

Monte Carlo Tree Search for Interpreting Stress in Natural Language

Kyle Swanson*, Joy Hsu*, Mirac Suzgun*

Department of Computer Science

Stanford University

{swansonk, joycj, msuzgun}@stanford.edu

Abstract

Natural language processing can facilitate the analysis of a person’s mental state from text they have written. Previous studies have developed models that can predict whether a person is experiencing a mental health condition from social media posts with high accuracy. Yet, these models cannot explain *why* the person is experiencing a particular mental state. In this work, we present a new method for explaining a person’s mental state from text using Monte Carlo tree search (MCTS). Our MCTS algorithm employs trained classification models to guide the search for key phrases that explain the writer’s mental state in a concise, interpretable manner. Furthermore, our algorithm can find both explanations that depend on the particular context of the text (e.g., a recent breakup) and those that are context-independent. Using a dataset of Reddit posts that exhibit stress, we demonstrate the ability of our MCTS algorithm to identify interpretable explanations for a person’s feeling of stress in both a context-dependent and context-independent manner.¹

1 Introduction

Disabilities associated with mental health conditions pose a significant challenge for many people around the world (Stauder et al., 2010; De Choudhury et al., 2013; Chen et al., 2018). To help people suffering from these conditions, it is crucial to identify those who are experiencing a mental health condition and understand the underlying causes.

Natural language processing (NLP) can help by analyzing a person’s mental state based on the text they have written. Previous studies (Turcan and McKeown, 2019; Demszky et al., 2020; Gjurković et al., 2020; Ansari et al., 2021) have demonstrated the ability of NLP models to process social media posts and predict stress, depression, and a range

r/Relationships: I can’t believe this. **My boyfriend just cheated on me** and then he bragged about it on twitter. What kind of a messed up person would do that? **I’m so angry with him** and I’m sure we’re going to **have a huge fight about this** when I see him tomorrow.

Figure 1: A fictitious example of text exhibiting stress in the relationships context and two explanations for that stress. The explanation in **blue** is context-dependent (specific to relationships) while the explanation in **red** is context-independent (general to any disagreement).

of emotions. These methods, however, are not able to explain *why* the person might be feeling the way they are, even if that information is clearly contained in the text analyzed by the model.

In this work, we seek to explain the underlying causes of a person’s mental state from their writing. We formulate such an explanation as a small set of phrases from the text that is sufficient to explain the person’s mental state. We wish to identify two complementary types of explanations: those that are particular to the situation the person is in, which we call *context-dependent*, and those that could appear across different contexts, which we call *context-independent*. Figure 1 shows an illustrating example. Identifying both types of explanations not only enhances our understanding of the underlying sources of a person’s mental state but also provides insights into how one’s mental state can be affected by general and specific causes.

To this end, we develop a novel Monte Carlo tree search (MCTS) algorithm that can effectively identify explanations that are either context-dependent or context-independent by leveraging the semantic capabilities of trained NLP models. We, both quantitatively and qualitatively, demonstrate the efficacy of this approach to explain a person’s mental state using a dataset of Reddit posts that exhibit stress (Turcan and McKeown, 2019).

*Denotes equal contribution.

¹Code and models are available at https://github.com/swansonk14/MCTS_Interpretability.

2 Related Work

Mental Health Prediction. Previous studies have tackled the task of mental health disability classification, using methods ranging from classical supervised techniques such as SVMs, logistic regression, Naive Bayes, MLPs, and decision trees to deeper models such as CNNs and GRUs (Turcan and McKeown, 2019; Gjurković et al., 2020; Ansari et al., 2021; Sampath et al., 2022). Other approaches utilize pre-trained, large language models with fine-tuning on specific mental health datasets (Ji et al., 2021; Matošević et al., 2021; Mauriello et al., 2021), which takes advantage of models trained on significantly larger datasets to speed up training and increase accuracy. Turcan and McKeown (2019) specifically focus on the task of stress prediction in Reddit posts, and they show that large BERT-based models outperform smaller models such as CNNs and logistic regression.

NLP Explainability. Explainability in NLP is an emerging topic of interest as language models have become larger and more accurate at the expense of reduced interpretability. Common methods for explainability include feature importance reporting across lexical or latent features (Danilevsky et al., 2020), model-agnostic approaches that extract post-hoc explanations (Ribeiro et al., 2016), and analogy-based explanations (Croce et al., 2019). Prior works have also focused on rationale identification (Lei et al., 2016) and text matching rationalization (Swanson et al., 2020), where models are designed to select small, interpretable segments of text when making predictions. Attention has also been used as a form of interpretability, but attention weights do not always correlate with impact on the model’s prediction, potentially limiting their usefulness (Serrano and Smith, 2019). In this work, we propose to use Monte Carlo tree search (Silver et al., 2016; Chaudhry and Lee, 2018; Jin et al., 2020; Albrecht et al., 2021; Yuan et al., 2021) as a post-hoc explainability method that can be applied to any model to flexibly identify multiple types of explanations for a model’s predictions.

3 The DREADDIT Dataset

The DREADDIT dataset (Turcan and McKeown, 2019) contains 3,553 Reddit posts that have human-annotated binary stress labels denoting whether a given text contains evidence of stress. Each post belongs to one of ten subreddits (e.g., “r/Relationships”), which we consider to be the con-

text of the post. The posts are split into 2,838 train posts and 715 test posts. Figures 8 and 9 (see Appendix) show the distributions of the stress labels and subreddit categories for the train and test sets.

4 Method

We assume that we have access to a training corpus $\mathcal{D}_{\text{train}}$ and a test corpus $\mathcal{D}_{\text{test}}$ to train and evaluate our models, respectively. The training corpus, $\mathcal{D}_{\text{train}} = (\mathbf{t}_i, \mathbf{s}_i, \mathbf{c}_i)_{i \in [1, n]}$, is a set of tuples, where each tuple contains a text $\mathbf{t}_i = \{t_i^1, \dots, t_i^{l_i}\} \in \mathbf{T}$ consisting of l_i tokens, its corresponding stress indicator $\mathbf{s}_i \in \mathbf{S} = \{0, 1\}$ denoting whether \mathbf{t}_i contains evidence of stress, and a context label $\mathbf{c}_i \in \mathbf{C}$ indicating the subreddit category the text belongs to. Similarly, we assume $\mathcal{D}_{\text{test}} = (\mathbf{t}_i, \mathbf{s}_i, \mathbf{c}_i)_{i \in [1, m]}$.

4.1 Classification of Stress and Context

We consider two types of classification tasks, namely binary stress classification and multi-class context (subreddit) classification. We refer to a model trained for the former task as a `stress` classifier, which can be thought of as a function mapping a piece of text $\mathbf{t} \in \mathbf{T}$ to a likelihood $p \in [0, 1]$. We refer to a model trained for the latter as a `context` classifier, which can be thought of as a function mapping a piece of text $\mathbf{t} \in \mathbf{T}$ to a probability simplex $\Delta^{|\mathbf{C}|-1}$.

We build simple stress and context prediction models using Bernoulli and Multinomial Naive Bayes, Support Vector Machine (Platt, 1999), and Multilayer Perceptron (Hinton, 1989). All of these models use vectors of word counts² as inputs. We also build large BERT-based models by adding a classification layer on top of the MentalRoBERTa model of Ji et al. (2021) and then fine-tuning the model on the training set.

4.2 Definition of an Explanation

An interpretable explanation for a person’s stress should consist of a small set of phrases from the full text that captures the core reasons behind the stress discussed within the text.

Formally, for a given piece of text in the corpus $\mathbf{t} \in \mathbf{T}$ that is labeled as stressed ($s = 1$), we define an *explanation* as a set of phrases $\mathbf{E} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_k\}$ where each phrase \mathbf{p}_j is a set of n_j contiguous tokens in the text, that is, $\mathbf{p}_j = \{t_l, t_{l+1}, \dots, t_{l+n_j-1}\}$ for some $l \in$

²We use `CountVectorizer` from `scikit-learn` fit on the training set with all default parameters.

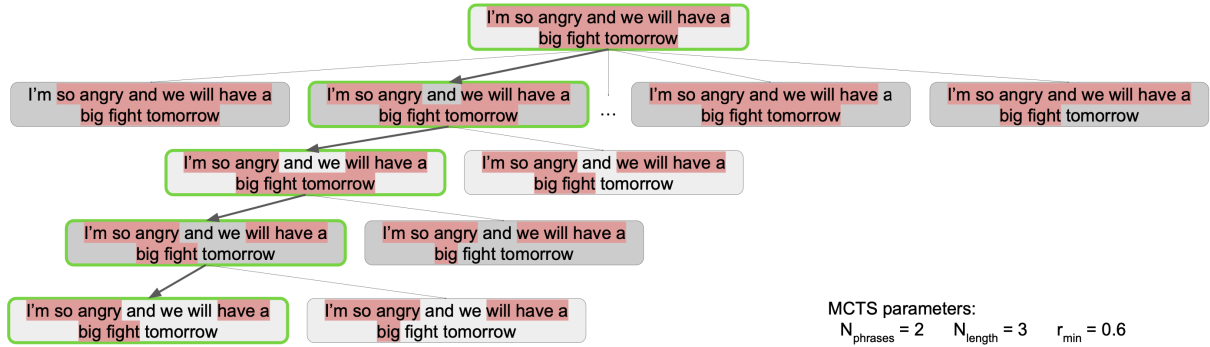


Figure 2: A portion of the tree of explanations searched by MCTS for an example text. Red indicates the text that is currently included in the explanation. The root of the tree is an explanation with a single phrase containing all the text. Each node in the tree can be expanded by removing the first or last token of a phrase or by removing a token in the middle of the phrase (constrained by certain MCTS parameters). Once a minimum number of tokens has been reached, the resulting explanation is given a reward based on the predictions of the stress and context models.

$\{1, 2, \dots, |\mathbf{t}| - n_j + 1\}$. Furthermore, the phrases must be non-overlapping, which means that $\mathbf{p}_j \cap \mathbf{p}_{j'} = \emptyset \quad \forall j \neq j' \in \{1, 2, \dots, |\mathbf{E}|\}$. In order to ensure interpretability, the explanation \mathbf{E} must satisfy three conditions.

a. Phrase count: $|\mathbf{E}| \leq N_{\text{phrases}}$, meaning the explanation must contain at most N_{phrases} phrases. Too many phrases would impede interpretability.

b. Phrase length: $|\mathbf{p}_j| \geq N_{\text{length}} \quad \forall j \in \{1, 2, \dots, |\mathbf{E}|\}$, meaning each phrase must have at least N_{length} tokens, preventing phrases that are too short to carry any meaning.

c. Proportion of tokens: $r_{\text{min}} \leq r(\mathbf{E}) \leq r_{\text{max}}$ where $r(\mathbf{E}) = \frac{1}{|\mathbf{t}|} \sum_{j=1}^{|\mathbf{E}|} |\mathbf{p}_j|$ is the proportion of tokens in the text that are included in the explanation and $0 \leq r_{\text{min}} \leq r_{\text{max}} \leq 1$ are lower and upper bounds on the proportion of tokens in the explanation. This constrains the overall verbosity of the explanation to a reasonable range.

4.3 Context-Dependent and Independent Explanations of Stress

We are interested in identifying two specific types of explanations for stress: one that depends on the context of the text and one that is independent of that context. We will refer to the context-dependent explanation as \mathbf{E}_{dep} and to the context-independent explanation as \mathbf{E}_{ind} .

In both cases, since the explanation must explain the stress in the text, the stress must be evident from just the text contained in the phrases of the explanation. We can verify this by using our stress classification model. Specifically, we want an explanation such that the average stress prediction across the phrases of the explanation is close to 1.

Hence for both \mathbf{E}_{dep} and \mathbf{E}_{ind} , we want

$$S(\mathbf{E}) = \frac{1}{|\mathbf{E}|} \sum_{j=1}^{|\mathbf{E}|} \text{stress}(\mathbf{p}_j) \approx 1$$

where $S(\mathbf{E})$ is the average stress across the phrases of the explanation.

However, the phrases of the context-dependent explanation \mathbf{E}_{dep} should indicate the context of the text while the context-independent explanation \mathbf{E}_{ind} should not. We enforce this by examining the entropy of the predictions of our context classification model. If the phrases of an explanation have low entropy, then the model is relatively sure of the context; hence, that explanation is context-dependent. If the entropy is high, then the model is unsure of the context and the explanation is context-independent. Formally, if we define

$$H(\mathbf{E}) = \frac{1}{|\mathbf{E}|} \sum_{j=1}^{|\mathbf{E}|} \text{entropy}(\text{context}(\mathbf{p}_j))$$

as the average Shannon entropy of the context predictions across phrases, we want $H(\mathbf{E}_{\text{dep}}) \approx 0$ and $H(\mathbf{E}_{\text{ind}}) \approx e_{\text{max}}$ where e_{max} is the maximum entropy (viz., entropy of a uniform distribution over contexts).

4.4 Finding Explanations with MCTS

We use the MCTS framework established in Silver et al. (2016), but we modify the search tree and the reward function to suite our purposes (see Figure 2). Each node in the tree represents an explanation $\mathbf{E} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_k\}$. The root of the tree represents the whole text piece as a single phrase, i.e., $\mathbf{E}_{\text{root}} = \{\mathbf{t}\}$. When the search is at a given

node in the tree, there are two options for expanding the next node: (i) remove the first or last token in any phrase, as long as the shortened phrase still contains at least N_{length} tokens, or (ii) remove a token in the middle of a phrase, thus breaking it into two phrases, as long as both resulting phrases have at least N_{length} tokens and the total number of phrases does not exceed $N_{phrases}$.

The search continues to expand nodes in the tree until either the current node cannot be expanded using either of the two rules above or the explanation at the current node contains too few tokens, i.e., $r(\mathbf{E}) \leq r_{min}$. This node serves as a leaf node and is given a reward equal to

$$R(\mathbf{E}) = S(\mathbf{E}) + I \cdot \alpha \cdot H(\mathbf{E})$$

for some $I \in \{-1, +1\}$ and $\alpha \geq 0$. We use $I = +1$ to select for high entropy (context-independent) explanations and $I = -1$ to select for low entropy (context-dependent) explanations. This reward is propagated back to all the nodes on the path from the root to this leaf node according to the update rules from Silver et al. (2016). After the search is complete, the best explanation $\hat{\mathbf{E}}$ is selected as

$$\hat{\mathbf{E}} = \underset{\mathbf{E}}{\operatorname{argmax}} R(\mathbf{E}) \text{ s.t. } r(\mathbf{E}) \leq r_{max},$$

which means $\hat{\mathbf{E}}$ is the explanation in the search tree that maximizes the reward while satisfying the condition on the maximum proportion of tokens. The other interpretability conditions are guaranteed by the rules of the search tree expansion.

5 Experiments

All of our experiments were run on the DREADDIT dataset. We report results of our stress and context classification models and share findings of our MCTS explanation algorithm.

5.1 Classification

As Table 1 illustrates, basic stress classification models, such as Naive Bayes classifiers, SVMs, and MLPs, performed reasonably on the test set of DREADDIT. The MentalRoBERTa^{FT} model for stress fine-tuned on the training set of DREADDIT for five epochs, however, was able to outperform all the other models, achieving an accuracy score of 82% and demonstrating the efficacy of the pre-training on mental health data³. Our results on the

³In contrast, the RoBERTa model trained from scratch achieved an accuracy score of almost 80%.

Model	Precision	Recall	F-1	Accuracy
Bernoulli NB	0.69	0.84	0.75	0.72
Multinomial NB	0.68	0.87	0.76	0.72
SVM	0.71	0.77	0.74	0.72
MLP	0.71	0.74	0.73	0.71
MentalRoBERTa ^{FT}	0.78	0.90	0.84	0.82

Table 1: Performances of stress classifiers on the test set of DREADDIT. While non-neural classifiers could not surpass 72% accuracy, the MentalRoBERTa^{FT} model fine-tuned on the DREADDIT train set yielded 82% accuracy. Here, the superscript ^{FT} denotes that the model was fine-tuned.

Model	Precision	Recall	F-1	Accuracy
Bernoulli NB	0.81	0.75	0.76	0.80
Multinomial NB	0.77	0.75	0.75	0.79
SVM	0.76	0.72	0.74	0.76
MLP	0.78	0.78	0.78	0.79
MentalRoBERTa ^{FT}	0.85	0.86	0.86	0.87

Table 2: Performances of context classifiers. We restricted our focus to three subreddits: “anxiety,” “assistance,” “relationships.” The fine-tuned MentalRoBERTa^{FT} model yielded the best results with 87% accuracy.

stress classification task are consistent with those of Turcan and McKeown (2019). Table 2 reports the performance of various models on the multi-class subreddit category classification. Here, we limited our attention to three categories, namely “anxiety,” “assistance,” and “relationships.” The Reddit posts in these categories embody various distinct everyday, financial, and interpersonal stress factors, but at the same time, they seem to have common (context-independent) stress elements. In this context classification task, all models were able to go beyond the 75% accuracy level, but MentalRoBERTa^{FT} yielded the highest accuracy.

5.2 Explainability

We demonstrate our MCTS approach to explainability using the same three categories as above. We use stress and context classification models implemented with Multinomial NB, MLP, and MentalRoBERTa^{FT}. For each of these models, we apply MCTS to identify explanations for each of the 166 test texts that is labeled as stressed and belongs to one of our three categories. We use the interpretability conditions $N_{phrases} = 3$, $N_{length} = 5$, $r_{min} = 0.2$, and $r_{max} = 0.5$ for all experiments⁴, and we use $\alpha = 10$ except where otherwise noted.

We quantitatively evaluate the explanations pro-

⁴These choices are arbitrary and could easily be changed.

		Original	Dependent	Independent
MNB	S	0.850 ± 0.317	0.706 ± 0.190	0.617 ± 0.124
	E	0.047 ± 0.140	0.274 ± 0.181	0.942 ± 0.086
MLP	S	0.725 ± 0.383	0.512 ± 0.194	0.546 ± 0.145
	E	0.214 ± 0.274	0.766 ± 0.163	1.067 ± 0.022
MRB	S	0.878 ± 0.324	0.830 ± 0.220	0.430 ± 0.273
	E	0.042 ± 0.124	0.019 ± 0.018	0.640 ± 0.171

Table 3: Stress (S) and context entropy (E) for original text, context-dependent explanation, and context-independent explanation for the Multinomial Naive Bayes (MNB), Multilayer Perceptron (MLP), and Mental RoBERTa (MRB) models. Results were generated through MCTS with stress and context entropy averaged over the test set. The Wilcoxon signed rank test (Wilcoxon, 1945) between dependent and independent entropy is $p < 0.0001$ for all models, indicating a very significant difference as desired.

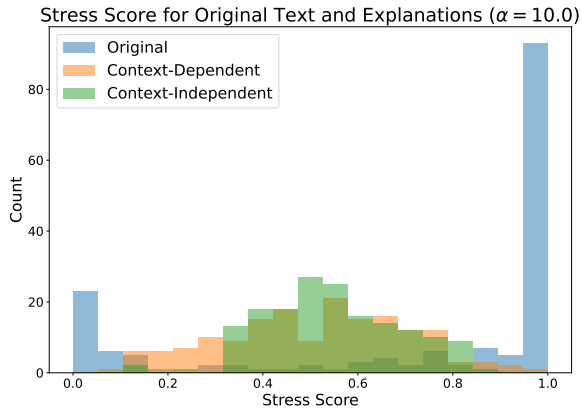


Figure 3: Histogram of stress scores for the original text and for the context-dependent and context-independent explanations extracted by our MCTS algorithm using an MLP model. Although stress is often higher in the original text than in the extracted explanations, the explanations still maintain a meaningful amount of stress.

duced by MCTS. In Table 3, we show the average stress and context entropy scores of the original text and of the context-dependent and context-independent explanations. Our method is able to maintain a reasonably high and consistent level of stress across the explanations while modulating the context entropy appropriately for the two different types of explanations. This indicates that our approach can identify both context-dependent and context-independent sources of stress.

Figures 3 and 4 further illustrate this result for the MLP model by showing the full distribution of stress and context entropy scores across the test examples. Figures 5, 6, and 7 in the Appendix show the stress and context entropy distributions for all three models and for different values of α . Lower α increases stress but decreases the differ-

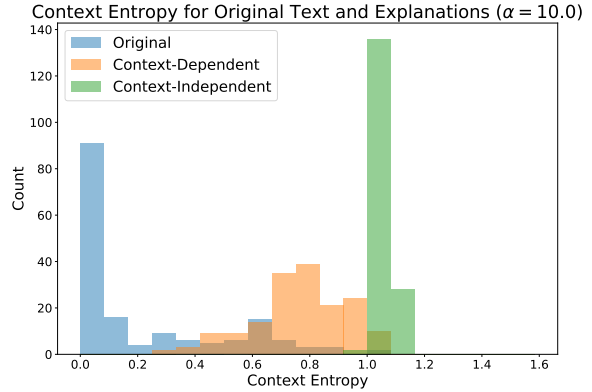


Figure 4: Histogram of context entropy for the original text and for the context-dependent and context-independent explanations extracted by our MCTS algorithm using an MLP model. The context-independent explanations clearly have much higher context entropy than the context-dependent explanations as desired.

ence in entropy between the two types of explanations while higher α decreases stress but increases the difference in entropy. This shows the flexibility of MCTS to select different types of explanations without retraining the classifiers.

Furthermore, we qualitatively demonstrate our approach. Tables 4, 5, and 6 in the Appendix show examples from each of the three subreddits that illustrate how our method captures different underlying sources of stress in an interpretable manner.

6 Conclusion

We propose a novel interpretability method for explaining stress in context-dependent and independent manners using Monte Carlo tree search. We demonstrate the effectiveness of our method by extracting both types of explanations from Reddit posts that exhibit stress. Although this work focuses on stress, our MCTS-based explanation framework is extremely flexible and can be applied to a wide variety of NLP models and prediction problems simply by specifying the appropriate reward function and interpretability conditions for the search tree. As in our work, the reward function can include multiple objectives with different weights, making it possible to extract a variety of explanations for added interpretability. Future work should further explore the range of explanations enabled by our framework. We hope that our explanation framework can improve understanding of the root causes of mental health conditions as well as provide interpretability for a variety of NLP tasks.

Acknowledgements

We would like to thank Margalit Glasgow, Masha Karelina, Megha Patel, Biscuit Russell, and Tayfun M. H. Mezarci for helpful comments and discussions. Swanson and Hsu gratefully acknowledge the support of the Knight-Hennessy Scholarship, Hsu gratefully acknowledges the support of the NSF GRFP, and Suzgun gratefully acknowledges the support of a Johann, Thales, Williams & Co. Graduate Fellowship. The authors also thank Dan Jurafsky for his support. The experiments presented in this paper were run on the Stanford NLP Cluster. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of Stanford University. All errors remain our own.

References

- Stefano V Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. 2021. Interpretable goal-based prediction and planning for autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1043–1049. IEEE.
- Gunjan Ansari, Muskan Garg, and Chandni Saxena. 2021. Data augmentation for mental health classification on social media. *arXiv preprint arXiv:2112.10064*.
- Muhammad Umar Chaudhry and Jee-Hyong Lee. 2018. Feature selection for high dimensional data using monte carlo tree search. *IEEE Access*, 6:76036–76048.
- Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660.
- Danilo Croce, Daniele Rossini, and Roberto Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. 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 4037–4046.
- Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.
- Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2020. Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*.
- Geoffrey Hinton. 1989. Connectionist learning procedures. *Artificial intelligence*, 40:185–234.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pages 4849–4859. PMLR.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lovro Matošević, Filip Sosa, and Marko Gašparac. 2021. Stressformers: Transferring knowledge for stress analysis in social media. *Text Analysis and Retrieval 2021 Course Project Reports*, page 56.
- Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.
- Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneshelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- Adrienne Stauder, Barna Konkoly Thege, Mónika Erika Kovács, Piroska Balog, Virginia P Williams, and Redford B Williams. 2010. Worldwide stress: different problems, similar solutions? cultural adaptation and evaluation of a standardized stress management program in hungary. *International Journal of Behavioral Medicine*, 17(1):25–32.
- Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. *arXiv preprint arXiv:2005.13111*.
- Elsbeth Turcan and Kathleen McKeown. 2019. Dreddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.
- Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.
- Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. 2021. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12241–12252. PMLR.

A Appendix

A.1 Additional Stress and Context Entropy Results

Figures 5, 6, and 7 show the stress and context entropy distributions of the original text and the context-dependent and context-independent explanations across the 166 stressed test examples in the “anxiety,” “assistance,” and “relationships” subreddits for the Multinomial Naive Bayes, Multilayer Perceptron, and MentalRoBERTa^{FT} models, respectively. For the Multinomial Naive Bayes and Multilayer Perceptron models, we experimented with $\alpha \in \{0.1, 1, 10\}$, with higher α weighting context entropy more than stress in the MCTS reward function. For the MentalRoBERTa^{FT} model, we used $\alpha = 10$.

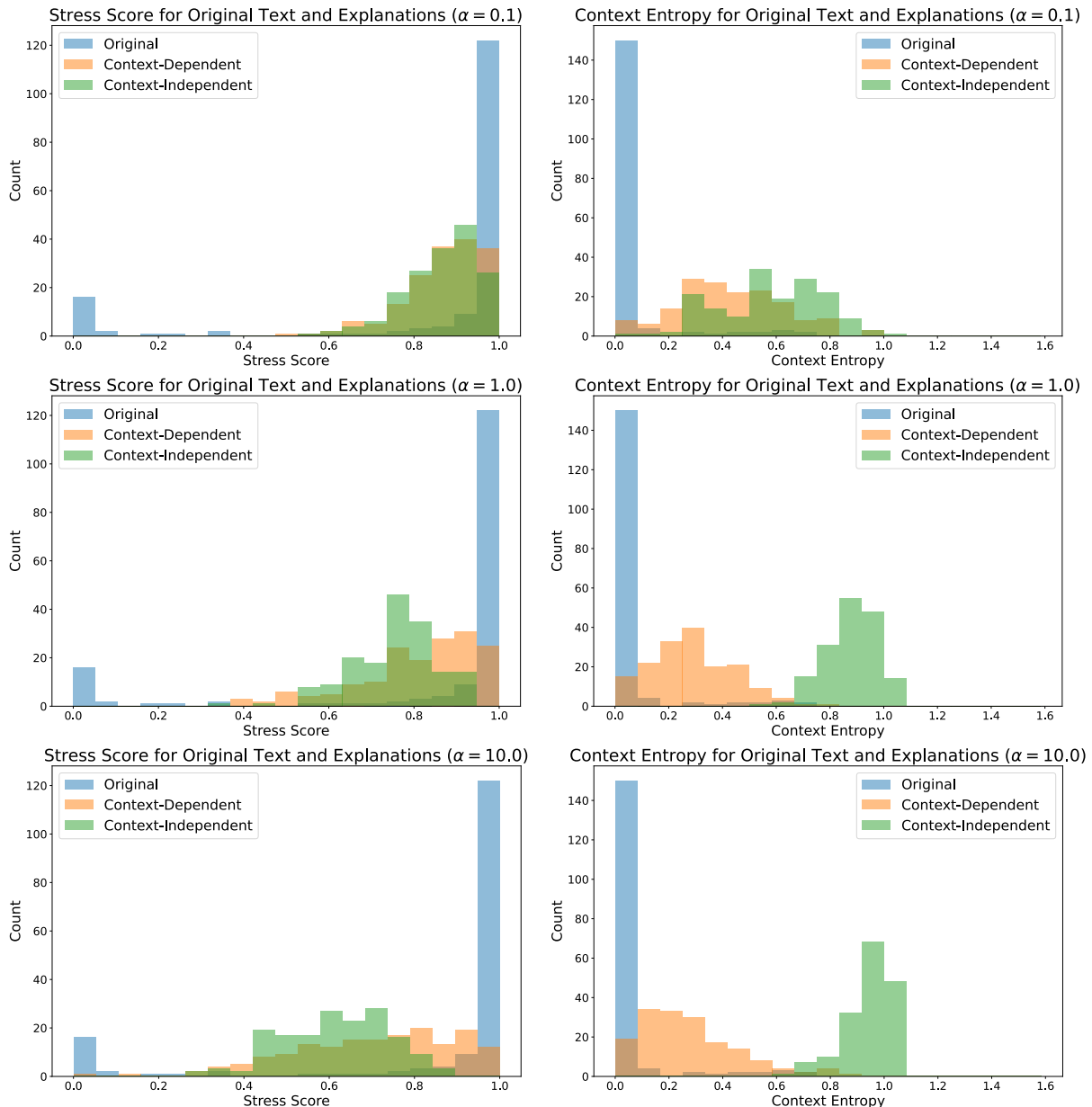


Figure 5: Histograms of stress and context entropy scores from the Multinomial Naive Bayes model for the original text and for the context-dependent and context-independent explanations extracted by our MCTS algorithm. The left column shows stress scores while the right column shows context entropy scores. From top to bottom, the rows show $\alpha = 0.1$, $\alpha = 1$, and $\alpha = 10$, where α controls the balance between stress and context entropy in the MCTS reward function. Higher α places less emphasis on stress and more emphasis on context entropy, resulting in a greater difference between context-dependent and context-independent entropy scores at the cost of lower stress.

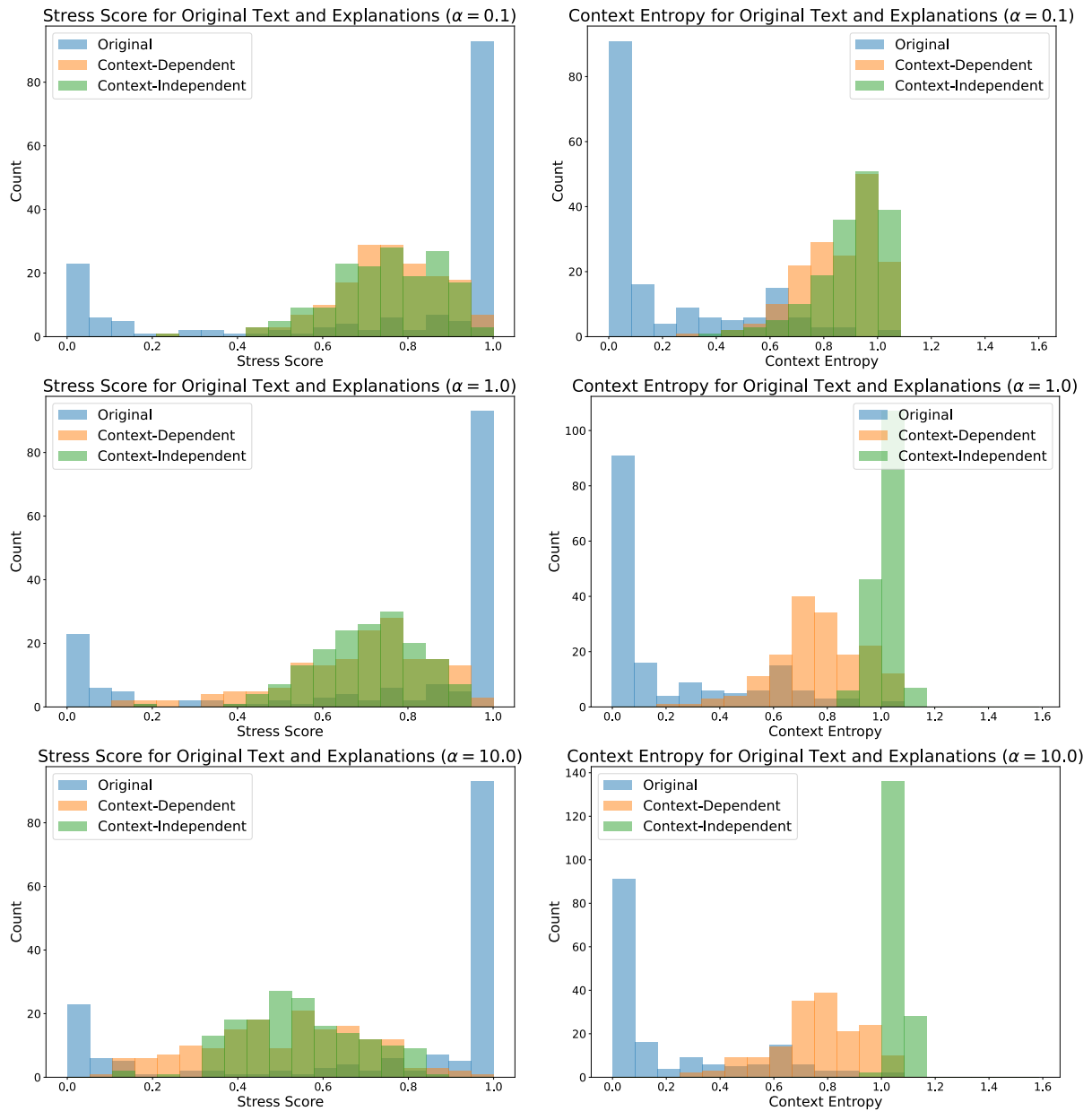


Figure 6: Histograms of stress and context entropy scores from the Multilayer Perceptron model for the original text and for the context-dependent and context-independent explanations extracted by our MCTS algorithm. The left column shows stress scores while the right column shows context entropy scores. From top to bottom, the rows show $\alpha = 0.1$, $\alpha = 1$, and $\alpha = 10$, where α controls the balance between stress and context entropy in the MCTS reward function. Higher α places less emphasis on stress and more emphasis on context entropy, resulting in a greater difference between context-dependent and context-independent entropy scores at the cost of lower stress.

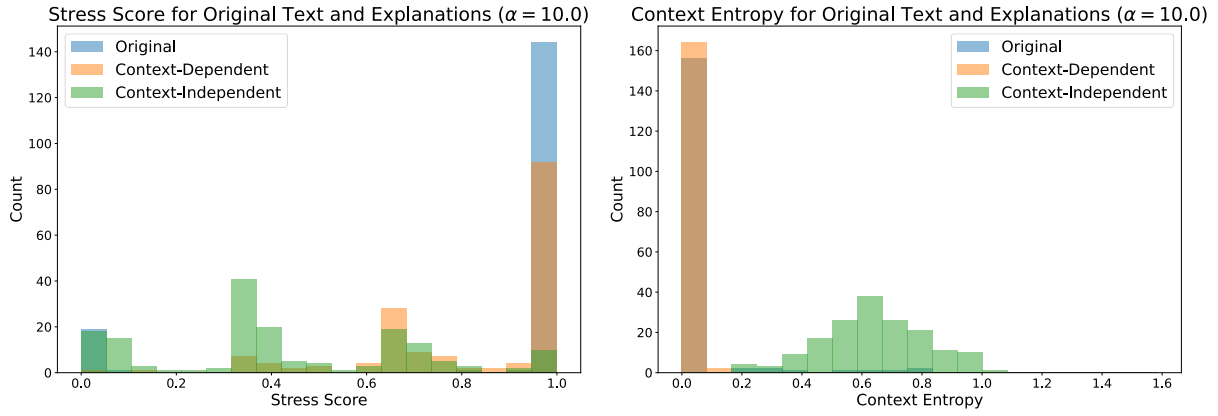


Figure 7: Histograms of stress and context entropy scores from the MentalRoBERTa^{FT} model for the original text and for the context-dependent and context-independent explanations extracted by our MCTS algorithm. The left plot shows stress scores while the right plot shows context entropy scores, both for $\alpha = 10$. Interestingly, the distributions are somewhat different from those of the Multinomial Naive Bayes (Figure 5) and Multilayer Perceptron (Figure 6) models. MentalRoBERTa^{FT} is capable of selecting different context-dependent and context-independent explanations as measured by entropy, but the model generally assigns more stress to context-dependent explanations than context-independent explanations, perhaps hinting at a meaningful difference between the types of explanations in terms of stress content.

A.2 Data Distribution

In Figure 8 and Figure 9, we show the data distribution of our stress and context (subreddit) labels.

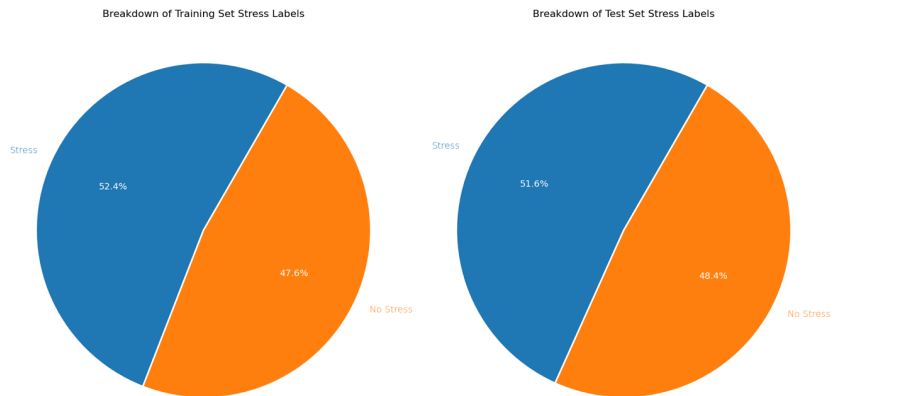


Figure 8: Training and test set stress label distribution.

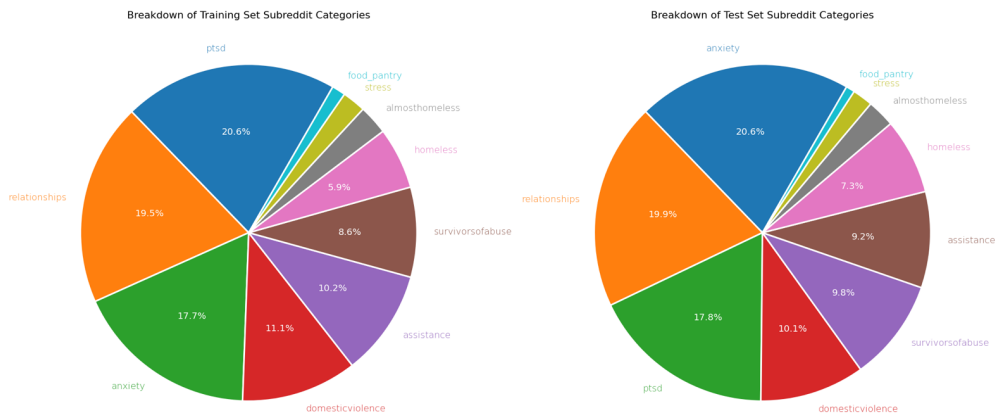


Figure 9: Training and test set subreddit label distribution.

A.3 MentalRoBERTa

MentalRoBERTa is a RoBERTa-based language model (Liu et al., 2019) that was pre-trained on a corpus of 13.7M sentences from Reddit that were posted on mental health-related subreddits, including, but not limited to, “r/Anxiety” and “r/Depression”. When training classifiers for stress and context classification tasks, we used the pre-trained MentalRoBERTa model on Hugging Face’s model repository, available at <https://huggingface.co/mental>, and fine-tuned the model on the DREADDIT dataset, using either the stress or context labels, for five epochs with a learning rate of 1e-4.

A.4 Qualitative Examples

In Tables 4, 5, and 6, we show qualitative examples of our MCTS method for explainability, with examples from each of three subreddits—“anxiety,” “assistance,” and “relationships”—from both the MLP and MentalRoBERTa^{FT} models.

Model	Category	Text (subreddit = “r/Anxiety”)	Stress	Entropy
MLP	Original	Lately I’ve just been having that terrible feeling in the pit of my stomach and also a feeling of nausea like I constantly need to throw up. I’m sleeping normal but still feeling so tired and drained and can’t really focus at work and because of that I feel like my work performance is slipping up. I am constantly afraid that I’m going to lose my job and that my manager hates me. This has been happening so much more frequently. About a week ago my doc gave me prozac (once a day) and xanax (only as needed) prescriptions and I feel like it’s helped with the bigger attacks and some dark thoughts but now its almost like just a little constant anxiety all the time and it sucks.	1.000	0.000
	Dependent	Lately I’ve just been having that terrible feeling in the pit of my stomach and also a feeling of nausea like I constantly need to throw up. I’m sleeping normal but still feeling so tired and drained and can’t really focus at work and because of that I feel like my work performance is slipping up. I am constantly afraid that I’m going to lose my job and that my manager hates me. This has been happening so much more frequently. About a week ago my doc gave me prozac (once a day) and xanax (only as needed) prescriptions and I feel like it’s helped with the bigger attacks and some dark thoughts but now its almost like just a little constant anxiety all the time and it sucks.	0.933	0.300
	Independent	Lately I’ve just been having that terrible feeling in the pit of my stomach and also a feeling of nausea like I constantly need to throw up. I’m sleeping normal but still feeling so tired and drained and can’t really focus at work and because of that I feel like my work performance is slipping up. I am constantly afraid that I’m going to lose my job and that my manager hates me. This has been happening so much more frequently. About a week ago my doc gave me prozac (once a day) and xanax (only as needed) prescriptions and I feel like it’s helped with the bigger attacks and some dark thoughts but now its almost like just a little constant anxiety all the time and it sucks.	0.489	1.045
	Dependent	Lately I’ve just been having that terrible feeling in the pit of my stomach and also a feeling of nausea like I constantly need to throw up. I’m sleeping normal but still feeling so tired and drained and can’t really focus at work and because of that I feel like my work performance is slipping up. I am constantly afraid that I’m going to lose my job and that my manager hates me. This has been happening so much more frequently. About a week ago my doc gave me prozac (once a day) and xanax (only as needed) prescriptions and I feel like it’s helped with the bigger attacks and some dark thoughts but now its almost like just a little constant anxiety all the time and it sucks.	0.998	0.006
MentalRoBERTa ^{FT}	Independent	Lately I’ve just been having that terrible feeling in the pit of my stomach and also a feeling of nausea like I constantly need to throw up. I’m sleeping normal but still feeling so tired and drained and can’t really focus at work and because of that I feel like my work performance is slipping up. I am constantly afraid that I’m going to lose my job and that my manager hates me. This has been happening so much more frequently. About a week ago my doc gave me prozac (once a day) and xanax (only as needed) prescriptions and I feel like it’s helped with the bigger attacks and some dark thoughts but now its almost like just a little constant anxiety all the time and it sucks.	0.670	0.627

Table 4: Qualitative examples from our MCTS explainability method for a post in the “r/Anxiety” subreddit. We show the full original text along with the context-dependent and context-independent explanations selected by MCTS using both the MLP and MentalRoBERTa^{FT} classifiers.

Model	Category	Text (subreddit = “r/Assistance”)	Stress	Entropy
	Original	I can’t ask my family because they don’t have the kind of money to help me. If anyone can help me even just a little bit, I would be ridiculously grateful. I just can’t even express what this has done to us. Yes, the bills are paid, but now we’re so anxious that we barely leave the house due to panic attacks. I’ve done things like ubereats but \$15 here and there isn’t even making a dent in what I need.	0.995	0.616
MLP	Dependent	I can’t ask my family because they don’t have the kind of money to help me. If anyone can help me even just a little bit, I would be ridiculously grateful. I just can’t even express what this has done to us. Yes, the bills are paid, but now we’re so anxious that we barely leave the house due to panic attacks. I’ve done things like ubereats but \$15 here and there isn’t even making a dent in what I need	0.723	0.640
	Independent	I can’t ask my family because they don’t have the kind of money to help me. If anyone can help me even just a little bit, I would be ridiculously grateful. I just can’t even express what this has done to us. Yes, the bills are paid, but now we’re so anxious that we barely leave the house due to panic attacks. I’ve done things like ubereats but \$15 here and there isn’t even making a dent in what I need.	0.584	1.064
Mental RoBERTa ^{FT}	Dependent	I can’t ask my family because they don’t have the kind of money to help me. If anyone can help me even just a little bit, I would be ridiculously grateful. I just can’t even express what this has done to us. Yes, the bills are paid, but now we’re so anxious that we barely leave the house due to panic attacks. I’ve done things like ubereats but \$15 here and there isn’t even making a dent in what I need.	0.999	0.005
	Independent	I can’t ask my family because they don’t have the kind of money to help me. If anyone can help me even just a little bit, I would be ridiculously grateful. I just can’t even express what this has done to us. Yes, the bills are paid, but now we’re so anxious that we barely leave the house due to panic attacks. I’ve done things like ubereats but \$15 here and there isn’t even making a dent in what I need.	0.478	0.518

Table 5: Qualitative examples from our MCTS explainability method for a post in the “r/Assistance” subreddit. We show the full original text along with the context-dependent and context-independent explanations selected by MCTS using both the MLP and MentalRoBERTa^{FT} classifiers.

Model	Category	Text (subreddit = “r/Relationships”)	Stress	Entropy
MLP	Original	We seem to be talking and accidentally being together more often in school, making what I think are feelings towards her only stronger. I can’t bring myself to bring this up with her because I’m scared that we will have a repeat of February again. I love her so much but I feel that if I have these feelings about other girls am I really devoted to her? This is in no way her fault, she has done nothing to deserve my questioning of my decision, this is my problem and mine alone. I am reluctant to bring this up with her because I’m worried that she might break up with me because I do truly still love her I’m just wondering if this other girl is a passing thought more focused than earlier and something I can overcome.	0.999	0.000
	Dependent	We seem to be talking and accidentally being together more often in school, making what I think are feelings towards her only stronger. I can’t bring myself to bring this up with her because I’m scared that we will have a repeat of February again. I love her so much but I feel that if I have these feelings about other girls am I really devoted to her? This is in no way her fault, she has done nothing to deserve my questioning of my decision, this is my problem and mine alone. I am reluctant to bring this up with her because I’m worried that she might break up with me because I do truly still love her I’m just wondering if this other girl is a passing thought more focused than earlier and something I can overcome.	0.734	0.437
	Independent	We seem to be talking and accidentally being together more often in school, making what I think are feelings towards her only stronger. I can’t bring myself to bring this up with her because I’m scared that we will have a repeat of February again. I love her so much but I feel that if I have these feelings about other girls am I really devoted to her? This is in no way her fault, she has done nothing to deserve my questioning of my decision, this is my problem and mine alone. I am reluctant to bring this up with her because I’m worried that she might break up with me because I do truly still love her I’m just wondering if this other girl is a passing thought more focused than earlier and something I can overcome.	0.510	1.043
	Dependent	We seem to be talking and accidentally being together more often in school, making what I think are feelings towards her only stronger. I can’t bring myself to bring this up with her because I’m scared that we will have a repeat of February again. I love her so much but I feel that if I have these feelings about other girls am I really devoted to her? This is in no way her fault, she has done nothing to deserve my questioning of my decision, this is my problem and mine alone. I am reluctant to bring this up with her because I’m worried that she might break up with me because I do truly still love her I’m just wondering if this other girl is a passing thought more focused than earlier and something I can overcome.	0.998	0.030
Mental RoBERTa ^{FT}	Independent	We seem to be talking and accidentally being together more often in school, making what I think are feelings towards her only stronger. I can’t bring myself to bring this up with her because I’m scared that we will have a repeat of February again. I love her so much but I feel that if I have these feelings about other girls am I really devoted to her? This is in no way her fault, she has done nothing to deserve my questioning of my decision, this is my problem and mine alone. I am reluctant to bring this up with her because I’m worried that she might break up with me because I do truly still love her I’m just wondering if this other girl is a passing thought more focused than earlier and something I can overcome.	0.712	0.444

Table 6: Qualitative examples from our MCTS explainability method for a post in the “r/Relationships” subreddit. We show the full original text along with the context-dependent and context-independent explanations selected by MCTS using both the MLP and MentalRoBERTa^{FT} classifiers.