# Discourse-Level Representations can Improve Prediction of Degree of Anxiety

**Swanie Juhng**[1], **Matthew Matero**[1], **Vasudha Varadarajan**[1], **Johannes C. Eichstaedt**[2]
**Adithya V. Ganesan**[1] and **H. Andrew Schwartz**[1]
[1]Department of Computer Science, Stony Brook University
[2]Department of Psychology, Stanford University
{sjuhng,mmatero,vvaradarajan,avirinchipur,has}@cs.stonybrook.edu
eichstaedt@stanford.edu

## Abstract

Anxiety disorders are the most common of mental illnesses, but relatively little is known about how to detect them from language. The primary clinical manifestation of anxiety is worry associated cognitive distortions, which are likely expressed at the discourse-level of semantics. Here, we investigate the development of a modern linguistic assessment for degree of anxiety, specifically evaluating the utility of discourse-level information in addition to lexical-level large language model embeddings. We find that a combined *lexico-discourse* model outperforms models based solely on state-of-the-art contextual embeddings (RoBERTa), with discourse-level representations derived from Sentence-BERT and DiscRE both providing additional predictive power not captured by lexical-level representations. Interpreting the model, we find that discourse patterns of causal explanations, among others, were used significantly more by those scoring high in anxiety, dovetailing with psychological literature.

## 1 Introduction

Anxiety disorders are one of the most prevalent mental health conditions, affecting an estimated 284 million people worldwide (Roth, 2018) and with an estimated financial burden of $46.6 billion annually in the U.S. alone (DeVane et al., 2005). This puts the impact of anxiety on par with depression (Guntuku et al., 2017; Mahdy et al., 2020), yet much less work in the NLP community has focused on detecting anxiety disorders as has been done for depressive disorders.

One of the key characteristics of anxiety disorders is cognitive distortion (Muran and Motta, 1993; Maric et al., 2011), or an illogical reasoning in dealing with life events (Kaplan et al., 2017). The primary window into such distortions is language, including one's own explanatory style – the way they reason about the occurrence of events (Peterson, 1991).

Explanatory style may not be well represented by single words or words in context (i.e., *lexical-level* features). For example, consider the *catastrophizing* statement (i.e., worrying that a bad event will lead to an extreme outcome) "*I'm sick. Now I'm going to miss my classes and fail them all.*" (Hazlett-Stevens and Craske, 2003). To see that "*fail them all*" is catastrophizing the event *"I'm sick"* requires understanding that the latter is a causal explanation for the expected falling behind. This is *discourse-level* information – semantics at the level of complete clausal statements or relating statements to each other (discourse relations) (Pitler et al., 2008).

Here, we propose a language-based assessment of anxiety utilizing both lexical-level and discourse-level representations. We first compare models that leverage discourse-level representations alone. We then propose a dual lexical- and discourse-level (*lexico-discourse*) approach and evaluate whether the combination of both types of representations leads to improved performance. Finally, we explore specific types of discourse relations that are thought to be associated with cognitive distortions, and look at their association with anxiety in order to illuminate what our lexico-discourse approach can pick up on at the discourse semantics level.

Our **contributions** include: (1) proposal of a novel user-level language assessment model that integrates both discourse-level and lexical-level representations; (2) empirical exploration of different discourse and lexical-level contextual embeddings and their value towards predicting the degree of anxiety as continuous values; (3) examination of the association between a person's anxiety and their discourse relation usage, finding that causal explanations are the most insightful for prediction; and (4) finding that to the best of our knowledge, this is the first model of anxiety from language specifically fit against a screening survey (rather than users self-declaring having experienced anxiety



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 2: Short Papers*, pages 1500–1511
July 9-14, 2023 ©2023 Association for Computational Linguistics

symptoms, or annotators perceiving the presence of the condition).

## 2 Related Work

Anxiety is characterized by disruptive feelings of uncertainty, dread, and fearfulness, and is generally defined as anticipation of future threats (Cohen et al., 2016). Researchers have recently been turning to social media language as a potential alternative source for mental health assessment, investigating, e.g., depression (Schwartz et al., 2014; Bathina et al., 2021; Kelley and Gillan, 2022), PTSD (Coppersmith et al., 2014; Benton et al., 2017b; Son et al., 2021), and suicide risk (Coppersmith et al., 2016; Mohammadi et al., 2019; Matero et al., 2019). Such an approach was also utilized in analyzing anxiety (Shen and Rudzicz, 2017; Tyshchenko, 2018; Guntuku et al., 2019; Budiyanto et al., 2019; Owen et al., 2020; Saifullah et al., 2021). Work towards this goal include Shen and Rudzicz (2017) who attempted to classify Reddit posts into binary levels of anxiety by lexical features and Guntuku et al. (2019) who explored Ngram associations with anxiety in Twitter users. Few have attempted to capture discourse-level information in such systems.

While some have focused on cognitive distortions in patient-therapist interactions (Simms et al., 2017; Burger et al., 2021; Shreevastava and Foltz, 2021), none have attempted to combine discourse-level information with more standard lexical-level embeddings in studying ecological (i.e., everyday, happening in the course of life) online language patterns. For mental health tasks, state-of-the-art systems have primarily relied on contextual word-level information from transformers like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) (Mohammadi et al., 2019; Matero et al., 2019). Furthermore, Ganesan et al. (2021) improved mental health task performance by reducing the dimensions of contextual embeddings to approximately $\frac{1}{12}$ of the original. Here, we seek to establish the role of the contextual embeddings as well as propose and evaluate a model that integrates discourse-level modeling with contextual embeddings, motivated by the ability of discourse relations to capture cognitive distortions.

## 3 Method

**Discourse-Level Embeddings.** We consider a variety of discourse-level embeddings, ranging from those capturing phrases or sentences to one capturing relations between clauses. *Sentence-BERT* (Reimers and Gurevych, 2019) is a variant of BERT that captures a whole sentence by optimizing for semantic similarity using siamese and triplet networks. *Phrase-BERT* (Wang et al., 2021) attempts to capture shorter phrasal semantics using contrastive learning with machine-generated paraphrases and mined phrases. Finally, *DiscRE* (Son et al., 2022) captures representations of the *relationship* between discourse units (i.e., clauses rooted with a main verb) using a weakly supervised, multi-task approach over bidirectional sequence models.

**Lexical Embeddings.** Amongst potential options for state-of-the-art auto-encoder language models, we consider BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Such selection is supported by empirical evidence; these two models have previously been found to result in top performance in related mental health assessment tasks (Matero et al., 2019; Ganesan et al., 2021). Beyond the fact that these models have lead to state-of-the-art performance in language understanding tasks, they are also known to capture *some* discourse information (Kishimoto et al., 2020; Liu et al., 2021). Thus, they form a very high benchmark to try to out-predict with discourse-level embeddings.

**Overall Model.** The architecture of our prediction models is laid out in Figure 1. Each model consists of a discourse submodel and lexical submodel, and the two following equations demonstrate the aggregation of representations in each submodel. $d, m, u$ each denotes discourse unit, message, and user.

The discourse submodel takes discourse units parsed from a message[1] to derive discourse-level embeddings, denoted as $e_u^d$ (Eq. 1), which are aggregated into message-level and then into a user-level embedding, $e_u$ (Eq. 2):

$$e_u^m = \text{compose}_{d \in m}(e_m^d) \tag{1}$$

$$e_u = \text{compose}_{m \in u}(e_u^m) \tag{2}$$

The lexical submodel takes the embeddings derived from the word-based transformer models as message-level representations and aggregates them to user-level. Compose is the embeddings aggregation function at each step, which can be mean, min, or max. Here we follow the practice from

[1]Discourse units are sentences for Sentence-BERT and clauses for DiscRE and Phrase-BERT.

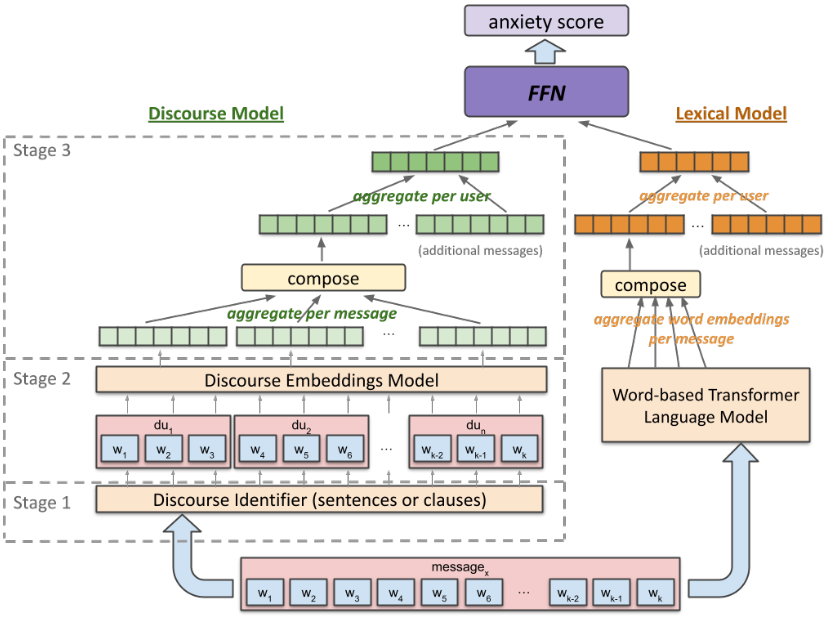


Figure 1: General architecture used for our anxiety assessment model. Depending on the model used, discourse units may be sentences or single clauses rooted by a main verb. The right-hand side, lexical model, follows the same approach as Ganesan et al. (2021) and Matero et al. (2021) for state-of-the-art assessment from contextual word embeddings.

Ganesan et al. (2021) and Matero et al. (2021) and use the mean.[2] Finally, the concatenation of the representations acts as input to our feed-forward network (FFN) that predicts the degree of anxiety.[3]

**Theoretically Relevant Discourse Dimensions.** Previous work has suggested open vocabulary (latent) embeddings of discourse relations (i.e., DiscRE, Sentence-BERT) are more powerful than explicitly defined relations (Son et al., 2022), thus we utilize models that score specific type of relations (e.g., causal explanation) as a means to *explain* what the embeddings and models are able to capture. We evaluate four discourse relations relevant to anxiety. *Causal explanations* are a statement of why an event happened. Using the model of Son et al. (2018) with F1 of approximately .87 over social media, we computed the percentage of the messages written by a user that contain causal explanation. *Counterfactuals* imagine what could have happened as an alternative to actual events. Using the model of Son et al. (2017), we calculate the proportion of the messages from each user that communicates counterfactual thoughts. Finally, *dissonance* refers to situations in which one's stated behavior or belief contradicts a prior belief; *consonance* is its opposite concept. We use the RoBERTa-based topic-independent classifier that evaluates whether a pair of messages composes dissonance (Varadarajan et al., 2022, 2023). Instead of assessing all pairs, we take two temporally adjacent messages (maximum distance of 2) to reduce computation time.

## 4 Dataset

Our primary dataset comprises 12,489 Facebook users who took a personality questionnaire, including assessment of anxiety, and consented to share their status updates for academic research (Stillwell and Kosinski, 2012). The anxiety assessment consists of the anxiety facet of the neuroticism factor (Johnson, 2014), which has shown to correlate with other measures of anxiety such as GAD-7 (Milić et al., 2019) and STAI (Teachman, 2006) as well as have high convergence with anxiety disorders themselves (Rector et al., 2012). Each user was asked the following five questions: *Get stressed out easily*, *Am not easily bothered by things* (inverse coded), *Am relaxed most of the time* (inverse coded), *Fear for the worst*, *Worry about things*. Users responded

[2] We also experimented with min, max, and combinations of the three as well as alternative compositions but found no benefit. Given we are focused primarily on integrating discourse-level information, we suggest future work explore more sophisticated aggregation and compositional methods.

[3] Using a single hidden layer of size 32 with *tanh* activation trained with a learning rate of 5e-3 and batch size of 500 users; Code available here: `https://github.com/swaniejuhng/lexico-discourse/`

| Inputs | MSE | MAE | $r_{dis}$ |
|---|---|---|---|
| sentiment lexicon | .799 | .722 | .110 |
| PB (Phrase-BERT) | .726 | .688 | .430 |
| SB (Sentence-BERT) | **.725** | **.686** | **.438** |
| DiscRE | .751 | .704 | .382 |

Table 1: Evaluation of baseline (sentiment lexicon) and our three discourse-level models. **Bold** represents best in column.

| Inputs | MSE | MAE | $r_{dis}$ |
|---|---|---|---|
| BERT L23 | .720 | .682 | .452 |
| BERT L21-24 | .717 | **.679** | .446 |
| **RoBERTa L23** | .717 | .683 | **.458** |
| RoBERTa L21-24 | **.714** | .680 | .453 |

Table 2: Performance of level-level representations (i.e., contextual word embedding models). We use standard extraction techniques for these models (second-to-last hidden layer and concatenation of top-4 hidden layers). **Bold** represents best in column.

on 1-5 Likert scales ("Very inaccurate." to "Very accurate."). The responses to these questions are averaged together to form a continuous variable which determines the degree of anxiety.

**Secondary Evaluation Data.** We also include an evaluation using another smaller dataset that was collected by the authors. It was collected from consenting participants and asked the same facet of anxiety questions. In this case, only the past 2 years of Facebook posts were used to build representations of each user to be used for prediction. This dataset is used only for evaluation, where training occurs over the previously described large Facebook set.

## 5 Results and Discussion

We evaluate our models by disattenuated Pearson correlation coefficient $r_{dis}$ (Spearman, 1987; Lynn et al., 2018) between the model predictions and anxiety scores derived from the survey as our main metric, but include mean squared error as well.

Table 1 displays the performances of the models trained solely on discourse-level representations as well as a sentiment lexicon baseline model (Mohammad and Turney, 2013). Models utilizing Phrase-BERT or Sentence-BERT yielded decent results, while the DiscRE-based is by itself somewhat less informative.

| Inputs | MSE | MAE | $r_{dis}$ |
|---|---|---|---|
| RB L23 | .717 | .683 | .458 |
| RB L23 + PB | .715 | .682 | .456 |
| RB L23 + SB | .711 | .680 | .466* |
| RB L23 + DiscRE | .714 | .681 | .464* |
| RB L23 + SB + PB | .712 | .680 | .462 |
| RB L23 + PB + DiscRE | .712 | .681 | .461 |
| **RB L23 + SB + DiscRE** | **.707** | **.678** | **.473*** |
| RB L23 + PB + SB + DiscRE | .710 | .679 | .465 |

Table 3: Final evaluation using our best lexical- and discourse- embeddings as an ensemble. **Bold** represents best in column. * indicates significant ($p < .05$) improvement over RB L23 model according to paired t-test on error.

| Inputs | MSE | MAE | $r_{dis}$ |
|---|---|---|---|
| base: mean | **.352** | **.486** | .0 |
| base: sentiment | .905 | .838 | .131 |
| RB L23 | 1.103 | .937 | .421 |
| **RB L23 + SB + DiscRE** | 1.047 | .912 | **.496** |

Table 4: Evaluation of our model on a different dataset. **Bold** represents best in column.

Table 2 compares BERT and RoBERTa using the embeddings from the second-to-last hidden layer (L23) and the top-4 hidden layers (L21-24). We choose the RoBERTa L23 embeddings to represent the performances of the contextual embeddings in the following experiments.

While Phrase-BERT performs well in isolation, Table 3 suggests utility did not increase when used alongside RoBERTa. Alternatively, the model that employed RoBERTa, Sentence-BERT, and DiscRE representations achieves the best performance among all. This implies the two discourse-level embeddings have non-overlapping utility that contextual embeddings lack.

In Table 4, we verified the performance of our models on the alternate, held-out Facebook dataset as described in Section 4. Our central finding, that utilizing discourse-level semantics improves performance, is replicated in this entirely new dataset with the model having RoBERTa L23 with Sentence-BERT and DiscRE having significantly lower error. The improvement is similar to the first dataset showing the generalization of our approach.

**Explaining Discourse Improvement.** We shine light on what the model is able to capture in terms of discourse-level information by finding whether theoretically-related dimensions of cognitive distortions are associated with the models'. Table 5

| **Discourse relation type** | Cohen's $\boldsymbol{d}$ |
|---|---|
| causal explanation | .695 |
| counterfactuals | .227 |
| dissonance | .229 |
| consonance | .231 |

Table 5: Association of theoretically related features, depicting how much our best model is picking up on each type of discourse relation. This depicts how specific discourse features are related to user-level anxiety and the type of discourse information that the open vocabulary embeddings can capture.

shows the Cohen's $d$ which was computed using the following equation,

$$d = \zeta_{high}\left(\frac{\text{posts}_{rel}}{\text{posts}_{all}}\right) - \zeta_{low}\left(\frac{\text{posts}_{rel}}{\text{posts}_{all}}\right) \quad (3)$$

*high* and *low* each indicates the group of users with predicted degree of anxiety higher or lower than median, and $\zeta$ is the "z-score" (mean-centered, standardized) of the proportions per user.

We see that all discourse dimensions were related to the score, but causal explanations, often related to overgeneralization, had the highest difference (e.g., "You know life is going to be permanently complicated when your in-laws start turning their backs on you like a domino effect."). This suggests that the causal explanation discourse relation may account for unique information to improve the overall results.

**Potential for Use in Practical Applications.** Other than use in medical settings, secondary use cases of our models include assessments from public entities such as public health officials, schools, and human resource department of companies to quantify levels of expressed anxiety.

## 6 Conclusion

Anxiety is one of the most prevalent mental health disorders, and the ability to more accurately assess it in a way that can capture cognitive distortions (i.e., via discourse-level features) could lead to improved diagnostics and treatment of the condition. We analyzed the effects of using both discourse- and lexical-level information within a single model for the assessment of degree of anxiety from Facebook status updates. We found benefit from the discourse-level information beyond lexical-level contextual embeddings (i.e., transformer language models) that have been found to produce state-of-the-art results for other mental health assessment tasks, motivating the idea that anxiety-based models can benefit from capturing not only contextual lexical information but also higher-level semantics at the level of thought patterns. Lastly, we examined the effect of theoretically relevant discourse relations in assessing anxiety, discovering that causal explanation is the most informative.

## 7 Ethics Statement

Our work is contributing to an area of research that requires valid assessments of mental health to robustly evaluate the progress the new approaches can make in order to ultimately improve mental health assessment (De Choudhury et al., 2013; Coppersmith et al., 2018; Zirikly et al., 2019; Son et al., 2021). The intention of this work for its stakeholders at this point in time, clinical psychology and the interdisciplinary area of NLP and psychology, is its use toward developing more accurate and validated techniques for the benefit of society and human well-being.

We view this work as a step toward an assessment tool that could be used alongside professional oversight from trained clinicians. In this interdisciplinary work, we aim to improve the state-of-the-art automatic assessment models. However, at this time, we do not enable use of our model(s) independently in practice to label a person's mental health states. Clinical diagnosis requires more information such as interviews and physical examinations in addition to surveys. In addition, use of such models for targeted messaging or any assessment based on private language without author consent is prohibited among our terms of use. This research has been approved by an independent academic institutional review board (IRB).

Before our models are used by trained clinicians, they must demonstrate validity in a clinical setting for the target clinical population. The study steps for said evaluation should be reviewed by an external ethical review board, and practice should follow clinical guidelines. Unlike an invasive medical device, the majority of measures used in psychiatry are not required to go through regulatory agency reviews (e.g., through the Food and Drug Administration (FDA) in the U.S.), but rather are indicated based on clinical practice guidelines after reliability and validity of these measures have been established in a large body of research. If future use cases of this technique seek to apply it as a marker or indicator for a specific condition, they may seek that the U.S. FDA officially declare it as a biomarker of the condition.

## 8 Limitations

This work has several key limitations. First, we have relied on evaluation against self-reported (questionnaires) assessment of anxiety. Self-reporting the degree of anxiety on a survey instrument is not entirely dependable in diagnostic accuracy. However, it has shown reliable associations with diagnoses, serving clinical assessment treatment purposes beyond diagnosis (Kroenke et al., 2001). For example, anxiety scores from self-reported surveys have been robustly associated with consequential real-world outcomes such as mortality (Kikkenborg Berg et al., 2014). Clinical evaluation of the assessments proposed in this work should be evaluated against clinical outcomes.

Furthermore, the sample may not fully reflect the language use of the general population as it is skewed towards young and female[4] and only focused on English spoken by those from the U.S. and U.K., although previous work suggests this dataset contains a diverse representation of socioeconomic status (Matz et al., 2019). Additionally, we do not focus on actual utilization of discourse relations in assessing anxiety, as the scope of this work limits us to showing the viability of modeling anxiety on a continuous scale and the importance of discourse information towards modeling it. Lastly, the strong associations of theoretical discourse relations come from models that themselves are not perfect, with F1 scores ranging from 0.770 for counterfactuals to 0.868 for causal explanations, though one might expect this error to lead to underestimates of correlation with anxiety.

With NLP increasingly working towards better human-focused applications (e.g., improving mental health assessment), we are presented with increasing considerations for human privacy as a trade-off with considerations for open data sharing. In this case, the data used was shared with consent only for academic research use. Open sharing of such data violates trust with research participants (and agreements with ethical review boards). These and additional issues are discussed at length in Benton et al. (2017a). While it would be ideal to release everything and preserve privacy, in this situation, we believe the fact that the unprecedented data is not universally available suggests an imperative for those with access to openly share our work as best possible within ethical guidelines. We are thus releasing aggregated anonymized features from the secondary evaluation dataset that allows one to qualitatively replicate the associations in our results while preserving the privacy of participants.

[4]The self-reported user age averaged 22.6 (SD 8.2), and over half (58.1%) marked their gender as female.

## References

Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4):458–466.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Setiyo Budiyanto, Harry Candra Sihombing, and Fajar Rahayu IM. 2019. Depression and anxiety detection through the closed-loop method using dass-21. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 17(4):2087–2097.

Franziska Burger, Mark A. Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: Extracting schemas from thought records. *PLOS ONE*, 16:1–24.

Scott D Cohen, Daniel Cukor, and Paul L Kimmel. 2016. Anxiety in patients treated with hemodialysis. *Clinical Journal of the American Society of Nephrology*, 11(12):2250–2255.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *Eighth international AAAI Conference on Weblogs and Social Media*.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical*

*Psychology*, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

C. Lindsay DeVane, Evelyn Chiao, Meg Franklin, and Eric J Kruep. 2005. Anxiety disorders in the 21st century: Status, challenges, opportunities, and comorbidity with depression. *AJMC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level nlp: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI Conference on Web and Social Media*, volume 13, pages 236–246.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49. Big data in the behavioural sciences.

Holly Hazlett-Stevens and Michelle G. Craske. 2003. The catastrophizing worry process in generalized anxiety disorder: A preliminary investigation of an analog population. *Behavioural and Cognitive Psychotherapy*, 31(4):387–401.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of Research in Personality*, 51:78–89.

Simona C Kaplan, Amanda S Morrison, Philippe R Goldin, Thomas M Olino, Richard G Heimberg, and James J Gross. 2017. The cognitive distortions questionnaire (cd-quest): validation in a sample of adults with social anxiety disorder. *Cognitive therapy and research*, 41(4):576–587.

Sean W Kelley and Claire M Gillan. 2022. Using language in social media posts to study the network dynamics of depression longitudinally. *Nature Communications*, 13(1):1–11.

Selina Kikkenborg Berg, Lau Caspar Thygesen, Jesper HASTRUP Svendsen, Anne Vinggaard Christensen, and Ann-Dorthe Zwisler. 2014. Anxiety predicts mortality in icd patients: results from the cross-sectional national copenhearticd survey with register follow-up. *Pacing and Clinical Electrophysiology*, 37(12):1641–1650.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

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*.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Nourane Mahdy, Dalia A Magdi, Ahmed Dahroug, and Mohammed Abo Rizka. 2020. Comparative study: different techniques to detect depression using social media. In *Internet of Things—Applications and Future*, pages 441–452. Springer.

Marija Maric, David A Heyne, Brigit M van Widenfelt, and P Michiel Westenberg. 2011. Distorted cognitive processing in youth: the structure of negative cognitive errors and their associations with anxiety. *Cognitive Therapy and Research*, 35(1):11–20.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammadzaman Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In

*Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 39–44.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2021. MeLT: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2959–2966, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandra C Matz, Jochen I Menges, David J Stillwell, and H Andrew Schwartz. 2019. Predicting individual-level income from facebook profiles. *PLOS ONE*, 14(3):e0214369.

Jakov Milić, Ivana Škrlec, Iva Milić Vranješ, Matea Podgornjak, and Marija Heffer. 2019. High levels of depression and anxiety among croatian medical and nursing students and the correlation between subjective happiness and personality traits. *International Review of Psychiatry*, 31(7-8):653–660.

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

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38, Minneapolis, Minnesota. Association for Computational Linguistics.

Elizabeth M. Muran and Robert W. Motta. 1993. Cognitive distortions and irrational beliefs in post-traumatic stress, anxiety, and depressive disorders. *Journal of Clinical Psychology*.

David Owen, Jose Camacho Collados, and Luis Espinosa-Anke. 2020. Towards preemptive detection of depression and anxiety in twitter. In *Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Christopher Peterson. 1991. The meaning and measurement of explanatory style. *Psychological Inquiry*, 2(1):1–10.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Neil A Rector, Robert Michael Bagby, Veronika Huta, and Lindsay E Ayearst. 2012. Examination of the trait facets of the five-factor model in discriminating specific mood and anxiety disorders. *Psychiatry Research*, 199(2):131–139.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

GA Roth. 2018. Global burden of disease collaborative network. global burden of disease study through 2017 (gbd 2017) results. *The Lancet*, 392:1736–1788.

Shoffan Saifullah, Yuli Fauziah, and Agus Sasmito Aribowo. 2021. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *arXiv preprint arXiv:2101.06353*.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pages 118–125.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60, Copenhagen, Denmark. Association for Computational Linguistics.

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

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez, and C. Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 508–512.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658, Vancouver, Canada. Association for Computational Linguistics.

Youngseo Son, Sean AP Clouston, Roman Kotov, Johannes C Eichstaedt, Evelyn J Bromet, Benjamin J Luft, and H Andrew Schwartz. 2021. World trade center responders in their own words: Predicting ptsd symptom trajectories with ai-based language analyses of interviews. *Psychological Medicine*.

Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. 2022. Discourse relation embeddings: Representing the relations between discourse segments in social media. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 45–55, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Charles Spearman. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471.

David Stillwell and Michal Kosinski. 2012. mypersonality project: Example of successful utilization of online social networks for large-scale social research.

Bethany A Teachman. 2006. Aging and negative affect: the rise and fall and rise of anxiety and depression symptoms. *Psychology and aging*, 21(1):201.

Yevhen Tyshchenko. 2018. Depression and anxiety detection from blog posts data. *Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia*.

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Luby, Christian C. Luhmann, and H. Andrew Schwartz. 2023. Transfer and active learning for dissonance detection: Addressing the rare-class challenge. In *Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Vasudha Varadarajan, Nikita Soni, Weixi Wang, Christian Luhmann, H. Andrew Schwartz, and Naoya Inoue. 2022. Detecting dissonant stance in social media: The role of topic exposure. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Association for Computational Linguistics.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

# A Appendix

| | | | | |
|---|---|---|---|---|
| i | hate | feel | so | sick |
| tired | i don't | i can't | anymore | me |
| i'm | my | hurts | sad | her |
| pain | she | wish | why | stupid |
| really | :( | want | alone | fucking |
| ugh | sleep | cry | feeling | i have |

Table 6: Top 30 Ngrams most associated with predicted anxiety score from our best model; extracted using DLATK (Schwartz et al., 2017).

For the main dataset, a 10-fold cross validation was used with a 9:1 split at the user-level for each fold on 11,773 users that wrote 2,077,115 messages, while 168,044 messages written by 716 users who took the full version of anxiety questionnaire were used for testing. Following the practice of Park et al. (2015) to ensure adequate representation of language, the test set also limited the users to those writing at least 1,000 words. On average, each user wrote approximately 180 messages, 298 sentences, and 581 clauses. The label of training subset has a mean of 2.983 and standard deviation of 0.915, whereas those of test set are 3.004 and 0.895.

The secondary evaluation dataset spans 165 users and 52,773 messages, the result of filtering for each user to have written 500 or more words total. Each user wrote around 320 messages, 674 sentences, and 1,045 clauses on average. The mean and standard deviation of the label are 3.769 and 0.593.

Table 6 shows Ngram (lexical-level) features associated with high scores: negative emotions ('hate', 'sick', 'tired', 'cry') as well as absolutes ('anymore') and negations ('I can't', 'I don't'). Notably, conjunctions are not present among the most distinguishing Ngrams, suggesting that many of the discourse relations are not explicitly signaled with connective words (e.g., "because", "while").

Although predicting anxiety as a continuous variable reflects recent work suggesting it should be treated on a spectrum, from a practical point of view, it is sometimes necessary to make a binary classification. We therefore evaluated classify-

| **Model** | F1 |
|---|---|
| baseline: most freq class | .354 |
| baseline: sentiment | .351 |
| **RB L23 + SB + DiscRE** | **.600** |

Table 7: Prediction accuracy for binary treatment of outcomes.

ing into low and high bins at the median (Table 7), showing that our model leveraging representations from RoBERTa, Sentence-BERT, and DiscRE again yields significant improvement compared to baseline and sentiment lexicon models.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☐ A1. Did you describe the limitations of your work?
*Left blank.*

☐ A2. Did you discuss any potential risks of your work?
*Left blank.*

☐ A3. Do the abstract and introduction summarize the paper's main claims?
*Left blank.*

☐ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☐ Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*Left blank.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Left blank.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Left blank.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Left blank.*

### C ☐ Did you run computational experiments?

*Left blank.*

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Left blank.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Left blank.*

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Left blank.*

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Left blank.*

**D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?**
*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Left blank.*