis, we will broaden our scope. We begin by identifying a specific, common type of negative emotional state (distress) for which there is little labeled data or work in the NLP community and which has an obvious negative impact on its experiencer if left unchecked. We believe that our models have the most potential to do good when focusing on negative and detrimental emotional experiences such as this.

⁵As we will also discuss in [Chapter 2](#), psychology research is far from unanimous on what emotions are basic or whether basic emotions even exist at all.

We additionally argue that the choice of emotion labels in NLP research is typically arbitrary when working in text; popularly applied psychology research was not designed for the text modality (for example, [Ekman \(1999\)](#) originally derived his famous basic emotion set by exploring which facial expressions people could identify from images), and other research develops emotion sets based on what emotions are available in the dataset, whether self-labeled or annotator-labeled (e.g., [Demszky et al. \(2020\)](#)). Therefore, when we create our own datasets for this work, particularly our distress dataset in [Chapter 3](#), we do not adhere to a specific basic emotion set.

Throughout this thesis, we will consistently focus on studying social media text as our emotional data. In our case, we most frequently gather data from Reddit,⁶ a social media website where users create posts and interact with others via comments, both in topic-specific subforums called subreddits. This structure means that individual subreddits can vary greatly in genre and content, from photo- and video-based subreddits (e.g., r/pics, r/aww, r/videos) to those that focus on long-form narratives including creative writing or personal stories (e.g., r/relationship_advice, r/nosleep). In our work on detecting and understanding emotional reactions, personal narrative is the most appropriate genre of text to study, since we can identify descriptive explanations of how the poster is feeling (as well as understand their causal stimuli). Self-motivated social media posts, furthermore, are different from, for example, clinical interviews, since posters are more likely to take a casual, everyday tone and express themselves in whatever way is most natural. We acknowledge that the use of a particular social media website does impose some limitations on the generalizability of our work, particularly restricting us to the particular demographics (gender identity, nationality, racial and ethnic identity, and so on) of the website’s users, and we hope that future work will be able to draw on a wider variety of human experiences.

⁶<https://www.reddit.com>

As we endeavor to detect emotional reactions in this data, a consistent theme of our work is that our models should be as explainable and interpretable as we can make them. In this thesis, we propose that using multitask learning, wherein a shared or similar set of parameters is encouraged to learn to perform multiple tasks, can improve explainability of our models. For example, in [Chapter 4](#), we use emotion detection as an auxiliary task for models whose primary task is distress detection. These models can predict the binary presence of distress in a document, but they have also been trained to predict the emotions in the document; because psychology research has connected distress and certain emotional states, they can be used as evidence for the distress prediction. Furthermore, as we show in the same chapter, multitask training changes the types of cues a model relies on to perform its primary task, making the predictions inherently more understandable to a human if we use an auxiliary task like emotion detection.

We will begin this part of the thesis by covering related work in [Chapter 2](#), including a high-level overview of some relevant psychological theory, NLP work in emotion and mental health, and multitask learning. Then, in [Chapter 3](#), we will introduce our Dreddit dataset for binary psychological distress detection along with some characterization of the different types of language and stressors contained within ([Turcan and McKeown, 2019](#)). In [Chapter 4](#), we will introduce three varieties of emotion-infused models that perform the distress detection task on this dataset; each one shares its language modeling parameters with an emotion classifier for relevant emotions, and we will show that applying multitask learning in this way guides the model towards performing distress detection in a more interpretable and emotional fashion ([Turcan et al., 2021b](#)). Finally, we will conclude with a discussion in [Chapter 5](#).

The contributions covered in this part of the thesis are as follows:

1. A dataset, Dreddit, for psychological distress detection ([Chapter 3](#)), and emotion-

infused models that incorporate emotion detection with existing emotion datasets ([Chapter 4](#)) and

2. An approach to the problem of distress detection that is inspired by psychology theory (including the definition of distress and its co-occurrence with particular emotions), enhancing the models' explainability ([Chapter 4](#)).

Chapter 2

Related Work for Detecting Emotional Reactions

Before we begin to present our work, we must situate ourselves in the context of existing research. As we tackle our goal of detecting and understanding emotional reactions in text, we draw on a vast literature from both psychology and natural language processing (NLP). First, we will give a very brief, high-level overview of how psychology research defines emotion, including various ways to categorize emotions into sets (Section 2.1). Then, we will describe NLP research on emotion that is most relevant to our work in this thesis (Section 2.2), focusing on work in emotion classification (§2.2.1). We will then provide some background on the definitions of psychological stress and distress, as well as NLP research into these concepts (Section 2.3). Finally, we will ground the methodology of much of our work by discussing the history of multitask learning in NLP (Section 2.4).

2.1 Emotion in Psychology

Emotion. The most fundamental question we can ask as we begin this work is, “What is an emotion?” While this question seems simple, decades of research in fields like psychology and neuroscience do not offer a singular answer. Much existing work has discussed this question in detail, attempting to formulate a definition (Kleinginna and Kleinginna, 1981; Ekman and Davidson, 1994; Cabanac, 2002; Kagan, 2007; Griffiths,

2008; Mulligan and Scherer, 2012). Research typically defines emotion as a mental state arising from some physiological and neurological response, usually situated with respect to some experience that has caused this response, and with some associated pleasure or displeasure. Emotions cause us to enact behaviors, which we can often understand from an evolutionary perspective (for example, fear may cause us to flee). Emotion may also cause us to hold cognitions that may feed into future attitudes, behaviors, and cognitions (if something has frightened us, we may dislike it and avoid it in the future). The concept of emotion can be examined from many perspectives on its origins and its effects on us as human beings, too many to cover in the scope of this thesis.

Today, the American Psychological Association defines an emotion as “a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event” (emo, 2023). We will adopt this as the psychological definition in this thesis, keeping in mind that there is a significant body of work it cannot cleanly incorporate. Furthermore, additional distinctions can be made between emotions and other concepts like feelings and moods based on factors including the length of time they persist, their inciting cause, or what process of the brain we are discussing (Damasio, 2004; Fox, 2018; Wong, 2016). Emotions are thought of as some of the shortest-term affective phenomena, with immediate causes and effects on the order of seconds or minutes.

Basic Emotions and Dimensional Models. With this understanding, we may then proceed to ask, “What emotions are there?”, and this, too, is a complex question with no scientific consensus. While we may think of many things as “emotions” in common parlance, including states like rage, relief, or tranquility, many long threads of research have attempted to identify minimal sets of emotions that can be distinctly identified across all humans regardless of their cultural background: so-called *basic emotions*. A multitude of such classification schemes have been proposed based on

various features like cross-cultural ability to identify facial expressions (Ekman, 1992), among others (Plutchik, 1980; Ekman, 1999; Parrott, 2001; James, 2007; Robinson, 2008; Cowen and Keltner, 2017), though no particular scheme is accepted above all others. Much competing research contends that emotion categories ought to be situated as points or regions in some organized continuous space, varying on axes like positive versus negative valence, dominance versus submission, or intensity of activation (Russell, 1980; Mehrabian, 1980; Scherer, 2005). Research is ongoing, and the discussion around what emotions are and whether a set of basic emotions exists continues to evolve.

Ekman’s original six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) form one of the smallest sets of basic emotions and are often incorporated into other, more comprehensive sets. A wide variety of other emotions have been defined by various research, including such states as guilt, love, pride, shame, relief, and longing (Shaver et al., 1987; Ekman, 1999). Models that place emotions into a structured space, such as Plutchik (1980), can describe many additional emotions as combinations of more basic emotions or somehow located between them (for example, Plutchik (1980) describes *love* as a combination of its basic emotions *joy* and *trust*). Dimensional models additionally introduce the idea that emotions can be related to one another, for example, in terms of intensity; one can begin to speak of *anger* versus *rage*.

2.2 Emotion in Natural Language Processing

In the field of natural language processing (NLP), research typically focuses on the emotion detection task, a classification problem with a discrete set of emotion labels, where one or several labels are assigned to a given document typically as describing the mental state of its author or a character it describes. Little work treats related concepts like mood (but see, e.g., De Choudhury et al. (2021); Tsakalidis et al. (2022)); NLP

researchers typically focus on emotion detection, which we will treat in §2.2.1, adopting one of the discrete basic emotion categorization schemes from psychology research (usually Ekman (1992) or Plutchik (1980)) or sometimes devising their own through intuition, empirical data, or annotations. Dimensional models also enjoy some focus, often the valence-arousal-dominance (VAD) model (Russell, 1980; Mehrabian, 1980), which positions emotions as points or regions in a real-valued three-dimensional space (e.g., Troiano et al. (2023)). Some prominent shared tasks and lexical resources have enabled a variety of work in this area (Mohammad et al., 2018; Whissel, 2009; Buechel and Hahn, 2017); we refer readers to recent surveys such as Nandwani and Verma (2021) and note that these continuous models are more popular in continuous domains like speech than in discrete domains like text (e.g., Yang and Hirschberg (2018)). Because emotion detection is a full and mature field, additional related problems such as emotion intensity and emotion cause have also been studied; we will discuss some of these in §2.2.2 and others, more fully, in Chapter 6.

2.2.1 Emotion Detection

Emotion detection is a widely studied problem across a range of languages, domains, and label sets (as revealed in multiple recent surveys on the topic: Alswaidan and Menai (2020); Mohammad (2020); Nandwani and Verma (2021); Chavan et al. (2023)). Work has been done in this area across a range of languages, genres, emotion label sets, and model architectures. Because this field is so vast and emotion classification is not the primary focus of this thesis, we will largely rely on work that has been done previously. Emotion classification has a long history, with a particularly large body of work in news headlines (e.g., Strapparava and Mihalcea (2010)) and Tweets or other microblogs (e.g., Abdul-Mageed and Ungar (2017)). Early work applied varied modeling techniques including rule-based models (e.g., Francisco and Gervás (2006)),

linear classifiers (e.g., [Alm et al. \(2005\)](#)), and lexicon-based methods (e.g., [Chaumartin \(2007\)](#)), while more recent work has made use of modern deep neural networks (e.g., [Yu et al. \(2018\)](#); [Abdillah et al. \(2020\)](#)), pre-trained language models (e.g., [Demszky et al. \(2020\)](#)), and large language models (LLMs) (e.g., [Peng et al. \(2023\)](#)). Most (though not all) work to date focuses on a small set of emotion labels, following basic emotion theories from psychology and being mindful of the challenges of classification with a large label set.

Our work in this thesis deals with unimodal English-language textual data in a classification setting. The work most relevant to ours focuses on personal narrative and online forums (e.g., [Khanpour and Caragea \(2018\)](#); [Hofmann et al. \(2020\)](#); [Casel et al. \(2021\)](#)). One dataset in particular, GoEmotions ([Demszky et al., 2020](#)), is created from Reddit comments and uses a large label set of 26 emotions; since much of the data we will examine in the first part of this thesis is sourced from Reddit, this dataset from the same genre of text is helpful to us.

Finally, of particular interest to this thesis, we wish to acknowledge the threads of work that aim to incorporate psychological theories of emotion into NLP models, not just the problem setting. Appraisal theory is one family of models describing how emotions are elicited; its core claim is that an emotional reaction to a stimulus results from how the experiencer subjectively interprets the stimulus (reviewed in [Scherer et al. \(2001\)](#)). For example, seeing a spider or insect might cause one observer to think they are dangerous and thus feel afraid, while a different observer might not evaluate them as dangerous and then experience other emotional reactions like calmness or curiosity. A line of work in emotion detection has drawn on this work to inform annotations and models across categorical ([Balahur et al., 2012](#); [Hofmann et al., 2020](#)) and dimensional ([Troiano et al., 2023](#)) emotion models, or even as a target task in itself ([Stranisci et al., 2022](#)). We are also aware of some work incorporating the Component Process Model

(Scherer et al., 2001), a superttheory that integrates cognitive appraisal among other features like physiological responses, to describe how emotions are elicited (Casel et al., 2021; Cortal et al., 2023).

2.2.2 Related Problems

We also wish to briefly acknowledge some additional NLP work aligned with the goals of this thesis, though not necessarily utilized directly in our work.

One widely-studied modification of the emotion classification problem is to evaluate not the binary presence or absence of an emotion, but its ordinal (e.g., low/medium/high) or real-valued (e.g., [0.0-1.0]) intensity. Some of the earliest work in the field of emotion processing aimed to understand the intensity of emotions (Strapparava and Mihalcea, 2010), and its study continues (e.g., Buechel and Hahn (2016)), including through more modern shared tasks (Mohammad and Bravo-Marquez, 2017). Although our work focuses on discrete expressions of emotion in text and we largely treat classification problems, intensity is an important component of understanding how emotion is being expressed and we hope to incorporate it in future work. Again, we refer readers to the relevant portions of recent surveys on emotion in NLP (e.g., Mohammad (2020)).

Much work has been done on understanding emotions in a structured way, including extracting the stimulus that has triggered the emotion along with various other semantic roles (such as the experiencer of the emotion). We will treat this body of work in Part II, where it aligns with our work. We wish to acknowledge more exploratory understanding of emotions as well, including efforts to organize emotions over time into narratives (e.g., Mohammad (2012); Kim et al. (2017); Panchendrarajan et al. (2021)) and understand the fine-grained syntactic and semantic effects of emotional language (e.g., Sintsova et al. (2018)). We believe that this type of work is invaluable and offers fresh new perspectives where it formulates new research problems and examines data

from new angles.

As a final note, we also distinguish emotion classification from the related and even more vastly studied field, sentiment analysis (see recent surveys, [Wankhade \(2022\)](#); [Poria et al. \(2023\)](#)), which studies opinions (positive or negative), usually towards something (such as a product for sale).

2.3 Mental Health and Psychological Distress

2.3.1 Psychology Research on Stress

Psychological Stress. The colloquial usage of the term “stress” equates it to anxiety or pressure, a feeling of uncertainty when some demand is put on a person who is not certain they are able to meet it. The technical definition used in psychology research largely agrees with it; Hans Selye, a prominent foundational researcher in this area, defines stress as “the nonspecific response of the body to any demand made upon it” ([Selye, 1976b](#)), while [Lazarus and Folkman \(1984\)](#) define stress as “a particular relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being”. These definitions position stress crucially as a reaction made in response to some *stressor* which is creating an actual or perceived threat ([Selye, 1978](#)). Selye’s work differentiated between *eustress* (desirable stress, a motivator) and *distress* (undesirable stress, a health risk) and stated that there was a certain amount of stress that is beneficial (and too much or too little can be harmful) ([Selye, 1976a](#)). An excess of psychological stress (usually specifically distress, if following Selye’s model) has been shown to contribute to a number of negative health outcomes ([Lupien et al., 2009](#); [Calcia et al., 2016](#); [Stewart et al., 2019](#); [Rozanski et al., 1999](#); [Harris et al., 2017](#)).

Stress Measurement. When psychology researchers measure and detect stress responses in humans, they often leverage a variety of physical signals such as the presence of the hormone cortisol in saliva (Allen et al., 2014), electroencephalogram (EEG) readings (Al-Shargie et al., 2016), and skin conductance or the galvanic skin response (Healey and Picard, 2005)), among others, and compare them to subjective self-reports of stress levels. A variety of psychological inventories, scales, and questionnaires focusing on long-term stress, such as the Perceived Stress Scale (PSS) (Cohen et al., 1983), have been developed for diagnostic and research purposes in a variety of situations. These scales may include self-reports of subjective feelings, standard lists of stressful life situations and events, and interviews with researchers or other professionals. We refer readers to recent surveys on this topic for a detailed treatment (Giannakakis et al., 2019; Crosswell, 2020).

2.3.2 NLP Research on Stress and Mental Health

Stress Detection. Stress detection methods focusing on biological and physical markers have the disadvantage of requiring additional hardware or needing human subjects to be physically present for measurement. Therefore, some work in NLP has tackled the problem of detecting stress from text, typically in various social media domains (Lin et al., 2017; Winata et al., 2018; Guntuku et al., 2018). This work often examines stress and behavioral patterns at the user level (Lin et al., 2017; Guntuku et al., 2018), finding that these high-level features and categories are important to classification performance as well as perhaps a more stable entity in text than moment-to-moment stress levels. We particularly note that Guntuku et al. (2018) invited social media users to fill out Cohen’s 10-item stress scale (Cohen et al., 1997), providing a user-reported stress measurement, while Lin et al. (2017) and Winata et al. (2018) use microblogs distantly labeled via relevant hashtags or phrases like “I am stressed”.

Related Tasks. A 2022 shared task focused on the detection of chronic stress [Weissenbacher et al. \(2022\)](#), focusing on the harmful effects of stress that specifically persists over a long period of time and collecting distantly supervised Twitter data by looking for keywords like “chronic” or “constant” to emphasize this aspect. Meanwhile, as with emotion detection, a small body of work has also focused on the related analytical task of identifying stressors or explaining the reasons for stress using NLP techniques ([Gopalakrishna Pillai et al., 2018](#); [Thukral et al., 2020](#); [Swanson et al., 2022](#)).

Modalities. Much computational work on detecting stress is multimodal or focuses on non-text modalities such as video or speech, where biological and physical information translates more directly (e.g., [Zuo et al. \(2012\)](#); [Kumar et al. \(2020\)](#); [Jaiswal et al. \(2020\)](#); also see [Slavich et al. \(2019\)](#)). Even of the work highlighted above, [Lin et al. \(2017\)](#) incorporate information from photos and [Winata et al. \(2018\)](#) expect their model to be applied as part of a pipeline that operates on speech data. We are also aware of a contemporaneous study to our work in [Chapter 4, Yao et al. \(2021\)](#), which presents similar insights to ours about the connectedness of emotion and stress and applies multitask models to stress prediction in a multimodal (text and speech) setting. In this thesis, we focus on work that can contribute to applications such as therapeutic chatbots, where non-text modalities may not always be available or desirable.

Other Mental Health Tasks. While our work in this thesis focuses on psychological distress, we also note that there is a significant related body of work investigating the detection and characterization of various mental health conditions from text, including depression ([Choudhury et al., 2013](#); [Coppersmith et al., 2015a](#); [Husseini Orabi et al., 2018](#); [Sampath et al., 2022](#)), suicide and suicidal ideation ([Wang et al., 2016](#); [Zirikly et al., 2019](#); [Yang et al., 2021](#); [Guzman-Nateras et al., 2022](#); [Izmaylov et al., 2023](#)), anxiety ([Coppersmith et al., 2015a](#); [Shen and Rudzicz, 2017](#); [Gruda and Hasan, 2019](#); [Jiang et al., 2020](#); [Cohan et al., 2018](#)), and various diagnostic disorders (e.g., Post-Traumatic

Stress Disorder (Coppersmith et al., 2015a,b), schizophrenia (Coppersmith et al., 2015a; Bar et al., 2019; Jiang et al., 2020; Zomick et al., 2019), and Borderline Personality Disorder (Coppersmith et al., 2015a; Jiang et al., 2020), among others). Some of this work has additionally leveraged similar insights to our work in psychological distress in [Chapter 4](#) and theorized that emotion detection and other related tasks can help inform computational models for these mental states and disorders (e.g., [Aragón et al. \(2019\)](#); [Li et al. \(2022\)](#)).

2.4 Multi-task Learning

Finally, we discuss the modeling and training paradigm we apply frequently in our work: multi-task learning. While many computational models in NLP have traditionally been trained and optimized to perform a single task (e.g., a model that only encodes a document and classifies the emotions therein), insights into the relatedness of different tasks have allowed researchers to leverage data from multiple different tasks in the same model. A key feature of this approach is that knowledge is shared between the parameters that perform different tasks, whether directly or indirectly, allowing one or more tasks to benefit from data other than their own training data. Multi-task learning has been applied to great effect across a range of problems in NLP (see recent surveys such as [Chen et al. \(2021\)](#); [Zhang et al. \(2023d\)](#)).

The most prominent type of multi-task learning in NLP is hard parameter sharing ([Caruana, 1993](#)), where multiple tasks share and update the same parameters, perhaps with some separate task-specific parameters added on in various ways. This is contrasted with soft parameter sharing (e.g., [Guo et al. \(2018b\)](#)), where different tasks have separate sets of parameters that are not required to be the same but rather encouraged to be close together by some constraint or regularization (for example, [Guo et al. \(2018b\)](#)

minimizes the Euclidean distance between models’ parameters using l_2 regularization in their loss construction). We will focus on hard parameter sharing in this work.

Multi-task learning is a broad field with many different possible architectures. Some of the most intuitive models have some collection of shared model parameters, often the input encoder, with small separate task-specific layers, often simply classification heads (e.g., [Liu et al. \(2019\)](#)) or sometimes a small stack of layers ([Collobert and Weston, 2008](#)). A popular modern variation of this style is to insert small “adapter” layers inside large pre-trained language models (often transformers), rather than at the end, and fine-tune these layers while keeping the rest of the pre-trained parameters static ([Houlsby et al., 2019](#); [Stickland and Murray, 2019](#); [Pilault et al., 2021](#)). Other models may form a hierarchical structure, feeding one module’s predictions for one task as input to another module that predicts a separate task on the same input (e.g., [He et al. \(2019\)](#); [Ma et al. \(2023\)](#)). Other architectures may separate global and task-specific local parameters in different ways, such as using global parameters as a memory rather than an encoder ([Liu et al., 2016](#)) or designing modules that are explicitly meant to learn various particular skills across tasks ([Feng et al., 2023](#)).

Finally, the most popular modern formulation of multi-task learning uses prompts to text generation models to indicate to the model which task the parameters are meant to perform for a given input (e.g., [Zhang et al. \(2023a\)](#)). It is becoming increasingly popular to use large language models (LLMs), which have been trained with language modeling objectives on massive corpora, as zero-shot learners to perform tasks they have never seen based solely on a prompt (e.g., [Zhang et al. \(2023b\)](#)). Most large language models are shown or assumed (as some models and data are proprietary) to be trained with simple language modeling objectives on plain text corpora (e.g., PaLM ([Chowdhery et al., 2022](#)), GPT-3 ([Brown et al., 2020](#)), GPT-4 ([OpenAI, 2023](#)), LLaMA ([Touvron et al., 2023](#))), although it is likely that some training data from

various popular tasks has made its way into the large Internet text dumps often used for these purposes. Meanwhile, inspired by these models' ability to perform few- or zero-shot learning via prompting, multiple language model bases have been developed by fine-tuning smaller language models on a multi-task dataset (e.g., T0 (Sanh et al., 2021), FLAN (Wei et al., 2021), Tk-Instruct (Wang et al., 2022)).

There are additionally several training methods for teaching a model to perform multiple tasks. Sometimes auxiliary tasks are used to improve the performance of one desired main task (e.g., Pujari et al. (2022)), while other times multiple tasks of equal interest may be optimized jointly (a famous example of this is the pre-training for language models like BERT (Devlin et al., 2019), where a masked language modeling task and a next sentence prediction task both contribute to developing a useful, flexible language model base). Transfer learning can additionally sometimes be seen as a sort of multi-task learning where a model is trained on multiple tasks in sequence rather than in parallel (Bengio, 2011).

Multi-task learning has been applied to many different problems across NLP, including argument mining (Cheng et al., 2020), stereotype detection (Pujari et al., 2022), metaphor detection (Le et al., 2020), sentence simplification (Guo et al., 2018a), sentiment classification (Zhang et al., 2017), representation learning (Xu et al., 2018), health informatics (Joshi et al., 2019), and emotion classification (Demszky et al., 2020), among many others. Modern work has often focused on training with a very large number of tasks at once (e.g., Padmakumar et al. (2022)). In this work we will use multi-task learning to improve the performance and interpretability of our models in distress detection, emotion detection, emotion cause detection (the related work for which we will discuss further in Chapter 6), and explanation generation.

Chapter 3

Detecting Distress in Social Media

To begin with, we turn to the problem of detecting stress, a universal human experience that can be a motivator in appropriate quantities, or lead to negative health outcomes in excess (Lupien et al., 2009; Calcia et al., 2016; Stewart et al., 2019; Rozanski et al., 1999; Harris et al., 2017). Because harmful levels of stress are common in our world today,¹ we find this to be an area with helpful applications across a range of domains (for example, in healthcare or in education). As with many affective phenomena, we find many expressions of stress on social media, where people constantly express themselves and their worries through text (Panger, 2017; Steinert and Dennis, 2022), and so we are well positioned to apply NLP techniques to this problem.

Stress is a subjective experience whose effects vary from person to person, and its definition has varied over time. Hans Selye, a prominent and foundational researcher in the field of stress, defined it as “the nonspecific response of the body to any demand made upon it” (Selye, 1976b); that is, stress is a reaction to any kind of expectation or change; this continues to be the accepted definition today (str, 2023a,b). The experiencer feels some amount of uncertainty to a new demand, which may include anything from the need to escape from a predator to the desire to be selected for a role in a musical by performing well in an audition. Selye additionally observes that this definition affords no value or valence connotation; things that are exciting and things

¹<https://www.apa.org/news/press/releases/stress/index?tab=2>

that are frightening can both induce stress. Therefore, he differentiates *eustress*, stress associated with desirable effects; and *distress*, stress associated with undesirable effects (Selye, 1976a). Selye asserts that these types of stress are in fact the same nonspecific response, and they can be differentiated by the type of stressor that causes them. For example, a first date would be likely to induce eustress while a dentist appointment might perhaps induce distress. We note that the same stressor may induce different reactions in different people and different situations—a job interview may be exciting or terrifying depending on how thoroughly the applicant has prepared and how attractive the position is to them.

In our work, we focus on detecting distress, as we view it as the location our research would be best situated to do the most social good. We define distress by its negative emotional reaction self-reported by the experiencer. In this chapter, we will describe our curation of the Dreddit dataset for stress detection in social media (Section 3.1), followed by our benchmarking setup (Section 3.2), the benchmark models’ performance (Section 3.3), and a discussion of the dataset’s characteristics and our results (Section 3.4).

The main contributions of this work (based on work published in Turcan and McKeown (2019)) include (1) the Dreddit dataset, a large unlabeled dataset of expressions of distress on Reddit; (2) an analysis of the dataset content and how it relates to known trends in psychological literature; and (3) benchmarking of non-neural and neural models against the dataset.

3.1 Dreddit

In order to obtain data for this work, we turn to the social media platform Reddit.² Reddit is a website characterized by topic-specific subforums called *subreddits* where users post messages relating to some subject such as a particular sports team, financial advice, or support for individuals experiencing a particular illness, and other users engage with them through comments and binary voting on the quality of the post. Reddit is home to a variety of support subreddits where users come together to share their experiences in a particular domain and offer advice and support to others (De Choudhury and De, 2014); this content allows us to observe natural expressions of many affective experiences, including various types of stress, across a variety of topics. Reddit data is typically long-form and not highly structured, which means that its posts provide ample context to study the nuances of different emotional states and situations. For our work in detecting distress, we collect and annotate the *Dreddit* dataset. We draw posts from a variety of different subreddits that contain expressions of distress in different situations and caused by different stressors.

3.1.1 Data Collection

In order to collect expressions of distress, we select categories of subreddits where we expect posters are likely to discuss topics related to distress. The categories are selected manually with some guidance from established stressful event scales and inventories (Holmes and Rahe, 1967), noting that some stressors may often be associated with eustress (e.g., pregnancy) and some may be difficult to find online (e.g., “beginning or end of school”):

- **Interpersonal conflict:** the abuse and social domains. Posters in the abuse

²<https://www.reddit.com/>

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 (often but not exclusively romantic) and seek advice for how to handle the situation.

- **Mental illness:** the 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:** the financial domain. Posters in the financial subreddits generally seek financial or material help from other posters.

Each of these domains includes one or more subreddits, since there is not often a consensus in the community on one central subreddit for a given topic, and some subreddits may focus on narrower aspects of a topic. For example, `r/almosthomeless` is an advice and support group for sharing stories and suggestions and in fact directly prohibits the solicitation of material aid, while `r/food_pantry` is a subreddit exclusively for connecting people in need of food with those who can donate to them. Meanwhile, `r/relationships` is perhaps the most well-known relationship advice subreddit and provides more than enough data for our purposes, but other such subreddits also exist with different formatting and focuses (`r/relationship_advice`, `r/dating_advice`, `r/AmITheAsshole`, etc.).

The statistics of our dataset are presented in [Table 3.1](#). We include ten subreddits across the five domains of abuse, social, anxiety, PTSD, and financial. Using the PRAW API³, we scrape all available posts on these subreddits between January 1, 2017 and November 19, 2018; in total, 187,444 unlabeled posts. As we will describe in [§3.1.2](#), we

³<https://github.com/praw-dev/praw>; note that because of Reddit’s 2023 policy changes (<https://www.reddit.com/r/reddit/comments/145bram/>), it may no longer be possible to replicate this data scraping with PRAW.

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 3.1: **Dreddit Data Statistics.** We include ten total subreddits from five domains in our dataset. Because some subreddits are more or less popular, the amount of data in each domain varies. We endeavor to label a comparable amount of data from each domain for training and testing.

assign binary distress labels to 3,553 segments of these posts to form a supervised and semi-supervised training set.

In Table 3.2, we provide examples of labeled segments from the various domains in our dataset. The samples are fairly typical; 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, making it a challenging dataset for contemporary models. The labeled data, with all identifiable names, emails, and hyperlinks removed, is currently made available online,⁴ while the larger corpus of unlabeled data is available upon request with the agreement not to make any personally identifying information (or models that are likely to have memorized personally identifying information) publicly available.

⁴<http://www.cs.columbia.edu/~eturcan/data/dreddit.zip>

Text	Domain	Label	Ann. Agreed
I only get it when I have a flashback or strong reaction to a trigger. I notice it sticks around even when I feel emotionally calm and can stick around for a long time after the trigger, like days or weeks. Its a new symptom I think. Also been having lots of nightmares again recently. Not sure what to do as I'm not currently in therapy, but I am waiting to be seen at a mental health clinic.	PTSD	stress	6/7 (86%)
Regardless, that didn't last long, maybe half a year. I released that apartment, and most of my belongings (I kept a few boxes of my things from the military, personal effects, but little else). Looking back, there were some signs of emotional manipulation here, but it was subtle... and you know how it is, love is blind. We got engaged. It was quite the affair.	abuse	not stress	5/5 (100%)
Our dog Jett has been diagnosed with diabetes and is now in the hospital to stabilize his blood sugar. Luckily, he seems to be doing well and he will be home with us soon. Unfortunately, his bill is large enough that we just can't cover it on our own (especially with our poor financial situation). We're being evicted from our home soon and trying to find a place with this bill is just too much for us by ourselves. To help us pay the bill we've set up a GoFundMe.	financial	stress	3/5 (60%)

Table 3.2: **Dreddit Data Examples.** Examples from our dataset with their domains, assigned labels, and number of annotators who agreed on the majority label (reproduced exactly as found, except that a link to the GoFundMe has been removed in the last example). Annotators labeled these five-sentence segments of larger posts.

3.1.2 Data Annotation

In order to train supervised and semi-supervised models, we annotate a subset of our data using Amazon Mechanical Turk⁵. We partition the posts into contiguous five-sentence chunks for labeling; we wish to annotate segments of the posts because this will help us understand what *sections* of the posts express stress, but we find through manual inspection that some amount of context is important. Providing the entire post as context would make it difficult for annotators to read and annotate since a single

⁵<https://www.mturk.com/>

post is on average 420 tokens long, so we follow this chunking approach to provide a useful amount of context for any given sentence. This type of data may be used in the future not only to *classify* the presence of stress, but also to *locate* its expressions in the text, even if they are diffused throughout the post.

Instructions: Below you are given snippets of text from social media. Please decide whether the person who wrote each snippet is, overall, stressed about what they're talking about. Let the text stand alone and try to decide without any additional context, but do remember these are parts of larger posts, so even if they are stressed, they may not be expressing it in this particular bit of text.

What does stress mean?: The Oxford English Dictionary defines stress as "a state of mental or emotional strain or tension resulting from adverse or demanding circumstances". This means that stress results from someone being uncertain that they can handle some threatening situation. We are interested in cases where that someone also feels negatively about it (sometimes we can find an event stressful, but also find it exciting and positive, like a first date or an interview).

- If you believe the writer **is stressed** and **has an overall negative attitude about it**, select "stress".
- If you believe the writer is not stressed, expresses stress but does not have an overall negative attitude about it, or has a negative attitude but you don't consider it stress (e.g., angry but not stressed), select "not stress".
- If you cannot tell—for example, because the snippet does not make sense, or is ambiguous—select "can't tell".

Remember to decide **based on the feelings the writer is expressing**, **NOT** whether you think the situation is or should be stressful.

Here are a few examples:

Sentence	Stress	Not Stress	Explanation
The interview process is super intimidating, like I don't even know how I'm gonna make it, but I'm so pumped for the job!	<input type="radio"/>	<input checked="" type="radio"/>	The writer is excited about the prospect of getting the job. Although the interview is a stressful event, they view it positively. If the writer dwelled on the difficulty of the interview, we would consider this stressful.
Over the past month, I've lost my appetite, have trouble sleeping, and feel like I'm shivering all the time. I hate it.	<input checked="" type="radio"/>	<input type="radio"/>	The writer is clearly exhibiting emotional symptoms of stress (nerves) and views it as a negative situation ("I hate it").
We ended up ending the relationship, and it sucks. I miss her a lot. But I hope she's doing well.	<input type="radio"/>	<input checked="" type="radio"/>	The writer is sad about the situation, but shows no stress—they are not dwelling on it or ruminating.
I have no idea what to even try. My anxiety just gets the best of me every single time and I can barely make conversation most of the time.	<input checked="" type="radio"/>	<input type="radio"/>	The writer is clearly exhibiting emotional symptoms of stress and feels lost.

Please label the following five text snippets and press the "Submit" button when you are finished.

Figure 3.1: **Annotation Instructions.** Instructions given to our human annotators on Amazon Mechanical Turk for the distress labeling task. Annotators were given definitions and examples and asked to label five datapoints at once.

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) in the same Human Intelligence Task (HIT) after taking a qualification test; Workers are

Abuse	Anxiety	Financial	PTSD	Social
55.75%	63.32%	39.47%	58.23%	44.24%

Table 3.3: **Dreaddit Distress by Domain.** Percentages of each Dreaddit domain that are labeled as “distress” (according to majority vote).

allowed to select “Stress”, “Not Stress”, or “Can’t Tell” for each segment. Annotators are given the instructions and examples shown in [Figure 3.1](#); note that, since this task is being performed by non-experts, we ask our annotators to label “stress” rather than “distress” and define it as “a state of mental or emotional strain or tension resulting from adverse or demanding circumstances”. We specifically ask Workers to decide whether the author is expressing both stress and a negative attitude about it (i.e., expressing distress), not whether the situation itself seems stressful to the annotators.

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 (soliciting up to 7 annotations if agreement is low) and include in each batch one of 50 “check questions” which have been previously verified by two in-house annotators and deemed (subjectively) straightforward by both. Our design does limit the amount of data we can feasibly label, but we hope to gain an understanding of how difficult this task is for humans by using a large number of annotators. After removing annotations which failed the check questions, and data points for which at least half of the annotators selected “Can’t Tell”, 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 and record the percentage of annotators who agreed. The resulting dataset is nearly balanced, with 52.3% of the data (1,857 instances) labeled as distress. The specific breakdown of distress percentage by domain is shown in [Table 3.3](#), where we see that distress does vary by domain.

Our agreement on all labeled data is $\kappa = 0.47$, using Fleiss’s Kappa ([Fleiss, 1971](#)),

considered “moderate agreement” by [Landis and Koch \(1977\)](#). We observe that annotators achieved perfect agreement on 39% of the data, and for another 32% the majority was 3/5 or less.⁶ This suggests that our data displays significant variation in how distress is expressed, which we explore in the next section.

3.1.3 Data Analysis

While all our data has the same genre and personal narrative style, we find distinctions among domains with which classification systems must contend in order to perform well, and distinctions between distressful and non-distressful data which may be useful when developing such systems. Posters in each subreddit express distress, but we expect that their different functions and stressors lead to differences in how they do so in each subreddit, domain, and broad category.

By domain. We examine the vocabulary patterns of each domain on our training data only, not including unlabeled data so that we may extend our analysis to the label level. First, we use the word categories from the Linguistic Inquiry and Word Count (LIWC) ([Pennebaker et al., 2015](#)), a lexicon-based tool that scores text according to 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 3.4](#). We first note that variety of expression depends on the domain and topic; for example, the variety in the expression of negative emotions is particularly low in the financial domain (with 1.54% of words being negative emotion (“negemo”) words and only 31% of “negemo” words used). We

⁶It is possible for the majority to be less than 3/5 when more than 5 annotations were solicited; for example, a 4/7 majority.

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 3.4: **LIWC Analysis by Domain.** Results from our 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.

also see clear topic shifts among domains: the interpersonal domains contain roughly 1.5 times as many social words, proportionally, as the others; and domains are stratified by their coverage of the anxiety word list (with the most in the mental illness domains and the least in the financial domain).

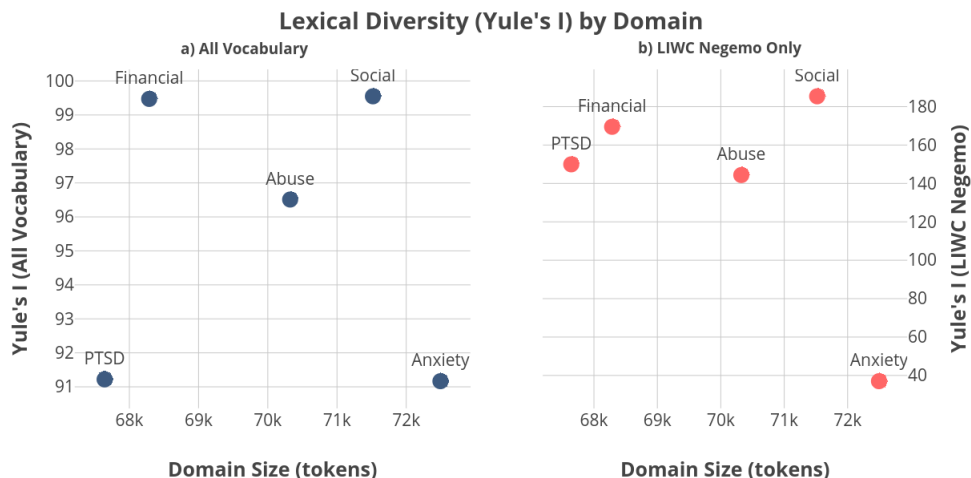


Figure 3.2: **Lexical Diversity by Domain.** Yule’s I measure (on the y-axes) is plotted against domain size (on the x-axes) and each domain is plotted as a point on two graphics. a) measures the lexical diversity of all words in the vocabulary, while b) deletes all words that were not included in LIWC’s negative emotion word list.

We also examine the overall lexical diversity of each domain by calculating Yule’s I measure (Yule, 1944). Figure 3.2 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. For example, posters in these domains discuss topics such as symptoms, medical care, and diagnoses. When we restrict our analysis to negative emotion words, this pattern persists only for anxiety; the PTSD domain has comparatively little lexical variety, but what it does have contributes to its variety of expression for negative emotions.

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

Table 3.5: **LIWC Analysis by Label.** Results from our analysis using LIWC word lists, with the same definitions as in Table 3.4. First-person pronouns (“1st-Person”) use the LIWC “I” word list.

By label. We perform similar analyses on data labeled distressful or non-distressful by a majority of annotators and present them in Table 3.5). We confirm some common results in the mental health literature, including that distressful data uses more first-person pronouns (perhaps reflecting increased self-focus) and that non-distressful data uses more social words (perhaps reflecting a better social support network).

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 3.6: **Complexity by Label.** Measures of syntactic complexity for stressful and non-stressful data.

Additionally, we calculate measures of syntactic complexity, including the percentage of words 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 3.6, we do see non-significant but persistent differences between distressful and non-distressful data, with distressful data being generally longer and more complex but also rated simpler by readability indices. These findings are intriguing and can be explored in future work.

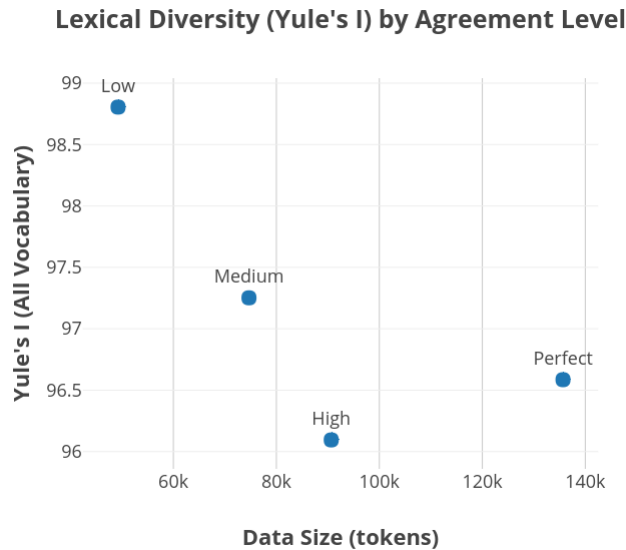


Figure 3.3: **Lexical Diversity by Agreement.** Yule’s I measure (on the y-axis) is plotted against domain size (on the x-axis) for each level of annotator agreement. Perfect means all annotators agreed; High, 4/5 or more; Medium, 3/5 or more; and Low, everything else.

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 3.3. While the amount of data and lexical variety seem to be related, Yule’s I measure controls for length, so we believe that this trend reflects a difference in the type of data that encourages high or low agreement.

3.2 Basic Modeling

3.2.1 Data Features

In order to train supervised models, we group the labeled segments by post and randomly select 10% of the posts ($\approx 10\%$ of the labeled segments) to form a test set. This ensures that while there is a reasonable distribution of labels and domains in the train and test set, the two do not explicitly share any of the same content. This results in a total of 2,838 train data points (51.6% labeled stressful) and 715 test data points (52.4% labeled stressful). Because our data is relatively small, we train our traditional supervised models with 10-fold cross-validation; for our neural models, we break off a further random 10% of the training data for validation and average the predictions of 10 randomly-initialized trained models.

In addition to the words of the posts (both as bag-of-n-grams and distributed word embeddings), we include features in three categories:

Lexical features. Average, maximum, and minimum scores for pleasantness, activation, and imagery from the Dictionary of Affect in Language (DAL) (Whissel, 2009); the full suite of 93 LIWC features; and sentiment calculated using the `Pattern` sentiment library (Smedt and Daelemans, 2012).

Syntactic features. Part-of-speech unigrams and bigrams, the Flesch-Kincaid Grade Level, and the Automated Readability Index.

Social media features. The UTC timestamp of the post; the ratio of upvotes to downvotes on the post, where an upvote roughly corresponds to a reaction of “like” and a downvote to “dislike” (**upvote ratio**); the net score of the post (**net score**) ($n_{\text{upvotes}} - n_{\text{downvotes}}$); and the total number of comments in the entire thread under the post.

3.2.2 Models

We first experiment with a suite of non-neural models, including Support Vector Machines (SVMs), logistic regression, Naïve Bayes, Perceptron, and decision trees. We tune the parameters for these models using grid search and 10-fold cross-validation, and obtain results for different combinations of input and features.

For input representation, we experiment with bag-of-n-grams (for $n \in \{1..3\}$), Google News pre-trained Word2Vec embeddings (300-dimensional) (Mikolov et al., 2013), Word2Vec embeddings trained on our large unlabeled corpus (300-dimensional, to match), and BERT embeddings trained on our unlabeled corpus (768-dimensional, the top-level [CLS] embedding) (Devlin et al., 2019). We experiment with subsets of the above features, including separating the features by category (lexical, syntactic, social) and by magnitude of the Pearson correlation coefficient (r) (Freedman et al., 2007) with the training labels. Finally, we stratify the training data by annotator agreement, including separate experiments on only data for which all annotators agreed, data for which at least 4/5 annotators agreed, and so on.

We finally experiment with neural models, although our dataset is relatively small. We train both a two-layer bidirectional Gated Recurrent Neural Network (GRNN) (Cho et al., 2014) and Convolutional Neural Network (CNN) (as designed in Kim (2014)) with parallel filters of size 2 and 3, as these have been shown to be effective in the literature on emotion detection in text (e.g., Xu et al. (2018); Abdul-Mageed and Ungar (2017)). Because neural models require large amounts of data, we do not cull the data by annotator agreement for these experiments and use all the labeled data we have. We experiment with training embeddings with random initialization as well as initializing with our domain-specific Word2Vec embeddings, and we also concatenate the best feature set from our non-neural experiments onto the representations after the recurrent

and convolutional/pooling layers respectively. Finally, we also apply BERT directly to our task, fine-tuning the pretrained BERT-base⁷ on our classification task for three epochs (as performed in [Devlin et al. \(2019\)](#) when applying BERT to any task).

Parameter Settings. We tune our traditional supervised models’ parameters using grid search, all as implemented in Python’s `scikit-learn` library ([Pedregosa et al., 2011](#)). Our best model uses unbalanced class weights, L2 penalty, and a constant term $C=10$, with other parameters at their default values. All cross-validation runs were initialized with the same random seed for comparability and reproducibility.

We train each of our neural models with the Adam optimizer ([Kingma and Ba, 2015](#)) for up to ten epochs with early stopping measured on the validation set. We apply a dropout rate of 0.5 during training in the recurrent layers and after the convolutional layers. We set our hidden sizes (i.e., the output of the recurrent and pooling layers) as well as our batch size to 128, and tune our learning rate to $5 \cdot 10^{-4}$; we set these parameters relatively small to try to work with our small data. We also experiment with scheduling the learning rate on plateau of the validation loss, and with pre-training the models on the Stanford Sentiment Treebank ([Socher et al., 2013](#)), a much larger sentiment dataset, to help combat the problem of small data, but this does not improve the performance of our neural networks.

3.3 Results

We present our results in [Table 3.7](#). Our best model is a logistic regression classifier with Word2Vec embeddings trained on our unlabeled corpus, high-correlation features (≥ 0.4 absolute Pearson’s r), 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

⁷Using the implementation available at <https://github.com/huggingface/pytorch-transformers>

Model	P	R	F
Majority baseline	0.5161	1.0000	0.6808
CNN + features*	0.6023	0.8455	0.7035
CNN*	0.5840	0.9322	0.7182
GRNN w/ attention + features*	0.6792	0.7859	0.7286
GRNN w/ attention*	0.7020	0.7724	0.7355
n-gram baseline*	0.7249	0.7642	0.7441
n-grams + features*	0.7474	0.7940	0.7700
LogReg w/ pretrained Word2Vec + features	0.7346	0.8103	0.7706
LogReg w/ fine-tuned BERT LM + features*	0.7704	0.8184	0.7937
LogReg w/ domain Word2Vec + features*	0.7433	0.8320	0.7980
BERT-base*	0.7518	0.8699	0.8065

Table 3.7: **Supervised Results.** Precision (P), recall (R), and F1-score (F) for our supervised models. Our best model achieves 79.80 F1-score on our test set, comparable to the state-of-the-art pretrained BERT-base model. In this table, “features” always refers to our best-performing feature set (≥ 0.4 absolute Pearson’s r). Models marked with a * show a significant improvement over the majority baseline (approximate randomization test, $p < 0.01$).

the majority baseline, the n-gram baseline, and the pre-trained embedding model (all 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, and we investigate the use of these features in the other model types. Particularly, we apply different architectures (GRNN and CNN) and different input representations (pretrained Word2Vec, domain-specific BERT).

We find that our logistic regression classifier described above achieves comparable performance to BERT-base (approximate randomization test, $p > 0.5$) with the added benefits of increased interpretability and less intensive training. Additionally, domain-specific word embeddings trained on our unlabeled corpus (Word2Vec, BERT) significantly outperform n-grams or pretrained embeddings, as expected, signaling the importance of domain knowledge in this problem.

We note that our basic deep learning models do not perform as well as our traditional supervised models or BERT, although they consistently, significantly outperform the

majority baseline. We believe this is due to a serious lack of data; our labeled dataset is orders of magnitude smaller than neural models typically require to perform well. We expect that neural models can make good use of our large unlabeled dataset. We believe that the superior performance of the pretrained BERT-base model (which uses no additional features) on our dataset supports this hypothesis as well.

		Agreement Threshold for Data			
		Any Majority	60% (3/5)	80% (4/5)	100% (5/5)
Features	None	75.40	76.31	78.48	77.69
	All	76.90	77.12	77.10	78.28
	LIWC	77.91	78.91	78.16	77.66
	DAL	75.58	77.06	78.05	77.06
	Lexical	76.42	77.92	77.54	77.88
	Syntactic	74.63	75.49	76.66	76.19
	Social	76.67	76.45	78.38	78.06
	$ r \geq 0.4$	77.44	78.76	79.80	78.52
	$ r \geq 0.3$	77.01	78.28	79.38	78.31
	$ r \geq 0.2$	77.53	78.61	79.02	78.28
	$ r \geq 0.1$	76.61	77.07	76.32	77.48

Table 3.8: **Feature Sets and Data Sets.** The results of our best classifier trained on different subsets of features and data. Features are grouped by type and by magnitude of their Pearson correlation with the train labels (no features had an absolute correlation greater than 0.5); data is separated by the proportion of annotators who agreed. Our best score (corresponding to our best non-neural model) is shown in bold.

In [Table 3.8](#), we examine the impact of different feature sets and levels of annotator agreement on our logistic regressor with domain-specific Word2Vec embeddings and find consistent patterns supporting this model. First, we see a tradeoff between data size and data quality, where lower-agreement data (which can be seen as lower-quality) results in worse performance, but the larger 80% agreement data consistently outperforms the smaller perfect agreement data. Additionally, LIWC features consistently perform well while syntactic features consistently do not, and we see a trend towards the quality of features over their quantity; those with the highest Pearson correlation with the train set (which all happen to be LIWC features) outperform sets with lower correlations,

which in turn outperform the set of all features. This suggests that distress detection relies on lexical information more than syntactic or social, and in particular, resources developed with psychological applications in mind, like LIWC, are very helpful.

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 stress, 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), and further work should be aware of them. We include some examples of these errors in [Table 3.9](#).

3.4 Discussion and Conclusions

Our examination of the Dreddit dataset tells us about how people express distress online in a variety of situations. As described in [§3.1.3](#), the topic a person is discussing (which in our case loosely corresponds to the stressor that is influencing them) and the way they are presenting themselves (e.g., to ask for help from others vs. to vent frustration) modulates the type of language they use. Lexical features that are overall useful to detecting distress, such as the amount of social words and explicit expressions of negative emotions, vary by domain when distress and non-distress are grouped together ([Table 3.4](#)).

We find distress detection to be a highly lexical problem. Across domains, people who are experiencing distress use more first-person pronouns, more negative emotion words, less positive emotion words, and less social words, focusing on themselves and their own (negative) feelings ([Table 3.5](#)). These trends reflect prior findings across

psychological and social science research. Individuals experiencing negative mental (and physical) health tend to focus on themselves as individuals and use first-person singular pronouns (Zimmermann et al., 2013; Lyons et al., 2018) while first-person plural pronouns emphasize group identity and are less prominent (Orvell et al., 2022). Use of negative emotion words to explicitly express these struggles is also well-documented (Lyons et al., 2018; Guo et al., 2022), and it is additionally known that a lack of close social relationships is related to negative mental (and physical) health outcomes (Umberson and Montez, 2010).

We note that the three most highly-correlated features in our training dataset reflect these trends as well (Section 3.3): a count of first-person pronouns by itself is highly informative, and the other two features are LIWC’s “summary” features, combining other counts in such a way as to show differences between them (Pennebaker et al., 2015). The emotional tone feature reflects the difference between positive and negative emotion words (Cohn et al., 2004) and the clout feature reflects the difference between first-person singular and plural pronouns (Kacewicz et al., 2014).

Additionally, while we find that there are non-significant but consistent trends presenting expressions of distress as longer, more complex, and prone to simpler words than non-distress (Table 3.6), syntactic features are not helpful for our classification models. We hypothesize that this may be due to the nature of our dataset; perhaps since expressions of distress are diffused throughout a post, a five-sentence chunk is not the correct granularity to observe effects of length and single sentences or, alternately, entire posts would be more illuminating.

Finally, we again draw attention to the effects of lexical diversity and interannotator agreement (Figure 3.3). It is a feature of highly-agreed-upon text in this dataset that it uses the same words repeatedly, whether distressful or non-distressful. Because this analysis is correlational, we cannot infer a specific causal relationship. However, it does

raise interesting new research questions about whether annotators may perform such annotation using a form of keyword searching.

The models we present in this chapter struggle with some finer aspects of distress detection. For example, since they rely heavily on lexical features, they can easily miss indirect expressions of distress, and they struggle to correctly interpret the writer’s framing and intentions (such as to identify that an author is telling a story about someone else who is in distress, and therefore even if they are using emotional words, this may not be evidence that the *author* is expressing distress). Our labeled dataset is also limited in size and drawn from one particular English source (Reddit), which imposes limitations on the demographics of people it can accurately represent.⁸ Finally, we note that the subjective nature of distress leads to variable interannotator agreement, and a more accurate framing of distress might be on a continuous scale.

Our contributions in this work include (1) the Dreddit dataset, a large unlabeled dataset of expressions of distress on Reddit (187,444 posts) with 3,553 segments labeled by multiple annotators for the binary presence of absence of distress; (2) an analysis of the content of the dataset by domain, label, and annotator agreement showing that it conforms to known trends in psychological literature and contains a variety of different expressions of distress; and (3) non-neural and neural models that achieve 80.65% F1 score on our dataset, displaying competence at the problem of distress detection but showing that it is a challenging problem with room for improvement.

⁸<https://www.alphr.com/demographics-reddit/>

Text	Gold Label	Agreement	Subreddit Name	Models Failed
Hello everyone, A very close friend of mine was in an accident a few years ago and deals with PTSD. He has horrific nightmares that wake him up and keep him in a state of fright. We live in separate provinces, so when he does have his dreams it is difficult to comfort him. Each time he calls, and I struggle with what to say on the phone.	Not Stress	60%	ptsd	Both
I asked the other day if they've set a date. He laughed in my face and said 'no' as if it were the most ridiculous thing he's ever heard. He comes home late, and showers immediately. Then, he showers every morning before he leaves. He doesn't talk to my mum and I, at all, and he's cagey and secretive about everything, to the point of hostility towards my sister.	Stress	60%	domesticviolence	BERT
If he's the textbook abuser, she is the textbook victim. She keeps giving him chances and accepting his apologies and living in this cycle of abuse. She thinks she's the one doing something wrong. I keep telling her that the only thing she is doing wrong is staying with this guy and thinking he will change. I tell her she does not deserve this treatment.	Not Stress	100%	domesticviolence	LogReg

Table 3.9: **Error Analysis Examples.** Examples of test samples our models failed to classify correctly. “BERT” refers to the state-of-the-art BERT-base model, while “LogReg” is our best logistic regressor.

Chapter 4

Multitask Models for Distress Detection

In [Chapter 3](#), we began studying the problem of psychological distress detection. We defined distress as a negative reaction to expectation or change and chose to focus on it because of the negative effects it can have in excess ([Lupien et al., 2009](#); [Calcia et al., 2016](#); [Stewart et al., 2019](#); [Rozanski et al., 1999](#); [Harris et al., 2017](#)) and its prevalence in the modern world.¹ This situates our work in an area where we believe it has great potential to do social good. Work in this area is particularly relevant because people worldwide lack access to mental health services for many reasons ([Coombs et al., 2021](#); [Wainberg et al., 2017](#)), including social stigma and a shortage of mental health professionals ([Satiani et al., 2018](#)).² We collected the Dreddit dataset, which is a dataset of segments of Reddit posts annotated for the binary presence or absence of distress, and benchmarked it, showing it to be a challenging dataset with many different types of stressors and situations represented.

With our data collected, we proceed to develop more sophisticated models to tackle this problem. For any model aiming to perform this task, we believe that two major aspects are vital to consider: psychological distress detection is a sensitive task in which mistakes by black-box models may lead to serious harms; and it is difficult to train robust, capable models for this task without an abundance of labeled data. We propose in this chapter that multitask learning can help solve both of these issues when the

¹<https://www.apa.org/news/press/releases/stress/index?tab=2>

²<https://www.apa.org/pubs/reports/practitioner/2022-covid-psychologist-workload>

additional tasks are selected according to domain-specific knowledge from psychology and associated fields.

The consequences of blindly trusting a model and mislabeling distress levels could be serious in a deployed application such as a therapeutic chatbot, where some users might not receive the help they need with the appropriate urgency.³ Research in AI and medicine suggests that for a high-stakes problem such as mental healthcare, it is imperative that users be able to trust a model’s predictions, which requires some understanding of how its decisions are made (Joyce et al., 2023; Chaddad et al., 2023).⁴ Furthermore, models that make decisions based on emotions that are known to co-occur with stress will be easier for humans to understand, and their mistakes will be more obvious. Therefore, in this chapter, it will be our focus to build explainable such models that draw on information such as emotion in order to make judgments about distress, and to understand how we can accomplish this influence of the models’ decision-making.

Meanwhile, given that the labeled data for this problem is quite small for training a supervised model, a common approach might be found in multi-task learning, where data labeled for tasks other than the primary problem are used to train the model (Caruana, 1993). We note that psychological distress has been shown to interact with emotion (Lazarus, 2006; Thoern et al., 2016; Levenson, 2019), a task that has far more publicly available labeled textual data. In a very traditional “hard parameter sharing” setting (such as Caruana (1993)), a model is trained to perform multiple tasks using a stack of shared parameters and some task-specific parameters, typically at the same time. This means that we would expect to have a single input with labels for multiple tasks. However, there are no currently available datasets labeled for both psychological distress and emotion, and therefore in this chapter we will present variations on traditional multitask learning. In this work, we experiment with a suite of three approaches to

³<https://www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness>

⁴<https://pewrsr.ch/3KqRDvn>

incorporating semantically related tasks—here, emotion classification—to create *emotion-infused* models capable of equally competent, but explainable, psychological distress detection as compared to a black-box model.

In this chapter, we will first describe the data we use to supplement our Dreddit dataset with emotion detection tasks (Section 4.1), then describe our three emotion-infused multitask models (Section 4.2), present their performance (Section 4.3), and finally conclude with a discussion of how multitask training changes the way the models make their predictions (Section 4.4).

The main contributions of this work (based on work published in Turcan et al. (2021a)) are as follows: (1) a model development paradigm that incorporates domain-specific knowledge from psychological fields of research in order to enhance machine learning models; (2) a suite of three approaches to emotion-infused models for predicting psychological distress, with performance comparable to our best models from Chapter 3; and (3) a framework for interpreting our models to show the impact of incorporating domain-specific knowledge in our multitask training.

4.1 Data

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

Table 4.1: **Distress and Emotion Data.** The datasets we use in this work and their relative sizes (in terms of total number of data points).

A comparison of all the datasets we use in this work can be seen in Table 4.1. We focus on our Dreddit dataset (Chapter 3) as the target task: binary distress classification

framed in terms of the author and the time—i.e., a post is labeled distressful only if the poster themselves is currently expressing distress.

Because this dataset is small for training a deep learning model, we also experiment with larger datasets to provide auxiliary information. We select the GoEmotions dataset (Demszky et al., 2020), which consists of 58,009 Reddit comments labeled by crowd workers with one or more of 27 emotions (or Neutral), for its large size and genre similarity to Dreddit. Here, 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 together into the Ekman 6 basic emotions (anger, disgust, fear, joy, sadness, surprise, neutral) (Ekman, 1992) (GoEmotions_{Ekman} or GoEmotions_E), and one into simple polarity (positive, negative, ambiguous, neutral) (GoEmotions_{sentiment} or GoEmotions_S). We run our experiments with each version of this dataset.

We also explore the use of another social media website, Vent. Vent is a platform more similar to Twitter or Tumblr than Reddit, where users post vents of any length, tag them as they like, and other users react to them or post comments. The benefit of Vent for this purpose is that posters self-identify some emotion they are feeling from a large list of pre-made emotions. The data we use is collected by Malko et al. (2021) (see also Malko et al. 2023).⁵ Guided by research that links chronic stress to mood disorders like depression as well as anxiety disorders (e.g., Khan and Khan (2017)), we hypothesize that expressions of fear and sadness may be associated with distress. Because Vent is very large and Dreddit is very small, we then select Vent data only from a specific set of emotions to test this hypothesis. We select Vent data that has been labeled with fear or sadness, which we expect to be most relevant to the domains gathered for Dreddit, as well as joy, for a contrast.

⁵Due to license and ethics policy restrictions, we do not make this data publicly available.

We note that this dataset is strictly single-class, whereas, as published, GoEmotions may have more than one emotion label per data point. In all, there are 1.6M vents in our dataset, much larger than Dreddit 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 originally labeled with fear, joy, or sadness, for a final set of 4,136 data points (3,342 of which are the train set). We call this subset GoEmotions_{FSJ}, and we compare it against Vent to see whether genre similarity or data size is more important in this multitask setting.

4.2 Methods

We present several different architectures for our emotion-infused models (§4.2.1) and compare them against our top-performing deep neural model from Chapter 3. Our emotion-infused models span three different ways to incorporate emotion information into our distress detection models, divided into multi-task learning and fine-tuning; each of the three approaches shares knowledge of the two tasks through the parameters of the BERT encoder, following the hard parameter sharing paradigm. In our setting, we consider distress detection to be our primary task and examine how the inclusion of emotion detection objectives during training changes the way this task is performed by our models.

4.2.1 Models

Multi^{Alt}. Our first multi-task models, which we refer to as Multi^{Alt}, are simply two single-task models sharing the same base BERT representation layers. The models are

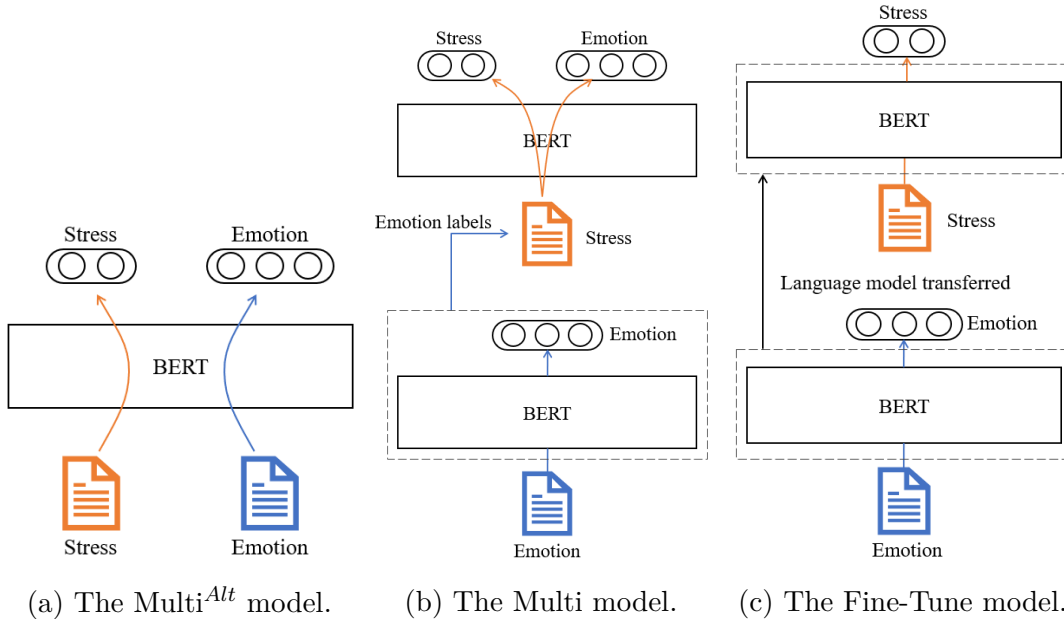


Figure 4.1: **Emotion-Infused Models.** The emotion-informed architectures we use in our experiments.

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 Dreddit data and the emotion task with the GoEmotions or Vent data.

We refer to the variants with a subscript, i.e., Multi^{Alt}_{GoEmotions_A} (i.e., *all* emotions from GoEmotions), Multi^{Alt}_{GoEmotions_E} (i.e., *Ekman* emotions only), Multi^{Alt}_{Vent} (i.e., the Vent data), etc. The Multi^{Alt} models can be seen in Figure 4.1a. One loss step for these models consists of only one dataset and task, so they are trained with the negative log-likelihood (NLL) loss for single-label tasks (Dreddit, Vent, GoEmotions_{FSJ}) and the binary cross-entropy (BCE) loss for multi-label tasks (GoEmotions_{A,E,S}).

Multi. We also experiment with a multi-task learning setup where we perform the two tasks at the same time on the *same input data*. We call this architecture Multi. However, because the Dreddit data 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 Dreddit. We then take these emotion labels to be “silver data” and

train on them alongside distress. The Multi model can be seen in [Figure 4.1b](#). 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 losses of the distress task and the emotion task are summed together for each batch with a tunable weight parameter, i.e., $\mathcal{L} = \lambda\mathcal{L}_{\text{distress}} + (1 - \lambda)\mathcal{L}_{\text{emotion}}$.

Fine-Tune. We experiment with models in which we first endow the BERT representation with knowledge of the emotion task by fine-tuning and then apply it to 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 parameters to initialize a BERT model that we continue to fine-tune on Dreddit. We denote these models as, e.g., Fine-Tune_{GoEmotions_A→Dreddit} for a model that was first trained on GoEmotions_{all} and then on Dreddit (for space, we will abbreviate Fine-Tune as FT). These fine-tuning models can be seen in [Figure 4.1c](#). These models are trained with the NLL and BCE losses as in the Multi^{Alt} models.

4.2.2 Baselines

We use as our primary baseline a re-implementation of our BERT-based classifier from [Chapter 3](#). We re-implement the model so that we can examine its variance and report an average of 3 runs with distinct random seeds. On average, the re-implementation scores lower than the single model reported, but with high variance. Because of this, we assume that our previous performance is from the high end of this variance and use our average score as our baseline in this work. This model is a pre-trained BERT language model (released as `bert-base-uncased` by [Wolf et al. \(2019\)](#); we use this same pre-trained language model as the basis for all our models)

followed by a dropout layer and a dense classification layer. We also report a recurrent neural network (RNN) model, which uses either a long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) or a gated recurrent unit (GRU) (Cho et al., 2014) in place of the transformer from BERT and is otherwise the same. These models are trained with the NLL and BCE losses as with the Multi^{Alt} models.

4.2.3 Training

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

We tune hyperparameters for all our models using Bayesian Optimization from the Python library `ax`.⁶ All models train the initial learning rate of the Adam optimizer and the dropout probability before the final classification layer; the Multi models also tune the loss weight parameter λ , and we also note that the RNN model tunes additional parameters such as the type of RNN unit, hidden dimension, etc. For all models, we tune parameters based on the F1 score on the Dreddit development set; we train an ensemble of three models with three different, fixed random seeds and average their performance for a given parameter setting. We report the mean and standard deviation of three models, with three different random seeds, trained with the best hyperparameters.

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

⁶<https://github.com/facebook/Ax>

with one CPU.

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

Table 4.2: **Hyperparameter Search Ranges.** Hyperparameter ranges for our models. BERT-based models tuned the first two; the Multi models additionally tuned λ ; the RNN additionally tuned the remainder.

Model	Learning Rate	P(dropout)
RNN	1.40×10^{-4}	0.86
BERT	4.27×10^{-5}	0.13
Multi ^{Alt} _{GoEmotions_A}	8.47×10^{-6}	0.40
Multi ^{Alt} _{GoEmotions_E}	1.08×10^{-5}	0.00
Multi ^{Alt} _{GoEmotions_S}	1.69×10^{-5}	0.00
Multi ^{Alt} _{GoEmotions_{FSJ}}	8.98×10^{-6}	0.00
Multi ^{Alt} _{Vent}	4.44×10^{-5}	0.00
Multi _{Dreaddit_S}	1.14×10^{-5}	0.00
Multi _{Dreaddit_{FSJ}}	1.79×10^{-5}	0.00
FT _{GoEmotions_A→Dr}	7.30×10^{-5}	0.05
FT _{GoEmotions_E→Dr}	1.35×10^{-5}	0.00
FT _{GoEmotions_S→Dr}	1.95×10^{-5}	0.09
FT _{GoEmotions_{FSJ}→Dr}	5.03×10^{-6}	0.03

Table 4.3: **Selected Hyperparameter Values.** Our models’ selected hyperparameters for Dreaddit.

Hyperparameter tuning was done the same way for every model, with Bayesian optimization as implemented by `ax`, with the F1 score on the Dreaddit development set as the criterion to optimize. Multi_{Dreaddit_S} and Multi_{Dreaddit_{FSJ}} were given 35 trials for time constraints; all other models were given 50 trials. All models were trained with a patience of 5 epochs and a tolerance of 0.0001 for dev set improvement, and allowed to run for a maximum of 20 epochs. All models tuned the initial learning rate and the

dropout probability, with the Multi models also tuning the lambda weight parameter between their two task losses. Additionally, the RNN model was initialized from scratch and additionally tuned the embedding dimension, hidden dimension, number of layers, and type of RNN. Our parameter search ranges are shown in [Table 4.2](#), and the selected hyperparameters for our models are given in [Table 4.3](#).

4.3 Results

Model	Binary F1	Accuracy
RNN	67.58 ± 1.22	68.86 ± 1.10
BERT	78.88 ± 1.09	79.11 ± 1.32
Multi ^{Alt} _{GoEmotions_A}	79.02 ± 0.35	79.72 ± 0.69
Multi ^{Alt} _{GoEmotions_E}	80.24 ± 1.39	81.07 ± 1.13
Multi ^{Alt} _{GoEmotions_S}	79.46 ± 1.05	79.86 ± 0.50
Multi ^{Alt} _{GoEmotions_{FSJ}}	79.17 ± 0.61	78.69 ± 1.86
Multi ^{Alt} _{Vent}	80.34 ± 1.39	79.67 ± 2.03
Multi _{Deaddit_{sentiment}}	78.97 ± 0.24	78.55 ± 0.07
Multi _{Deaddit_{FSJ}}	78.90 ± 0.59	78.55 ± 0.07

Table 4.4: **Multitask Distress Detection Performance.** Results of our multitask models. The best result under each metric is bold.

We report the results of our multi-task models in [Table 4.4](#)⁷. 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, Multi^{Alt}_{Vent} performs best. With regards to GoEmotions, the 28-way classification of GoEmotions_A naturally leads to lower numerical performance than the tasks with smaller numbers of classes, and we expect that GoEmotions_S may group too many distinctly labeled emotions together under the same emotion labels; it seems GoEmotions_E is the happy medium for this model. We also note that the Multi^{Alt}_{Vent} and Multi^{Alt}_{GoEmotions_E} models

⁷We 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.

perform equally well, which indicates that the genre mismatch is not an issue for this problem, or that Vent has a similar enough genre to Reddit that it does not affect the results. Somewhat surprisingly, $\text{Multi}_{GoEmotions_{FSJ}}^{Alt}$ does not do as well as $\text{Multi}_{Vent}^{Alt}$; however, the GoEmotions data is much smaller than Vent, especially when subsampled to select specific emotions.

Model	Binary F1	Accuracy
BERT	78.88 ± 1.09	79.11 ± 1.32
$\text{FT}_{GoEmotions_A \rightarrow Dreddit}$	76.40 ± 0.50	76.83 ± 0.40
$\text{FT}_{GoEmotions_E \rightarrow Dreddit}$	79.44 ± 0.29	79.53 ± 0.46
$\text{FT}_{GoEmotions_S \rightarrow Dreddit}$	79.75 ± 0.52	80.61 ± 0.40
$\text{FT}_{GoEmotions_{FSJ} \rightarrow Dr}$	80.25 ± 0.24	80.98 ± 0.20

Table 4.5: **Fine-Tuning Distress Detection Performance.** Results of our fine-tuning models. The best result under each metric is bold.

We further report the results of our fine-tuning models in [Table 4.5](#). Because we expect that genre similarity should play a larger role when the secondary task can offer no direct training signal during the primary task fine-tuning, we evaluate on GoEmotions here and not Vent. Here, we observe that our best model, $\text{Fine-Tune}_{GoEmotions_{FSJ} \rightarrow Dreddit}$, scores at least one standard deviation above BERT. We see higher increases in performance for the simpler classification problems in GoEmotions_S and GoEmotions_{FSJ} and worsened performance for GoEmotions_A, suggesting that in the sequential paradigm, more complex problems are not able to interact appropriately with the main task and instead interfere.

Dataset	Macro F1
GoEmotions_A	48.98
GoEmotions_E	62.16
GoEmotions_S	69.65
GoEmotions_{FSJ}	91.87

Table 4.6: **Fine-Tuning Emotion Detection Performance.** Performance of our fine-tuning BERT models on the different GoEmotions labelings and datasets.

We also report the performance of the fine-tuning BERT models we trained on GoEmotions in order to label Dreddit with emotion in [Table 4.6](#); these results track well with the fine-tuning results reported by [Demszky et al. \(2020\)](#). Because these models are intermediates used for labeling, we report the F1 scores of the single model we actually used for labeling, although we tuned their parameters with an average of 3 different instances as with all other models. Many-way classification problems have much more opportunity for error and noise in an already-noisy process of labeling unlabeled data, so we use only the two best-performing GoEmotions models, which are those trained on the fewest-label datasets, GoEmotions_S and GoEmotions_{FSJ} , for our Multi models.

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

4.4 Analysis

We perform three different analyses to probe our trained models and discover what information they learn to use. For our Multi^{Att} 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.

	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 4.7: **Learned Task Correlations.** Correlations of the gold labels for each dataset with labels predicted by the other classifier in a Multi^{Alt} model. GoEmotions_{FSJ} (abbreviated for space as GoEmo_{FSJ}) is starred because its emotion data is not multi-label and therefore the correlation ratio η is used instead of the coefficient of determination R^2 (which is used for the other, multilabel GoEmotions variants).

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

Table 4.8: **Learned Task Correlations Per Emotion Label.** 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.

4.4.1 Multitask Knowledge

We perform an analysis of our Multi^{Alt} models to see what information they learn about emotion.⁸ We take the development sets of each of the datasets (Dreaddit and GoEmotions) and predict their labels under the other task (i.e., emotion for Dreaddit and vice-versa). We report the correlation of these predicted labels with the gold labels in Table 4.7⁹. In this case, the GoEmotions_{FSJ} variant is a single-label three-way classification problem, so we report the correlation ratio η (Fisher, 1925). 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 4.8 and include the

⁸We did perform an equivalent analysis on the Multi models, which shows similar trends, but as Multi^{Alt} shows better performance, we omit it for space.

⁹We also note the possibility that different combinations of emotions are relevant to distress; however, not enough of our data is labeled with multiple emotion labels (4% of Dreaddit’s silver labels from GoEmotions_S, 9% of GoEmotions_E) to test this hypothesis in this work.

All and Ekman sets as well as the Vent data in the appendix.

We observe that our multi-task models generally learn a moderate correlation between the 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, which makes intuitive sense. 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 stress. 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.

4.4.2 LIWC Analysis

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

We first analyze the unigrams our various models use to perform distress classification using LIME. LIME accepts an input from our development set, perturbs it in the bag-of-unigrams space, and runs one of our classifiers on each perturbation to calculate the importance of various unigrams; through this process, we acquire the 10 unigrams with the highest magnitude output by LIME for each development example and consider them “explanations”. We thus have 2,760 individual unigram explanations for the entire development set to analyze.

We then use the word lists from LIWC 2015’s 72 psychological categories to see what types of words each classifier tends to use to make decisions of distress vs. non-distress.

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

Table 4.9: **Rationale Prevalence by LIWC Category.** A comparison of how often several of our models rely on words from several LIWC categories to make their decisions, according to LIME. These numbers represent the percentage of available LIWC words each model selected in the top 10 LIME explanations for the entire dev set. Dr is Dreddit, and GE is GoEmotions.

An abbreviated list of results, showing our best models from each category, is shown in Table 4.9.¹⁰ We observe small but consistent effects suggesting that, in comparison to the basic BERT model, our emotion-enhanced models broadly learn to use the following information:

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

Cognitive processes. All models show some increase in using words related to cognitive processes as compared to BERT; however, its subcategory Certainty, which includes words about absoluteness such as *never*, *obvious*, and *clearly*, shows larger changes. For example, Multi_{Dreddit_{FSJ}} uses Certainty twice as often as BERT. These cognitive words seem to target the mental aspects of distress. Rumination and a focus

¹⁰The full table does not fit comfortably on a page, and so is available at <https://github.com/eturcan/emotion-infused-explanations>

on absoluteness are known signs of anxiety disorders, an extreme form of chronic distress (Nolen-Hoeksema et al., 2008; Miranda and Mennin, 2007).

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

4.4.3 Salient Words

We then investigate the data itself for highly significant words using the measure of relative salience proposed by Mohammad (2012):

$$\text{RelativeSalience}(w|T_1, T_2) = \frac{f_1}{N_1} - \frac{f_2}{N_2} \quad (4.1)$$

That is, it measures the importance of a token w in two different corpora T_1, T_2 by subtracting their two relative frequencies (where f_1, f_2 are the counts of token w in each corpus and N_1, N_2 are the total tokens in each corpus). We compute this measure for all words in the Dreddit training data, taking our two corpora to be the subsets labeled distress and not-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 4.10 with the full list of words available in the appendix. We examine the words and divide them into related groups in

	Category	Emotion Words
Distress	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-Distress	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 4.10: **Example Relative Salience Terms.** Some examples of words identified by relative salience on the Dreddit training data as indicative of distress or not-distress. We group the words by hand into semantically meaningful categories for ease of understanding.

order to understand what types of information should theoretically be most important to classifying the data. For example, we see that different sets of function words are actually among the most important for both classes, with words like conjunctions typically appearing more indicative of distress (which echoes our previous finding that distressful 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 positive 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 then look at the intersection between relative salience and LIME explanations, counting how many LIME explanations are highly salient words for distress or non-distress; results are shown in Table 4.11. We see that both the BERT baseline and our emotion-infused models show similar patterns of behavior, including a higher proportion of distress salient words ($\sim 32\%$) than non-distress salient words ($\sim 16\%$). Most, though not all, emotion-infused models learn to increase the proportion of non-distress words and decrease the proportion of distress words (e.g., $\text{Multi}_{Vent}^{Alt}$ reduces distress words by 2% and increases non-distress words by 3.5%). We note that the Dreddit dataset is nearly balanced (non-distress is the minority class by a slim margin; see §3.1.2) and

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

Table 4.11: **Rationale Prevalence by Relative Salience.** A comparison of how often several of our models rely on words identified as salient for distress or non-distress to make their decisions, according to LIME. These numbers represent the percentage of available relative salience words each model selected in the top 10 LIME explanations. So that the table fits on a page, Dreaddit is Dr, and GoEmotions is GE.

that from manual examination, similar categories of words are identified as most salient for distress and non-distress (Table 4.10).

We can hypothesize several things from these results. First, we observe high-intensity active words among the negative sentiment words identified for distress (*cry, hate*) as opposed to lower-intensity passive words for non-distress (*thank, hope, nice*). The nature of our problem means that distress is a very concentrated category of text, while non-distress is simply “everything else” and may have little cohesive identity. It is possible that salient words for distress are simply more distinctive or unified in some way that reflects in our models’ embedding space. Second, because we have seen our emotion-infused models rely on other classes of words (Table 4.9), it is possible that our emotion-infused models begin to rely on features other than collections of common words to identify distress. Perhaps they broaden their space to include similar words that may not individually be as common, or perhaps they look at combinations of words together. Further research is needed to test hypotheses such as these.

4.5 Discussion and Conclusions

In this work, we have examined how multitask learning can influence the way models make their predictions and allow these predictions to be more interpretable and understandable. We believe that it is key to select the tasks in this paradigm intelligently, perhaps according to domain knowledge; for example, guided by psychological literature suggesting that the experiences of distress and emotions correlate (Lazarus, 2006; Thoern et al., 2016; Levenson, 2019), here we select emotion detection as a task to help inform our distress detection.

Empirical results show that emotion and distress labels correlate reliably, although they are not entirely predictive of one another (Table 4.7, Table 4.8), matching with our intuition and prior psychological research. Models can therefore rely on similar types of information or cues to predict both tasks effectively, and we identify emotion as a good candidate problem for our purposes. We note that a model that can perform multiple tasks inherently gains a measure of human-understandability when the tasks are related; we may be able to look at an experienced emotion as an explanation for the distress prediction.

When we incorporate emotion classification into our distress prediction models, we find observable changes in how these models make predictions about distress as compared to single-task models, without sacrificing performance on the target task. In this case, we see that multitask models trained to perform emotion classification use intuitively emotional words at a higher rate as rationales for their distress predictions (Table 4.9). These include not only natural matches like positive and negative affective words, but also words that support less obvious connections, such as personal pronouns, cognitive processes like certainty and rumination, words that echo helplessness or community, and biological processes (Table 4.10, Table 4.11).

We limit our models in this chapter to using one additional task, emotion classification, to improve the interpretability of distress prediction, but intuitively many other types of tasks may be related, such as event detection (to identify stressors), personality or other demographic detection (since expressions and experiences of distress may vary from person to person), or any of a number of natural language understanding tasks (to give the model more grounding in what might be a distressing situation). As we have noted in [Chapter 2](#), biological signals including voice and facial expressions are also known strong indicators of distress (including, e.g., voice pitch, voice intensity, blinking; see [Giannakakis et al. \(2019\)](#); [Slavich et al. \(2019\)](#) for a more detailed treatment). We will treat event detection throughout [Part II](#), and leave these other tasks to future work. We also note that this work was completed before the development of large language models (LLMs) such as GPT-3 ([Brown et al., 2020](#)); while we utilize these models for similar problems in [Chapter 8](#), future work might revisit this problem with these improved model capabilities.

As in [Chapter 3](#), the data we use in this work limits the demographics of people our models may accurately represent. Additionally, while GoEmotions is drawn from the same website as Dreddit, Vent is a primarily Australian social media platform which may skew more heavily towards the LGBT+ community than Reddit,¹¹ meaning that the emotional experiences of users in this dataset may not match those found in our Dreddit dataset. Additionally, psychology research suggests that emotional expression and identification interact with various demographic categories such as gender and ethnicity, including, e.g., evidence for varying amounts of expressivity by race, gender, and age, as well as interactions between culture and the frequency of certain emotions ([Chaplin, 2015](#); [Mesquita et al., 2016](#); [Fan et al., 2021](#)). This means that our conclusions about the interactions of distress and emotion, too, may be limited by the demographics

¹¹https://www.huffpost.com/archive/au/entry/vent-the-australian-social-media-app-giving-millennials-a-safe_n_9676384

of our data.

Our contributions in this work include (1) a methodology for incorporating domain-specific knowledge into classifiers for a related task, namely, using multi-task learning to train the classifier to perform both tasks; (2) three specific approaches to executing this methodology to create emotion-infused models: the Multi^{Att}, Multi, and Fine-Tune models; and (3) a framework for analyzing and interpreting these models to show how multitask learning changes their decision-making process by identifying psychologically relevant categories of words using LIWC (such as affective words as well as cognition, drives, or biology) and showing that the multitask models rely more often on these types of words to make decisions about distress prediction than the single-task models.

Chapter 5

Discussion for Detecting Emotional Reactions

In this section of the thesis, we have presented work that collects a new dataset for the task of psychological distress detection across a variety of stressful situations (Chapter 3) and develops emotion-infused models on that data, using the related task of emotion detection to create more interpretable model behavior (Chapter 4). We have developed models that detect a certain kind of emotional state and can support their predictions with more basic emotional states and psychological processes (e.g., the presence of sentiment-bearing words or words conveying certainty); this is a single proof of concept that can be expanded to other emotional states by drawing on the proper psychology research for those emotional states. Looking back at our goals, we can say that duplicating such a model paradigm for various emotional states of interest would create a robust and explainable system for answering the question, “What am I feeling?” As we move forward into the second part of the thesis, we will focus on explaining not the predictions of the model, but the emotional reaction itself as we detect and explain the causes of emotional reactions. Before we move on, however, we will briefly close this part of the thesis with a discussion of our progress as well as future work that might help us attain our goals.

Two key themes in this section of the thesis include explainability and multitask learning. First, understanding the predictions of so-called “black-box models” (those whose parameters are not readily interpretable to humans, usually deep neural networks)

is difficult and can be approached from many different angles (see a recent survey on the subject, [Dwivedi et al. \(2023\)](#)). As the usage of AI becomes more widespread, including for sensitive and important tasks, we believe that it is vital to understand how our models are making their predictions so that we can identify when they may be mistaken. While the capabilities of modern language models are impressive, this does not mean that all of their outputs must be true or logical. Just as a patient might explain their rationale for a suggested diagnosis while still seeking the advice of a doctor (“My foot hurts when I put pressure on it, so I think I may have strained a muscle when I fell; do you agree?”), we envision AI models that can work in tandem with experts by making helpful predictions but also explaining their logic. Here, we make a first attempt at such models by training them to be abler to perform multiple relevant tasks at once.

Meanwhile, multitask learning has demonstrated great promise in helping develop capable, adaptable models for many applications (e.g., [Sanh et al. \(2021\)](#); [Wei et al. \(2021\)](#); [Wang et al. \(2022\)](#)). We believe that training a model on a variety of different tasks can help it make connections and express its stored knowledge in different ways (which we will continue to discuss in [Part II](#)). Since the work in this part of the thesis was completed, a proliferation of massively multitask datasets and models has arisen, leading to improved robustness, especially for smaller models such as those above.¹ In our work, we demonstrate ways in which multitask learning can be built into a model and increase explainability, helping models support their predictions in ways understandable by humans.

In summary, our current work has demonstrated a bespoke model that detects one type of emotional reaction using support from multitask learning. We can now

¹We note a separate phenomenon, *data contamination*, which affects large language models and their massive training datasets. This occurs when the training text data includes the content of datasets later used for evaluation, and can contribute to LLMs’ perceived few- or zero-shot performance ([Li and Flanigan, 2023](#)); see also <https://hitz-zentroa.github.io/lm-contamination/blog/>. This seems more of a concern for models such as the GPT family which are trained primarily on language modeling objectives, but research is ongoing.

envision several core directions for future work in this area. We hope that future work can explore drawing on knowledge from other fields to support the prediction of other emotional reactions, investigate how model capabilities for this task have changed with the advent of LLMs, and understand how emotional reactions may depend on differing demographic information and ensure that people from a much wider range of groups are accurately represented in emotional training data.

5.0.1 Contributions

The contributions covered in this part of the thesis are as follows:

A dataset and emotion-infused models for the problem of psychological distress detection. We present a dataset, Dreddit, for the new problem of psychological distress detection, along with multi-pronged analysis of its content (Chapter 3). We show that Dreddit contains expressions of distress as well as counterexamples from varying domains and note some features which may be correlated with distress, including emotional wording, sentence complexity, and first-person pronouns (a trend which is supported by psychology literature). We benchmark contemporaneous state-of-the-art models against this dataset and proceed to develop three types of emotion-infused models that draw on psychology research linking distress and particular common emotions to yield more understandable results (Chapter 4). Our Multi, Multi^{Alt}, and Fine-Tune models use visibly different keywords to make their predictions about distress as compared to a single-task baseline, focusing more on words representing emotional, cognitive, and social processes.

An approach to the problem of psychological distress detection that is inspired by psychological theory. We first define psychological distress by drawing on established psychological theory (Selye, 1976b) (Chapter 3). We observe trends in our data that we can corroborate with published psychology research, such as the idea that

people experiencing distress tend to focus on the self and use more first-person singular pronouns (Zimmermann et al., 2013; Lyons et al., 2018). Then, when developing our emotion-infused models, we draw on work that suggests that distress is correlated with negative emotions (Lazarus, 2006; Thoern et al., 2016; Levenson, 2019) to develop our multitask models. Finally, when analyzing the results of our emotion-infused models, we note that we can support their behavior with published work, such as relating words about “certainty” to distress (Nolen-Hoeksema et al., 2008; Miranda and Mennin, 2007). In summary, we leverage research from this field in both our model creation and our analysis of results, and we find that we can create more explainable models as well as produce empirical support for various psychological theories.

5.0.2 Limitations and Future Work

We acknowledge that the work we present in this thesis is limited in various ways stemming from all aspects of its implementation. First, and perhaps most importantly, our models are trained on a limited amount of data from a specific source and cannot necessarily accurately represent emotional experiences outside of that training data. As we have previously noted, our Dreddit dataset is composed of English-language data from the social media website Reddit, where the userbase leans heavily towards educated white non-Hispanic males in the United States.² This means that individuals of other demographics are not likely to be well-represented in our work. Additionally, while personal narrative is an intuitive way to represent emotional experiences, this means that our models may have trouble generalizing to different genres, such as microblogs (which have far less context). We hope that, in the future, data of various genres accurately representing the emotional reactions of a wide range of people can continue to be gathered and made available for training models like ours.

²<https://www.alphr.com/demographics-reddit/>

Second, we note that the work in this part of the thesis was done before the advent of large language models (LLMs), and therefore our model performance is limited by the types of models we chose at the time. Modern LLMs have shown exceptional performance across a wide range of tasks (Liang et al., 2023), and it is possible that they may be able to demonstrate enhanced performance and explainability for ours. One of the benefits of smaller models like ours over LLMs is that we can build expert knowledge explicitly into the design of the model; however, LLMs do possess knowledge of a wide range of fields through their massive training data and are capable of generating fluent justifications for their predictions. We will investigate using LLMs to perform more complicated emotional tasks in [Chapter 8](#), although not to tackle the same problems as in the first part of this thesis.

Third, the work we have presented in this section of the thesis focuses on one particular type of emotional reaction, psychological distress. There are, naturally, many other types of emotional reactions, and many of them can be supported by their own theories from psychology and other fields (for example, in the field of psychotherapy, anger is often framed as a *secondary emotion*, meaning that it follows some other more vulnerable emotion like fear or sadness to protect the experiencer, e.g., Pascual-Leone et al. (2013)³). Building a bespoke model to detect and explain each different emotional state we wish to study can help us test various theories of emotion empirically, but there is also room to build modern models that can detect a wide range of emotions and explain themselves with different expert knowledge.

³The theory presented describes primary and secondary emotions, but it does not uniquely assign each emotion to one category. Anger is used as an example here because it is frequently described as a secondary emotion in the field, but it may also be understood as a primary emotion in some contexts.

Part II

Understanding Emotions and Their Causes

Introduction to Understanding Emotions and Events

While detecting emotional reactions by themselves is an important problem, the overall goal of improving our mental health relies on understanding the reasoning for our emotions and then deciding what changes to make. We focus our efforts in this second part of this thesis on understanding a key component of emotional reactions: the stimuli that cause them. Stimuli can take many forms, including entities (for example, a spider causing fear), various types of events (a birthday party causing joy), and many more complicated situations (perhaps stress is due to a combination of overlapping factors all at once more than any single factor alone, such as financial strain plus physical injury and interpersonal conflict). For simplicity, in this thesis, we will primarily consider situations in which we can identify a singular stimulus, although our work in [Chapter 8](#) will touch on more complex situations as well.

Our goal in this work is to contribute to a more detailed understanding of emotional reactions beyond simple identification. One method for reaching this more detailed understanding is to impose a structure on the emotional reaction—in prior work, this is often done by imposing relations between the emotion and other entities in the document and identifying those entities. Certain roles can be associated with a given emotional reaction, such as the cause or experiencer, and can be extracted from a document in various ways, such as span tagging. For example, in the sentence “Alice was frustrated with the long line at the bakery”, we can identify the emotional state *frustrated* and say that *Alice* is its experiencer and *the long line at the bakery* is its cause;

presenting these pieces of information together gives us a more thorough and actionable understanding of the situation than does the emotional state by itself. The cause of an emotion can be viewed in this framework, where the stimulus of an emotion is some *thing* in the text (whether a noun phrase entity, some type of verb event, or an arbitrary span of text); in [Chapter 7](#), we will take a span-tagging approach to identifying the cause of an emotion.

Another approach to understanding emotional reactions is to take a less structured view of the input document and aim to generate a free-text explanation of the emotional state described. Whereas a structured description of an event may be useful to another machine application that can expect input in a particular format, a free-text explanation may be more useful to a human—for example, a therapist responsible for multiple patients who may need a refresher when each patient returns. We can imagine interpretative summaries of an input document that explain what emotional reaction is being felt, why it has happened, how the experiencer is dealing with their emotions, and much more. While this type of long-form generation problem has historically been challenging for small models, the advent of LLMs and their remarkable generation capabilities has recently opened up new doors in this area. In [Chapter 8](#), we will explore the capabilities of LLMs and smaller distilled models to generate this type of explanation.

As in [Part I](#), we will ground our work in this part in multitask learning as a way to improve the explainability and performance of our models. In [Chapter 7](#), the connectedness of our target tasks motivates our decision to perform them jointly or in sequence, using one task’s predictions as input information for the other; in this setting, we find that multitask learning improves model performance. Meanwhile, in [Chapter 8](#), we draw on current literature and fine-tune a distilled generative model with multiple relevant auxiliary tasks, examining their effects on the emotional explanations

it generates while mimicking an LLM. Our work in this part of the thesis draws on a broader selection of auxiliary tasks than that in [Part I](#) and examines the potential for multiple auxiliary tasks to influence a model at once.

This part of the thesis will begin with an overview of related work in [Chapter 6](#), including a discussion of NLP research about emotion cause detection (and other semantic roles related to emotion causes) and about emotional explanation generation.⁴ In [Chapter 7](#), we will present three varieties of joint models that perform the emotion detection and emotion cause tagging tasks in tandem by relying on commonsense reasoning as well as different multitask paradigms ([Turcan et al., 2021a](#)). Then, in [Chapter 8](#), we will present our work on generating emotion-focused explanations using distilled models, where we define a new, more complex problem for understanding emotional reactions and experiment with multitask learning to improve models' capability to perform it (we intend to publish this work in the future). Finally, we will conclude with a discussion in [Chapter 9](#).

The contributions covered in this part of the thesis are as follows:

- Multitask and commonsense-infused models for joint emotion detection and emotion cause extraction in news headlines, showing performance gains ([Chapter 7](#));
- A dataset for the new problem of emotion-focused explanation, and a characterization of the abilities of distilled generation models (based on large language models) for this problem ([Chapter 8](#)); and
- An approach to these problems that focuses on incorporating external knowledge using multitask learning ([Chapter 7](#), [Chapter 8](#)) and commonsense reasoning ([Chapter 7](#)) to enhance performance.

⁴We refer the reader to our first section of related work, [Chapter 2](#), for our broad coverage of psychology research on emotion as well as multitask learning in NLP, since these features are common to both parts of the thesis.

Chapter 6

Related Work for Understanding Emotion and Events

6.1 Emotion-Cause Detection

As we noted in [Chapter 2](#), most modern psychology research accepts that emotions usually have a stimulus that incites them (along with some other key features such as behaviors that accompany them), regardless of the definition of emotion used ([emo, 2023](#); [Lazarus and Folkman, 1984](#); [Scherer, 2005](#)).¹ As it is, of course, impossible to directly measure what emotion someone is feeling in response to any kind of stimulus, psychology researchers must match specific stimuli to specific emotions (usually words, pictures, or videos when in a laboratory setting) ultimately through self-report and common sense ([Pascual-Leone et al., 2015](#)).²

In NLP, much work has looked at identifying the causes of emotions in text, sometimes along with other salient role information such as the experiencer of the emotion or its target. This work contributes to a more structured understanding of emotions as compared to simply detecting their presence or ranking their intensity by taking into account emotions' inherent relationships to other entities in the world of their experiencers.

¹Although, of course, research is hardly unanimous on the causal nature of the stimulus; see for example [Schachter and Singer \(1962\)](#).

²Certain measurable biological processes may of course be linked to certain emotions, such as elevated heart rate and fear, but these are not ironclad (e.g., elevated heart rate may also plausibly be linked to anger or excitement) and therefore also ultimately rely on self-report and common sense.

Past and current work has examined many types of semantic roles associated with emotions (e.g., Oberländer et al. (2020)), such as the experiencer (the person who feels the emotion), target (the thing the emotion is detected towards, if one exists—e.g., one can be angry *at* someone else), cause (the thing that elicits the emotion), and so on.³ The *cause* of an emotion is usually defined in agreement with the colloquial usage of the word, such that it is equivalent to the emotion stimulus from psychology research—it is the person or event that elicits an emotional reaction. Therefore, since the emotion cause has a clear antecedent in psychology literature and is a core component of defining an emotion, we choose to focus on this particular semantic role in our research. The emotion cause extraction problem has largely been treated in Chinese and English, with some work in additional languages such as Italian (Russo et al., 2011).

A large body of work has examined this problem as a clause-wise classification task based on a Chinese dataset published by Gui et al. (2016). In this task, a single sentence which contains some emotional reaction and its cause is broken up into multiple clauses, and the goal is usually to identify which clause contains the expression of an emotion and which contains its cause (there may be multiple emotion-cause pairs in one sentence, although this is not common in the data). A wide variety of models and methods have been applied to this dataset (Fan et al., 2020; Chen et al., 2018; Xia et al., 2019; Xia and Ding, 2019; Wei et al., 2020; Ding et al., 2020; Yuan et al., 2020) as well as other similar clause-level datasets (e.g., Chen et al. (2010)). The original authors took a span-extraction approach to the problem, noting that some of their causes were expressed as verb phrases and others as noun phrases. In this setting, each individual word in the input is classified with a tag, often in a scheme like IOB (Ramshaw and Marcus, 1995), and a span of contiguous tags⁴ identifies the part of the

³Note that a semantic role in linguistics is typically defined as a relationship between a noun phrase and a governing verb; these emotional semantic roles can be seen as akin to arguments in event extraction, if the reader prefers.

⁴In the IOB scheme, a B tag (“beginning”) followed by any number of I tags (“inside”) forms a

sentence expressing the cause. However, the clause-level task is much more tractable and has remained more popular.

Work dealing with English data in this area popularly annotates a handful of semantic roles, not just cause, and often deals with small datasets annotated by multiple human annotators. While some early work treated emotion cause detection as a classification problem (e.g., [Mohammad et al. \(2014\)](#)), much later work instead frames this task as a span extraction problem ([Ghazi et al., 2015](#); [Kim and Klinger, 2018](#); [Poria et al., 2020](#); [Bostan et al., 2020](#); [Oberländer et al., 2020](#)). [Oberländer and Klinger \(2020\)](#) additionally published work (contemporary to ours in [Chapter 7](#)) in which they compare the span-level and clause-level settings and argue that the span-level formulation is more appropriate for capturing emotion causes in English. Most existing work publishes small, human-annotated datasets for this problem from domains like novels and news headlines, and as a result, a significant challenge in this area is adjudicating various human-annotated spans into one correct gold cause. We would argue that there is frequently no single correct gold cause for a given emotion in text, making this a challenging problem for traditional supervised machine learning models and evaluation metrics.

6.1.1 Related Problems

As noted in [§2.2.2](#), emotion analysis in NLP is closely linked to sentiment analysis; the most important difference is that sentiment is typically treated as a binary or a categorical or continuous scale from “negative” to “positive” while emotions have a broader, multidimensional label space. Just as emotions have a stimulus or cause, sentiment is often framed as an attitude *towards* something or some aspect of something—

for example, a sentiment can be held towards a mobile phone, or more specifically span separated from the rest of the sentence, tagged with O (“outside”)

towards aspects of a phone such as its battery life, screen resolution, and so on. This problem is called *aspect-based sentiment analysis* (e.g., [Pontiki et al. \(2014\)](#); [Sun et al. \(2019\)](#); [Xu et al. \(2020a,b\)](#)) and while it is distinct from emotion-cause detection, we wish to acknowledge that there are similarities between the two.

The target or aspect of a sentiment may be seen as analogous to the target of an emotion, which can arguably be the same as its cause. For example, Alice may be angry *at* Bob (making Bob the target of the emotion *angry*) because of the cause “Bob ate her last red velvet cupcake”, but perhaps Alice may also be angry *at* “the shortage of replacement red velvet cupcakes at the store”, which is both the cause and the target. In the first case, sentiment target is not the same as emotion cause, while in the second, they are equivalent (if we assume that anger necessarily implies a negative sentiment). With regards to aspect-based sentiment, one might see the aspect in question as a cause for the sentiment; for example, if I say, “the battery life of my new phone is abysmal and I wish I hadn’t bought it,” my sentiment towards the battery life (aspect) of the phone (target) is negative, and the cause for my negative sentiment towards the phone is the battery life. However, overall, these are two separate fields with distinct bodies of work and treatment in the literature, and in this thesis we will focus on emotion cause rather than any form of sentiment analysis. We refer readers to recent surveys of sentiment analysis for a discussion of these and similar problems ([Poria et al., 2023](#)).

Affective Events. Another related NLP task is affective events ([Ding and Riloff, 2016](#)), which are events with some emotional effect on their experiencer—that is, this is a way of looking at the stimulus of an emotion specifically as an event. The focus is on identifying the emotion associated with an event (a classification problem), rather than identifying the event that has caused an emotion as in much of the literature on emotion cause detection. Some of this work has focused on applying psychological theory to gain a deeper understanding of the events ([Ding and Riloff, 2018](#)), and various other

work has applied a variety of techniques to improve the affect classification, including data augmentation via modern LLMs (Saito et al., 2019; Zhuang et al., 2020; Zhuang and Riloff, 2023).

6.2 Emotional Summaries and Explanations

Existing work has also begun to branch out from a low-level syntactic formulation of emotion causes into more high-level generative framing. Rather than an extraction problem which yields a span of text or a label, we can treat emotion cause detection as a generation or especially summarization problem that creates a natural language explanation of the cause, creating room to integrate additional aspects like multiple documents or external world knowledge to explain nuances of the cause. Just as with syntactic emotion cause extraction, terminology in this field is disparate and evolving; when we present our work in this area in [Chapter 8](#), we will call our output an *explanation* of the emotion, though some similar work may discuss *emotional summaries* or *emotion triggers*.

We note at this point that while we do generally strive to create models that are understandable and explainable, this particular work aims to generate explanations of a different kind. In our work on this subject, we focus on explaining the content of our inputs, rather than how the models make their predictions. We can consider there to be some work that bridges the gap between these two fields, in that it trains models to explain their predictions by explaining why they are true (e.g., the well-known chain-of-thought or step-by-step prompting (Wei et al., 2022)). However, overall, this field is distinct from generating explanations of emotion causes, and we refer the reader to a survey on the topic of explainable models for more information (Lai et al., 2021).

Because this area is new and growing, there is little related work at the time of

writing. The most similar work to ours in this field is that of [Zhan et al. \(2022\)](#), who generate abstractive summaries of emotion triggers on Reddit about the COVID-19 pandemic, jointly with identifying the emotion expressed. [Sosea et al. \(2023\)](#) further extend this work by integrating emotion detection directly into extractive summarization for the same dataset. We will discuss work in a similar setting in [Chapter 8](#), where we use a broader range of triggering events, an additional event extraction task, and different definitions of the text to be generated.

We also note several related but distinct problems studied in this area. Some work focuses on emotional summaries as a summarization problem first, integrating an emotional focus (e.g., [Zhou et al. \(2023\)](#)). Adjacent fields such as aspect-based sentiment analysis have also begun to incorporate summarization (e.g., [Musto et al. \(2019\)](#)); again, the goals and frameworks of these fields are distinct from our own, though their methodologies may be relevant in interesting ways. Finally, this type of emotion-focused summarization is also studied in non-text or multimodal settings (e.g., [Tu et al. \(2020\)](#)).

Chapter 7

Emotion-Cause Pair Extraction

Having examined several emotion and emotion-adjacent classification problems in text ([Part I](#)), we begin our more detailed exploration of their causes by considering the extraction of causes that are mentioned directly in the text. Emotion is commonly defined in psychology research as a reaction to an experience ([emo, 2023](#)); *something* is typically understood to have caused or elicited an emotional reaction. If the cause of an emotional reaction is explicitly given in the text, we can frame the question of “What has caused this emotional reaction?” as an information extraction problem and apply NLP techniques. We will begin with relatively concrete causes that can be expressed as a short span of text, usually describing an event; while not all emotional reactions necessarily have such a cause (for example, perhaps the experience causing an emotion is an ongoing situation or is not understood by the experiencer), many interesting and important reactions do, such as those reported in news headlines or given in microblogs.¹

Prior work in NLP has treated the emotion-cause detection problem as a classification task with the goal of selecting which clause of a sentence expresses the emotion and which expresses the cause (see [Chapter 6](#) for a detailed discussion). Bolstered by contemporary work ([Bostan et al., 2020](#); [Oberländer and Klinger, 2020](#)), we argue that a more flexible approach is needed to capture causes; for example, in English,

¹We will turn to causes that do not need to be stated in such a short, explicit manner in [Chapter 8](#).

an emotion cause can be expressed syntactically as a clause (*as troops open fire on protestors*), noun phrase (*1,000 non-perishable food donations*), or verb phrase (*jumped into an ice-cold river*). Therefore, we approach emotion-cause detection in this work as a span-tagging problem, which has the flexibility to capture any span of words as a cause. We also believe that understanding the cause and understanding the emotional reaction are mutually beneficial aims, and therefore we handle both emotion recognition and emotion cause tagging jointly as our objectives in this work.

In order to tackle this problem, we propose several ways in which to approach these two tasks. First, these two tasks should not be independent; because the cause is the trigger for the emotion, knowledge about what the cause is should narrow down what emotion may be expressed, and vice versa. Therefore, we present a multi-task learning framework to model them jointly. Second, considering that common-sense knowledge plays an important role in understanding implicitly expressed emotions and the reasons for those emotions (see, for example, similar work in emotion detection (Ghosal et al., 2020)), we additionally propose that a grounding in natural language understanding should be beneficial to semantic tasks such as these, and we explore the use of common-sense knowledge via adapted knowledge models (COMET, Bosselut et al. (2019)) for both tasks. A key feature of our approach is to combine these adapted knowledge models (i.e., COMET), which are specifically trained to use and express common-sense knowledge, with pre-trained language models such as BERT (Devlin et al., 2019).

In this chapter, we will begin by describing the dataset we use for this work (Section 7.1) and proceed to present each of our three types of emotion-cause detection models (Section 7.2): single-task (§7.2.1), multitask (§7.2.2), and adapted knowledge models (§7.2.3). We will present our results (Section 7.3), give an analysis of our model outputs (Section 7.4), and finally conclude with a discussion (Section 7.5).

Our contributions in this work (based on work published in [Turcan et al. \(2021b\)](#)) are as follows: (1) an under-studied formulation of the emotion-cause detection problem as a sequence tagging problem; (2) a set of three types of models that perform the emotion classification and emotion cause tagging tasks jointly while using commonsense knowledge with improved performance; and (3) analysis to gain insight into both model performance and the contents of the dataset we use.

7.1 Data

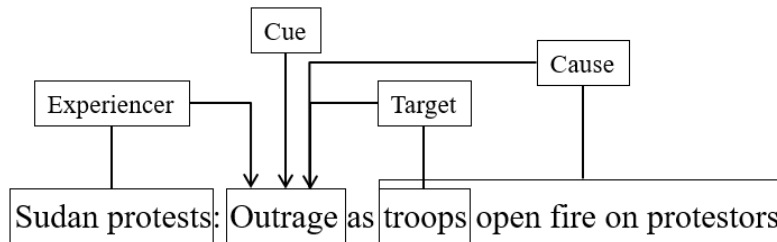


Figure 7.1: **Semantic Roles in GoodNewsEveryone.** An example of the semantic roles annotated by [Bostan et al. \(2020\)](#)

For our experiments, we use the GoodNewsEveryone corpus ([Bostan et al., 2020](#)), 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 7.1](#).² We focus on the emotion detection and cause tagging tasks in this work. At the time of this work, GoodNewsEveryone was the largest English dataset labeled for both of these tasks.

In our 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

²As formulated in this dataset, the emotions detected are not those of any author or reader, but of actors mentioned in the text.

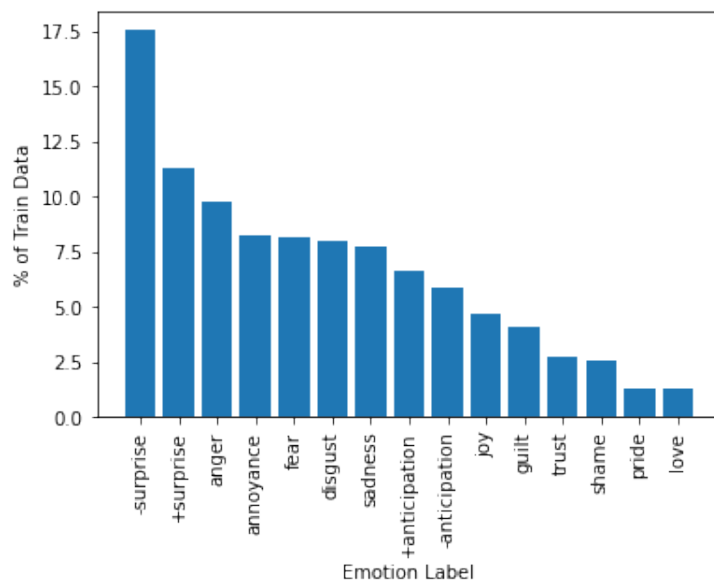


Figure 7.2: **GoodNewsEveryone Label Distribution.** Distribution of adjudicated emotion labels in the GoodNewsEveryone train data, as a percentage of the data points. “Positive” and “Negative” are abbreviated as + and -.

capturing subtle nuances of emotion, many external resources focus on a smaller set of emotions. We also note that the label distribution of this dataset heavily favors the more basic emotions, as shown in Figure 7.2. 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 for our experiments. We randomly split the remaining data (2,503 data points) into 80% train, 10% development, and 10% test.

7.2 Modeling

An important feature showcased by the GoodNewsEveryone dataset is that causes of emotions can be expressed through different syntactic constituents such as clauses, verb phrases, or noun-phrases. Thus, we approach the cause detection problem as

a sequence tagging problem using the IOB scheme (Ramshaw and Marcus, 1995): $\mathcal{C} = \{\text{I-cause}, \text{O}, \text{B-cause}\}$. Our approach is supported by results from Oberländer and Klinger (2020) and Yuan et al. (2020), whose work was published contemporaneously with ours (Turcan et al., 2021b); this work shows that modeling emotion cause detection as a sequence tagging problem is better suited than a clause classification problem. We tackle the emotion detection task as a seven-way classification task with $\mathcal{E} = \{\text{anger}, \text{disgust}, \text{fear}, \text{joy}, \text{sadness}, \text{negative surprise}, \text{positive surprise}\}$.

7.2.1 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),³ which we fine-tune on our data, as the basis of this model. Our preprocessing strategy for all of our models consists of the pretrained BERT vocabulary and WordPiece tokenizer (Wu et al., 2016)⁴ from Huggingface (Wolf et al., 2019). Therefore, for a sequence of n WordPiece tokens, our input to the BERT model is a sequence of $n+2$ tokens, $X = [[\text{CLS}], x_1, x_2, \dots, x_n, [\text{SEP}]]$, where each x_i is from a finite WordPiece vocabulary and [CLS] and [SEP] are BERT’s begin and end tokens. Passing X through BERT yields a sequence of vector hidden states $H = [h_{[\text{CLS}]}, h_1, h_2, \dots, h_n, h_{[\text{SEP}]}]$ with dimension $d_{\text{BERT}} = 768$. For emotion classification, we pool these hidden states and allow hyperparameter tuning to select the best type: selecting the [CLS] token ($h_f = h_{[\text{CLS}]}$), 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):

³We use BERT-BASE-UNCASED. We experimented with BERT-BASE-CASED, but it underperformed as the headlines incorporated into GoodNewsEveryone come from different news sources and have different capitalization styles.

⁴In the tagging setting, we ignore all tags predicted for subword tokens and use only the tag of the first subword.

$$h_f = \sum_{i=1}^n \alpha_i h_i \quad (7.1)$$

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:

$$e = \text{softmax}(W_e h_f + b_e) \quad (7.2)$$

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 directly on each hidden state:

$$c_i = \text{softmax}(W_c h_i + b_c) \quad (7.3)$$

with $c_i \in \mathbb{R}^{|\mathcal{C}|}$ and for trainable parameters $W_c \in \mathbb{R}^{|\mathcal{C}| \times d_{BERT}}$ and $b_c \in \mathbb{R}^{|\mathcal{C}|}$. We refer to both of these single-task models as BERT; if the task is not clear from the context, we will refer to the emotion detection model as BERT_E and the cause tagging model as BERT_C. Our training loss for emotion classification as well as emotion cause tagging is the mean negative log-likelihood (NLL) loss per minibatch of size b :

$$\text{NLL}_{\text{emo}} = -\frac{1}{b} \sum_j \sum_k y_{jk} \log e_{jk} \quad (7.4)$$

$$\text{NLL}_{\text{cause}} = -\frac{1}{b} \sum_i \sum_j \sum_k y_{ijk} \log c_{ijk} \quad (7.5)$$

where j is the index of the sentence in the minibatch, k is the index of the label being considered (emotion labels for NLL_{emo} and IOB tags for $\text{NLL}_{\text{cause}}$), i is the index of the i^{th} token in the j^{th} sentence in the minibatch, $y_{jk} \in \{0, 1\}$ is the gold probability of the k^{th} emotion label for the j^{th} sentence, $y_{ijk} \in \{0, 1\}$ is the gold probability of

the k^{th} cause tag for the i^{th} token in the j^{th} sentence, and e_{jk} and c_{ijk} are the output probabilities of the k^{th} emotion label and of the k^{th} cause label for the i^{th} token, both for the j^{th} sentence.

7.2.2 Multi-Task Models

Our hypothesis is that the emotion detection and cause tagging tasks are closely related and can inform each other; therefore, we propose three multi-task learning models to test this hypothesis. 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 lambda parameter: $\lambda \text{NLL}_{\text{emo}} + (1 - \lambda) \text{NLL}_{\text{cause}}$, using NLL_{emo} and $\text{NLL}_{\text{cause}}$ from Equation 7.4 and Equation 7.5.

Multi. The first model, Multi, is the classical multi-task learning framework with hard parameter sharing, where both tasks share the same 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 our emotion scores e and cause tag scores c from the same set of hidden states H .

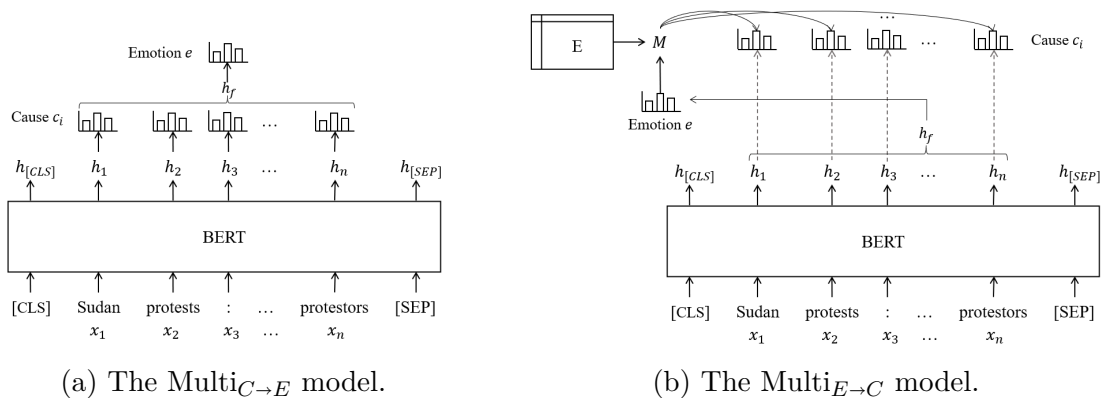


Figure 7.3: **Multi-Task Models.** Our multi-task models for concurrent emotion detection and emotion cause tagging.

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→E}. We assume that if a certain text span is given as the cause of an emotion, it should be possible to classify that emotion correctly while looking only at the words of the cause span. Therefore, we propose the Multi_{C→E} model, the architecture of which is illustrated in [Figure 7.3a](#). 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 7.3](#). 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 softmax over the sequence 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 7.1](#), where $\alpha_i = \frac{\exp P(\text{Cause}|x_i)}{\sum_{j=1}^n \exp P(\text{Cause}|x_j)}$, and emotion classification as in [Equation 7.2](#). For the Multi_{C→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→C}. Next, we hypothesize that knowledge of the predicted emotion should help us identify salient cause words. The Multi_{E→C} model first performs emotion classification, which results in a probability distribution over predicted emotion labels, as in the $BERT_E$ model and [Equation 7.2](#). We additionally keep an emotion embedding matrix E , where $E[i]$ is a learnable representation of the i -th emotion label (see [Figure 7.3b](#)) with dimension d_e (in our experiments, we set $d_e = 300$). 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 emotion 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 $W_{c'} \in \mathbb{R}^{|C| \times (d_{BERT} + d_e)}$ and $b_{c'} \in \mathbb{R}^{|C|}$ are trainable parameters. In the $\text{Multi}_{E \rightarrow C}$ model, we again do teacher forcing and use the gold emotion labels before doing the sequence tagging for cause detection (i.e., e is a one-hot vector where the gold emotion label has probability 1 and all other emotion labels have probability 0). At inference time, the model will use the predicted emotion distribution instead.

7.2.3 Adapted Knowledge Models

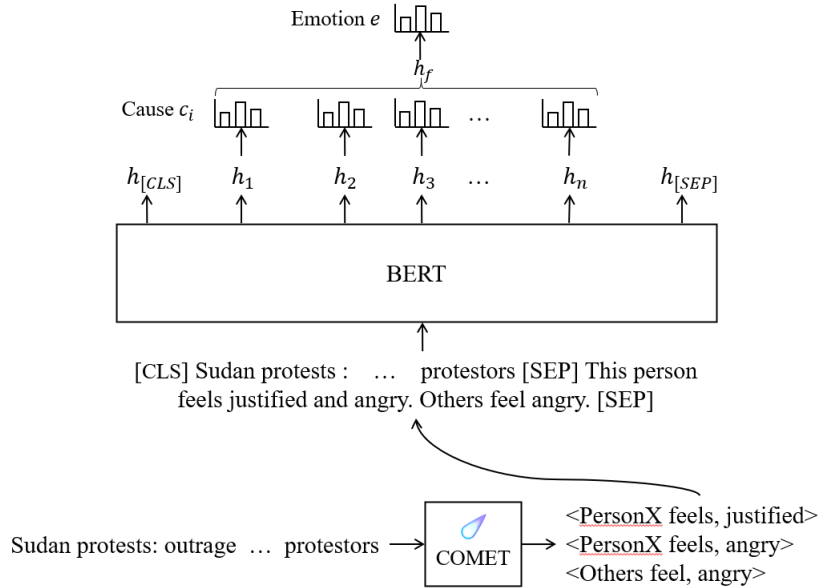


Figure 7.4: **Adapted Knowledge Model Architecture.** The architecture of our proposed $\text{Multi}_{C \rightarrow E}^{\text{COMET}}$ model.

Contemporary work has shown that fine-tuning pre-trained language models such as GPT-2 on *knowledge graph tuples* such as ConceptNet (Li et al., 2016) 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 common-sense 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, ATOMIC’s authors present the example of $\langle \text{PersonX pays PersonY a compliment, xReact, PersonX will feel good} \rangle$. xReact refers to the feelings of the primary entity in the event, and oReact refers to the feelings of others (in this instance, oReact yields “PersonY will feel flattered”). 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 7.4). To use this knowledge model in our task, 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 COMET-ATOMIC (using the model weights released by the authors), 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 $X = [[\text{CLS}], x_1, x_2, \dots, x_n, [\text{SEP}], z_1, z_2, \dots, z_m, [\text{SEP}]]$, 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 7.4).

This approach allows us to incorporate information from COMET into all our single- and multi-task BERT-based models; the example shown in Figure 7.4 is our $\text{Multi}_{C \rightarrow E}$ model. We refer to the COMET variants of these models as: $\text{BERT}^{\text{COMET}}$ (single-task models) and $\text{Multi}^{\text{COMET}}$, $\text{Multi}_{C \rightarrow E}^{\text{COMET}}$, $\text{Multi}_{E \rightarrow C}^{\text{COMET}}$ for the three multi-task models.

7.2.4 Experimental Setup

Evaluation Metrics For emotion classification, we report macro-averaged F1 and accuracy. For cause tagging, we report exact span-level F1 (which we refer to as *span F1*), as developed for named entity recognition (e.g., [Tjong Kim Sang and De Meulder \(2003\)](#)), where a span is marked as correct if and only if its type and span boundaries match the gold exactly.⁵

Training and Hyperparameter Selection The classification layers are initialized randomly from a uniform distribution over $[-0.07, 0.07]$ ⁶, and all the parameters are trained on our dataset for up to 20 epochs, with early stopping based on the performance on the validation data (macro F1 for emotion, span F1 for cause). All models are trained with the Adam optimizer ([Kingma and Ba, 2015](#)). We highlight again that for our $\text{Multi}_{C \rightarrow E}$ and $\text{Multi}_{E \rightarrow C}$ models, we use teacher forcing during training to avoid cascading training error. Because the subset of the data we use is relatively small, we follow best practices for dealing with neural models on small data and select hyperparameters and models using the average performance of five models with different fixed random seeds on the development set. We then base our models’ performance on the average of the results from these five runs (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 two sets of hyperparameters for 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. The hyperparameters we tune are dropout in our

⁵Our cause tagging task has only one type, “cause”, as GoodNewsEveryone is aggregated such that each data point has exactly one emotion-cause pair. We note that this problem formulation leaves open the possibility of multiple emotion-cause pairs.

⁶The default initialization from the `gluon` package: <https://mxnet.apache.org/versions/1.7.0/api/python/docs/api/gluon/index.html>

linear layers, initial learning rate of the optimizer, COMET relation type, lambda weight for our multi-task models, and the type of pooler for emotion classification (enumerated in §7.2.1).

Parameter Name	Type	Range or Values
pooler	Categorical	[cls, mean, max, attention]
learning rate	Continuous	$[10^{-6}, 10^{-4}]$
dropout	Continuous	[0, 0.9]
lambda	Continuous	[0.1, 0.9]
comet_relations	Categorical	[xReact, oReact, both]

Table 7.1: **Emotion Detection and Cause Extraction Hyperparameter Search Ranges.** Our hyperparameter search ranges for all models.

Model	Target Task	Parameter Name	Parameter Value
BERT _E	Emotion	pooler	cls
		dropout	0.8999992513311351
		lr	$2.0872134970009262 \times 10^{-5}$
BERT _C	Cause	dropout	0.04011659404129298
		lr	$9.609926650689472 \times 10^{-5}$
BERT _E ^{COMET}	Emotion	pooler	cls
		dropout	0.6467089448672897
		lr	$3.548213539029209 \times 10^{-5}$
		comet_relations	both
BERT _C ^{COMET}	Cause	dropout	0.8806119007595122
		lr	$9.913585728926367 \times 10^{-5}$
		comet_relations	xReact

Table 7.2: **Single-Task Model Hyperparameters.** The selected hyperparameter values for our single-task models.

We include descriptions of our hyperparameter tuning setup and the selected hyperparameters for each of our models in Table 7.1; we note that single-task cause models (BERT_C and COMET_C) do not tune the pooler, all single-task models do not tune the lambda parameter, and all non-common-sense models do not tune comet_relations. The parameters selected by all of our models can be seen in Table 7.2, Table 7.3, and Table 7.4. All of our models are trained with minibatches of size $b = 32$.

Model	Target Task	Parameter Name	Parameter Value
Multi	Emotion	pooler dropout lr lambda	mean 0.1438975482079587 $2.170218150294524 \times 10^{-5}$ 0.3736515054477897
	Cause	pooler dropout lr lambda	cls 0.8929935089177194 $9.929740332732521 \times 10^{-5}$ 0.6103686494768474
Multi _{E→C}	Emotion	pooler dropout lr lambda	max 0.2511612834815036 $3.179072019077849 \times 10^{-5}$ 0.4938386162506444
	Cause	pooler dropout lr lambda	max 0.763419047616446 $8.680439371509037 \times 10^{-5}$ 0.1341940851689314
Multi _{C→E}	Emotion	pooler dropout lr lambda	max 0.8138762283528274 $4.2586257586160994 \times 10^{-5}$ 0.8531247637209994
	Cause	pooler dropout lr lambda	mean 0.6992099059226856 $9.859155309987275 \times 10^{-5}$ 0.4855821360212248

Table 7.3: **Multi-Task Hyperparameters.** The selected hyperparameter values for our multi-task BERT models.

We used Bayesian optimization as implemented by Amazon SageMaker⁷ to tune these parameters, giving the learning rate a logarithmic scale and the dropout and lambda a linear one and allowing 75 iterations of parameter choice before selecting the setting with the best performance on the development set. Each individual instance of each model consisted of five different restarts with five distinct random seeds.

⁷<https://aws.amazon.com/sagemaker/>

Model	Target Task	Parameter Name	Parameter Value
Multi ^{COMET}	Emotion	pooler dropout lr lambda comet_relations	max 0.22350077887111716 $3.137385699389837 \times 10^{-5}$ 0.7676911585403968 both
	Cause	pooler dropout lr lambda comet_relations	mean 0.8891347000216091 $8.123006047625093 \times 10^{-5}$ 0.1 both
Multi ^{COMET} _{E→C}	Emotion	pooler dropout lr lambda comet_relations	mean 0.1372637910712323 $3.0408118480380588 \times 10^{-5}$ 0.8968243966922735 both
	Cause	pooler dropout lr lambda comet_relations	max 0.5319636087561394 $7.581334242472624 \times 10^{-5}$ 0.10896064677810494 both
Multi ^{COMET} _{C→E}	Emotion	pooler dropout lr lambda comet_relations	cls 0.7359624181177503 $1.9853909769532754 \times 10^{-5}$ 0.7947522633173147 both
	Cause	pooler dropout lr lambda comet_relations	max 0.01896406469706125 $8.360862387915605 \times 10^{-5}$ 0.14588492191321054 oReact

Table 7.4: **Multi-Task Adapted Knowledge Model Hyperparameters.** The selected hyperparameter values for our multi-task COMET models.

7.3 Results

We present the results of our models in Table 7.5.⁸ We see that the overall best model for each task is a multi-task adapted knowledge model, with $\text{Multi}_{C \rightarrow E}^{\text{COMET}}$ performing best for emotion (which is a statistically significant improvement over BERT by the paired t-test, $p < 0.05$) and $\text{Multi}^{\text{COMET}}$ 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.

	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→E}	37.74 ± 2.12	38.74 ± 2.07	39.08 ± 3.73
Multi _{E→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 _{C→E} ^{COMET}	39.26* ± 1.13	40.79 ± 2.17	38.68 ± 1.36
Multi _{E→C} ^{COMET}	37.44 ± 1.37	38.58 ± 1.44	36.27 ± 1.31

Table 7.5: **Emotion Detection and Emotion Cause Tagging Performance.** 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 ($p < 0.05$).

We wish to draw attention to the specific contribution of synthesizing common-sense reasoning and multi-task models: $\text{Multi}_{C \rightarrow E}$ alone does not outperform BERT on either cause or emotion, but $\text{Multi}_{C \rightarrow E}^{\text{COMET}}$ outperforms both BERT and $\text{Multi}_{C \rightarrow E}$ on both tasks. For cause, we also see additional benefits of common-sense reasoning alone: BERT^{COMET} outperforms BERT (multi-task modeling alone, Multi, also outperforms

⁸Oberländer and Klinger (2020) report an F1 score of 34 in this problem setting on this dataset, but on a larger subset of the data (as they do not limit themselves to the Ekman emotions) and so we cannot directly compare our work to theirs.

BERT for this task) and Multi^{COMET} outperforms Multi. These results speak to the differences between the two tasks, suggesting that common-sense reasoning and cause information may be complementary for emotion detection, but that for cause tagging, common-sense reasoning and emotion information may overlap, perhaps with emotion information as a subset of that provided by common-sense reasoning. The common-sense reasoning we have used in this task (xReact and oReact from ATOMIC) is expressed as possible emotional reactions to an input string, so this makes intuitive sense.

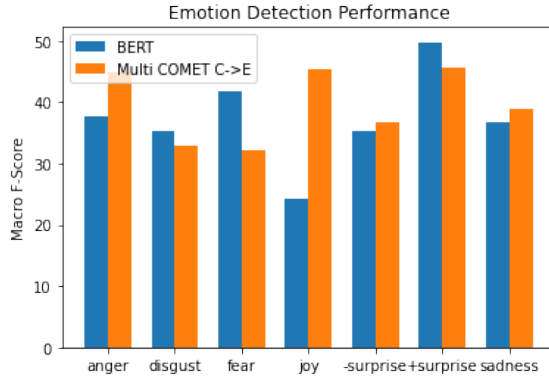


Figure 7.5: **Emotion Detection Performance by Label.** Performance of the BERT and Multi^{COMET}_{C→E} models on emotion classification.

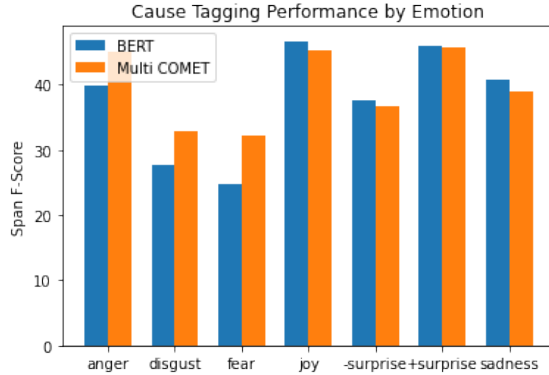


Figure 7.6: **Emotion Cause Tagging Performance by Label** Performance of the BERT and Multi^{COMET} models on cause tagging, broken down by emotion.

Finally, we also present per-emotion results for our best model for each task (Multi^{COMET}_{C→E} for emotion and Multi^{COMET} for cause) against the single-task BERT

baselines in [Figure 7.5](#) and [Figure 7.6](#); 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 $\text{Multi}_{C \rightarrow E}^{\text{COMET}}$ consistently improves over BERT by a significant margin on joy and to a lesser extent on anger and sadness. Meanwhile, for cause tagging, $\text{Multi}^{\text{COMET}}$ improves over BERT on anger, disgust, and fear, while yielding very similar performance on the rest of the emotions.

7.4 Analysis

In order to understand the impact of common-sense reasoning and multi-task modeling in our models, we provide several types of analysis. In addition to our results in [Section 7.3](#), we include examples of our various models’ outputs showcasing the impact of our methods ([§7.4.1](#)) as well as an analysis of the dataset itself, which we find to contain variable labels from multiple annotators ([§7.4.2](#)).

7.4.1 Example Outputs

We provide some example outputs from our systems for both cause and emotion in [Table 7.6](#); the various Multi models have been grouped together for readability and because they often produce similar outputs, but the outputs for every model are available in [Appendix A](#). In the first example, the addition of COMET to BERT informs the model enough to choose the gold emotion label; in the third and fourth, either COMET or multi-task learning is enough to help the model select key words that should be included in the cause (*return* and *triple shooting*). We also particularly note the second example, in which multi-task learning is needed both for the BERT and $\text{BERT}^{\text{COMET}}$ models to be able to correctly predict the gold emotion. This suggests

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

Table 7.6: **Example Outputs.** 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→C}, or Multi_{C→E} (details on selection and results for each individual model available in appendix; most multi-task models gave similar outputs).

that for cause, both common-sense reasoning and emotion classification may carry overlapping useful information for cause tagging, while for emotion, different instances may be helped more by different aspects of our models.

7.4.2 Label Agreement

Upon inspection of the GoodNewsEveryone data, we discover significant variation in the emotion labels produced by annotators as cautioned by the authors in their original publication.⁹ From our inspection of the development data, we see recurring cases where different annotators give directly opposing labels for the same input, depending on how they interpret the headline and whose emotions they choose to focus on. For example, our development set includes the following example: *Simona Stuns Serena at Wimbledon: Game, Set and “Best Match” for Halep*. The gold adjudicated emotion

⁹While the authors selected data according to agreement on the emotion labeling task, they found that in only 75% of cases do at least 3 annotators agree, with diminishing numbers for larger annotators.

label for this example is *negative surprise*, but annotators actually included multiple primary and secondary emotion labels including *joy*, *negative surprise*, *positive surprise*, *pride*, and *shame*, which can be understood as various emotions felt by the different entities and even reader. For this input, COMET suggests xReact may be *happy* or *proud* and oReact may be *happy*—these reactions are likely most appropriate for tennis player Simona Halep, or a reader who is a fan of hers, but not the only possible emotion that can be inferred from the headline.

Metric	BERT	BERT ^{COMET}	Multi	Multi _{E→C}	Multi _{C→E}	Multi ^{COM}	Multi ^{COMET} _{E→C}	Multi ^{COMET} _{C→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 7.7: **Model Accuracy on Alternative Emotion Labels.** Comparison of gold accuracy and non-gold (¬gold) accuracy for our emotion classification models.

Inspired by the variation in the data, we compute also models’ accuracy using the human annotations that did not agree with the gold (i.e., a predicted emotion label is correct if it was suggested by a human annotator but was not part of a majority vote to be included in the gold). We denote this ¬Gold, and we compare the performance of our models with respect to Gold and ¬Gold. We present the results of this analysis in [Table 7.7¹⁰](#). In this table, a higher ¬Gold accuracy means that the model is more likely to produce emotion labels that were not the gold but were suggested by some annotator. First of all, we note that all models have a relatively high ¬Gold accuracy (about half the magnitude of their gold accuracy); we believe this reflects the wide variety of annotations given by the annotators. We see a tradeoff between the Gold and ¬Gold accuracy, and we note that generally the single-task models have higher ¬Gold accuracy and the COMET-enhanced multi-task models have higher Gold accuracy. This suggests that our language models have general knowledge about emotion already, but that

¹⁰Note that we perform this analysis on just one of our ensemble of five models, so the accuracy numbers do not exactly correspond to those in [Table 7.5](#).

applying common-sense knowledge helps pare down the space of plausible outputs to those that are most commonly selected by human annotators. (Recall that this dataset was annotated by taking the most frequent of the annotator-provided emotion labels.) Furthermore, since the multi-task models have higher Gold accuracy and lower \neg Gold accuracy than the single-task models, this suggests that also predicting the cause of an emotion causes the model to narrow down the space of possible emotion labels to only those that are most common.

7.5 Discussion and Conclusion

Our multitask and adapted knowledge models show consistently increased performance on both the emotion recognition and emotion cause tagging tasks (Table 7.5), making use of the connections between these two tasks as well as the clarifying influence of common-sense reasoning. Both of these factors contribute to increased performance in isolation, and combining them in the same model improves performance further, particularly for accurate emotion recognition. We find that our best models improve differently on different sets of emotions (Figure 7.5, Figure 7.6), typically improving on medium-data labels like anger, while the baseline already performs well on the highly-represented surprise labels. By examining some sample outputs (§7.4.1), we can see potential mechanisms by which common-sense reasoning and multitask learning contribute to our models' performance, helping them refine guesses among similar emotion labels or target the correct entities as emotion causes, rather than just their effects. Overall, our results indicate that emotion recognition and emotion cause tagging are complex problems that can benefit from a variety of relevant information about situational understanding and the human experience.

An interesting discussion that arises from this work is about the annotation of

emotional tasks for machine learning. As we see in this work (§7.4.2), human annotators can and do often disagree about the emotions being expressed in a given situation. Emotional states are complex; a person may be able to name multiple, even conflicting, emotions at the same time about different aspects of the same situation. The more distinct emotion labels a dataset includes, the finer the distinctions between them, and the more likely it is that a person would select multiple labels to describe their emotions at a given time. For example, in the complete GoodNewsEverywhere dataset, the emotions *trust*, *love*, and *joy* are certainly distinct, but we can easily imagine many situations in which they might co-occur (spending time with beloved family or friends, for example). Our recommendation from this work is that emotion is not inherently a single-label problem; being able to describe the different facets of a person’s emotional reaction gives more complete and useful information.

In the GoodNewsEveryone dataset used in this work, annotators are asked to label a set of emotions for one experiencer about one situation, although there may be multiple experiencers mentioned in the text and each may be attributed different emotions about the same or even different situations. With data such as news headlines, the reader themselves may even be considered an experiencer (which would explain the high prevalence of surprise labels in this dataset), opening the door to an even broader range of potential emotional reactions depending on how the reader perceives the described situation. Thus, although the authors did adjudicate one “correct” answer, they also provided multiple options offered by the annotators for each label and urged researchers to consider the entire distribution of annotator labels in their work.

This highlights the difficulty of annotating data for emotional tasks like these. While we are able to develop annotation schemes that yield singular answers and create classification models that learn them competently, doing so erases a significant portion of the full picture of emotional understanding possible from text. Conversely, treating

emotional reactions as a distribution or multi-output problem may provide a more comprehensive emotional picture, but obtaining a reliable and stable gold target in this case may require a significant financial and time investment that is not feasible. Moving forward, researchers will need to develop new strategies for gathering and labeling emotional data—two possible directions may be to leverage the labeling capabilities of large language models (Dubois et al., 2023) or perhaps to incorporate structured label spaces that use the relationships between emotions to define richer types of labels for individual datapoints.

We note some limitations of the models we present in this chapter. Our data comes from a highly structured domain, news headlines, which is likely to present the relevant semantic roles clearly and concisely; we will treat input where causes and emotions are not necessarily stated this way in Chapter 8, although many domains remain to be addressed. Additionally, the inherently subjective, multifaceted nature of emotional tasks places inherent limitations on the work we can do in a discrete setting. Finally, our incorporation of common-sense reasoning is also limited to ATOMIC’s emotion relations, xReact and oReact; it is possible that additional forms of common-sense would improve performance on this and many other tasks in different ways.¹¹

Our contributions in this work include the following:

1. a formulation of the emotion-cause detection problem as a sequence tagging problem, a formulation that little research has used, allowing our models the flexibility to capture emotion causes expressed as various semantic constructions;
2. a set of models that perform the emotion classification and emotion cause tagging tasks jointly and incorporate commonsense knowledge, including:
 - (a) single-task BERT,

¹¹We do note that we experimented with a selection of other ATOMIC relations and found these to perform best in this work, but this area is open to future work.

- (b) Multi,
- (c) our sequential multitask models $\text{Multi}_{E \rightarrow C}$ and $\text{Multi}_{C \rightarrow E}$, and
- (d) our adapted knowledge model variants of i-iii incorporating COMET)

with $\text{Multi}_{E \rightarrow C}^{\text{COMET}}$ performing best on emotion recognition and $\text{Multi}^{\text{COMET}}$ performing best on emotion cause tagging; and

3. analysis of our models and data showing how commonsense knowledge and multitask learning can improve predictions of these tasks and how these features interact with a variously human-labeled dataset to produce humanlike predictions.

Chapter 8

Generating Emotion-Focused Explanations

While it is useful to be able to pinpoint the event that causes an emotional reaction, extracting emotion causes as text spans (as we have just discussed in [Chapter 7](#)) is a very structured problem that provides no additional information. As we have noted several times, there are many possible extra factors to consider when analyzing an emotional reaction to a triggering event—for example, semantic roles such as the target, or additional knowledge such as the relationships between participants or relevant social norms. Recall that we hope to deploy our models in settings related to mental health, where this additional nuance and context can be very important for healthcare workers to decide what resources or treatments may be most effective for a particular case. In this chapter, we will move from a structured tagging problem to a less structured but more informative *explanation generation* problem with the aim of creating a more well-rounded understanding of emotional reactions and the events that trigger them. We create a new dataset of Reddit posts, *Generating Emotional Descriptions* (GEDdit), which draws from Dreddit as well as other emotional subreddits, and we produce explanations that perform three simultaneous tasks: emotion identification, triggering event extraction, and explanation generation.

The advent of large language models (LLMs) has enabled a host of new generation problems in NLP, and so we would expect them to be a vital tool for this problem. While open-source LLMs have become more common, they can still require many resources

to run, and the size of an LLM makes it difficult to fine-tune if it cannot perform a given task in a zero-shot or in-context setting. In this work, we leverage the capability of LLMs to generate data for our problem, where human annotation is likely to be costly and time-consuming, but in order to create a customizable model that actually tackles the problem of explanation generation, we turn to distilled models. We aim to benchmark small language models on this task, training them to mimic the outputs of LLMs, and explore various facets of their capabilities. In developing these distilled models, we are inspired by previous work which incorporates multitask learning (such as [Dankers et al. \(2019\)](#) and including our own work from [Chapter 4](#) and [Chapter 7](#)). We leverage a variety of related tasks, including question answering, paraphrase detection, figurative language and metaphor understanding, and emotion detection, to enhance our distilled models' outputs and show that they benefit performance across all three of our target tasks.

In this chapter, we first begin by defining our problem ([Section 8.1](#)) and the criteria that a good explanation should meet ([§8.1.1](#)). We then introduce the data we use in this work ([Section 8.2](#)), including our pipeline from human seed annotations ([§8.2.1](#)), to LLM data generation ([§8.2.2](#)), to test set curation ([§8.2.3](#)). We then proceed to describe our methods ([Section 8.3](#)), including a suite of distilled models for this problem ([§8.3.1](#)) as well as our evaluation procedures ([§8.3.2](#)). Finally, we present our results ([Section 8.4](#)) and conclude with a discussion ([Section 8.5](#)).

Our contributions in this work (based on work yet to be published) are as follows: (1) the new problem of emotional explanation generation, including multiple criteria by which an explanation should be judged; (2) a new dataset, GEDdit, for this problem consisting of Reddit posts from various emotional subreddits, along with a seed set of human explanations and a larger set of silver GPT-3 generated explanations; and (3) a suite of distilled models incorporating multitask learning for this dataset, along with an

examination of their capabilities across three tasks (explanation generation, emotion identification, and triggering event extraction).

8.1 Problem Definition

Our goal in this work is to generate free-form explanations of a person’s emotional state and what has caused it. There can be many potential facets to such an explanation; for example, a person tasked with writing an explanation may include details about what support network the author has and how that affects their emotions, how much control they feel they have over the situation, whether the situation is shocking or has happened before, and so on. Let us take the following (artificial) story as an example document.

“My (17F) sister May (22F) is driving me crazy. So, she has a dog (Rocky) and we live with our parents. The thing is, she never takes care of Rocky at all. She doesn’t take him for walks, she doesn’t brush him...and I end up taking care of the dog all the time. I sort of felt bad for Rocky initially and sometimes I would take him with me when I went out running, and I guess my parents and May just assume I’m the one who takes care of him now? He’s not my dog and I’m actually super busy (I do track, I’m trying to apply for college, never mind seeing my boyfriend ever).

I’ve tried to tell my mom that it’s not my job to take care of Rocky but we just got into an argument about it and nothing changed. I feel really bad thinking about what it’ll be like for him when I go away to college but I’m so tired of all this.”

The author is clearly expressing frustration about her situation: feeling as though she is being made to take care of an animal that is not hers during a high-stress period

in her life. She expresses resentment towards her sister and her dog, as well as pity towards the dog and a myriad of other emotions we may be able to infer, such as anger or guilt. The author seems to feel as though she is in an unjust situation and has little control over it. It is likely that her anxiety over things like her school applications is exacerbating her frustration and resentment. To an outside observer, the situation may seem rather low-stakes, but the author is clearly very upset about it. Overall, she gives the impression that she is having trouble standing up for herself.

In this post, therefore, we can discuss many facets of the situation: multiple emotions the author is feeling (frustration, resentment, anger, guilt); the cause for these emotions (being obliged to take care of an animal she does not have time for); secondary stressors that interact with the cause (track and field, college applications, the author's boyfriend); the author's implied appraisals of the situation (she lacks control over her life, her family does not care about her feelings); potential coping or resolution strategies the author has tried (talking to her mother); and so on.

There is also some important information missing from the story: the author mentions speaking with her mother about the issue, but not with her sister, who she seems to regard as the source of the problem. She does not disclose what she discussed with her mother that led to an argument, or what her relationships with any of her family members are like outside this particular situation. Our picture of any event from the perspective of only one person is always incomplete; sometimes we can reasonably speculate on missing information, but we must always be aware of what assumptions we are making.

8.1.1 Criteria for an Explanation

In this preliminary work, we aim to benchmark current models' abilities to generate the type of explanation we describe above. Because we must develop data from scratch,

we choose to begin by defining three basic components of the explanation: a list of **emotional states** the author is experiencing, the triggering **event** that has caused them, and a long-form, natural language **explanation** that may encapsulate any or all of the other potential facets that may be present. When we evaluate the quality of our explanations, we select a handful of initial criteria for what makes a “good” explanation, though we are interested in future work that examines explanations on a more critical level. There may be a very large space of quality outputs that focus on various aspects of the explanation, give more or less detail, and so on; therefore, in this work, we focus on more general criteria.

First, it is important that an explanation does not hallucinate information that is not present, including details as minor as the demographic information of the author; the explanation should be **factually consistent** with the input. It is acceptable for an explanation to speculate about the reasons for and possible outcomes of the situation, as humans do, though we prefer that speculation be marked with words like “might” and not presented as fact. Second, and in direct tension with the first criterion, we would like our explanations to be **informative**. Simple restatement of the input is unlikely to be helpful to a professional trying to assist the posters; the “summaries” we generate should have some useful or insightful commentary on the post and be informative to the reader. We define an uninformative explanation to be a simple summary of the input and an informative explanation to introduce some new information not explicitly stated in the input, whether through insightful summarization, inclusion of world knowledge, or any other creative generation. In order for the summary to agree with both the factually consistent and informative criteria at once, the creative generation must not factually contradict the input.¹ Finally, since the explanations are generated by models (and in our case, smaller language models that are not as capable as LLMs), we need

¹The creative generation may or may not be **reasonable**, that is, logically sound to a human reader, but this is a criterion for future work.

them to be grammatically **fluent** and coherent.

We provide some manually constructed examples of factual consistency and informativeness, along with rationales for each, for an anonymized post in [Table 8.1](#) (fluency is a criterion that has been discussed in much previous work and is generally self-explanatory to annotators).

8.2 Data

In order to create a new dataset for this problem, we once again turn to the social media website Reddit.² We hand-select several subreddits, or topic-specific subforums, where posters are anecdotally likely to write the kinds of long-form narrative posts we are interested in to serve as our dataset. We initially recruit student social workers to develop our explanations and write sample explanations for a small set of posts (§8.2.1). Then, we use these explanations as seed data to prompt GPT-3 (Brown et al., 2020) to generate explanations for the remainder of our dataset (§8.2.2). Finally, we also generate our test set from this data, using human judgments to select a quality subset of the data (§8.2.3). The final dataset, consisting of all these pieces, is called *Generating Emotional Descriptions*, or GEDdit.

To gather our data, we begin by identifying the anxiety and interpersonal domains from the Dreddit dataset we created in [Chapter 3](#) as highly emotional and narrative in style and therefore appropriate for this new work. As discussed in that chapter, the anxiety domain contains posts from r/stress and r/anxiety, and the interpersonal domain contains posts from r/relationships. We also identify additional negative affective states which may provide a wider variety of data: sadness (r/sad), grief (r/GriefSupport, r/widowers, r/PetLoss, r/Miscarriage, r/SuicideBereavement,

²<https://www.reddit.com/>

Example input

Got a whole week of exams starting tomorrow and I'm freaking out! Last time, my grades took a hit, and it turned into a week of non-stop anxiety attacks, ending with a major panic attack. To top it off, a bunch of people were disappointed, making me feel even worse. These tests are deciding my whole year's grades. Any words of wisdom or good vibes would be much appreciated!

Factually consistent	Factually inconsistent
<p>The poster is anxious about their important upcoming exam week and recalls that their last exam week turned out badly. They may be worried this exam week will also go badly and negatively impact their final grades.</p>	<p>The poster is anxious about her upcoming science presentation and worries that her mother will be disappointed if she scores poorly.</p>
<p>Rationale: The explanation correctly summarizes events and does not make any unmarked assumptions.</p>	<p>Rationale: The explanation makes unfounded assumptions such as the gender of the poster or that the poster's mother will be disappointed, and incorrectly recalls details (a presentation instead of exams).</p>
Informative	Not informative
<p>This person is worried about their school exams; their previous exam week went badly for them and they are worried it may happen again. They say these tests will decide the whole year's grades, which could put them at a big disadvantage (such as needing to repeat classes) if they score badly. The poster does seem overall hopeful as they ask for advice or well-wishes.</p>	<p>The poster is worried about their school exams because the tests are important and their last exam week went badly, resulting in anxiety and disappointment. They are seeking wisdom and good vibes.</p>
<p>Rationale: The explanation makes some inferences explicit (e.g., "they are worried it may happen again"), incorporates world knowledge (overall failing grades may necessitate repeating classes), and comments on the high-level tone of the poster (anxious but hopeful).</p>	<p>Rationale: The explanation does nothing but summarize the input using mostly the same words.</p>

Table 8.1: **Emotional Explanation Criteria Examples.** Manually constructed positive and negative examples of factually consistent and informative explanations for a sample social media post. Note that the input is from our dataset but paraphrased in order to minimize searchability.

Domain	Subreddit Name	Total Posts	Avg Tokens/Post
anger	r/Anger	1,407	242
	r/angry	593	261
	Total	2,000	248
anxiety	r/anxiety	1,986	233
	r/stress	14	214
	Total	2,000	233
grief	r/bereavement	7	234
	r/COVIDgrief	3	194
	r/grief	51	220
	r/GriefSupport	736	234
	r/Grieving	32	230
	r/Miscarriage	330	237
	r/Petloss	344	246
	r/SuicideBereavement	150	235
	r/widowers	347	235
	Total	2,000	236
sadness	r/sad	2,000	234
interpersonal	r/relationships	2,000	278
	All	10,000	246

Table 8.2: **GEDdit Data Statistics**. Data contained in our GEDdit dataset, broken down by subreddit. The average length of a post in each subreddit, in tokens (rounded to the nearest integer) according to Huggingface’s pretrained `google/flan-t5-large` tokenizer, is given as well.

r/grief, r/Grieving, r/bereavement, r/COVIDgrief), and anger (r/Anger, r/angry). We gather the anxiety and interpersonal domains from our full (labeled and unlabeled) Dreddit dataset and randomly sample the three new domains from all of Reddit using the Pushshift Reddit dataset (Baumgartner et al., 2020);³ we set a target of 10,000 datapoints and sample equally from all sources to obtain our data for this work. When gathering data, we target posts between 100 and 300 tokens in order to strike a balance between having a detailed narrative and being able to fit a post comfortably into an LLM prompt. The statistics are shown in Table 8.2. When partitioning the data, we reserve 10% (1,000 posts) for development and further reserve 200 posts for creating a

³Note that with 2023 changes to Reddit and Pushshift’s API policies, it may no longer be possible to gather exactly the same data.

test set (which we will describe in §8.2.3), leaving us with 8,800 datapoints for training. All splits are sampled randomly but stratified such that the source domains remain in consistent proportions in each split.

Emotional Content

Subreddit Name	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Positive	Negative
r/Anger	6.87	3.13	3.71	4.43	2.17	4.30	1.82	3.18	4.98	9.9
r/angry	3.88	3.19	3.25	3.09	2.60	3.21	1.54	3.70	5.62	7.08
All Anger	5.98	3.15	3.57	4.03	2.30	3.97	1.74	3.34	5.17	9.11
r/anxiety	4.03	5.38	2.24	5.71	2.44	5.02	1.67	3.65	5.65	8.34
r/stress	2.32	4.08	2.12	3.95	2.41	3.90	0.97	4.22	6.76	5.88
All Anxiety	4.02	5.37	2.24	5.70	2.44	5.01	1.66	3.65	5.66	8.33
r/bereavement	3.20	3.41	2.14	3.50	2.79	4.49	1.57	3.18	5.89	6.86
r/COVIDGrief	3.74	5.91	2.60	4.63	3.29	5.12	2.54	3.59	5.18	6.62
r/grief	2.78	4.33	2.38	4.06	4.13	5.63	2.20	4.38	6.88	7.18
r/GriefSupport	2.80	3.90	2.04	3.75	3.32	5.19	1.92	4.15	6.03	6.79
r/Grieving	2.85	4.43	1.89	3.18	3.51	4.94	2.61	3.60	6.10	6.51
r/Miscarriage	2.03	4.09	2.58	3.79	3.52	4.45	1.56	4.09	6.34	6.42
r/Petloss	2.30	3.90	1.90	3.46	3.79	4.44	1.97	3.72	6.19	6.30
r/SuicideBereavement	3.07	3.83	2.00	3.67	3.67	5.01	2.11	4.29	6.23	6.70
r/widowers	2.44	4.28	1.83	3.22	3.82	4.31	1.95	3.99	6.52	5.80
All Grief	2.55	4.01	2.07	3.61	3.57	4.78	1.91	4.04	6.23	6.47
r/sad	2.85	3.90	2.24	3.37	3.69	4.28	1.80	4.19	6.31	6.38
r/relationships	2.03	3.63	1.48	2.22	3.23	2.68	1.50	4.09	6.14	4.47
All	3.49	4.01	2.32	3.79	3.05	4.15	1.72	3.86	5.90	6.95

Table 8.3: **Emotions in GEDdit.** Scores are broken down by domain and further by subreddit. For the eight Plutchik basic emotions and two polarities (positive, negative), we present the average percentage of words in a single post that are marked by the NRC Emolex (Mohammad and Turney, 2013) for that emotion. Note that r/sad and r/relationships each make up their own domain.

In order to gain a rough overview of the emotional content of our dataset, we use existing emotion lexicons (the NRC Emotion-Word Association Lexicon, or EmoLex (Mohammad and Turney, 2013), and NRC Valence-Arousal-Dominance Lexicon (Mohammad, 2018)). EmoLex offers binary emotion associations for the eight Plutchik basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) (Plutchik, 1980) as well as binary polarity (positive and negative) for individual words, while the NRC-VAD Lexicon provides real-valued valence, arousal, and dominance scores (following the circumplex model (Russell, 1980)) for individual words on a scale

of 0-1.⁴ Therefore, we calculate two types of metrics for a particular partition of our dataset (a subreddit or domain), one type for each lexicon. Using EmoLex, for each post we tally the percentage of non-stopwords (according to the static list of Bird et al. (2009), the Natural Language Toolkit) associated with each emotion or polarity category. We present the average of this percentage across all posts in a partition as a score for that partition. Then, using the NRC-VAD Lexicon, we collect all VAD-bearing non-stopwords in a post and calculate their average magnitude. The average of this average magnitude across all posts in a partition is presented for that partition. The results of this analysis can be seen in Table 8.3 and Table 8.4.

Subreddit Name	Valence	Arousal	Dominance
r/Anger	0.538	0.496	0.514
r/angry	0.548	0.484	0.499
All Anger	0.541	0.492	0.509
r/anxiety	0.563	0.480	0.505
r/stress	0.580	0.474	0.518
All Anxiety	0.563	0.480	0.505
r/bereavement	0.611	0.443	0.523
r/COVIDGrief	0.582	0.455	0.494
r/grief	0.588	0.453	0.507
r/GriefSupport	0.589	0.452	0.508
r/Grieving	0.601	0.456	0.509
r/Miscarriage	0.580	0.458	0.504
r/Petloss	0.587	0.441	0.489
r/SuicideBereavement	0.588	0.459	0.508
r/widowers	0.603	0.450	0.507
All Grief	0.589	0.451	0.504
r/sad	0.594	0.457	0.505
r/relationships	0.616	0.457	0.520
All	0.581	0.468	0.509

Table 8.4: **VAD in GEDdit**. Scores are broken down by domain and further by subreddit. We present the average magnitude of all posts in that category, where the magnitude of a single post is defined as the average magnitude of all VAD-bearing words in the post. Note that r/sad and r/relationships each make up their own domain.

⁴A score of 1 indicates positive valence, maximum arousal or activity, and maximum dominance or power.

We can see some expected trends, such as a high percentage of angry words in r/Anger and a high percentage of fearful words in r/anxiety. We also note several interesting outliers⁵. r/COVIDGrief has a high percentage of anger words, much higher than any other grief-related subreddit, suggesting unique circumstances about deaths from the COVID-19 pandemic that elicit anger. The anger domain also contains relatively high percentages of disgust words as compared to the rest of the corpus, suggesting that the two emotions are correlated in some way. All domains and subreddits score similarly on arousal and dominance, suggesting similar intensities of emotion across the corpus. We also note that all domains, despite having been selected for negative emotional reactions, contain approximately 2-3% joy words and 4-6% positive words, indicating some emotional nuance in our data and lexicons.

8.2.1 Human Annotations

In order to generate the types of emotional explanations we want, we first solicit seed annotations from student social workers, with the intention of using them to prompt a LLM to generate similar explanations on a large dataset. After running a five-item pilot study with computer science students, we recruit three social work students to annotate a selection of 30 randomly-queried posts from the chosen subreddits (avoiding the posts used in the dataset itself), resulting in 90 unique annotations. Each annotator annotates each post, meaning that each post has three unique explanations. The nature of our dataset does not lend itself to agreement statistics, and, in fact, we hope that these varying annotations will lead to variety in our outputs downstream.

Annotators are provided with the social media post and asked to label the emotions felt by the poster, the event that has caused the emotional reaction, and an explanation

⁵Note that, aside from the percentage of angry words in r/Anger, all of these values lie within one standard deviation of the mean as calculated from the entire corpus; these observations are not meant to carry statistical weight.

of the emotional reaction. Annotators are provided with the Ekman basic emotion set (Ekman, 1992) as a prompt but also allowed and encouraged to provide their own emotion labels freely; this is intended to cultivate variety in the data. Annotators are allowed a high degree of freedom in designing the explanations, simply being asked to describe why the emotional reaction has occurred and what is interesting or useful in the field of social work to discuss with respect to the post; we intend for this to impart real-world utility to the explanations.

Post: Got a whole week of exams starting tomorrow and I'm freaking out! Last time, my grades took a hit, and it turned into a week of non-stop anxiety attacks, ending with a major panic attack. To top it off, a bunch of people were disappointed, making me feel even worse. School pressure is no joke, and now these tests are deciding my whole year's grades. Honestly, I'm terrified and clueless about what to do. Any words of wisdom or good vibes would be much appreciated!

Emotions: fear

Event: starting exam week at school

Explanation: The author is terrified of exam week at school because of how it has gone for them in the past. Previously, tests have not only affected the author's grades a lot but also caused severe emotional distress and brought social shame for them. They feel helpless because they feel it is unavoidable that the exam week will once again trigger severe anxiety for them and will result in poor academic performance and disappointment from their community.

Figure 8.1: **Example Annotation.** A sample datapoint and explanation from GEDdit. Please note that the source datapoint has been paraphrased (using ChatGPT) so as to limit searchability.

We present a sample annotation in Figure 8.1. The annotator summarizes or rephrases much of the post, but also provides some insight (for example, making explicit the author's fear that this exam week will be the same as the last one, which they only express implicitly in the post).

8.2.2 Data Generation by Prompting

The annotation process for a free-response task such as ours may be time-consuming and expensive to an extent that it prohibits collecting enough data to train a computational model. Meanwhile, LLMs have demonstrated great proficiency in performing a range of similar language tasks in a few- or zero-shot setting (Liang et al., 2023). Therefore, in this work, we choose to generate explanations for our dataset using an LLM, in our case, GPT-3 (Brown et al., 2020) (chosen for its cost). We generate explanations for all of GEDdit in this way, including the test set; we will describe in §8.2.3 how we ensure its quality.

For each input we provide to GPT-3, we include natural-language instructions as well as two in-context examples sampled from our seed annotations.⁶ We reproduce a full example prompt here, noting that for the purposes of this dissertation we have paraphrased all social media posts so as to limit their searchability. Additionally, for illustrative purposes, each part of the prompt is labeled in italics and square brackets; these labels are not provided to the model. However, all other aspects of the prompts, including the explanations provide by our annotators, remain unchanged; aside from the wording of the posts, this is a real prompt fed to GPT-3 to generate a silver explanation for this datapoint. The prompt consists of instructions, two full examples separated by blank lines, and finally a target datapoint for which an explanation should be generated, terminating in a prompt (“Emotions:”).

[Instructions]

Read the following social media posts and describe the emotions felt by the people who wrote them, including what is causing the emotions and how

⁶Two examples were selected because the inputs are very long, but we wanted the model to have more than one example to mitigate potential copying behavior.

the author is handling them. Be careful to describe the emotions as the author seems to feel them, and not as you think they should. Try not to assume demographic information we do not know about the author. Add as much detail as you need to fully convey your thoughts.

[Example 1]

Post: Just wanted to send a huge thank you to everyone here. Dealing with a miscarriage has been the toughest journey for me, but reading all your posts and comments has been a huge source of comfort. Finally got back into therapy yesterday after over a year, and when my therapist asked how I've been, I couldn't hold back the tears. She asked if I'd done any grief counseling, and I told her no, but that I found an amazing supportive community online. Miscarriage can feel isolating, but the support I've found here has been truly incredible. Just wanted to express my gratitude.

Emotions: joy, sadness

Event: Receiving support after miscarriage

Explanation: The author is devastated and grieving as they recently suffered losing their unborn child. They mention crying when trying to go to therapy, how hard the process has been, and even further, that they feel alone in their experience. However, the author also expresses their gratitude to the community for the supportive messages that have helped them through the process. Considering the impact of a miscarriage, receiving helpful support from strangers online was likely surprising and poignant, making the author feel happier and better able to grieve.

[Example 2]

Post: Got a whole week of exams starting tomorrow and I'm freaking out! Last time, my grades took a hit, and it turned into a week of non-stop anxiety attacks, ending with a major panic attack. To top it off, a bunch of people were disappointed, making me feel even worse. School pressure is no joke, and now these tests are deciding my whole year's grades. Honestly, I'm terrified and clueless about what to do. Any words of wisdom or good vibes would be much appreciated!

Emotions: fear

Event: starting exam week at school

Explanation: The author is terrified of exam week at school because of how it has gone for them in the past. Previously, tests have not only affected the author's grades a lot but also caused severe emotional distress and brought social shame for them. They feel helpless because they feel it is unavoidable that the exam week will once again trigger severe anxiety for them and will result in poor academic performance and disappointment from their community.

[Target Post]

Post: I absolutely adore my college campus – it feels more like home than my actual home sometimes. But whenever I have to leave, I start feeling pretty down and dealing with intrusive thoughts. Managed to stay on campus for fall semester, but now that I'm home for the holidays, I'm feeling pretty blue and missing my crew a ton. Anyone else ever feel like this? And any tips on how to shake off the blues? Should I distract myself or try tackling

this head-on?

Emotions:

GPT-3 begins generating its explanation with the very next token. We provide the LLM with the same instructions given to our human annotators⁷ as well as two examples from our set of human-annotated posts to demonstrate the task using in-context learning (Radford et al., 2019; Brown et al., 2020). So that no one example or annotator’s style biases the dataset, we sample the examples uniformly without replacement anew for each datapoint we give to the LLM. We set the temperature to 0.7 to encourage creativity.

8.2.3 Test Set Generation

As part of our evaluation, which we will describe in detail in §8.3.2, we wish to have a small test set that can be used for human evaluation. However, as part of a well-rounded analysis of the results, we also wish to calculate some reference-based metrics (e.g., ROUGE (Lin, 2004)) on this test set, necessitating that it have some “gold” explanations to use for comparison. We opt to use the generated explanations of GPT-3 as our “gold” explanations when calculating these reference-based metrics. However, because these explanations are still model outputs, we pare down our 200 dedicated test examples using annotator-based filtering to decide which outputs are good or bad quality and therefore should or should not remain in the test set.

We recruit seven NLP graduate students (all native English speakers) to annotate the test set. To mitigate the bias of any one annotator pair, we assign annotator pairs to each datapoint uniformly, that is, ensuring that each annotator is matched equally often with each other annotator. For each datapoint, each annotator is asked to provide

⁷We did experiment with various types of prompts; these humanlike instructions produced the best outputs based on qualitative examination.

a yes-or-no answer to the following questions, corresponding to our criteria for a good explanation in §8.1.1:

1. Is the explanation factually consistent with the social media post?
2. Is the explanation informative; that is, does it provide some new information now explicitly stated in the post? This new information may be world knowledge, commonsense knowledge, speculation, or simply making explicit something that the original poster only implied.
3. Is the explanation fluent?
4. Is the provided list of emotions reasonable given the social media post?
5. Is the provided triggering event reasonable given the social media post?

An explanation is only accepted if both annotators answer “yes” to every question. Because raw answers from the annotators show a significant amount of disagreement, we further have annotators discuss their answers until they agree on a final adjudicated answer for each question. This results in a final test set of 72 quality explanations.

8.3 Methods

In order to tackle this problem, the obvious answer with modern models is to use an LLM such as the one that generated the training data. However, most LLMs suffer from a number of drawbacks, including an unwieldy number of parameters for fine-tuning, proprietary training data, and high cost. Therefore, in this work, we seek to characterize the capability of smaller generative models to produce the types of explanations we are interested in; we fine-tune pre-trained Flan-T5-Large models (Chung et al., 2022) on the training data produced by GPT-3 to examine their capabilities. We present variants

of these models trained to mimic every component of the output at once (Flan-T5_{GPT}) versus each component individually (Flan-T5_{Split}). We then leverage multitask learning to enhance our models’ performance by training them on a range of additional tasks, including question answering, paraphrase detection, figurative language and metaphor understanding, and emotion detection (Flan-T5_{Split}+datasets).

8.3.1 Models

Distilled Model

The base for each of our models is a T5 model (Raffel et al., 2019) using Flan instruction fine-tuning (Chung et al., 2022), which we select because of its performance on limited-data tasks and its similar framing to the data generation prompts given to GPT-3. All of our presented models use the `google/flan-t5-large` pre-trained checkpoint hosted by Huggingface (Wolf et al., 2019). Our data consists of three separate tasks: emotion identification, event extraction, and explanation generation. In its most basic form, our model learns to perform all three of these tasks simultaneously, the same as GPT-3, and is trained on the plaintext output of the LLM (line breaks and task headers included). We also provide the model with instructions for the task, as the Flan fine-tuning data is usually formatted; in our case, because of space considerations, we provide an abbreviated set of instructions (“Read the following social media post and describe the emotions felt by the person who wrote it.”) and do not provide any in-context examples, as multiple social media posts will rapidly become too large for the model.⁸ We call this basic variant of our model Flan-T5_{GPT}, as it aims to mimic the output from GPT-3 completely.

⁸We did try teaching Flan-T5 the task without a prompt, but it was not capable of learning the task format.

Multitask Model

We hypothesize that a smaller generative model will have an easier time learning to perform each task individually rather than as a block of output. Providing the model with separate tasks with separate prompts should allow the model to ignore formatting⁹ and place some amount of importance on the actual content of the instructions, which can be completely ignored in the case of Flan-T5_{GPT} because they are static across all inputs. Thus, we also create a model variant, Flan-T5_{Split}, with the same architecture but modified training.

A single training datapoint for this model is split into three datapoints, one for each task (emotion, event, and explanation). Each task is presented to the model with a different prompt or set of instructions (which can be seen in [Appendix B](#)). We continue framing each individual task as a generation task and use the same language modeling head as in the basic model, relying on the prompts to dictate which task is being performed.

External Tasks

Finally, we also hypothesize that a model like Flan-T5, which has been fine-tuned with a variety of different tasks, may be able to benefit from the inclusion of various pre-existing tasks such as paraphrasing or emotion detection which may indirectly provide it with additional training signal and mitigate overfitting or overwriting of prior task knowledge. We select a number of tasks which may be relevant to fluency and creativity: question answering (Natural Questions ([Kwiatkowski et al., 2019](#)), SQuAD v2 ([Rajpurkar et al., 2018](#))), paraphrase detection (Quora Question Pairs,¹⁰ Para-NMT

⁹Although T5 is able to produce the expected format consistently, some amount of parameters must still be used to learn it.

¹⁰First published as part of GLUE ([Wang et al., 2018](#)) and available originally from <https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>.

Task	Dataset	Present in Flan?	Type
figurative language understanding	BIG-bench Metaphor Boolean	No	generation
figurative language understanding	BIG-bench Understanding Metaphor	No	generation
summarization	BIG-bench Understanding Fables	No	generation
paraphrasing	BIG-bench Rephrase	No	generation
suicide risk detection	BIG-bench Suicide Risk	No	generation
figurative language understanding	FLUTE	No	generation
emotion detection	GoEmotions	No	multi-label classification
question answering	Natural Questions	Yes	generation
paraphrasing	Para-NMT	Yes	regression
paraphrasing	Quora Question Pairs	Yes	generation
question answering	SQuAD	Yes	generation

Table 8.5: **External Tasks.** External tasks and datasets used in the Flan-T5_{Split}+dataset models.

(Wieting and Gimpel, 2018)), figurative language and metaphor understanding (FLUTE (Chakrabarty et al., 2022), BIG-bench tasks (Srivastava et al., 2023)), and emotion detection (GoEmotions (Demszky et al., 2020), BIG-bench tasks). A selection of these tasks are present in the Flan fine-tuning data and are therefore being re-presented to the model during our fine-tuning phase, and some tasks are new to the model. The full list of tasks is shown in Table 8.5.

We train various model variants, Flan-T5_{Split}+dataset, adding one or more external tasks to the training data of the Flan-T5_{Split} model. We also express models trained with entire categories of additional tasks this way, for example, Flan-T5_{Split}+QA, which includes both Natural Questions and SQuAD. Note that most datasets are expressed as generation tasks, to match the style of Flan fine-tuning; some datasets are naturally generative, such as question answering datasets, and some are easily adaptable (e.g., BIG-bench’s boolean metaphor task requires a simple true or false answer, which can be expressed as asking the model a yes or no question). Two datasets require additional modeling heads: Para-NMT, which is a regression task asking the model to rate the similarity of two potential paraphrases, and GoEmotions, which is a multi-label classification task.

8.3.2 Evaluation

When evaluating our models, we utilize a combination of automatic metrics and human judgments. Automated metrics are useful for benchmarking performance; we then also use more qualitative human judgments to obtain a more detailed picture of our models’ capabilities.

Automatic Evaluation

Our automatic metrics are divided into several categories: extractiveness measurements, reference-based metrics, and factuality measurements. Many of these metrics are typically used in a summarization context, framing our generated explanation as a sort of emotional summary of the post. To judge models’ extractiveness, we calculate **extractive fragment coverage** (the percentage of words in the explanation that are present in the source post) and **extractive fragment density** (the average length of an extractive fragment copied into the explanation from the source post) as formulated by [Grusky et al. \(2018\)](#).¹¹ In general, we would like our explanations to summarize rather than completely restate the post, so a lower score on these metrics is preferable; however, it is not necessarily our goal to *minimize* them, because extractiveness is not inherently undesirable.¹² We also compare the generated explanations to their silver GPT-3 generated counterparts using reference-based metrics: **ROUGE** (-1, -2, and -L) ([Lin, 2004](#)) and **BERTScore** ([Zhang et al., 2020](#)); a higher score on these metrics indicates that the model is behaving like GPT-3 (and like its training data), but again, it is not necessarily desirable to maximize them, since the space of acceptable potential

¹¹We totally ignore function words from a pre-generated list of English stopwords ([Montani et al., 2023](#)) when calculating these metrics so as to avoid penalizing the model for their use.

¹²For example, an explanation could directly quote from the post to make a point, and the names of entities or events described in the post (e.g., family members) should probably be directly quoted in the explanation instead of using synonyms or paraphrases.

explanations is quite broad. Finally, we use two metrics to evaluate the explanation’s faithfulness or factual consistency with the source post: **QAFactEval** (Fabbri et al., 2022) and **SummaC** (Laban et al., 2022); these metrics are generally straightforwardly desirable to maximize.

Because the majority of our models perform each of our three tasks (explanation generation, emotion identification, event extraction) separately, we evaluate each of the tasks separately using all of the metrics above, treating each as a separate generation task. The nature of the emotion task in particular (outputting a list of emotion words) makes it most suited to evaluation by ROUGE-1 (essentially unigram overlap between the reference and the prediction) or BERTScore (which incorporates similarity), but we report all metrics for emotion for completeness. We also report classical precision, recall, and F1 scores for emotion words (comparing the set of predicted emotions against the set of reference emotion words), though we find a similarity metric like BERTScore most appropriate because the potential set of emotion labels is bound only by the model’s vocabulary.

While human evaluation (or LLM evaluation) remains the standard for evaluating high-order generative tasks such as this one, obtaining these judgments can be costly and/or time-consuming. We use these automatic metrics to obtain a broader picture of many models’ performance without needing to solicit human judgments for each model variation we present.

Human Evaluation

In order to gain a deeper understanding of the strengths and weaknesses of our models, we conduct a human evaluation of a small sample of test set outputs from two selected models. We feel that the emotion and event tasks can be accurately represented by automatic metrics because of their short length and uncomplicated objective, so we

INSTRUCTIONS [\(hide\)](#)

Thank you for helping with our evaluation task! In this project, we are evaluating hypothetical technology that may be able to assist professionals who deal with clients' or patients' mental health, such as counselors or therapists. This technology **does not yet exist in the real world**, and we are examining the performance of some prototypes. In this task, the hypothetical technology is intended to read social media posts and write an explanation of the author's emotional state, which might be helpful for someone like a triage counselor who is assisting many clients at once.

You will be shown a social media post about some emotional topic as well as an explanation of the post that attempts to explain the emotions the poster is expressing and why. Your job is to answer some questions about the quality of the explanation, which may have been generated automatically or may have been written by a human.

Sometimes machine-generated text does not make sense or gets factual details wrong. An ideal explanation would accurately describe the situation and offer some kind of helpful or insightful commentary. You are encouraged to use your own judgment when deciding whether an explanation is good or bad, although there will also be questions on specific aspects of the explanation.

A **factually consistent** explanation does not contradict the original post (for example, by attributing actions to the wrong person). It may speculate on things that are not stated (for example, relationships between people or their demographics), but it should not simply present speculation as fact. It should flag speculation; for example, it may say "may" or "might" instead of "is".

An **informative** explanation does something more than simply restate the post. It may bring in some background knowledge ("because society places a lot of importance on X, the author may be feeling Y") or make some insightful summarization choices ("the relationship seems abusive because X, Y, and Z", "the author is probably afraid of this happening again in the future"). Basically, some kind of information should be presented that is not just explicitly said in the post. This criterion may be rather fuzzy, and you are encouraged to use your own judgement or reach out to me if you are unsure.

If you believe a free-response question does not apply (e.g., asked what you dislike about an explanation but there is nothing), you may leave the box blank or enter something like "N/A".

Content Warning: Please be aware that some of the content in this annotation task may be triggering. This data may include topics such as discussions of miscarriage, death of a child or pet, various mental health topics, and strong emotions such as anger and sadness. If you have any concerns or encounter any data that is problematic for you, you can stop annotating and [contact me](#) at any time. This will not affect your compensation.

Figure 8.2: **Human Evaluation Instructions.** The annotation instructions for our human evaluation of our generated emotional explanations. Annotators could collapse the HTML element containing the instructions for readability, and a link to contact the primary author was provided at the bottom.

ANNOTATION

Please use the following social media post for your annotation.

\$(post)

This is the given explanation for the author's emotions.

\$(explanation)

1. Overall, how would you rate the quality of the explanation?

Very bad quality Bad quality Good quality Very good quality

2. What do you like about the explanation, if anything?

Enter your answer.

3. What do you dislike about the explanation, if anything?

Enter your answer.

4. Is the explanation factually consistent with the post?

Yes No

5. If the explanation is not factually consistent, what parts of the explanation contradict the post?

Enter your answer.

6. How informative is the explanation? To what extent does it tell you information not explicitly stated in the post?

Not at all Not very much Somewhat Significantly

7. What new information or insight does the explanation give, if any?

Enter your answer.

8. Imagine that you are a professional such as a counselor, and someone has presented to you as potentially in need (or not in need!) of mental health support and guidance to appropriate resources. You are given this social media post and the explanation as part of the triage process. How useful would this explanation be to you in deciding whether, or how urgently, to provide this support, as compared to just having the post?

Not at all useful Not very useful Somewhat useful Very useful

9. (Optional) Do you have any other comments or thoughts on this explanation?

Enter your answer.

Figure 8.3: **Human Evaluation Interface.** The annotation interface for our human evaluation of our generated emotional explanations. Student social workers completed annotations for 21 items each using this interface on Amazon Mechanical Turk.

focus our human evaluation efforts on the longer and more complex explanation task. For this evaluation, we recruit a set of four student social workers (distinct from the set of annotators who created our seed annotations in §8.2.1) to evaluate a random sample of 21 items from the test set. We structure the annotations in the same manner as those in §8.2.3 to create our test set: each annotator is matched equally with each of the three other annotators to avoid any systematic bias (selecting 21 items allows us to ensure that each annotator pair is matched exactly seven times). Annotators are paid \$20/hr for their work according to institutional standards.

We present the instructions and interface for our human evaluation in Figure 8.2 and Figure 8.3, respectively. Because our annotators are researchers in social work rather than computer science, we first bring the instructions before a focus group of our interdisciplinary grant collaborators for feedback, which allows us to frame the annotations in more understandable ways. For example, a layperson may not understand that automatically generated text may sometimes hallucinate or become grammatically incoherent, and it is helpful to give detailed context about the current state of our technology and the situations in which it may be used in the future.¹³

We ask annotators to evaluate an explanation on all the criteria from §8.1.1, including both numerical ratings and free responses. Given the general fluency of LLMs, we fold fluency into an overall “quality” metric, which we evaluate on a 4-point Likert scale. Informativeness is also ranked on a 4-point Likert scale, while factual consistency is a binary judgment. We also add a new criterion in question 8, **utility**, evaluating whether the explanations in their current state would be useful in a real-world setting of the type we envision (also evaluated on a 4-point Likert scale). Annotators were asked to explain in their own words what they did or did not like about the explanations and

¹³So as not to bias our annotators, we did allow them to think some explanations may have been written by a human, though in fact all the evaluated data was automatically generated. The nature of the data was revealed after the annotations were complete, as part of a debriefing.

Model	Extractiveness		Reference-Based				Factuality	
	Overlap	Density	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score	SummaC	QAFactEval
Flan-T5 GPT	0.64	3.32	0.46	0.18	0.30	0.39	0.27	1.50
Flan-T5 Split	0.47	2.75	0.46	0.17	0.29	0.39	0.25	0.96
Flan-T5 Split + QA	0.53	3.34	0.47	0.19	0.31	0.39	0.26	1.30
Flan-T5 Split + Metaphor	0.46	2.85	0.46	0.17	0.30	0.36	0.25	0.89
Flan-T5 Split + Paraphrase	0.43	2.96	0.44	0.17	0.29	0.36	0.25	0.74
Flan-T5 Split + Emotion	0.44	2.75	0.44	0.16	0.28	0.35	0.26	0.91
Flan-T5 Split + All	0.48	3.14	0.44	0.17	0.29	0.35	0.26	0.96

Figure 8.4: **Explanation Automatic Evaluation Results.** The results of our models on the test set, for the explanation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsubsection 8.3.2](#) for discussion).

why an explanation a) was not factually consistent or b) was informative.

8.4 Results and Analysis

8.4.1 Automated Metrics

We present the results of our automated metrics on the test set in [Figure 8.4](#) (explanation), [Figure 8.5](#) (emotion), and [Figure 8.6](#) (event). For readability, and because we expect their behavior to be most stable on the explanation task, we present the models with groups of additional tasks only and show results on the individual multitask models in the appendix. Additionally, we render these tables as images and color the cells corresponding to each metric in a gradient such that the least desirable score is red and the most desirable score is green (refer to [subsubsection 8.3.2](#) for a discussion of whether a higher or lower score is best for each metric).

We see that overall, **Flan-T5_{Split}+QA** and **Flan-T5_{GPT}** perform best on the explanation task. This is according to our automated metrics, which means that these models mimic the behavior of GPT-3 well. Flan-T5_{Split}+QA performs very slightly better on reference-based metrics, and Flan-T5_{GPT} performs slightly better with

Model	Extractiveness		Reference-Based							Factuality	
	Overlap	Density	Precision	Recall	F1	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SummaC	QAFactEval
Flan-T5 GPT	0.35	0.03	0.67	0.54	0.59	0.40	0.16	0.36	0.37	0.26	0.01
Flan-T5 Split	0.32	0.00	0.53	0.50	0.50	0.62	0.21	0.55	0.76	0.27	0.02
Flan-T5 Split + QA	0.34	0.00	0.62	0.50	0.54	0.66	0.27	0.60	0.78	0.27	0.05
Flan-T5 Split + Metaphor	0.35	0.00	0.64	0.56	0.58	0.68	0.27	0.60	0.78	0.26	0.06
Flan-T5 Split + Paraphrase	0.34	0.03	0.55	0.50	0.51	0.63	0.27	0.59	0.76	0.26	0.04
Flan-T5 Split + Emotion	0.38	0.00	0.65	0.49	0.55	0.66	0.31	0.62	0.78	0.27	0.00
Flan-T5 Split + All	0.39	0.00	0.70	0.52	0.58	0.69	0.31	0.63	0.79	0.27	0.00

Figure 8.5: **Emotion Automatic Evaluation Results.** The results of our models on the test set, for the emotion task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsubsection 8.3.2](#) for discussion).

respect to factuality. We also note that Flan-T5_{Split}, as well as models that include QA tasks, are highly extractive; other combinations of tasks such as the paraphrase tasks can reduce extractiveness. Interestingly, the addition of QA tasks does not improve factuality, nor does the addition of any other task reduce factuality as compared to Flan-T5_{GPT}.

For emotion detection, overall, Flan-T5_{Split}+All performs best. As we discussed in [subsubsection 8.3.2](#), we feel ROUGE-1 and BERTScore are the most appropriate metrics by which to measure emotion detection, though we present all metrics for completeness. Flan-T5_{Split}+Emotion, as expected, also performs quite well. Flan-T5_{Split}+All is highly extractive, meaning that it produces emotions directly mentioned by the author in the post; Flan-T5_{Split}, Flan-T5_{Split}+Paraphrase, and surprisingly, Flan-T5_{Split}+QA perform more abstractively. Density for almost every model is zero, which is expected because the output is a list of emotion words; surprisingly, Flan-T5_{GPT} and Flan-T5_{Split}+Paraphrase achieve a small but nonzero score on density. As expected, factuality scores carry little meaning for evaluating this task, particularly under SummaC. This is also the task for which the Flan-T5_{GPT} baseline performs the worst; while it performs well on the other two tasks, this shows that T5-large may still have some difficulty outputting multiple tasks at once, unlike LLMs.

Model	Extractiveness		Reference-Based				Factuality	
	Overlap	Density	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SummaC	QAFactEval
Flan-T5 GPT	0.69	2.44	0.33	0.17	0.31	0.30	0.41	2.21
Flan-T5 Split	0.60	2.18	0.28	0.14	0.27	0.28	0.43	2.38
Flan-T5 Split + QA	0.64	2.31	0.35	0.17	0.33	0.30	0.40	2.13
Flan-T5 Split + Metaphor	0.50	1.95	0.28	0.11	0.25	0.28	0.38	1.53
Flan-T5 Split + Paraphrase	0.54	1.66	0.28	0.12	0.27	0.27	0.39	1.44
Flan-T5 Split + Emotion	0.55	1.66	0.27	0.10	0.25	0.28	0.42	1.79
Flan-T5 Split + All	0.54	2.14	0.26	0.11	0.24	0.25	0.42	2.09

Figure 8.6: **Event Automatic Evaluation Results.** The results of our models on the test set, for the explanation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsection 8.3.2](#) for discussion).

For the event task, we see that Flan-T5_{Split}+QA and Flan-T5_{GPT} perform best. Overall, with respect to grouped models, Flan-T5_{Split}+QA performs most like GPT-3 according to the reference-based metrics; Flan-T5_{GPT} also performs very well on reference-based metrics, and both these models are highly extractive. Flan-T5_{Split}+Metaphor, Flan-T5_{Split}+Paraphrase, and Flan-T5_{Split}+Emotion are more abstractive. Surprisingly, Flan-T5_{Split} performs best on factuality. We note that actually the best overall scores on the event task come from adding single additional tasks at once (particularly Para-NMT, though QQP, NQ, and SQuAD also perform well).¹⁴ Because our main focus is on explanation generation, none of the additional tasks we use are focused on event extraction, so it is possible that for this particular task they interfere with one another during training.

Interestingly, we do not observe a substantial difference between the performance of tasks already present in Flan’s instruction-based fine-tuning (QA and paraphrasing tasks) and tasks that are new to the model. Flan’s data is very large, so it is possible that the specifics of the tasks have not been fully retained by the time we finish our task-specific fine-tuning.

¹⁴See [Appendix C](#)

Model	Quality	Factuality	Informativeness	Utility
Flan-T5 _{GPT}	2.71 ± 0.83	0.51 ± 0.50	2.15 ± 0.93	2.76 ± 1.05
Flan-T5 _{Split} +QA	2.41 ± 0.94	0.29 ± 0.45	2.07 ± 0.92	2.46 ± 1.04

Table 8.6: **Sample Human Annotation Results.** The collected results of our human evaluation. Each cell shows the mean ± standard deviation of the collected scores under each metric. The mean of the best model under each metric is bolded. No differences are significant under the approximate randomization test.

8.4.2 Human Evaluation

The two models we select for human evaluation are Flan-T5_{GPT}+QA (a baseline) and Flan-T5_{Split}+QA, because we observe the best performance under the automatic metrics for these two models on the explanation task, and they represent two different training styles. We also hope that the addition of QA tasks will change the model’s behavior on metrics like factuality. We present the results of this human evaluation in [Table 8.6](#). In this table, the reported scores are the average across all evaluated datapoints from the given model; factuality is binary, where 0 is “not factually consistent” and 1 is “factually consistent”, and the remaining criteria range from 1 to 4, where 1 is “not at all” or “very bad” and 4 is “significantly” or “very good”.

Overall, the two models perform quite similarly; while Flan-T5_{GPT}, which mimics the behavior of GPT-3, actually performs slightly better across most metrics, the differences are not significant and the variance is quite high (per metric, the observed variance for each model is similar). On average, both models score between 2 and 3 on the Likert metrics, which means that the annotators overall rank the explanations as average or neutral (since 2 is “bad” and 3 is “good”). Surprisingly, the largest difference is in factual consistency, where Flan-T5_{GPT} outperforms Flan-T5_{Split}+QA (again, not significantly); this matches the automatic evaluation results according to QAFactEval. The models overall perform worst on informativeness, which is a reasonably subjective criterion—it makes sense that it would be difficult for a human to judge and for a model

to achieve. When annotators provide their free thoughts on the model outputs, they generally praise both models for accurately identifying emotions and events even in a complex post and selecting important phrases from the input post to include in their summary, while noting that the models both tend to miss important context and are not often informative.

These results highlight that human evaluation of a good explanation is a difficult and unsolved problem. The results of this evaluation and the automatic evaluation do not conclusively select one model as the best overall performer, but show that different approaches have advantages over others in different ways, for example:

- When performing multitask fine-tuning, models trained on QA tasks (Flan-T5_{Split}+QA, Flan-T5_{Split}+All) are more skilled at generating GPT-like outputs (i.e., they score highly on reference-based metrics).
- Models trained on QA tasks also generally generate extractive explanations that quote the input.
- Models trained on emotional tasks (Flan-T5_{Split}+Emotion, Flan-T5_{Split}+All) can more accurately identify emotions.
- Flan-T5_{GPT}, trained to mimic the style of GPT-3 completely, appears to generate explanations that are more subjectively useful to humans, though it struggles with emotion detection.

8.4.3 Qualitative Analysis

In this section, we present several sample outputs from our models in individual tables for readability. We caution the reader that in this section we will explicitly show social media posts that deal with topics including death of a pet, grief, social anxiety,

Input Post
<p>Today, I unexpectedly lost my beloved Yorkshire Terrier, my baby girl for 13 years, to a sudden heart attack. I wasn't there when it happened, and coming home to the news was devastating. She'd been healthy until a bout of ovarian cancer last year, swiftly treated with surgery. Her passing was so sudden, I'm struggling to accept it. She's been my closest friend since I was 7, and she has just been the most important companion all that time. I'm heartbroken that I couldn't be there for her in her final moments. Saying goodbye was agonizing, especially knowing she couldn't hear me. I'm overwhelmed with grief and unsure how to begin healing. Wishing her peaceful rest.</p>
Flan-T5_{Split}+All Output
<p>Emotions: grief, guilt, sadness Event: losing a baby to heart attack Explanation: The author is feeling immense guilt and regret for not being able to do something to save their baby, who was on their way home when they were a child. They are also feeling regret for not being able to do something to save their baby, and for not being able to say goodbye to their baby. The author is also feeling a sense of sadness for the loss of their baby, and for not being able to be with them when they needed them.</p>
Flan-T5_{Split}+Emotion Output
<p>Emotions: grief, sorrow, guilt Event: Loss of a pet Explanation: The author is feeling a deep sense of grief and sadness over the loss of their beloved pet. They feel guilty for not being able to do anything to prevent the death, and for not being able to be there for their pet in its final moments. They are also feeling a deep sense of loss, as their pet had been a part of their life since they were a young child. The author expresses a deep sense of loss and is struggling to find a way to cope with their grief.</p>

Table 8.7: **Emotion vs. All Multitask Model Example.** A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.

and abusive relationships. For the purposes of this discussion section, the posts are anonymized by automatic paraphrasing and summarizing,¹⁵ so they are not reproduced exactly as found, but any key features of the input that are discussed in the model outputs will remain. Note that this does complicate visualizing the models’ copying behavior in these examples, though where salient we have tried to make it clear that the outputs use the same words as the inputs.

First, in [Table 8.7](#), we show two outputs for the same post from our Flan-T5_{Split}+Emotion and Flan-T5_{Split}+All model, demonstrating potential interference by the additional tasks. For this post, the +Emotion model is factually consistent with the input and correctly identifies both the emotional reaction and its triggering event, while the +All model misunderstands the input, not connecting the phrase “baby girl” to a pet dog. The +Emotion model output is overall more coherent and informative, and makes explicit connections such as “[The author is] feeling a deep sense of loss, as their pet had been part of their life since they were a young child,” which is stated less explicitly in the source post. We also note that the +Emotion model has chosen *sorrow* where the +All model has chosen *sadness* as an emotion label, sorrow being a less common term with arguably more appropriate connotations for grief. In this case, it seems that adding additional tasks may have hampered the model’s ability to generate a high-quality emotional explanation.

Next, in [Table 8.8](#), we show two outputs from our Flan-T5_{Split}+Emotion model and our Flan-T5_{GPT} model that demonstrate the tradeoff between factual consistency and informativeness. In this instance, we see that Flan-T5_{GPT} first of all makes a factual mistake (misattributing the author’s line about “especially someone I was interested in” and conflating this person with the other woman in the narrative), where the +Emotion model, while somewhat disfluent, is factually consistent with the input. The two models

¹⁵Thanks to ChatGPT, with some manual editing.

Input Post
I'm struggling a bit with making friends at my new job. I tend to speak without filtering, and it's caused some awkward moments. One girl I've been getting to know has only been talking about her miserable life, and I unintentionally called her out for it. It didn't go over well with others, especially someone I was interested in. I worry I'm trying too hard to have something relevant to say and might be coming off as annoying. I want to engage with people, but I worry it's better if I keep my mouth shut. Any advice that's not just "think before you speak"?
Flan-T5_{GPT} Output
<p>Emotions: insecurity, fear, anxiety</p> <p>Event: Accidentally calling out a girl</p> <p>Explanation: The author is feeling insecure and anxious because they have been working up the courage to ask out a girl, and they feel like they have been trying too hard to input something relevant to the other people, which may have come off as annoying. The author is afraid that if they keep their mouth shut, they may not be able to talk to people.</p>
Flan-T5_{Split}+Emotion Output
<p>Emotions: anxiety, worry</p> <p>Event: Accidentally making a mistake at work</p> <p>Explanation: The author is feeling anxious and fearful because they have a problem with saying things without thinking about the people they are at work with. They are worried that they may have messed up their words and that they may have messed things up. They are also feeling insecure because they are trying to talk to people, but they feel like it would be better to keep their mouth shut. They are trying to talk to people, but they feel like it would be better to just keep their mouth shut.</p>

Table 8.8: **Multitask Emotion vs. Combined Model Example.** A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.

Input Post
I've been with this guy for five years, and I love him desperately. He's a pastor, confident, and outgoing, but our relationship has been toxic and abusive, filled with lies. Despite leaving, he came back claiming he's changed. There's been some improvement, but I still don't fully trust him. He's controlling and lies all the time, and I'm considering leaving, but I'm torn. He sells himself well, talking about how rich and self-sufficient and attractive he is. He's very confident, and I admire that about him. I don't think I can do better than him. But my friends don't support us being together. How do I move on from someone I admire like this?
Flan-T5_{Split}+All Output
Emotions: confusion, fear, sadness Event: Relationship with a man they care about Explanation: The author is feeling a range of emotions due to the fact that they have been in a relationship with a man they care about for a while. They feel fear and sadness due to the fact that they have been in an abusive relationship for a long time, and they don't know how to move on. They also feel confusion because they don't know if they are missing on something great or if they are missing out on something great. The author is also feeling hopelessness because they don't know how to move on and don't know how to move on.
Flan-T5_{Split}+QA Output
Emotions: confusion, guilt, sadness Event: Being in a toxic toxic relationship Explanation: The author is feeling confused and uncertain about how to move on from the relationship they have been in for five years. They feel guilty for not being able to trust the man they have been in a relationship with for a long time, and for not being able to trust him enough to talk about his accomplishments. They are also worried that they may be missing out on something great, and are unsure if they can trust him enough to talk about his accomplishments. The author is feeling overwhelmed and uncertain about how to move on from the relationship, and is asking for advice on how to do so.

Table 8.9: **QA vs. All Multitask Model Example.** A pair of example outputs from our models. Note that as always within this chapter, the source social media post is paraphrased to minimize searchability.

make different choices in what segments of the input to emphasize; the output of Flan-T5_{split} is overall shorter and picks the very salient segment about “trying too hard to input something relevant to the other people”, while the +Emotion model carefully paraphrases the entire post. Overall, in this instance Flan-T5_{GPT} is more creative with its output, which in this instance leads to hallucinations and factual mistakes.

Finally, in [Table 8.9](#), we give two examples from our Flan-T5_{split}+QA and Flan-T5_{split}+All models. One interesting difference between these two models is their choice of triggering event, with the +QA model correctly (if disfluently) using the word “toxic” to describe the relationship, while the +All model describes it as a “relationship with a man they care about” (which is also true, according to the first sentence of the post, though probably less salient to the reader). We see the +All model make some insightful choices, such as suggesting that the length of the abusive relationship specifically leads to fear and sadness. The +QA model introduces some factual inconsistencies (e.g., nowhere in the post does the author mention not being able to talk about their partner’s accomplishments), which is surprising given that the +QA model scores well on QA-based factuality metrics. In this instance, it is difficult to say that one output is higher quality than the other in all respects.

Overall, we see that each of our models has its own advantages and its own weaknesses. No one multitask approach seems conclusively best to train a distilled model to perform multiple concurrent tasks. We do see some evidence confirming that this is a difficult task for distilled models to learn to mimic the outputs of an LLM, as some prior research has suggested ([Gudibande et al., 2023](#)); in the previous examples, we find that Flan-T5-Large is easily able to learn the desired format as well as the general style, though the content and internal logic can be somewhat disjointed. We conclude that further research is needed to understand how multitask learning can best assist small models with problems like these. For example, with a small number of tasks, it is

possible that rearranging the multitask curriculum or selecting different tasks may be useful (e.g., [Varshney et al. \(2022\)](#); [Zhang et al. \(2023d\)](#)).

8.5 Discussion and Conclusion

In this chapter of the thesis, we introduce a new task, explanation generation, suitable for the dawning era of more nuanced affective NLP with LLMs. Through dialogue with social science researchers, we define criteria for a useful and good explanation (§8.1.1) and develop a dataset, GEDdit, sourced from Reddit and supplemented with expert-written seed examples (§8.2.1) and LLM-generated explanations (§8.2.2). We train a suite of distilled Flan-T5-Large models to imitate the outputs of GPT-3, including multitask training with a variety of different tasks (§8.3.1), and we discuss appropriate evaluation metrics for our models (§8.3.2). While the results of our experiments in this chapter are inconclusive, we demonstrate that relatively small models are able to make good progress on our explanation generation task (Figure 8.4) while simultaneously performing two additional tasks, emotion detection and triggering event extraction (Figure 8.5, Figure 8.6).

We find the explanation generation task to be a challenging task for models to perform and also for humans to quantify and evaluate. Outputs of models that perform well under automated metrics (reference-based and otherwise) do not always appear high-quality under manual inspection (§8.4.3), supporting prior research that automated metrics do not necessarily capture human intuitions about quality (e.g., [Callison-Burch et al. \(2006\)](#)). We find that different arrangements of multitask learning lead to improved performance according to automatic evaluation for each of our tasks.

- According to reference-based metrics, Flan-T5_{Split+QA} performs best on the explanation generation task, while Flan-T5_{GPT} generates the most factual expla-

nations.

- According to almost all metrics, Flan-T5_{Split}+All performs best on the emotion detection task, while Flan-T5_{Split}+Metaphor scores highest on QAFactEval.
- Flan-T5_{Split}+QA scores highest on the reference-based metrics for event extraction, while Flan-T5_{Split} performs best according to factuality metrics.
- There is no strong consensus on which models perform most extractively or most abstractively, though Flan-T5_{Split}+Metaphor, Flan-T5_{Split}+Paraphrase, and Flan-T5_{Split}+Emotion consistently perform fairly abstractively across tasks.

Overall, according to automatic and human evaluation, our models generally excel at accurately summarizing the emotional state described in the input and rely on copying appropriate segments of the input, though they struggle with the more creative and thoughtful aspects of explanation generation, such as making generalizations or incorporating world knowledge (which we describe in this work as *informativeness*).

As with all work, our work is limited by the problem formulation, data, and models we choose. Because of cost limitations, the data we use in this work is relatively small. Since this work was begun, newer LLMs have been released, including multiple that are open-source (e.g., LLaMa (Touvron et al., 2023), Mistral (Jiang et al., 2023)), which may be able to label more data more efficiently, or even perform the explanation generation problem on a large scale. As with all of our work in this thesis, the choice of social media data from Reddit additionally limits the demographics to which we can say our work applies.

Our contributions in this work include (1) a formulation of the new problem of emotional explanation generation, including multiple criteria by which an explanation should be judged (factual consistency, informativeness, and fluency); (2) a new dataset

for this problem consisting of 10,000 Reddit posts from various emotional subreddits, along with a seed set of 90 human explanations and a larger set of silver GPT-3 generated explanations; and (3) a suite of distilled models incorporating multitask learning for this dataset, along with an examination of their capabilities across three tasks (explanation generation, emotion identification, and triggering event extraction).

Chapter 9

Discussion for Understanding Emotions and Events

In this section of the thesis, we have presented two types of models designed for a deeper understanding of emotional reactions: joint emotion-cause detection models supplemented with commonsense reasoning ([Chapter 7](#)) and distilled generative models trained to explain an emotional reaction in a free-text way ([Chapter 8](#)). These two problems are related but distinct in both scope and execution: we begin by using sequence tagging to identify the cause of a single emotion as a text span, and we then proceed to broaden our scope to the more general problem of explaining a holistic emotional reaction with generated text. These systems also perform emotion detection, albeit with the help of black-box models rather than expert knowledge thus, we present unified systems for emotional state detection and explanation. In contrast to [Part I](#) of this thesis, in this part we explore the use of LLMs, which are typically highly capable but minimally interpretable, to understand and explain emotional reactions. We believe that the next promising research direction for future work is to combine the model philosophies from the two parts of this thesis to create capable models that can perform multiple tasks with the competence of LLMs but the explainability that comes from integrating domain and expert knowledge into their architecture. To conclude this thesis, we will begin by reflecting on the themes discussed in this second part.

In [Part II](#) of the thesis, we continue to revisit the themes of explainability and multitask learning found throughout [Part I](#), and we also explore a third theme: the

inherent fuzziness of emotional tasks. As we discussed in [Part I](#), expert research does not even necessarily agree on the definition of an emotion, let alone any adjacent concepts. We find that in NLP, where much of our research is empirical, affective problems are difficult to operationalize and evaluate because emotion is inherently subjective and inextricably linked to the person experiencing it. We will discuss each of these themes in turn.

First, we focus in this part of the thesis on different architectures. In the first part of the thesis, we primarily used parameter sharing and off-the-shelf post-hoc explainability frameworks to understand our models’ behavior ([Chapter 4](#)), and in this part, we examine two distinct modeling paradigms. In [Chapter 7](#), we develop more sophisticated multitask frameworks that perform tasks sequentially, feeding the predictions from one task in as input for the other, and we create our models to incorporate parameters whose function should be intuitive to humans. Our models in this section follow a philosophy of surfacing important predictions at multiple steps to make their processes more transparent (e.g., using natural language commonsense derived from COMET). Then, as we proceed to develop models that make use of LLMs ([Chapter 8](#)), we demonstrate a very different and largely non-explainable type of model. While LLMs are capable of performing more complex tasks such as writing explanations for a given emotional reaction, it is currently opaque how they arrive at their outputs, and we hope that future work will make interpretability and explainability a focus.

Second, as in [Part I](#), our models in the second part of this thesis leverage multitask learning—in this case, as a way to share knowledge between different tasks and improve model performance. This is in contrast to the first part of the thesis, where we use multitask learning primarily to improve models’ explainability. In this second part, we experiment with different mechanisms of incorporating multiple tasks ([Chapter 7](#)) and incorporating many tasks at once ([Chapter 8](#)), a more detailed treatment which leads

to improved performance. We continue to prioritize explainability (for example, using COMET and sequential multitask learning) and believe there is ample room for future work to examine the effects of multitask learning on the explainability of larger models with more target tasks.

Third, we wish to remark on the inherent variability and fuzziness of emotional problems in NLP. As we saw in [Chapter 7](#), emotion labeling is a task with high variability even for human annotators. Human emotional states are not often able to be fully described by a single word. We believe that emotion detection is not inherently a single-label classification problem; it is likely best described by a vastly multilabel regression problem where people feel multiple different things at different intensities at once. As in any branch of science, we can develop theoretical models of emotion that make some simplifications and are still useful, but as the capability of our computational models increases, we believe NLP researchers should increasingly focus on understanding emotion as a more complex problem. We make some initial strides toward this goal in this thesis, and we find that even defining more complex emotional problems and the criteria that would characterize a good output is challenging and deserving of dedicated research in its own right. Different research develops different names and definitions for the same or similar problems (for example, is it the “cause” of an emotion or its “stimulus”, and is this interchangeable with the “target”?), and while this is desirable because it inspires and facilitates many distinct threads of research, we call for systematic examination of different emotional tasks in NLP, their similarities and differences, and how they may or may not interact.

9.1 Contributions

The contributions covered in this part of the thesis are as follows:

Multitask and commonsense-infused models for joint emotion detection and emotion cause extraction. We present three varieties of multitask model in [Chapter 7](#) (Multi, Multi $C \rightarrow E$, and Multi $E \rightarrow C$), along with adapted knowledge model variants of each that incorporate commonsense reasoning using COMET. We train and evaluate these models on emotion detection and emotion cause tagging for news headlines, with each model sharing knowledge differently between the two tasks. A combination of multitask learning and commonsense reasoning yields the best performance on both tasks, outperforming single-task baselines as well as either technique alone.

A dataset for the new problem of emotion-focused explanation, and a characterization of the abilities of distilled generation models. We define the new problem of emotional explanation generation ([Chapter 8](#)), including basic criteria for a good explanation, and collect a dataset for this problem from Reddit, Generating Emotional Descriptions (GEDdit). We collect expert human annotations for a small seed set of data and label a larger corpus with silver outputs generated by a LLM; we then proceed to train several varieties of distilled models based on T5 to replicate its behavior, including Flan-T5 GPT , Flan-T5 $Split$, and Flan-T5 $Split$ +dataset variants for multiple different kinds of external datasets. Results show that Flan-T5 GPT , Flan-T5 $Split$ +QA, and Flan-T5 $Split$ + *All* perform best at this task at a macro level, but that overall, this is a complex problem where no one distilled model is superior to others on automated metrics and human evaluations.

An approach to these problems that is inspired by psychology theory and incorporates domain knowledge where possible. As we develop our joint emotion detection and emotion-cause extraction models in [Chapter 7](#), we draw on the external knowledge model COMET in order to enhance our models' performance on both tasks. Meanwhile, when we turn to generating emotional explanations ([Chapter 8](#)), we rely on experts in social science to help us develop our problem definition, create our seed

annotations, and evaluate the quality of our outputs, giving us a clearer picture of how our models can actually be useful in a real-world setting.

9.2 Limitations and Future Work

As with any practical, empirical work, the definitions, data, and techniques we choose impose certain limitations on our results. A significant limitation that applies to our work in [Part I](#) also applies here: data representation. While we use more and different sources of data in this part of the thesis than just Dreddit, any selection of data imposes limitations and bias upon models it is used to train. Our work in [Chapter 8](#) used GEDdit, which is still sourced from Reddit, albeit more recent data than Dreddit. The GoodNewsEveryone dataset we use in [Chapter 7](#) draws on more varied sources (including a variety of United States news sources), but news headlines are quite short with little context (making accurate emotion detection difficult), and we have no readily-accessible information about the demographics of the people represented therein. As with [Part I](#), we continue to hope that future work will consider the emotional experiences of all people as expressed in many different genres.

In developing our models, we are struck with the capability of LLMs to perform complex emotional tasks, as opposed to smaller models. Models based on smaller language models like BERT demonstrate competence at lower-level tasks such as sequence tagging, even if the underlying problem is emotionally complex in nature (such as identifying expressions of emotions and their causes ([Chapter 7](#))). However, we believe that generation is a much more promising format for understanding emotional reactions, as paragraph-level output gives our models the space and freedom to include details and explanations for their conclusions. We hope that future work will explore the utility of LLMs for complex emotional problems like these while increasing their

explainability and transparency. In particular, we believe there is room to incorporate the kind of expert knowledge we leverage in [Part I](#)—for example, by prompting LLMs to perform relevant side tasks in addition to the target task.

In this work, we also largely limit ourselves to understanding the cause of emotional reactions, motivated by the widespread acceptance in psychological theory that an emotion has a stimulus. However, as we have discussed ([Chapter 6](#)), NLP research has examined many other facets of emotions, such as identifying the experiencer or target. We can also imagine models which focus on explaining other facets of the reaction, such as what steps a person is taking to handle a negative emotional reaction or how the emotional reaction is influenced by or influences other people’s emotional states. The types of free-text explanations we produce in [Chapter 8](#) would be flexible enough to accommodate many of these additional focuses, and we believe that the logical next step in this research would be to explore developing different *types* of emotional explanations focusing on different aspects of the emotional reactions.

In summary, we have developed models that can flexibly perform emotion detection for a variety of emotional states and also concurrently perform tasks centering on understanding these emotions in a deeper way. There are many exciting new directions for research in this area, including exploring a range of different understanding-related tasks (such as identifying other semantic roles or explaining relevant social connections in an emotional situation), developing LLM-based models that incorporate concrete aspects of expert knowledge such as psychological theory to increase their explainability and transparency, and ensuring that people across different demographics (gender, race, nationality, and so on) are accurately represented in training data and future models.

Chapter 10

Conclusion

Finally, with the emotion-infused distress detection models of [Part I](#) and the broader, more flexible explanation models of [Part II](#), we arrive at the conclusion to this thesis. In [Part I](#), we began by detecting emotional reactions, presenting both a new dataset for the new problem of psychological distress detection (based on definitions from psychology literature) and a suite of models incorporating further psychology research and leveraging emotion detection to create more human-understandable model predictions for this problem. In [Part II](#) we sought a deeper understanding of the reasons for emotional reactions, first using multitask models with commonsense reasoning to perform joint emotion-cause detection, and then using models that generate free-text explanations of an emotional reaction and its cause. As we set out to do in our introduction, we have developed computational models that begin to answer the ambitious questions of “What am I feeling?” and “Why am I feeling this?”

Throughout this work, we have used multitask learning and drawn on domain knowledge in our models to improve their explainability and performance. We advocate strongly for the creation of theory-guided models in various areas moving forward. While we have developed models for limited use cases (for example, theory-guided models for one type of emotional reaction, distress), as we will discuss in [Section 10.2](#), future work has many potential avenues to pursue. We believe that it is possible to fuse our bespoke theory-guided modeling approaches with the adaptive power of LLMs

to create models capable of understanding a wide range of emotional experiences in detail and of supporting their conclusions with human-understandable evidence drawn from research and knowledge in fields like psychology.

This thesis is completed at a very exciting time in NLP, when the rise of LLMs has dramatically increased the capabilities of our models and enabled a host of new tasks in a generative setting. Over the course of the thesis, our work evolves from binary distress detection in [Chapter 3](#) to free-text explanation generation in [Chapter 8](#) thanks in part to the evolution of new models. This progression has allowed us a deeper understanding of emotional reactions from the beginning to the end of the thesis, and future research will be capable of many new and exciting tasks. We can envision tools that produce coherent emotional narratives from a series of texts, provide personalized access to mental health resources based on individual factors, or interact with users to build a collaborative understanding of their current emotional state that is meaningful for the users. We hope that further research in this area will take full advantage of powerful modern models while keeping explainability at the forefront of its design goals.

10.1 Contributions

We once again summarize the contributions of this thesis:

A dataset for psychological distress detection ([Chapter 3](#)), and emotion-infused models that incorporate emotion detection with existing emotion datasets ([Chapter 4](#)). We present our dataset, Dreddit, for the new problem of psychological distress detection, along with multi-pronged analysis of its content ([Chapter 3](#)). We show that Dreddit contains expressions of distress as well as counterexamples from various domains and uncover some features which may be correlated with distress, including emotional wording, sentence complexity, and first-person pronouns.

We benchmark contemporaneous state-of-the-art models against this dataset, finding it to be challenging, and proceed to develop three types of emotion-infused models that draw on psychology research linking distress and particular common emotions to yield more understandable results (Chapter 4). Our Multi, Multi^{Alt}, and Fine-Tune models use visibly different keywords to make their predictions about distress as compared to a single-task baseline, focusing more on words representing emotional, cognitive, and social processes.

Multitask and commonsense-infused models for joint emotion detection and emotion cause extraction in news headlines, showing performance gains (Chapter 7). To perform the tasks of emotion detection and emotion cause tagging concurrently, we present three varieties of multitask model (Multi, Multi_{C→E}, and Multi_{E→C}), along with adapted knowledge model variants of each that incorporate commonsense reasoning using COMET. We train and evaluate these models on emotion detection and emotion cause tagging for news headlines, with each model sharing knowledge differently between the two tasks. A combination of multitask learning and commonsense reasoning yields the best performance on both tasks (Multi_{C→E}^{COMET} for emotion and Multi^{COMET} for cause), outperforming single-task baselines as well as either technique alone.

A dataset for the new problem of emotion-focused explanation, and a characterization of the abilities of distilled generation models (based on large language models) for this problem (Chapter 8). We define the new problem of emotional explanation generation, including basic criteria for a good explanation, and collect GEDdit, a dataset for this problem from Reddit. We collect expert human annotations for a small seed set of data and label a larger corpus with silver outputs generated by a LLM. We then proceed to train several varieties of distilled models based on T5 to replicate its behavior: Flan-T5_{GPT}, Flan-T5_{Split}, and Flan-T5_{Split}+dataset,

for a variety of additional datasets and tasks including metaphorical language understanding, paraphrasing, question answering, and emotion identification. We evaluate our explanations using both automatic metrics (for extractiveness, similarity to GPT-3 through reference-based metrics, and factuality) and human evaluation of their quality. Results show that this is a difficult problem to perform and evaluate, but that overall, Flan-T5_{GPT}, Flan-T5_{Split}+QA, and Flan-T5_{Split}+All perform well.

An overarching approach that is inspired by psychology theory and incorporates expert knowledge into our models where possible, enhancing explainability and performance. We are motivated by theory as we develop our datasets and models in this thesis. We define the problem of psychological distress detection throughout [Part I](#) by drawing on established psychological theory ([Selye, 1976b](#)). In [Chapter 8](#), we also use seed explanations written by domain experts, student social workers, to define a good target explanation of an emotional reaction. We design our models with insight from published psychology research, including our multitask models in [Chapter 4](#) for distress detection and [Chapter 7](#) for joint emotion-cause detection. Finally, as we analyze trends in our data and results, we support them with research from other fields (for example, the idea that people experiencing distress tend to focus on the self and use more first-person singular pronouns ([Zimmermann et al., 2013](#); [Lyons et al., 2018](#)), which we observe in [Chapter 3](#)). This approach makes our models more understandable and their predictions more explainable by drawing on domain knowledge and it also results in improved performance throughout [Part II](#).

10.2 Limitations and Future Work

In our previous discussions ([Chapter 5](#), [Chapter 9](#)), we have laid out several limitations of this work, chiefly including that we have limited ourselves to a specific emotional

state in [Part I](#) (psychological distress), that we have limited ourselves to a particular additional aspect of emotion in [Part II](#) (cause, as opposed to target, experiencer, etc.), and that our choice of data limits the types of people whose emotional states we can learn to represent. We also note that our work is unimodal, but there is a proliferation of image, video, and audio data that carries much emotional meaning, often paired with text in a multimodal setting (such as social media). Our work furthermore focuses only on English data; however, the majority of people worldwide express their emotional experiences in a multitude of other languages, which we cannot assume simply align with English. We also acknowledge that while responsible and explainable AI is a focus of our work, there are myriad other ethical considerations involved in developing and deploying models for mental health in the real world beyond those we consider here, such as data privacy and the responsibilities of AI models that are deployed *without* human oversight.

The first key direction we envision for future work is to combine the domain knowledge of our emotion detection models with the power and capability of modern large models. In [Part I](#), we developed bespoke small-scale models for a particular type of emotional reaction, supported by psychological theory about the nature of distress. However, since this work was done, much larger models have been deployed and attain top performance across a range of NLP tasks; the prevailing trend is towards flexible models that can perform many tasks at once, often with little or no extra training. These LLMs typically perform well, but because they are so massive, their inner workings are largely opaque to human interpretation.

Our focus in [Part I](#) of the thesis was on developing explainable, responsible models, which we tackled by incorporating domain knowledge into small-scale models for a specific type of emotional reaction. Moving into [Part II](#) and particularly [Chapter 8](#), we developed broader models to detect and explain a variety of emotional states, which

in this work meant that they did not incorporate knowledge about particular theories. Because we believe it is paramount that AI be deployed alongside (not in place of) humans and that humans understand how AI makes its predictions (and when it may be mistaken), we propose a synthesis of these approaches. The size of LLMs and the number of tasks they are expected to perform generally make it infeasible to modify existing architectures to incorporate domain knowledge for any particular task. We do believe that emotional reactions in addition to distress and theories relating to them should be studied in detail as we did throughout [Part I](#), as computational models can uncover strong empirical evidence for or against theory. However, our architectures in [Chapter 4](#) simply applied multitask learning with a related task, leaving the underlying model’s form intact. Therefore, it should be feasible to incorporate domain knowledge to arbitrary tasks (including detection of myriad emotional reactions) by simply fine-tuning existing models in intelligent ways.¹

Furthermore, it is a mistake to trust the outputs of models without checking them, even though the generations of LLMs are typically convincing and high-quality. We believe that future models should be trained to produce evidence and thought processes for their predictions (which has also previously been shown to improve performance, e.g., [Chakrabarty et al. \(2022\)](#); [Wei et al. \(2022\)](#)). Future models trained to provide real citations alongside their predictions, perform multiple related structured tasks at once (such as with our explanations in [Chapter 8](#)), or provide information about what trends or gaps in their training data are relevant at prediction time may have great potential to become more grounded, trustworthy, and understandable. In the world of detecting emotional reactions, this may include things like detecting multiple semantic roles at once (cause, experiencer, etc.), giving the model’s best guess about the experiencer’s

¹We do also advocate for careful consideration of how future models are designed, including particular modules that can be readily interpreted. However, this is out of scope of our current research.

demographics² or about their personality, or providing the names of relevant concepts or explicitly discussing the emotional reaction in the context of a relevant theory.

The second major strand of future work we hope receives focus is increasing the variety of people whose experiences can be understood by computational models in an unbiased and inclusive way. This is a significant challenge because humans are not unbiased and not always inclusive, and so any data we generate and any research we perform is subject to our biases. People can express and perceive emotion differently in social settings depending on aspects like the ethnicity and gender of the person expressing and the person perceiving (e.g., [Kring and Gordon \(1998\)](#); [LaFrance et al. \(2003\)](#); [Soto and Levenson \(2009\)](#)).³ Many further pieces of demographic information can influence a person’s identity and how they experience the world, such as age, religion, nationality, sexual orientation, and so on, intersecting to create each individual’s unique experience. This means that the best judge of what a person is feeling will always be the person themselves. While it may not be possible for a single model (or a single person!) to fully understand every possible experience, in order to give our emotion models the best chance of *understanding* the experiences of different people, we believe it is vital they are *trained* on the experiences of different people.

One important first step in this direction will be to examine the current capabilities and biases of the models on which we rely for tasks like these. We call for studies on the accuracy of our models (particularly LLMs) on emotional tasks, additionally broken down by the demographics of the person whose experience is being analyzed. Here, again, we may find it enlightening to understand whether our models can correctly predict what kind of person wrote the text they are analyzing. Moving forward from simply benchmarking LLMs’ impressive performance, analyzing the emotional capabilities of

²In a demonstrative sense to make hidden biases explicit, rather than as a target task the model should strive to perform well.

³We are not aware of any research examining whether these visual effects persist in written text but hope this can be investigated.

LLMs is a new and developing field among many other facets of study and we hope that future work will keep diversity one of its primary goals in this area.

Of course, we also advocate for the development and collection of data that represents different people when we train our emotional models. While we may expect that the training data of LLMs contains much relevant data by virtue of their size, we both do not know the data composition of proprietary LLMs and cannot yet say the same about the smaller curated datasets we use for fine-tuning and evaluation. Gathering this data will likely be expensive and time-consuming, particularly because we believe we should also value participants' privacy when dealing with demographic information, but we hope that the community will agree that it is important.

In conclusion, in this thesis we have developed limited prototype models for both detecting and understanding emotional reactions in social media. In [Part I](#), we focused on developing models to detect a particular type of emotional reaction, psychological distress; and in [Part II](#), we broadened our scope to develop models that identify the causes of and generate explanations for a variety of emotional reactions. In our work, we have utilized both multitask learning and expert knowledge (including psychology research and commonsense knowledge databases) to design models with enhanced explainability and performance for our tasks. As the field of NLP moves forward with future work, we hope that we will develop models that 1) are underpinned by theory and expert knowledge and 2) represent the experiences of many different types of people from many different demographic groups.

References

- Emotion. In *APA Dictionary of Psychology*. American Psychological Association, 2023. URL <https://dictionary.apa.org/emotion>.
- Stress. In *APA Dictionary of Psychology*. American Psychological Association, 2023a. URL <https://dictionary.apa.org/stress>.
- Stressor. In *APA Dictionary of Psychology*. American Psychological Association, 2023b. URL <https://dictionary.apa.org/stressor>.
- Jiddy Abdillah, Ibnu Asror, and Yanuar Wibowo. Emotion classification of song lyrics using bidirectional LSTM method with GloVe word representation weighting. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4:723–729, 08 2020. doi: 10.29207/resti.v4i4.2156.
- 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. doi: 10.18653/v1/P17-1067. URL <https://www.aclweb.org/anthology/P17-1067>.
- Fares Al-Shargie, Masashi Kiguchi, Nasreen Badruddin, Sarat C. Dass, and Ahmad Fadzil Mohammad Hani. Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomedical Optics Express*, 7(10):3882–3898, October 2016. doi: 10.1364/BOE.7.003882.
- Andrew P. Allen, Paul J. Kennedy, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. Biological and psychological markers of stress in humans: Focus on the trier social stress test. *Neuroscience & Biobehavioral Reviews*, 38:94–124, January 2014. doi: 10.1016/j.neubiorev.2013.11.005.
- Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, page 579–586. Association for Computational Linguistics, 2005. doi: 10.3115/1220575.1220648. URL <https://doi.org/10.3115/1220575.1220648>.
- Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62:2937 – 2987, 2020. URL <https://api.semanticscholar.org/CorpusID:254143430>.
- Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y Gómez. Detecting depression in social media using fine-grained emotions. In Jill Burstein, Christy Doran, and Tamar Solorio, editors,

- 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 1481–1486, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1151. URL <https://aclanthology.org/N19-1151>.
- 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. URL <http://arxiv.org/abs/1409.0473>.
- Alexandra Balahur, Jesus M. Hermida, and Andres Montoyo. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1):88–101, 2012. doi: 10.1109/T-AFFC.2011.33.
- Kfir Bar, Vered Zilberstein, Ido Ziv, Heli Baram, Nachum Dershowitz, Samuel Itzikowitz, and Eiran Vadim Harel. Semantic characteristics of schizophrenic speech. In Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys, editors, *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 84–93, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3010. URL <https://aclanthology.org/W19-3010>.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset, 2020.
- Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW’11, page 17–37. JMLR.org, 2011.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc.", 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. doi: 10.18653/v1/p19-1470. URL <https://doi.org/10.18653/v1/p19-1470>.
- Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception.

- In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odiijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1554–1566. European Language Resources Association, 2020. URL <https://www.aclweb.org/anthology/2020.lrec-1.194/>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Sven Buechel and Udo Hahn. Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence, ECAI’16*, page 1114–1122, NLD, 2016. IOS Press. ISBN 9781614996712. doi: 10.3233/978-1-61499-672-9-1114. URL <https://doi.org/10.3233/978-1-61499-672-9-1114>.
- Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *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. URL <https://aclanthology.org/E17-2092>.
- Michel Cabanac. What is emotion? *Behavioural Processes*, 60(2):69–83, 2002. ISSN 0376-6357. doi: [https://doi.org/10.1016/S0376-6357\(02\)00078-5](https://doi.org/10.1016/S0376-6357(02)00078-5). URL <https://www.sciencedirect.com/science/article/pii/S0376635702000785>.
- Marilia A. Calcia, David R. Bonsall, Peter S. Bloomfield, Sudhakar Selvaraj, Tatiana Barichello, and Oliver D. Howes. Stress and neuroinflammation: a systematic review of the effects of stress on microglia and the implications for mental illness. *Psychopharmacology*, 233(9):1637–1650, May 2016. doi: 10.1007/s00213-016-4218-9.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In Diana McCarthy and Shuly Wintner, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL <https://aclanthology.org/E06-1032>.

- 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.
- Felix Casel, Amelie Heindl, and Roman Klinger. Emotion recognition under consideration of the emotion component process model. In Kilian Evang, Laura Kallmeyer, Rainer Osswald, Jakub Waszczuk, and Torsten Zesch, editors, *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany, 6–9 September 2021. KONVENS 2021 Organizers. URL <https://aclanthology.org/2021.konvens-1.5>.
- Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors*, 23:634, 01 2023. doi: 10.3390/s23020634.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.481. URL <https://aclanthology.org/2022.emnlp-main.481>.
- Tara M. Chaplin. Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, 7(1):14–21, 2015. doi: 10.1177/1754073914544408. URL <https://doi.org/10.1177/1754073914544408>. PMID: 26089983.
- François-Régis Chaumartin. UPAR7: A knowledge-based system for headline sentiment tagging. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 422–425, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/S07-1094>.
- Disha Chavan, Esha Anvekar, Megha Dandapat, Vaibhav Bichave, and Jayashree Jagdale. Machine learning applied in emotion classification: A survey on dataset, techniques, and trends for text based documents. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1697–1703, 2023. doi: 10.1109/ICACCS57279.2023.10112801.
- Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview, 2021.
- Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL <https://aclanthology.org/C10-1021>.

- 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. doi: 10.18653/v1/D18-1066. URL <https://www.aclweb.org/anthology/D18-1066>.
- Liyang Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.569. URL <https://aclanthology.org/2020.emnlp-main.569>.
- Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014. URL <http://aclweb.org/anthology/D/D14/D14-1179.pdf>.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013. URL <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351>.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson,

- Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.
- Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-1126>.
- Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 3rd edition, 2015.
- Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385–396, 1983. ISSN 00221465. URL <http://www.jstor.org/stable/2136404>.
- Sheldon Cohen, Ronald C. Kessler, and Lynn Underwood Gordon, editors. *Measuring stress: A guide for health and social scientists*. Oxford University Press, 1997. ISBN 1998-07054-000.
- Michael A. Cohn, Matthias R. Mehl, and James W. Pennebaker. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, 15: 687–93, October 2004. doi: 10.1111/j.0956-7976.2004.00741.x.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL <https://doi.org/10.1145/1390156.1390177>.
- Nicholas C Coombs, Wyatt E Meriwether, James Caringi, and Sophia R Newcomer. Barriers to healthcare access among u.s. adults with mental health challenges: A population-based study. *Social Science & Medicine - Population Health*, 15, September 2021.
- Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado, June 5 2015a. Association for Computational Linguistics. doi: 10.3115/v1/W15-1201. URL <https://aclanthology.org/W15-1201>.

- Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado, June 5 2015b. Association for Computational Linguistics. doi: 10.3115/v1/W15-1204. URL <https://aclanthology.org/W15-1204>.
- Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. Emotion recognition based on psychological components in guided narratives for emotion regulation. In Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.latechclfl-1.8. URL <https://aclanthology.org/2023.latechclfl-1.8>.
- Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. doi: 10.1073/pnas.1702247114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1702247114>.
- Alexandra D. Crosswell. Best practices for stress measurement: How to measure psychological stress in health research. *Health Psychology Open*, (2), 2020. doi: 10.1177/2055102920933072.
- Antonio R. Damasio. Emotions and feelings: A neurobiological perspective. In Antony S. R. Mainstead, Nico Frijda, and Agneta Fischer, editors, *Feelings and emotions: The Amsterdam symposium.*, pages 49–57. Cambridge University Press, 2004. doi: <https://doi.org/10.1017/CBO9780511806582.004>.
- Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1227. URL <https://aclanthology.org/D19-1227>.
- Munmun De Choudhury and Sushovan De. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80, May 2014. doi: 10.1609/icwsm.v8i1.14526. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14526>.
- Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! exploring human emotional states in social media. *Proceedings of the*

International AAAI Conference on Web and Social Media, 6(1):66–73, Aug. 2021. doi: 10.1609/icwsm.v6i1.14279. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14279>.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL <https://www.aclweb.org/anthology/2020.acl-main.372>.

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. URL <https://www.aclweb.org/anthology/N19-1423>.

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. doi: 10.18653/v1/N18-1174. URL <https://www.aclweb.org/anthology/N18-1174>.

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. doi: 10.18653/v1/2020.acl-main.288. URL <https://www.aclweb.org/anthology/2020.acl-main.288>.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback, 2023.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3561048. URL <https://doi.org/10.1145/3561048>.

- Paul Ekman. Are there basic emotions? *Psychological Review*, 99(5):550–553, 1992. doi: <https://doi.org/10.1037/0033-295X.99.3.550>.
- Paul Ekman. *Handbook of cognition and emotion*, chapter Basic emotions, pages 45–60. John Wiley & Sons Ltd., 1999. doi: [doi:10.1002/0470013494.ch3](https://doi.org/10.1002/0470013494.ch3).
- Paul Ekman and Richard J. Davidson, editors. *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.
- Gemma Espejo, Wade Reiner, and Michael Wenzinger. Exploring the role of artificial intelligence in mental healthcare: Progress, pitfalls, and promises. *Cureus*, Sep 2023. doi: [10.7759/cureus.44748](https://doi.org/10.7759/cureus.44748).
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi: [10.18653/v1/2022.naacl-main.187](https://doi.org/10.18653/v1/2022.naacl-main.187). URL <https://aclanthology.org/2022.naacl-main.187>.
- 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. doi: [10.18653/v1/2020.acl-main.342](https://doi.org/10.18653/v1/2020.acl-main.342). URL <https://www.aclweb.org/anthology/2020.acl-main.342>.
- Yingruo Fan, Jacqueline Lam, and Victor Li. Demographic effects on facial emotion expression: An interdisciplinary investigation of the facial action units of happiness. *Scientific Reports*, 11, 03 2021. doi: [10.1038/s41598-021-84632-9](https://doi.org/10.1038/s41598-021-84632-9).
- Zhangyin Feng, Yong Dai, Fan Zhang, Duyu Tang, Xiaocheng Feng, Shuangzhi Wu, Bing Qin, Yunbo Cao, and Shuming Shi. Skillnet-x: A multilingual multitask model with sparsely activated skills, 2023.
- Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.
- Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: [10.1037/h0031619](https://doi.org/10.1037/h0031619).
- Elaine Fox. Perspectives from affective science on understanding the nature of emotion. *Brain and Neuroscience Advances*, 2:2398212818812628, 2018. doi: [10.1177/2398212818812628](https://doi.org/10.1177/2398212818812628). URL <https://doi.org/10.1177/2398212818812628>. PMID: 32166161.

- Virginia Francisco and Pablo Gervás. Automated mark up of affective information in english texts. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, pages 375–382, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-39091-6.
- David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.
- Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *CICLing*, 2015.
- Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.224. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.224>.
- 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*, 13(1):440–460, 2019.
- Reshmi Gopalakrishna Pillai, Mike Thelwall, and Constantin Orasan. What makes you stressed? finding reasons from tweets. In Alexandra Balahur, Saif M. Mohammad, Veronique Hoste, and Roman Klinger, editors, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 266–272, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6239. URL <https://aclanthology.org/W18-6239>.
- 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. ISSN 0883-6612. doi: 10.1007/s12160-009-9081-z. URL <https://doi.org/10.1007/s12160-009-9081-z>.
- Paul E Griffiths. *What emotions really are: The problem of psychological categories*. University of Chicago Press, 2008.
- Dritjon Gruda and Souleiman Hasan. Feeling anxious? perceiving anxiety in tweets using machine learning. *CoRR*, abs/1909.06959, 2019. URL <http://arxiv.org/abs/1909.06959>.
- Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American*

- Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL <https://aclanthology.org/N18-1065>.
- Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms, 2023.
- 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. doi: 10.18653/v1/D16-1170. URL <https://www.aclweb.org/anthology/D16-1170>.
- 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. URL <http://arxiv.org/abs/1811.07430>.
- Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Dynamic multi-level multi-task learning for sentence simplification. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA, August 2018a. Association for Computational Linguistics. URL <https://aclanthology.org/C18-1039>.
- Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Soft layer-specific multi-task summarization with entailment and question generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia, July 2018b. Association for Computational Linguistics. doi: 10.18653/v1/P18-1064. URL <https://aclanthology.org/P18-1064>.
- Tianyou Guo, Zhihao Zhang, Alyx Taylor, Daniel L. Hall, Albert S. Yeung, Arthur F. Kramer, and Liye Zou. Association of social support with negative emotions among Chinese adolescents during Omicron-related lockdown of Shenzhen City: The roles of rumination and sleep quality. *Frontiers in Psychiatry*, 13, 2022. ISSN 1664-0640. doi: 10.3389/fpsy.2022.957382. URL <https://www.frontiersin.org/articles/10.3389/fpsy.2022.957382>.
- Luis Guzman-Nateras, Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. Event detection for suicide understanding. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1952–1961, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.150. URL <https://aclanthology.org/2022.findings-naacl.150>.

- M D Haque and Sabirat Rubya. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth and uHealth*, 11, May 2023. doi: 10.2196/44838.
- Melissa L. Harris, Christopher Oldmeadow, Alexis Hure, Judy Luu, Deborah Loxton, and John Attia. Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling. *PLoS One*, 12(2), February 2017.
- Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1048. URL <https://aclanthology.org/P19-1048>.
- J.A. Healey and R.W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005. doi: 10.1109/TITS.2005.848368.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. Appraisal theories for emotion classification in text. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.11. URL <https://aclanthology.org/2020.coling-main.11>.
- Thomas H. Holmes and Richard H. Rahe. The social readjustment rating scale. *Journal of Psychosomatic Research*, 1967.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Larousilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751, 2019. URL <http://arxiv.org/abs/1902.00751>.
- Ahmed Hussein Orabi, Prasadith Buddhitha, Mahmoud Hussein Orabi, and Diana Inkpen. Deep learning for depression detection of Twitter users. In Kate Loveys, Kate Niederhoffer, Emily Prud’hommeaux, Rebecca Resnik, and Philip Resnik, editors, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0609. URL <https://aclanthology.org/W18-0609>.

- Daniel Izmaylov, Avi Segal, Kobi Gal, Meytal Grimland, and Yossi Levi-Belz. Combining psychological theory with language models for suicide risk detection. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2430–2438, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.184. URL <https://aclanthology.org/2023.findings-eacl.184>.
- Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. MuSE: a multimodal dataset of stressed emotion. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1499–1510, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://www.aclweb.org/anthology/2020.lrec-1.187>.
- William James. *The Principles of Psychology*, volume 2. Cosimo, Inc., April 2007.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. Mistral 7b, 2023.
- Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online, November 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.louhi-1.16>.
- Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. Does multi-task learning always help?: An evaluation on health informatics. In Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay, editors, *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 151–158, Sydney, Australia, 4–6 December 2019. Australasian Language Technology Association. URL <https://aclanthology.org/U19-1020>.
- Dan Joyce, Andrey Kormilitzin, Katharine Smith, and Andrea Cipriani. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6, 01 2023. doi: 10.1038/s41746-023-00751-9.
- Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014. doi: 10.1177/0261927X13502654. URL <https://doi.org/10.1177/0261927X13502654>.

- Jerome Kagan. *What is Emotion?: History, Measures, and Meanings*. Yale University Press, 2007.
- Sarah Khan and Rafeeq Alam Khan. Chronic stress leads to anxiety and depression. 2017. URL <https://api.semanticscholar.org/CorpusID:8632459>.
- Hamed Khanpour and Cornelia Caragea. Fine-grained emotion detection in health-related online posts. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1147. URL <https://aclanthology.org/D18-1147>.
- 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. URL <https://www.aclweb.org/anthology/C18-1114>.
- Evgeny Kim, Sebastian Padó, and Roman Klinger. Investigating the relationship between literary genres and emotional plot development. In Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2203. URL <https://aclanthology.org/W17-2203>.
- Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014. URL <http://arxiv.org/abs/1408.5882>.
- 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. doi: 10.21236/ada006655.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- Paul R. Kleinginna and Anne M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, 1981. doi: 10.1007/BF00992553.
- Ann M Kring and Albert H Gordon. Sex differences in emotion: expression, experience, and physiology. *Journal of personality and social psychology*, 74(3):686, 1998.

- 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.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL <https://aclanthology.org/2022.tacl-1.10>.
- Marianne LaFrance, Marvin A Hecht, and Elizabeth Levy Paluck. The contingent smile: a meta-analysis of sex differences in smiling. *Psychological bulletin*, 129(2): 305, 2003.
- Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies, 2021.
- J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977. doi: 10.2307/2529310.
- Richard S. Lazarus. *Stress and emotion: A new synthesis*. Springer Publishing Company, 1st edition, 2006.
- Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.
- Duong Le, My Thai, and Thien Nguyen. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8139–8146, Apr. 2020. doi: 10.1609/aaai.v34i05.6326. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6326>.
- Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVlyder, Michel Walter, Sofian Berrouguet, and et al. Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5), May 2021. doi: 10.2196/15708.
- Robert W. Levenson. Stress and illness: A role for specific emotions. *Psychosomatic Medicine*, 81(8):720–730, October 2019. doi: 10.1097/PSY.0000000000000736.

- Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore, 2023.
- Chuyuan Li, Chloé Braud, and Maxime Amblard. Multi-task learning for depression detection in dialogs. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 68–75, Edinburgh, UK, September 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.sigdial-1.7>.
- 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. doi: 10.18653/v1/P16-1137. URL <https://www.aclweb.org/anthology/P16-1137>.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- 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, September 2017. ISSN 1041-4347. doi: 10.1109/TKDE.2017.2686382.
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Deep multi-task learning with shared memory for text classification. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 118–127, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1012. URL <https://aclanthology.org/D16-1012>.

- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL <https://aclanthology.org/P19-1441>.
- Sonia J. Lupien, Bruce S. McEwen, Megan R. Gunnar, and Christine Heim. Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature Reviews Neuroscience*, 10(6):434–445, June 2009. doi: 10.1038/nrn2639.
- Minna Lyons, Nazli Deniz Aksayli, and Gayle Brewer. Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87:207–211, 2018. ISSN 0747-5632. doi: <https://doi.org/10.1016/j.chb.2018.05.035>. URL <https://www.sciencedirect.com/science/article/pii/S0747563218302619>.
- Shirong Ma, Yinghui Li, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. Progressive multi-task learning framework for chinese text error correction, 2023.
- Anton Malko, Cecile Paris, Andreas Duenser, Maria Kangas, Diego Molla, Ross Sparks, and Stephen Wan. Demonstrating the reliability of self-annotated emotion data. In Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik, editors, *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 45–54, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.clpsych-1.5. URL <https://aclanthology.org/2021.clpsych-1.5>.
- Anton Malko, Andreas Duenser, Maria Kangas, Diego Mollá-Aliod, and Cecile Paris. Message similarity as a proxy to repetitive thinking: Associations with non-suicidal self-injury and suicidal ideation on social media. *Computers in Human Behavior Reports*, 11:100320, 2023. ISSN 2451-9588. doi: <https://doi.org/10.1016/j.chbr.2023.100320>. URL <https://www.sciencedirect.com/science/article/pii/S2451958823000532>.
- John J McGrath, Ali Al-Hamzawi, Jordi Alonso, Yasmin Altwajjri, Laura H Andrade, Evelyn J Bromet, Ronny Bruffaerts, José Miguel Caldas de Almeida, Stephanie Chardoul, Wai Tat Chiu, Louisa Degenhardt, Olga V Demler, Finola Ferry, Oye Gureje, Josep Maria Haro, Elie G Karam, Georges Karam, Salma M Khaled, Viviane Kovess-Masfety, Marta Magno, Maria Elena Medina-Mora, Jacek Moskalewicz, Fernando Navarro-Mateu, Daisuke Nishi, Oleguer Plana-Ripoll, José Posada-Villa, Charlene Rapsey, Nancy A Sampson, Juan Carlos Stagnaro, Dan J Stein, Margreet ten Have, Yolanda Torres, Cristian Vladescu, Peter W Woodruff, Zahari Zarkov, Ronald C Kessler, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso,

- Yasmin A. Altwaijri, Laura Helena Andrade, Lukoye Atwoli, Corina Benjet, Evelyn J. Bromet, Ronny Bruffaerts, Brendan Bunting, José Miguel Caldas de Almeida, Graça Cardoso, Stephanie Chardoul, Alfredo H. Cía, Louisa Degenhardt, Giovanni De Girolamo, Oye Gureje, Josep Maria Haro, Meredith G. Harris, Hristo Hinkov, Chi yi Hu, Peter De Jonge, Aimee N. Karam, Elie G. Karam, Georges Karam, Alan E. Kazdin, Norito Kawakami, Ronald C. Kessler, Andrzej Kiejna, Viviane Kovess-Masfety, John J. McGrath, Maria Elena Medina-Mora, Jacek Moskalewicz, Fernando Navarro-Mateu, Daisuke Nishi, Marina Piazza, José Posada-Villa, Kate M. Scott, Juan Carlos Stagnaro, Dan J. Stein, Margreet Ten Have, Yolanda Torres, Maria Carmen Viana, Daniel V. Vigo, Cristian Vladescu, David R. Williams, Peter Woodruff, Bogdan Wojtyniak, Miguel Xavier, and Alan M. Zaslavsky. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, 10(9):668–681, 2023. ISSN 2215-0366. doi: [https://doi.org/10.1016/S2215-0366\(23\)00193-1](https://doi.org/10.1016/S2215-0366(23)00193-1). URL <https://www.sciencedirect.com/science/article/pii/S2215036623001931>.
- Albert Mehrabian. *Basic dimensions for a general psychological theory : implications for personality, social, environmental, and developmental studies*. Cambridge: Oelgeschlager, Gunn & Hain, 1980.
- Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. The cultural construction of emotions. *Current Opinion in Psychology*, 8:31–36, 2016. ISSN 2352-250X. doi: <https://doi.org/10.1016/j.copsyc.2015.09.015>. URL <https://www.sciencedirect.com/science/article/pii/S2352250X15002419>. Culture.
- 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. Curran Associates Inc., 2013. URL <http://dl.acm.org/citation.cfm?id=2999792.2999959>.
- 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.
- Saif Mohammad and Felipe Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot, editors, *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. doi: 10.18653/v1/W17-5205. URL <https://aclanthology.org/W17-5205>.
- Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, 2014.

- 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. doi: 10.1016/j.dss.2012.05.030. URL <https://doi.org/10.1016/j.dss.2012.05.030>.
- Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.
- Saif M. Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *CoRR*, abs/2005.11882, 2020. URL <https://arxiv.org/abs/2005.11882>.
- Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.
- Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.
- Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. explosion/spaCy: v3.7.2: Fixes for APIs and requirements, October 2023. URL <https://doi.org/10.5281/zenodo.10009823>.
- Kevin Mulligan and Klaus R. Scherer. Toward a working definition of emotion. *Emotion Review*, 4(4):345–357, 2012. doi: 10.1177/1754073912445818. URL <https://doi.org/10.1177/1754073912445818>.
- Cataldo Musto, Gaetano Rossiello, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Combining text summarization and aspect-based sentiment analysis of users’ reviews to justify recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys ’19*, page 383–387, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347024. URL <https://doi.org/10.1145/3298689.3347024>.
- Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), August 2021. doi: 10.1007/s13278-021-00776-6.
- Susan Nolen-Hoeksema, Blair E. Wisco, and Sonja Lyubomirsky. Rethinking rumination. *Perspectives on Psychological Science*, 3:400–424, September 2008. doi: <https://doi.org/10.1111/j.1745-6924.2008.00088.x>.
- Laura Ana Maria Oberländer and Roman Klinger. Token sequence labeling vs. clause classification for English emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70, December 2020.

- Laura Ana Maria Oberländer, Kevin Reich, and Roman Klinger. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 119–128, 2020.
- OpenAI. Gpt-4 technical report, 2023.
- Ariana Orvell, Susan A. Gelman, and Ethan Kross. What “you” and “we” say about me: How small shifts in language reveal and empower fundamental shifts in perspective. *Social and Personality Psychology Compass*, 16(5):e12665, 2022. doi: <https://doi.org/10.1111/spc3.12665>. URL <https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/spc3.12665>.
- Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. Exploring the role of task transferability in large-scale multi-task learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.183. URL <https://aclanthology.org/2022.naacl-main.183>.
- Rrubaa Panchendrarajan, Wynne Hsu, and Mong Li Lee. Emotion-aware event summarization in microblogs. In *Companion Proceedings of the Web Conference 2021, WWW ’21*, page 486–494, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3452311. URL <https://doi.org/10.1145/3442442.3452311>.
- Galen Thomas Panger. *Emotion in Social Media*. PhD thesis, University of California, Berkeley, 2017.
- W. Gerrod Parrott, editor. *Emotions in Social Psychology: Key Readings in Social Psychology*. Psychology Press, 2001.
- Antonio Pascual-Leone, Phoenix Gillis, Terence Singh, and Cristina A. Andreescu. Problem anger in psychotherapy: An emotion-focused perspective on hate, rage, and rejecting anger. *Journal of Contemporary Psychotherapy*, 43(2), 2013. doi: 10.1007/s10879-012-9214-8. URL <https://doi.org/10.1007/s10879-012-9214-8>.
- Antonio Pascual-Leone, Sabine C. Herpertz, and Ueli Kramer. Experimental Designs and the ‘Emotion Stimulus Critique’: Hidden Problems and Potential Solutions in the Study of Emotion. *Psychopathology*, 49(1):60–68, 12 2015. ISSN 0254-4962. doi: 10.1159/000442294. URL <https://doi.org/10.1159/000442294>.

- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W. Schuller. Customising general large language models for specialised emotion recognition tasks, 2023.
- James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. 2015. URL https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf.
- Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088, 2018. URL <http://arxiv.org/abs/1811.01088>.
- Jonathan Pilault, Amine El hattami, and Christopher Pal. Conditionally adaptive multi-task learning: Improving transfer learning in {nlp} using fewer parameters & less data. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=de11dbHzAMF>.
- Robert Plutchik. Chapter 1 - A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: <https://doi.org/10.1016/B978-0-12-558701-3.50007-7>. URL <https://www.sciencedirect.com/science/article/pii/B9780125587013500077>.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2004. URL <https://www.aclweb.org/anthology/S14-2004>.
- Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations, 2020.
- 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*, 14(1):108–132, 2023. doi: 10.1109/TAFFC.2020.3038167.

- Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. Reinforcement guided multi-task learning framework for low-resource stereotype detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6703–6712, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.462. URL <https://aclanthology.org/2022.acl-long.462>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL <https://api.semanticscholar.org/CorpusID:160025533>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL <http://arxiv.org/abs/1910.10683>.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL <https://aclanthology.org/P18-2124>.
- Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040, 1995. URL <http://arxiv.org/abs/cmp-lg/9505040>.
- 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. doi: 10.1145/2939672.2939778. URL <https://doi.org/10.1145/2939672.2939778>.
- David L. Robinson. Brain function, emotional experience and personality. *Netherlands Journal of Psychology*, 64(4):152–168, 2008. doi: 10.1007/BF03076418.
- Alan Rozanski, James Blumenthal, and Jay Kaplan. Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. *Circulation*, 99(16):2192–2217, 1999.
- James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39, December 1980. ISSN 1939-1315. doi: <https://doi.org/10.1037/h0077714>.

- 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. URL <https://www.aclweb.org/anthology/W11-1720>.
- Jun Saito, Yugo Murawaki, and Sadao Kurohashi. Minimally supervised learning of affective events using discourse relations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5758–5765, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1581. URL <https://aclanthology.org/D19-1581>.
- Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. Findings of the shared task on detecting signs of depression from social media. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.51. URL <https://aclanthology.org/2022.ltedi-1.51>.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multi-task prompted training enables zero-shot task generalization. *CoRR*, abs/2110.08207, 2021. URL <https://arxiv.org/abs/2110.08207>.
- 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. URL <http://arxiv.org/abs/1811.00146>.
- Anand Satiani, Julie Niedermier, Bhagwan Satiani, and Dale P Svendsen. Projected workforce of psychiatrists in the united states: A population analysis. *Psychiatric Services*, 69, June 2018.

- Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69(5), 1962.
- Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. doi: 10.1177/0539018405058216. URL <https://doi.org/10.1177/0539018405058216>.
- Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001. ISBN 0-19-513007-3.
- Hans Selye. I - History and general outline of the stress concept. In *Stress in Health and Disease*, pages 3–34. Butterworth-Heinemann, 1976a. ISBN 978-0-407-98510-0. doi: <https://doi.org/10.1016/B978-0-407-98510-0.50009-4>. URL <https://www.sciencedirect.com/science/article/pii/B9780407985100500094>.
- Hans Selye. *Stress without Distress*, pages 137–146. Springer US, Boston, MA, 1976b. ISBN 978-1-4684-2238-2. doi: 10.1007/978-1-4684-2238-2_9. URL https://doi.org/10.1007/978-1-4684-2238-2_9.
- Hans Selye. *The Stress of life, rev.* McGraw Hill, 1978.
- R.J. Senter and E.A. Smith. Automated readability index. November 1967. URL <https://apps.dtic.mil/dtic/tr/fulltext/u2/667273.pdf>.
- Phillip R. Shaver, Judith C. Schwartz, Donald Kirson, and Catherine O'Connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52 6:1061–86, 1987. URL <https://api.semanticscholar.org/CorpusID:38465861>.
- Judy Hanwen Shen and Frank Rudzicz. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3107. URL <https://www.aclweb.org/anthology/W17-3107>.
- Valentina Sintsova, Margarita Bolívar Jiménez, and Pearl Pu. Modeling the impact of modifiers on emotional statements. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 71–89, Cham, 2018. Springer International Publishing. ISBN 978-3-319-77116-8.
- George M Slavich, Sara Taylor, and Rosalind W Picard. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22(4):408–413, July 2019.
- Tom De Smedt and Walter Daelemans. Pattern for Python. *Journal of Machine Learning Research*, 13:2063–2067, 2012.

- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642, 2013. URL <https://aclanthology.info/papers/D13-1170/d13-1170>.
- Tiberiu Sosea, Hongli Zhan, Junyi Jessy Li, and Cornelia Caragea. Unsupervised extractive summarization of emotion triggers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9550–9569, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.531. URL <https://aclanthology.org/2023.acl-long.531>.
- Jose Angel Soto and Robert W Levenson. Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. *Emotion*, 9(6):874, 2009.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova,

Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muenighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert

Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

Steffen Steinert and Matthew James Dennis. Emotions and digital well-being: On social media’s emotional affordances. *Philosophy & Technology*, 35(2), April 2022. doi: 10.1007/s13347-022-00530-6.

Jeremy G. Stewart, Grant S. Shields, Erika C. Esposito, Elizabeth A. Cosby, Nicholas B. Allen, George M. Slavich, and Randy P. Auerbach. Life stress and suicide in adolescents. *Journal of Abnormal Child Psychology*, 47(10):1707–1722, October 2019.

Asa Cooper Stickland and Iain Murray. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/stickland19a.html>.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. APPReddit: a corpus of Reddit posts annotated for

- appraisal. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.406>.
- Carlo Strapparava and Rada Mihalcea. Annotating and identifying emotions in text. In Giuliano Armano, Marco de Gemmis, Giovanni Semeraro, and Eloisa Vargiu, editors, *Intelligent Information Access*, volume 301, pages 21–38. Springer Berlin, Heidelberg, 2010. doi: 10.1007/978-3-642-14000-6_2. URL https://doi.org/10.1007/978-3-642-14000-6_2.
- Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1035. URL <https://www.aclweb.org/anthology/N19-1035>.
- Kyle Swanson, Joy Hsu, and Mirac Suzgun. Monte Carlo tree search for interpreting stress in natural language. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 107–119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.ltedi-1.12. URL <https://aclanthology.org/2022.ltedi-1.12>.
- Hanna A. Thoern, Marcus Grueschow, Ulrike Ehlert, Christian C. Ruff, and Brigit Kleim. Attentional bias towards positive emotion predicts stress resilience. *PLoS ONE*, 11(3), March 2016. doi: 10.1371/journal.pone.0148368.
- Sachin Thukral, Suyash Sangwan, Arnab Chatterjee, and Lipika Dey. Identifying pandemic-related stress factors from social-media posts – Effects on students and young-adults. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpCOVID19-2.23. URL <https://aclanthology.org/2020.nlpCOVID19-2.23>.
- Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL <https://www.aclweb.org/anthology/W03-0419>.

- Roberto Tornero-Costa, Antonio Martinez-Millana, Natasha Azzopardi-Muscat, Ledia Lazeri, Vicente Traver, and David Novillo-Ortiz. Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review. *JMIR Mental Health*, 10, Feb 2023. doi: 10.2196/42045.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- Enrica Troiano, Laura Oberländer, and Roman Klinger. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72, March 2023. doi: 10.1162/coli_a_00461. URL <https://aclanthology.org/2023.cl-1.1>.
- Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clpsych-1.16. URL <https://aclanthology.org/2022.clpsych-1.16>.
- Guoyun Tu, Yanwei Fu, Boyang Li, Jiarui Gao, Yu-Gang Jiang, and Xiangyang Xue. A multi-task neural approach for emotion attribution, classification, and summarization. *IEEE Transactions on Multimedia*, 22(1):148–159, 2020. doi: 10.1109/TMM.2019.2922129.
- Elsbeth Turcan and Kathy McKeown. Dreddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6213. URL <https://aclanthology.org/D19-6213>.
- Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.230. URL <https://aclanthology.org/2021.naacl-main.230>.
- Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3975–3989, Online, August 2021b.

Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.348. URL <https://aclanthology.org/2021.findings-acl.348>.

Debra Umberson and Jennifer Montez. Social relationships and health: A flashpoint for health policy. *Journal of health and social behavior*, 51 Suppl:S54–66, 10 2010. doi: 10.1177/0022146510383501.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Let the model decide its curriculum for multitask learning. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta, editors, *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 117–125, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.13. URL <https://aclanthology.org/2022.deeplo-1.13>.

Milton L. Wainberg, Pamela Scorza, James M. Shultz, Liat Helpman, Jennifer J. Mootz, Karen A. Johnson, Yuval Neria, Jean-Marie E. Bradford, Maria A. Oquendo, and Melissa R. Arbuckle. Challenges and opportunities in global mental health: A research-to-practice perspective. *Current Psychiatry Reports*, 19, April 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://aclanthology.org/W18-5446>.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. SuperNaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL <https://aclanthology.org/2022.emnlp-main.340>.

- Yufei Wang, Stephen Wan, and Cécile Paris. The role of features and context on suicide ideation detection. In Trevor Cohn, editor, *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 94–102, Melbourne, Australia, December 2016. URL <https://aclanthology.org/U16-1010>.
- Mayur Wankhade. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–3780, 2022. doi: 10.1007/s10462-022-10144-1. URL <https://doi.org/10.1007/s10462-022-10144-1>.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL <https://arxiv.org/abs/2109.01652>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL <https://arxiv.org/abs/2201.11903>.
- 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. doi: 10.18653/v1/2020.acl-main.289. URL <https://www.aclweb.org/anthology/2020.acl-main.289>.
- Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In Graciela Gonzalez-Hernandez and Davy Weissenbacher, editors, *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.smm4h-1.54>.
- 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, October 2009. doi: 10.2466/PR0.105.2.509-521.
- John Wieting and Kevin Gimpel. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1042. URL <https://aclanthology.org/P18-1042>.

- 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. URL <http://arxiv.org/abs/1805.12307>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.
- Muk Yan Wong. The mood-emotion loop. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 173(11):3061–3080, 2016. ISSN 00318116, 15730883. URL <http://www.jstor.org/stable/44122239>.
- The World Health Organization. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, 2022.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL <http://arxiv.org/abs/1609.08144>.
- Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. *CoRR*, abs/1906.01267, 2019. URL <http://arxiv.org/abs/1906.01267>.
- Rui Xia, Mengran Zhang, and Zixiang Ding. RTHN: A RNN-transformer hierarchical network for emotion cause extraction. *CoRR*, abs/1906.01236, 2019. URL <http://arxiv.org/abs/1906.01236>.
- Hu Xu, Bing Liu, Lei Shu, and Philip Yu. DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1725–1731, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.156. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.156>.
- Hu Xu, Lei Shu, Philip Yu, and Bing Liu. Understanding pre-trained BERT for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/

- 2020.coling-main.21. URL <https://www.aclweb.org/anthology/2020.coling-main.21>.
- Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. Emo2Vec: Learning generalized emotion representation by multi-task training. In Alexandra Balahur, Saif M. Mohammad, Véronique Hoste, and Roman Klinger, editors, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 292–298. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-6243. URL <https://doi.org/10.18653/v1/w18-6243>.
- Chenghao Yang, Yudong Zhang, and Smaranda Muresan. Weakly-supervised methods for suicide risk assessment: Role of related domains. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1049–1057, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.133. URL <https://aclanthology.org/2021.acl-short.133>.
- Zixiaofan Yang and Julia Hirschberg. Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks. In *Proc. Interspeech 2018*, pages 3092–3096, 2018. doi: 10.21437/Interspeech.2018-2397.
- Yiqun Yao, Michalis Papakostas, Mihai Burzo, Mohamed Abouelenien, and Rada Mihalcea. MUSER: MULTImodal stress detection using emotion recognition as an auxiliary task. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2714–2725, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.216. URL <https://aclanthology.org/2021.naacl-main.216>.
- Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1137. URL <https://aclanthology.org/D18-1137>.
- Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3568–3573, Online, November 2020. Association for Computational Linguistics. doi:

10.18653/v1/2020.emnlp-main.289. URL <https://www.aclweb.org/anthology/2020.emnlp-main.289>.

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

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. Why do you feel this way? summarizing triggers of emotions in social media posts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.642. URL <https://aclanthology.org/2022.emnlp-main.642>.

Honglun Zhang, Liqiang Xiao, Yongkun Wang, and Yaohui Jin. A generalized recurrent neural architecture for text classification with multi-task learning. *CoRR*, abs/1707.02892, 2017. URL <http://arxiv.org/abs/1707.02892>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=SkeHuCVFDr>.

Wenzheng Zhang, Chenyan Xiong, Karl Stratos, and Arnold Overwijk. Improving Multitask Retrieval by Promoting Task Specialization. *Transactions of the Association for Computational Linguistics*, 11:1201–1212, 09 2023a. ISSN 2307-387X. doi: 10.1162/tacl_a_00597. URL https://doi.org/10.1162/tacl_a_00597.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance, 2023b.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren’s song in the ai ocean: A survey on hallucination in large language models, 2023c.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia, May 2023d. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.66. URL <https://aclanthology.org/2023.eacl-main.66>.

- Yongxin Zhou, Fabien Ringeval, and François Portet. Evaluating emotional nuances in dialogue summarization, 2023.
- Yuan Zhuang and Ellen Riloff. Eliciting affective events from language models by multiple view co-prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3189–3201, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.199. URL <https://aclanthology.org/2023.findings-acl.199>.
- 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. doi: 10.18653/v1/2020.emnlp-main.452. URL <https://www.aclweb.org/anthology/2020.emnlp-main.452>.
- Johannes Zimmermann, Markus Wolf, Astrid Bock, Doris Peham, and Cord Benecke. The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of Research in Personality*, 47(3):218–225, 2013. ISSN 0092-6566. doi: <https://doi.org/10.1016/j.jrp.2013.01.008>. URL <https://www.sciencedirect.com/science/article/pii/S0092656613000160>.
- Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys, editors, *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3003. URL <https://aclanthology.org/W19-3003>.
- Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. Linguistic analysis of schizophrenia in Reddit posts. In Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys, editors, *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3009. URL <https://aclanthology.org/W19-3009>.
- 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). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/594_Paper.pdf.

Appendix A: Emotion-Cause Joint Extraction Extended

Examples

We include the output of all models for our four selected examples in §7.4.1 in Table A.1, Table A.2, Table A.3, and Table A.4.

Mexico reels from shooting attack in El Paso

fear

Model	Output
BERT	negative surprise
BERT ^{COMET}	fear
Multi	negative surprise
Multi _{C→E}	negative surprise
Multi _{E→C}	negative surprise
Multi ^{COMET}	fear
Multi ^{COMET} _{C→E}	fear
Multi ^{COMET} _{E→C}	fear

Table A.1: **All Outputs for Emotion-Cause Example 1.** Full model outputs for our first provided example from Chapter 7.

Insane video shows Viking Sky cruise ship thrown into chaos at sea

fear

Model	Output
BERT	negative surprise
BERT ^{COMET}	negative surprise
Multi	fear
Multi _{C→E}	negative surprise
Multi _{E→C}	fear
Multi ^{COMET}	fear
Multi ^{COMET} _{C→E}	fear
Multi ^{COMET} _{E→C}	negative surprise

Table A.2: **All Outputs for Emotion-Cause Example 2.** Full model outputs for our second provided example from Chapter 7.

Durant **could return for Game 3**

positive surprise

Model	Output
BERT	for game
BERT ^{COMET}	could return for game
Multi	could return for game
Multi _{C→E}	could return for game
Multi _{E→C}	could return for game
Multi ^{COMET}	could return for game
Multi ^{COMET} _{C→E}	could return for game
Multi ^{COMET} _{E→C}	could return for game

Table A.3: **All Outputs for Emotion-Cause Example 3.** Full model outputs for our third provided example from [Chapter 7](#).

Dan Fagan: **Triple shooting near New Orleans School yet another sign of city’s crime problem**

negative surprise

Model	Output
BERT	school yet another sign of city’s crime
BERT ^{COMET}	: triple shooting near new orleans school yet another sign of city’s crime
Multi	shooting near new orleans school yet another sign of city’s crime
Multi _{C→E}	: triple shooting near new orleans school yet another sign of city’s crime
Multi _{E→C}	: triple shooting near new orleans school yet another sign of city’s crime
Multi ^{COMET}	: triple shooting near new orleans school yet another sign of city’s crime
Multi ^{COMET} _{C→E}	: triple shooting near new orleans school yet another sign of city’s crime
Multi ^{COMET} _{E→C}	: triple shooting near new orleans school yet another sign of city’s crime

Table A.4: **All Outputs for Emotion-Cause Example 4.** Full model outputs for our fourth provided example from [Chapter 7](#).

Appendix B: Emotional Explanation Prompts

Task	Instructions or Prompt
Emotion Identification (Ours)	What emotions is this person feeling?
Event Extraction (Ours)	What event has triggered this person’s emotions?
Explanation Generation (Ours)	Explain the feelings this person is experiencing.
BIG-bench Metaphor Boolean	Is the second sentence the correct interpretation of the first?
BIG-bench Understanding Metaphor	What is the literal interpretation of this metaphor?
BIG-bench Understanding Fables	What is the moral of this fable?
BIG-bench Rephrase	Rephrase the given sentence so that it retains its meaning, but contains the given keyword.
BIG-bench Suicide Risk	What level of suicide risk can we identify from this text?
FLUTE	Does this premise entail or contradict this hypothesis? Why?
GoEmotions	What emotions is this comment expressing?
Natural Questions	Read the following document and answer a question about it in a complete sentence.
Para-NMT	On a scale of 0-1, about how similar are these two sentences?
Quora Question Pairs	Are these two questions asking the same thing?
SQuAD	Read the following document and answer a question about it in a complete sentence. If the question cannot be answered from the text, explain this and offer a plausible guess.

Table B.1: **Task Prompts.** Individual instructions or prompts given to the Flan-T5_{split}+dataset models in [Chapter 8](#).

A full list of prompts and instructions for each of our tasks, as used in the Flan-T5_{split}+dataset series of models in [Chapter 8](#), is shown in [Table B.1](#). Note that classification and regression tasks include classification or regression heads rather than the language modeling head, but we still pass the model instructions so as to match the rest of its training.

Appendix C: Full Individual Model Results for Explanation Generation Models

The full results of our models from [Chapter 8](#) are shown in [Figure C.1](#), [Figure C.2](#), and [Figure C.3](#). These tables include models trained with each one of the individual tasks that compose the group models, and we do find that in many cases their performance matches or exceeds the performance of the grouped models.

Model	Extractiveness		Reference-Based				Factuality	
	Overlap	Density	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SummaC	QAFactEval
Flan-T5 GPT	0.64	3.32	0.46	0.18	0.30	0.39	0.27	1.50
Flan-T5 Split	0.47	2.75	0.46	0.17	0.29	0.39	0.25	0.96
Flan-T5 Split + BB-Metaphor-Boolean	0.42	2.82	0.44	0.16	0.29	0.35	0.25	0.83
Flan-T5 Split + BB-Understanding-Metaphor	0.50	3.17	0.43	0.16	0.29	0.34	0.26	1.17
Flan-T5 Split + BB-Rephrase	0.48	3.07	0.43	0.16	0.28	0.34	0.26	0.83
Flan-T5 Split + BB-Suicide-Risk	0.55	3.22	0.42	0.16	0.28	0.32	0.27	1.01
Flan-T5 Split + BB-Fables	0.48	3.11	0.45	0.17	0.30	0.37	0.27	1.03
Flan-T5 Split + FLUTE	0.55	3.22	0.43	0.16	0.28	0.33	0.27	1.20
Flan-T5 Split + GoEmotions	0.51	3.11	0.45	0.17	0.29	0.35	0.26	0.93
Flan-T5 Split + Para-NMT	0.52	3.10	0.45	0.18	0.30	0.36	0.26	1.02
Flan-T5 Split + Quora-Question-Pairs	0.45	3.01	0.44	0.16	0.29	0.35	0.25	0.89
Flan-T5 Split + Natural-Questions	0.51	2.87	0.43	0.17	0.29	0.35	0.26	1.04
Flan-T5 Split + SQuAD	0.51	3.00	0.44	0.16	0.29	0.35	0.25	0.92
Flan-T5 Split + QA	0.53	3.34	0.47	0.19	0.31	0.39	0.26	1.30
Flan-T5 Split + Metaphor	0.46	2.85	0.46	0.17	0.30	0.36	0.25	0.89
Flan-T5 Split + Paraphrase	0.43	2.96	0.44	0.17	0.29	0.36	0.25	0.74
Flan-T5 Split + Emotion	0.44	2.75	0.44	0.16	0.28	0.35	0.26	0.91
Flan-T5 Split + All	0.48	3.14	0.44	0.17	0.29	0.35	0.26	0.96

Figure C.1: **Explanation Automatic Evaluation Results (Full)**. The results of all of our models on the test set, for the explanation generation task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsection 8.3.2](#) for discussion).

Model	Extractiveness		Reference-Based							Factuality	
	Overlap	Density	Precision	Recall	F1	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SummaC	QAFactEval
Flan-T5 GPT	0.35	0.03	0.67	0.54	0.59	0.40	0.16	0.36	0.37	0.26	0.01
Flan-T5 Split	0.32	0.00	0.53	0.50	0.50	0.62	0.21	0.55	0.76	0.27	0.02
Flan-T5 Split + BB-Metaphor-Boolean	0.34	0.00	0.62	0.49	0.54	0.65	0.28	0.60	0.77	0.27	0.03
Flan-T5 Split + BB-Understanding-Metaphor	0.33	0.00	0.60	0.52	0.55	0.66	0.27	0.59	0.78	0.27	0.05
Flan-T5 Split + BB-Rephrase	0.35	0.00	0.57	0.48	0.51	0.63	0.27	0.57	0.76	0.26	0.17
Flan-T5 Split + BB-Suicide-Risk	0.37	0.00	0.65	0.50	0.55	0.67	0.28	0.59	0.78	0.27	0.09
Flan-T5 Split + BB-Fables	0.34	0.00	0.64	0.57	0.59	0.69	0.31	0.62	0.79	0.27	0.00
Flan-T5 Split + FLUTE	0.31	0.00	0.60	0.58	0.57	0.67	0.31	0.60	0.78	0.26	0.03
Flan-T5 Split + GoEmotions	0.34	0.00	0.64	0.55	0.58	0.68	0.30	0.61	0.79	0.26	0.03
Flan-T5 Split + Para-NMT	0.37	0.00	0.62	0.51	0.55	0.66	0.25	0.59	0.78	0.27	0.03
Flan-T5 Split + Quora-Question-Pairs	0.35	0.00	0.61	0.50	0.54	0.66	0.28	0.61	0.78	0.28	0.03
Flan-T5 Split + Natural-Questions	0.35	0.00	0.62	0.50	0.54	0.65	0.27	0.60	0.77	0.27	0.03
Flan-T5 Split + SQuAD	0.36	0.00	0.68	0.53	0.58	0.69	0.29	0.63	0.80	0.27	0.03
Flan-T5 Split + QA	0.34	0.00	0.62	0.50	0.54	0.66	0.27	0.60	0.78	0.27	0.05
Flan-T5 Split + Metaphor	0.35	0.00	0.64	0.56	0.58	0.68	0.27	0.60	0.78	0.26	0.06
Flan-T5 Split + Paraphrase	0.34	0.03	0.55	0.50	0.51	0.63	0.27	0.59	0.76	0.26	0.04
Flan-T5 Split + Emotion	0.38	0.00	0.65	0.49	0.55	0.66	0.31	0.62	0.78	0.27	0.00
Flan-T5 Split + All	0.39	0.00	0.70	0.52	0.58	0.69	0.31	0.63	0.79	0.27	0.00

Figure C.2: **Emotion Automatic Evaluation Results (Full)**. The results of all of our models on the test set, for the emotion detection task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsection 8.3.2](#) for discussion).

Model	Extractiveness		Reference-Based				Factuality	
	Overlap	Density	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	SummaC	QAFactEval
Flan-T5 GPT	0.69	2.44	0.33	0.17	0.31	0.30	0.41	2.21
Flan-T5 Split	0.60	2.18	0.28	0.14	0.27	0.28	0.43	2.38
Flan-T5 Split + BB-Metaphor-Boolean	0.52	1.65	0.29	0.12	0.27	0.23	0.41	1.83
Flan-T5 Split + BB-Understanding-Metaphor	0.59	1.83	0.28	0.13	0.27	0.24	0.39	1.81
Flan-T5 Split + BB-Rephrase	0.59	2.05	0.28	0.12	0.27	0.28	0.45	1.92
Flan-T5 Split + BB-Suicide-Risk	0.57	1.99	0.27	0.12	0.26	0.26	0.41	1.89
Flan-T5 Split + BB-Fables	0.53	1.87	0.29	0.15	0.28	0.28	0.38	1.95
Flan-T5 Split + FLUTE	0.60	2.12	0.29	0.13	0.28	0.30	0.38	1.59
Flan-T5 Split + GoEmotions	0.55	1.84	0.30	0.16	0.29	0.27	0.42	1.86
Flan-T5 Split + Para-NMT	0.65	2.10	0.36	0.19	0.34	0.33	0.45	1.78
Flan-T5 Split + Quora-Question-Pairs	0.51	1.49	0.33	0.14	0.30	0.31	0.40	1.60
Flan-T5 Split + Natural-Questions	0.59	2.08	0.34	0.19	0.32	0.32	0.41	2.19
Flan-T5 Split + SQuAD	0.59	1.66	0.35	0.19	0.33	0.32	0.42	1.76
Flan-T5 Split + QA	0.64	2.31	0.35	0.17	0.33	0.30	0.40	2.13
Flan-T5 Split + Metaphor	0.50	1.95	0.28	0.11	0.25	0.28	0.38	1.53
Flan-T5 Split + Paraphrase	0.54	1.66	0.28	0.12	0.27	0.27	0.39	1.44
Flan-T5 Split + Emotion	0.55	1.66	0.27	0.10	0.25	0.28	0.42	1.79
Flan-T5 Split + All	0.54	2.14	0.26	0.11	0.24	0.25	0.42	2.09

Figure C.3: **Event Automatic Evaluation Results (Full)**. The results of all of our models on the test set, for the event extraction task. The cells have been colored from red (undesirable score) to green (desirable score), and the best score in each column is **bolded** (see [subsection 8.3.2](#) for discussion).