Unraveling Anxiety from Individual and Societal Perspectives using Artificial Intelligence

A Dissertation Proposal presented

by

Swanie Juhng

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

Stony Brook University

March 2025

Stony Brook University

The Graduate School

Swanie Juhng

We, the dissertation proposal committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation proposal.

H. Andrew Schwartz – Dissertation Advisor
Associate Professor, Department of Computer Science

Niranjan Balasubramanian – Chairperson of Defense
Associate Professor, Department of Computer Science

Yifan Sun
Assistant Professor, Department of Computer Science

Ryan Boyd
Assistant Professor, Department of Behavioral and Brain Sciences,
University of Texas at Dallas

This dissertation is accepted by the Graduate School

Eric Wertheimer

Dean of the Graduate School

Abstract of the Dissertation

Unraveling Anxiety from Individual and Societal Perspectives using Artificial Intelligence

by

Swanie Juhng

Doctor of Philosophy

in

Computer Science

Stony Brook University

2025

Anxiety disorders are characterized by persistent and excessive form of fear and worry that interferes with daily functioning, distinguishing it from the adaptive anxiety that helps individuals respond to challenges. Despite affecting millions worldwide and costing a significant public health burden, anxiety disorders still remain underdiagnosed than actual prevalence due to lack of understanding and stigmatization. Leveraging machine learning (ML) and natural language processing (NLP) approaches can help bridge this gap by enabling scalable and accessible mental health assessments, offering a data-driven understanding of anxiety from individual and societal perspectives, and shedding light on societal stigmas toward mental health conditions. At the same time, advancing ML and NLP techniques for anxiety research presents unique technical challenges, such as effectively modeling linguistic markers of anxiety and ensuring interpretability in mental health predictions.

This dissertation investigates anxiety from both individual and societal perspectives using artificial intelligence. First, we explore individual manifestations of anxiety through three methodological advancements: (1) integrating contextual and discourse-level embeddings to improve language-based anxiety prediction using Facebook posts and self-reported surveys; (2) enhancing cognitive dissonance detection in Twitter dataset with transfer learning and active learning; and (3) developing longitudinal representation learning approaches that achieve both predictive utility and interpretability of adolescent psychopathology. Finally, we propose a work that extends our analysis to societal dimension of anxiety by identifying and categorizing social norms expressed in large-scale social media data and examining their associations with anxiety. By combining data-driven methods with psychological insights, this work studies anxiety from various angles – capturing both individual experiences and societal influences – offering a step toward a more comprehensive understanding of its causes and manifestations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Anxiety disorders are characterized by persistent, excessive, and sometimes irrational fear and worry that interfere with daily functioning, distinguishing them from the adaptive anxiety that helps individuals respond to challenges. While moderate level of anxiety is often a natural response to potential threats and can play a functional role in motivating vigilance and preparation [129, 22], pathological anxiety is debilitating, leading to chronic distress and significant impairment in daily life [6]. Individuals diagnosed with generalized anxiety disorder (GAD), for example, struggle with uncontrollable worry, restlessness, irritability, and sleep disturbance [2].

Anxiety disorders are also a widespread public health concern, affecting 284 million people worldwide [135] and imposing an economic cost of approximately $46.6 billion annually in the U.S. alone [35]. Nonetheless, anxiety disorders are underdiagnosed compared to its prevalence [71] partly due to lack of understanding and stigmatization [18, 157].

Leveraging machine learning (ML) and natural language processing (NLP) approaches can address these gaps by enabling scalable and accessible mental health assessments [143], offering a data-driven understanding of anxiety from not only individual but also societal perspectives [37], and shedding light on societal stigmas toward mental health conditions [49]. At the same time, advancing ML and NLP techniques for anxiety research presents unique technical challenges, such as effectively modeling linguistic markers of anxiety and ensuring interpretability in mental health predictions [86].

Anxiety is not a monolithic experience; its causes and expressions vary widely across individuals, shaped by cognitive processes, personal histories, and environmental contexts [147]. This variability necessitates an approach that captures the diverse ways in which anxiety manifests at the individual level.

At the same time, anxiety does not exist in isolation – it is inevitably intertwined with social norms and external pressures. Individuals experiencing anxiety are often influenced by societal expectations, implicit behavioral rules, and perceived obligations. Understanding anxiety, therefore, also requires an exploration of the broader societal structures that contribute to its prevalence and intensity.

Motivated by this need for a comprehensive perspective, this dissertation examines anxiety from both individual and societal viewpoints using artificial intelligence. Specifically, Chapters 2, 3, and 4 focus on understanding **individual manifestations** of anxiety through language-based prediction using discourse-level information, cognitive dissonance detection, and longitudinal modeling of adolescent psychopathology. Chapter 5 proposes extending this investigation to the **societal level** by analyzing the role of social norms in shaping anxiety, leveraging social media as a lens to quantify these influences.

In Chapter 2, we studied language-based anxiety prediction of the Facebook users utilizing both lexical-level and discourse-level representation derived from their status updates. The primary clinical manifestation of anxiety is worry associated with cognitive distortions, which are likely expressed at the discourse-level of semantics. We investigated 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.

In Chapter 3, we systematically explored various active learning approaches to enhance detection of cognitive dissonance from Twitter data. Cognitive dissonance is a phenomenon that happens when two elements of cognition (i.e., thoughts, experiences, actions, beliefs) within a person do not follow one another or are contrary, and it is known to be highly correlated with anxiety disorders. While the phenomenon is common enough to occur on a daily basis and expressed in language, it is still relatively rare among the myriad of other relationships between beliefs that occur across random selections of linguistics expression, thus making the automatic detection of it a rare-class problem. Active learning has in general been proposed to alleviate such challenges, but choice of selection strategy, the criteria by which rare-class examples are chosen, has not been systematically evaluated. We proposed and investigated active learning solutions to the rare class problem of dissonance detection through utilizing models trained on closely related tasks and the evaluation of acquisition strategies, including a proposed *probability-of-rare-class* (PRC) approach. We performed these experiments for a specific rare class problem: collecting language samples of cognitive dissonance from social media. We found that PRC is a simple and effective strategy to guid annotations and ultimately imrpove model accuracy.

In Chapter 4, we presented two approaches towards longitudinal representational learning for adolescent mental health. Mental health onset at childhood or adolescence is common for certain disorder such as anxiety and depression, and they can significantly impact well-being across the lifespan. Understanding their developmental trajectories is critical for early diagnosis, intervention, and treatment. However, mental disorders are highly heterogeneous and often co-occur, necessitating longitudinal predictive models that not only identify conditions but also capture shared mental health trajectories. Existing longitudinal representation approaches mostly prioritize predictive accuracy at the expense of interpretability. We developed and compared two longitudinal and interpretable modeling approaches, a novel *Multi-wave Integration for Multi-domain Encoding* (MIME) and an adaptation of *Agglomerative Clustering* for longitudinal data, to derive

representations that balance predictive utility and interpretability from longitudinal mental health data. On the predictive side, both methods outperform baselines at forecasting development of depression, anxiety, ADHD, and substance / alcohol use disorders. Furthermore, we found our proposed two approaches to be interpretable based on human judgments of coherence over an intrusion task with random clustering as a baseline. By bridging predictive utility and interpretability, this work provides a foundation for more clinically meaningful longitudinal modeling of mental health conditions.

Finally, in Chapter 5 we propose an approach of obtaining categories of social norms, the variance in their trends across different ethnicities, time, and regions, and their association with anxiety using large-scale social media. Social norms serve as shared standards of acceptable behavior within social groups, regulating interactions and fostering societal stability. While adherence to norms can facilitate social cohesion, the pressure to conform may also induce anxiety, particularly when norms conflict with personal values or promote harmful behaviors. Although social psychology research has examined specific anxiety-inducing norms, there remains a gap in computational approaches that systematically analyze these norms and their impact on mental health. In this work, we propose a data-driven method to comprehensively explore social norms expressed in social media and their relationship to anxiety. Our approach focuses on social expectations imposed by the people exercising influence over a person and examines trends across ethnicities, time, and geographic regions in the United States. By leveraging machine learning and natural language processing techniques, this study aims to provide comprehensive insights into the role of various social norms in shaping anxiety.

By combining data-driven methods with psychological insights, this dissertation studies anxiety from various angles – capturing both individual experiences and societal influences – offering a step toward a more comprehensive understanding of its causes and manifestations.

1.1 Conclusion

Chapter 2

Understanding Language of Anxiety with Discourse-Level Information

2.1 Introduction

One of the key characteristics of anxiety disorders is cognitive distortion [114, 100], or an illogical reasoning in dealing with life events [68]. The primary window into such distortions is language, including one's own explanatory style – the way they reason about the occurrence of events [123].

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.*" [62]. 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) [125].

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

2.2 Related Work

Researchers have recently been turning to social media language as a potential alternative source for mental health assessment, investigating, e.g., depression [139, 11, 72], PTSD [29, 15, 154], and suicide risk [30, 110, 102]. Such an approach was also utilized in analyzing anxiety [144, 166, 53, 20, 120, 136]. Work towards this goal include Shen and Rudzicz [144] who attempted to classify Reddit posts into binary levels of anxiety by lexical features and Guntuku et al. [53] 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 [148, 21, 146], 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 [36] and RoBERTa [93] [110, 102]. Furthermore, Ganesan et al. [48] 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.

2.3 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* [132] is a variant of BERT that captures a whole sentence by optimizing for semantic similarity using siamese and triplet networks. *Phrase-BERT* [171] attempts to capture shorter phrasal semantics using contrastive learning with machine-generated paraphrases and mined phrases. Finally, *DiscRE* [151] captures representations of the *relationship* between discourse units (i.e., clauses rooted with a main verb) using a weakly supervised, multitask approach over bidirectional sequence models.

Lexical Embeddings Amongst potential options for state-of-the-art auto-encoder language models, we consider BERT [36] and RoBERTa [93]. 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 [102, 48]. Beyond the fact that these models have lead to state-of-the-art performance in language understanding tasks, they

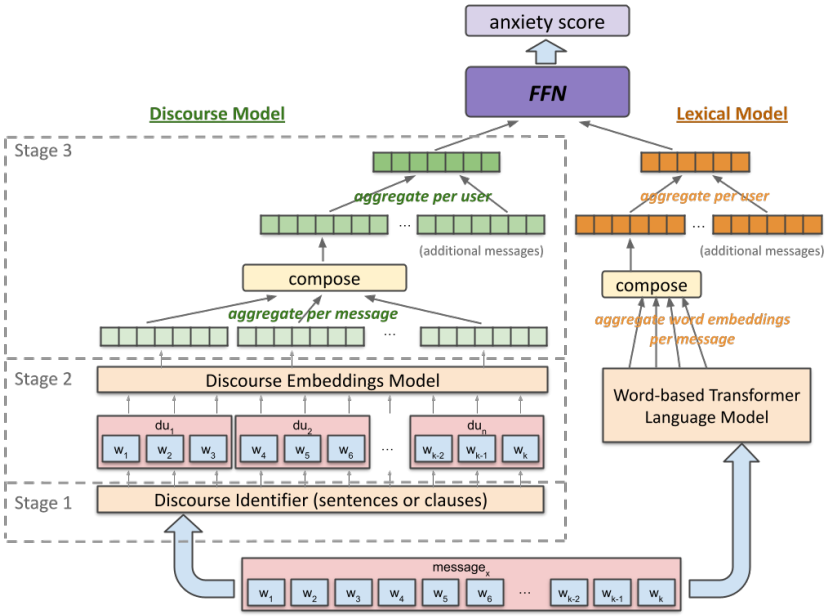


FIGURE 2.1: General architecture used for our anxiety assessment model. Depending on the model used, discourse units may be sentences or single clauses rooted by a main verb. The right-hand side, lexical model, follows the same approach as Ganesan et al. [48] and Matero et al. [101] for state-of-the-art assessment from contextual word embeddings.

are also known to capture *some* discourse information [78, 92]. 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 2.1. 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. 2.1), which are aggregated into message-level and then into a user-level embedding, e_u (Eq. 2.2):

$$e_u^m = \text{compose}_{d \in m}(e_m^d) \tag{2.1}$$

$$e_u = \text{compose}_{m \in u}(e_u^m) \tag{2.2}$$

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

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 Ganesan et al. [48] and Matero et al. [101] 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 [151], 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, Bayas, and Schwartz [150] 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. [153], 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 [168, 169]. Instead of assessing all pairs, we take two temporally adjacent messages (maximum distance of 2) to reduce computation time.

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

2.4 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 [159]. The anxiety assessment consists of the anxiety facet of the neuroticism factor [66], which has shown to correlate with other measures of anxiety such as GAD-7 [107] and STAI [164] as well as have high convergence with anxiety disorders themselves [131]. 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 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.

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. [121] 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.

Secondary Evaluation Data We also include an evaluation using another smaller dataset that was collected by the authors. It was collected from consenting participants

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 2.2: Performance of lexical-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.

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.

This 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.

2.5 Results and Discussion

We evaluate our models by disattenuated Pearson correlation coefficient r_{dis} [155, 94] between the model predictions and anxiety scores derived from the survey as our main metric, but include mean squared error as well.

Table 2.1 displays the performances of the models trained solely on discourse-level representations as well as a sentiment lexicon baseline model [109]. Models utilizing Phrase-BERT or Sentence-BERT yielded decent results, while the DiscRE-based is by itself somewhat less informative.

Table 2.2 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

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 2.3: 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 2.4: Evaluation of our model on a secondary evaluation dataset. **Bold** represents best in column.

L23 embeddings to represent the performances of the contextual embeddings in the following experiments.

While Phrase-BERT performs well in isolation, Table 2.3 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 2.4, we verified the performance of our models on the alternate, held-out Facebook dataset as described in Section 2.4. 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.

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 2.5: Top 30 Ngrams most associated with predicted anxiety score from our best model; extracted using DLATK [138].

Table 2.5 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 classifying into low and high bins at the median (Table 2.6), showing that our model leveraging representations from RoBERTa, Sentence-BERT, and DiscRE again yields significant improvement compared to baseline and sentiment lexicon models.

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' prediction. Table 2.7 shows the Cohen's d which was computed using the following equation,

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

TABLE 2.6: Prediction accuracy for binary treatment of outcomes.

Discourse relation type	Cohen's ***d***
causal explanation	.695
counterfactuals	.227
dissonance	.229
consonance	.231

TABLE 2.7: 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.

$$d = \zeta_{high}\left(\frac{\text{posts}_{rel}}{\text{posts}_{all}}\right) - \zeta_{low}\left(\frac{\text{posts}_{rel}}{\text{posts}_{all}}\right) \tag{2.3}$$

high and *low* each indicates the group of users with predicted degree of anxiety higher or lower than median, and ζ 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.

2.6 Conclusion

The ability to more accurately assess anxiety 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.

2.7 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 [34, 31, 181, 154]. 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.

2.8 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 [83]. For example, anxiety scores from self-reported surveys have been robustly associated with consequential real-world outcomes such as mortality [76]. 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 [103]. Additionally, we do not focus on actual utilization of discourse relations in assessing anxiety, as the scope of this work

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

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, Coppersmith, and Dredze [14]. 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.

2.9 Limitations

We use RoBERTa-base models trained on a single 12GB memory GPU (we used a NVIDIA Titan XP graphics card) for our experiments. Obtaining annotations for cognitive dissonance are limited by the availability of annotators and is not easily scalable in crowdsourcing platforms due to the required training and expertise in identifying dissonance. Due to this limitation, only two iterations of the AL loop for each setting were feasible for experiments. The transfer learning experiments in this paper were limited to two similar tasks, but there might be other tasks that could further improve or exceed the zero-shot performance of the models to cold start the active learning.

We focus on fine-tuning and active learning selection strategies to improve performance of rare-class classification for a specific task: dissonance detection across discourse units. Therefore, further work would be necessary to determine if the findings extend to other tasks. Additionally, the results may be different for other languages or time intervals of data collection. The performance of the neural parser on splitting tweets into discourse units can produce parses that are imperfect but the annotators and our systems worked off its output regardless to keep the process consistent. An improved discourse parser may also lead to improved annotator agreement and/or classifier accuracy. The dataset that we release from this paper, which contains labels of expressions of some cognitive states, was constructed using criteria that may not be fully objective.

2.10 Ethics Statement

The dataset for annotation was created from public social media posts with all usernames, phone numbers, addresses, and URLs removed. The research was approved by an academic institutional ethics review board. All of our work was restricted to document-level information; no user-level information was used. According to Twitter User Agreement, no further user content is required to use the publicly available data.

The detection of dissonance has many beneficial applications such as understanding belief trends and study of mental health from consenting individuals. However, it also could be used toward manipulative goals via targeted messaging to influence beliefs potential without users' awareness of such goals, a use-case that this work does not intend. Further, while we hope such models could be used to help better understand and assess mental health, clinical evaluations would need to be conducted before our models are integrated into any mental health practice.

Chapter 3

Enhancing Cognitive Dissonance Detection with Transfer and Active Learning

3.1 Introduction

Cognitive dissonance occurs during everyday thinking when one experiences two or more beliefs that are inconsistent in some way [57]. Often expressed in language, dissonance plays a role in many aspects of life, for example affecting health-related behavior such as smoking [24] and contributing to the development of (and exit from) extremism [33]. It is also known to have high correlation with anxiety disorders [162]. However, while the phenomenon is common enough to occur on a daily basis, dissonance is still relatively rare among the myriad of other relationships between beliefs that occur across random selections of linguistic expressions and thus makes the automatic detection of it a rare-class problem.

Despite recent advances in modeling sequences of words, rare-class tasks – when the class label is very infrequent (e.g., $< 5\%$ of samples) – remain challenging due to the low rate of positive examples. Not only are more random examples necessary to reach a substantial amount of the rare class (e.g., 1,000 examples to reach just 50 examples), but also it is easy for human annotators to miss the rare instances where dissonance is

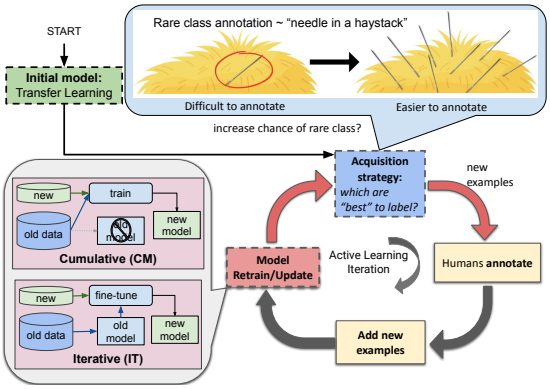


FIGURE 3.1: Demonstration of the active learning (AL) loop in general. Our paper examines the three highlighted steps: (i) Bootstrapping with TL model, (ii) Acquisition strategy, and (iii) Model update.

present. Here, we develop and address the challenges of creating a resource for language-based assessment of dissonance.

Active learning using large language models presents both new opportunities and challenges. On the one hand, large language models (LLMs) offer unmatched representations of documentations, able to achieve state-of-the-art language understanding task performance with transfer learning, often only with a few iterations of fine-tuning [93]. On the other hand, representations are high-dimensional, and models trained or fine-tuned with only a small number of examples are prone to overfitting, especially when there is a large class imbalance as in rare-class problems. While LLMs have enabled attempts to tackle increasingly complex semantic challenges across a growing list of tasks, getting annotated examples for such problems can become a bottleneck due to its time- and labor-intensiveness [175]. Since data-centric improvements for more novel tasks can provide a faster path than model-centric improvements [117], active learning can be a way forward to be both data-centric and address bottlenecks in label acquisition – it aims to reduce annotation costs as well as alleviate the training data deficiency that large language models face.

However, while active learning has been studied for multiple natural language tasks [145, 89], little is known about active learning acquisition strategies for LM-based approaches, especially for rare-class problems. *High data imbalance* coupled with *very less*

training data poses the challenge of "absolute rarity" [1], as in our task of dissonance detection. We address this problem by using a novel combination of evaluating the ordering of transfer learning from similar tasks to cold-start the active learning loop, and by acquiring with a relatively simple acquisition strategy focused on *probability-of-rare-class* (PRC) to increase the rare class samples.

Our contributions include: (1) a novel systematic comparison of five common acquisition strategies for active learning for a rare class problem[1]; (2) a systematic comparison of two different approaches to handling AL iterations for LLMs – cumulative and iterative fine-tuned model updates – finding the cumulative approach works best; (3) evaluating annotation costs of a rare-class task, finding that minimum annotation cost does not necessarily lead to better models, especially in realistic scenarios such as *absolute rarity*; and (4) release of a novel dataset[2] for the task of identifying cognitive dissonance in social media documents.

3.2 Related Work

Active learning in NLP has been largely studied as a theoretical improvement over traditional ML for scarce data. In this work, we specifically investigate *pool-based* active learning, or picking out samples to annotate from a larger pool of unlabeled data, and particularly data for a *rare-class* problem where LMs are not well-understood yet.

Acquisition Strategies Sampling strategies for active learning can be broadly classified into three: uncertainty sampling [142, 170, 116], representative (or diversity) sampling [28, 140, 51], and the combination of the two [179]. The uncertainty sampling strategies that employ classification probabilities, Bayesian methods such as variational ratios [45], and deep-learning specific methods [64] often use epistemic (or model) uncertainty. We choose maximum entropy to represent the uncertainty sampling, since it is

[1]Code: https://github.com/humanlab/rare-class-AL
[2]Dataset: https://github.com/humanlab/dissonance-twitter-dataset

usually on par with more elaborated counterparts [165]. As a popular diversity sampling baseline to compare against, we pick select CoreSet [140]. The state-of-the-art methods combine these two strategies in novel ways, such as using statistical uncertainty in combination with some form of data clustering for diversity sampling [180, 7]. Our work uses Contrastive Active Learning [99] to represent this strategy.

On the other hand, Karamcheti et al. [69] and Munjal et al. [113] claim there is rather small to no advantage in using active learning strategies, because a number of samples might be collectively outliers, and existing strategies contribute little to discover them and instead harm the performance of subsequent models. Researchers recently have also focused on the futility of complex acquisition functions applied to difficult problems and argued that random acquisition performs competitive to more sophisticated strategies, especially when the labeled pool has grown larger [140, 38]. Furthermore, a large-scale annotation of randomly sampled data could be less expensive than ranking data to annotate in each round of active learning, if there is not much advantage (i.e., such as capturing rare classes) in using a specific strategy.

Rare Class AL There has been a growing number of applications of active learning in data imbalance and rare class problems. Such works include [81, 25, 40] which proposed frameworks to improve model performance with data imbalance but failed to check the feasibility and costs in a real-world, active annotation setting where not only is rare class very infrequent (4%) but very few (< 70) examples of the rare class exist due to small dataset size ("absolute rarity"). They also fail to compare against a simple, rare class probability of the model. While some work in the pre-LLM era use probability outputs of a classifier (certainty-based sampling) which is similar to the proposed PRC, they claim to work better in conjunction with co-selection using other uncertainty sampling strategies, and that certainty-based sampling alone performs poorly in terms of increasing rare-class samples [88]. Many studies also focus on rare class *discovery*, or finding outlying samples that do not fall under the existing categories [63, 55, 60]. This is different from our task

which focuses on the *detection* of a rare class.

3.3 Task

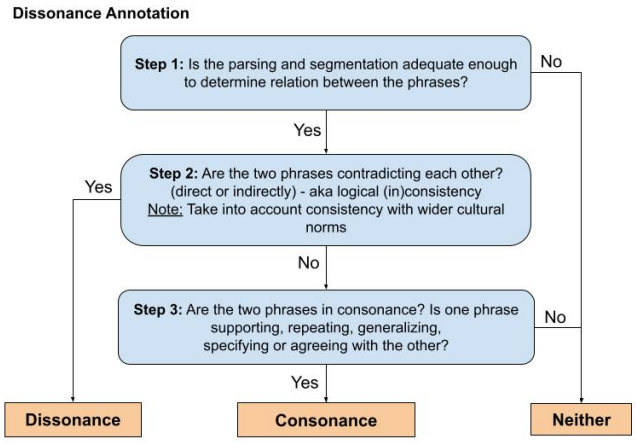
Cognitive dissonance is a phenomenon that happens when two elements of cognition (i.e., thoughts, experiences, actions, beliefs) within a person do not follow one another or are contradictory, and consonance is when one belief follows from the other [58]. Cognitive dissonance raises psychological discomfort, encouraging a person to resolve the dissonance. As the magnitude of dissonance increases, the pressure to resolve it grows as well [59, 104].

Social psychology has used this human tendency to resolve dissonance to understand important psychological processes such as determinants of attitudes and beliefs, consequences of decisions, internalization of values, and the effects of disagreement among persons [58]. Dissonance is also related to anxiety disorders [67], relevant to understanding extremism and predicting cognitive styles of users. Our approach to annotating cognitive dissonance on social media is motivated by the two-stage annotation approach described in [168]. To the best of our knowledge this is the first social media dataset for cognitive dissonance.

3.4 Methods

3.4.1 Annotation and Dataset

Following the definition of cognitive dissonance in §3.3, we treat discourse units as semantic elements that can represent beliefs. A discourse unit consists of words or phrases that have a meaning [126]– and then cognitive dissonance is analogous to a discourse relation between two discourse units. Recent work [152] represents discourse relations in a continuous vector space, motivating us to look at cognitive dissonance, too, as a relationship between two "thought" discourse units.



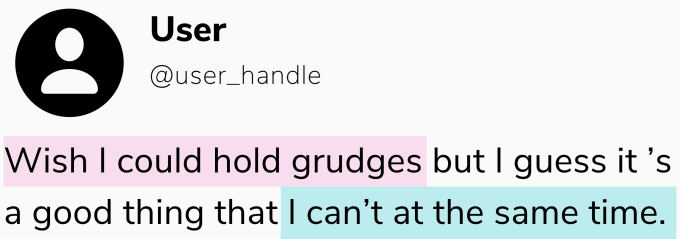


FIGURE 3.2: Above: Flowchart describing the steps for the annotators to label tweets as DISSONANCE, CONSONANCE, or NEITHER.
Below: An example of a pair of THOUGHT segments in a tweet annotated as dissonance.

We build a dissonance dataset by first sampling posts between 2011 and 2020 on Twitter. The tweets were parsed into discourse units using the parser by Wang, Li, and Yang [172] which uses the PDTB framework. [3]

Each discourse unit in a document is initially annotated into THOUGHT or OTHER. [4] A THOUGHT is a discourse unit describing the author's own beliefs, experiences and actions and are potential elements to be in dissonance. OTHER comprises of anything else, from meaningless phrases to coherent beliefs that belong to someone other than the author. For the annotation of dissonance, pairs of THOUGHT units from each tweet are extracted, and then annotated to compose CONSONANCE, DISSONANCE or NEITHER according to the framework described in Figure 3.2 – a three-class annotation. This framework was developed from annotator training to spot examples of dissonance, followed by discussion with a cognitive scientist.

Among a random selection of tweets, the natural frequency of the DISSONANCE class is around 3.5%. The annotations were carried out by a team of three annotators, with the third annotator tiebreaking the samples disagreed by the first two annotators.

Initial Set ($iter_0$) This dataset is used to select the best transfer model to effectively cold-start the AL loop. We start with a total of 1,901 examples of dissonance task annotations, which we split into a training set of 901 examples (henceforth, $iter_0$) with 43 examples of dissonance (4.77%) picked randomly from discourse-parsed tweets. We create initial development and test sets with 500 examples each. They were created such that all the THOUGHT pairs that were a part of a single tweet belong to the same set.

Final Development and Test Datasets We gather additional 984 annotations for development set and 956 annotations for test set in addition to the previously mentioned

[3] PDTB [128] and RST [98] are the two major frameworks for discourse parsing; we use the former for this work since PDTB is lexically grounded and identifies discourse relations using lexical cues. While the RST framework could be helpful since rhetorical relations are viewed as cognitive entities [163], the complex relationships defined with RST's nested structures can complicate our search for cognitive dissonance samples at the preliminary stage of data collection.

[4] This was a simpler, large-scale annotation to pick out discourse units describing author's own beliefs. We do not go into the details of this specific annotation since it is not pertinent to this work.

500 for each, summing up to 1,484 development examples (*dev*) and 1,456 test examples (*test*) with around 10% dissonance examples in each, to account for increased frequency of occurrence of the rare class after incorporating novel acquisition strategies.

3.4.2 Modeling

Architecture

A RoBERTa-based dissonance classifier is used consistently across all the experiments in this paper: for any two THOUGHT segments belonging to a single post, the input is in the form of "[CLS] $segment_1$ [SEP] $segment_2$ [SEP]". We take the contextualized word embedding $\mathbf{x} \in \mathbb{R}^d$ of [CLS] in the final layer and feed it into the linear classifier: $y = \text{softmax}(W\mathbf{x} + \mathbf{b})$, where $W \in \mathbb{R}^{d \times 2}, \mathbf{b} \in \mathbb{R}^2$ is a learned parameter. We trained the model parameters with cross entropy loss for 10 epochs, using AdamW optimizer with the learning rate of 3×10^{-5}, batch size of 16, and warm up ratio of 0.1. To avoid overfitting, we use early stopping (patience of 4) with the AUC score. We run the AL experiments on the datasets delineated in §3.4.1. While the annotations are for three classes (Figure 3.2), the models used for AL across all strategies classify labels to binary level (dissonance or not dissonance), as we are focused specifically on the dissonance class – while dissonance is rare, it is also essential to perform well in detecting this class.

AL strategies

Since our annotation process brought about only a small incremental improvement for performance on the rare class, yet contributed much to modeling the dominant classes, we hypothesized that using probability of the rare class as an acquisition strategy in active learning could work just as well as other strategies that are based on diversity and uncertainty sampling. We ran our analyses over four other common acquisition strategies by picking out the top 10% (300 out of an unannotated data pool containing

3,000 examples). We limit to only four other strategies because of the annotation costs and limited time.

Prc For a rare, hard class, we use a binary classifier that outputs the probability of rare class learned from the samples encountered so far. This is a computationally inexpensive and simple method that could be easily surpassed by other complex AL strategies but was surprisingly found to be the most effective in this study. The examples from the data pool that are predicted to have the highest probability of the rare class by the classification model from previous iteration are selected.

Random As a baseline, we randomly sample examples from the data pool, which reflects the natural distribution of classes. Random method has been considered to be a solid baseline to compare against, as many AL strategies do not merit when the annotation pool scales up and collective outliers are missed, as explained in §3.2.

Entropy We use predictive entropy as the uncertainty-based sampling baseline to compare against. While Least Confident Class (LCC) is a popular strategy to capture samples based on uncertainty, it is calculated based on only one class, working best for binary classification and provides merit within balanced classes, whereas predictive entropy is a generalized form of LCC, and a more popular variant [45].

Cal Contrastive Active Learning [99] is a state-of-the-art approach that chooses data points that are closely located in the model feature space yet predicted by models to have maximally different likelihoods from each other. This method is relevant to the task at hand because in rare class problems, it is often difficult for a model to learn the decision boundary around the rare class due to the low number of such samples. Thus we focus on a method that tries to pick out samples at the decision boundary of the rare class.

	Random	Entropy	CoreSet	Cal	Prc
Random	×	12.15%	11.52%	10.83%	11.02%
Entropy		×	64.68%	76.33%	87.98%
CoreSet			×	58.67%	61.65%
Cal				×	82.98%

Table 3.1: % overlap in the samples picked out by the base model for the five strategies described in 3.4.2. Probability of rare class (PRC) has a significant overlap with a state-of-the-art approaches, implying that for the rare class problem, PRC is a computationally inexpensive, alternative acquisition approach.

CoreSet An acquisition method that has worked well as a diversity sampling method is CoreSet [140]. This method uses a greedy strategy to sample a subset of data that is most representative of the real dataset, i.e., the larger data pool that we sample from.

Model Update

To the best of our knowledge, the question of model update in an AL loop has not been explored. We explore two fine-tuning approaches to update the model following annotation of new samples in each round of the active learning loop – cumulative (Cm) and iterative (It). Figure 3.1 provides a visual explanation of the two approaches.

Cumulative (Cm) At each round of the AL loop, the 300 newly annotated samples are combined with the previous ones as the input to fine-tune the classification model from a base pretrained language model.

Iterative (It) At each round of the AL loop, the 300 newly annotated samples are used to further fine-tune the model trained during the previous loop.

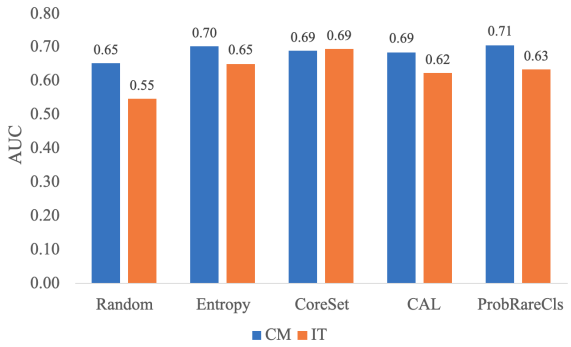


FIGURE 3.3: AUC for the five strategies for IT and CM model updates. This shows that the CM model update always performs equally with or better than the IT update.

	IT					CM				
Strategy	**F1-macro**	**F1-Dis**	**Prec-Dis**	**Rec-Dis**	**AUC**	**F1-macro**	**F1-Dis**	**Prec-Dis**	**Rec-Dis**	**AUC**
RANDOM	0.556	0.175	0.119	**0.336**	0.546	0.640	0.362	0.397	0.334	0.652
ENTROPY	0.632	0.351	0.401	0.318	0.650	0.649	**0.398**	0.540	0.315	0.702
CORESET	**0.652**	0.397	0.513	0.329	0.694	0.635	0.375	0.523	0.292	0.688
CAL	0.612	0.306	0.331	0.321	0.623	0.644	0.383	0.497	0.313	0.685
PRC	0.616	0.322	0.371	0.309	0.633	0.633	0.382	**0.580**	0.285	**0.706**

TABLE 3.2: Comparison of five annotation strategies for iterative (IT) and cumulative (CM) approaches for 2 class classification. The metrics are averaged over two iterations of active learning, with 300 new examples annotated in each iteration (adds between 3-10% samples of dissonance in each round, depending on the strategy). **Bold** represents the best for each reported metric. The performance of CM approach exceeds that of IT across most acquisition functions . While the performance on adding 10 to 30 samples of dissonance is not expected to cause large jumps in performance, note that using the PRC strategy leads to significant gain in performance in detecting the dissonance class.

3.5 Results

3.5.1 Acquisition Strategies

Table 3.1 shows the overlap of samples picked out in each iteration from the same larger data pool for the model at iteration 0 (base model). RANDOM has the lowest overlaps with all the other strategies. We also find that there is a significant overlap ($> 80\%$) in the samples between ENTROPY or CAL, the state-of-the-art approach, and PRC. CAL has a higher overlap with ENTROPY rather than CORESET, showing that samples deemed to be both highly informative and contrastive by the model are also usually likely to be

dissonant. This is contrastive to the prior literature revealing that poor calibration of large language models often renders the models to rarely be uncertain of their outcomes [54]. All strategies except RANDOM have $> 55\%$ overlap with each other. This implies that diversity- and uncertainty-based methods are not as different from each other as they theoretically are and inclined to pick similar samples – hinting that a lot of diversity-based sampling measures mostly pick highly informative samples as well. Furthermore, PRC tends to choose samples that the "state-of-the-art" model also picks in rare-case scenarios, indicating that it could be a computationally inexpensive alternative.

Table 3.2 shows the results averaged over two rounds of active annotation and learning for five strategies with two types of model updates. While the performance for dissonance class across all strategies do not seem to boost much in a single round of active learning (since adding 300 new annotations adds only between 10-30 dissonance examples in each round), Figure 3.3 shows that the CM approach always performs better than IT. IT could help models generalize to new domains during transfer learning, but it may not add a lot of value when data is collected in the same domain in each iteration of the AL loop. This could be because IT biases the model towards the distribution of the latest sample set due to the effects of catastrophic forgetting [176] while CM implicitly balances all batches of data.

The performance of RANDOM-CM strategy lags behind the rest of the CM strategies. The other strategies perform better than RANDOM but one strategy does not offer significant advantages over another, further confirming the observation from Table 3.1 that the AL strategies have a significant overlap and could be choosing very similar samples.

3.5.2 Qualitative Evaluation of Annotation Costs

Table 3.3 displays the results of a study on the quality of annotation, measuring subjective difficulty and time taken. We sampled 300 examples from a data pool of 3,000 unannotated examples for each strategy so that the experiment is consistent with the

unlabeled pool size used across other experiments for each of the strategies. Of these 300, we picked 125 (from each strategy) to get annotated for their difficulty on a scale of 0-5. This number was chosen based on balancing having enough examples per strategy for meaningful statistics while not taking too much of annotator's time and effort. The annotations were conducted on a simple annotation app that records the time taken to produce the first label an annotator decides on (i.e., any corrections to the label wouldn't count towards the time calculation). The Pearson correlation between the average time taken and the average difficulty value was 0.41.

	Rare %	Time (s)	Subj. diff.
RANDOM	3.20	11.96	-0.065
ENTROPY	6.80	12.78	0.035
CORESET	6.00	11.89	0.039
CAL	4.80	11.88	-0.045
PRC	**7.60**	**13.55**	**0.071**

TABLE 3.3: Evaluation of annotation difficulty by selection strategy. Rare % is how much the rare class (dissonance) was selected; Time is per instance and subj diff is z-scored subjective rating of difficulty. Our PRC approach selects the most rare class instances but also results in more costly annotations in terms of time and subjective ratings.

Annotation cost (in terms of time taken to annotate) is known to increase when employing active learning strategies compared to that of a random baseline [141]. We find that PRC picks out the "most difficult" samples, and takes almost a second longer to annotate than average (average time taken: 12.59s), followed by ENTROPY and CORESET strategies – this complies with ENTROPY picking the most uncertain samples and CORESET executing diversity sampling and representing the data better, thus increasing the number of dissonance samples. The subjective difficulty reported is the average z-score of difficulty scores picked by the annotators. This is done to normalize the variability of subjective ratings. The inter-rater reliability for the entire exercise was measured using the Cohen's κ for two annotators, which was calculated to be 0.37 (fair agreement), with an overlap of 66%.

In general, we found that PRC addresses the rare-class challenge better than the other AL strategies. We also found that both the ENTROPY and CORESET strategies substantially increase the number of dissonant examples, thus partially addressing the needle-in-haystack problem.

3.5.3 A final dataset: Putting it all together.

We release two versions of train data: small and big; along with the development and test data (see §3.4.1) The *small* set comprises the 2,924 examples which were use for the active learning experiments discussed previously. Building on our learnings from the active learning experiments, we created a second (*big*) data set with 6,649 examples that includes the small plus an additional 3,725 examples derived over more rounds of active learning restricted to the PRC or ENTROPY strategies. It contains 692 dissonant samples, comprising 10.40% among all.

3.6 Conclusion

In this work, we have systematically studied approaches to key steps of active learning for tackling a rare-class modeling using a modern large language-modeling approach. While transformer-based systems have enabled greater accuracies with fewer training examples, data acquisition obstacles still persist for rare-class tasks – when the class label is very infrequent (e.g., $< 5\%$ of samples). We examined pool-based active annotation and learning in a real-world, rare class, natural language setting by exploring five common acquisition strategies with two different model update approaches. We found that a relatively simple acquisition using the probability of rare class for a model could lead to significant improvement in the rare class samples. We also qualitatively analyzed the data samples extracted from each data acquisition strategy by using subjective scoring and timing the annotators, finding PRC to be the most difficult to annotate, while also remaining the best method to improve rare class samples and model performance. Our final dataset of

9,589 examples (*Big* train + dev + test) is made available along with an implementation of the PRC method and our state-of-the-art model for cognitive dissonance detection.

Chapter 4

Longitudinal Representation Learning for Adolescent Mental Health

4.1 Introduction

Mental health conditions experienced in childhood and adolescence exert significant influence on well-being throughout adulthood [112], with the most prevalent disorders such as anxiety, depression, and attention-deficit hyperactivity disorder (ADHD) often showing onset during adolescence [12, 43, 134]. However, mental disorders are heterogeneous [82] and often co-occurring as well as have complex trajectories of onset [80]. To better understand the onset of mental health conditions, models that can balance complexities with interpretability are needed.

In this work, we introduce and evaluate two longitudinal clustering techniques: (a) a novel *Multi-wave Integration for Multi-domain Encoding* (MIME) and (b) *Agglomerative Clustering*, to derive representations that can forecast and provide interpretability from longitudinal mental health data. We generate longitudinal representations from rigorously recorded psychological metrics and diagnoses of youths between ages 3 to 15, then examine the representations' utility to forecast diagnoses at age 18.

Importantly, we utilize the Stony Brook Temperament Study Dataset [79] containing standardized and rigorously assessed mental health conditions, which is novel for the domain of unsupervised learning. While many works in the field have utilized EHR-based

approaches, there is no standardization or consistency to the assessments and diagnoses in health records [149]. There is a variance in the quantity and quality of information clinicians collect, who provides that information, and what their diagnostic conclusions are. Semi-structured diagnostic interviews were developed to address this issue – to standardize the collection of information used to make diagnoses [10, 156]. The Stony Brook Temperament Study dataset follows this optimal practice, ensuring consistency and reliability in mental health assessments.

Our **main contributions** include: (1) the proposal of a novel task for generating longitudinal representations of mental health conditions; (2) the exploration of the ability of those derived representations to forecast future conditions; (3) the introduction of a longitudinal dataset of standardized mental health assessments [79] which avoids limitations of EHR data [10, 156], and (4) an interpretation of the suggested representations and their relation to existing psychological theories and research. We release both interpretable models openly for further research and literature on psychopathology.

4.2 Background

Adolescent Mental Health Mental health is shaped by a continuum of life experiences, with conditions experienced in childhood affecting later life [112]. Onset at childhood or adolescence is common for certain disorder such as anxiety [12], depression [43], ADHD [134], and conduct disorders [39]. Untreated conditions in youth could lead to serious outcomes and comorbidities, or co-occurrences of disorders, later in life [9], and hence early diagnosis and intervention of mental disorders for young populations is particularly critical.

Related Work Researchers have long attempted to predict mental health conditions of children or adolescents such as anxiety [23], depression [56], ADHD [97], and autism

[167] to name a few. Several works have further aimed for interpretability in their models [74, 115].

Also, various approaches were used in the area to understand or enhance prediction of mental disorders. Some have utilized dimensionality reduction methods such as PCA and t-SNE to understand mental health conditions such as Alzheimer's disease [16, 106], depression [111], ADHD [108], bipolar disorder [105], and trauma recovery [160]. The use of autoencoders has recently gained some attention among psychology researchers for extracting features to enhance personality assessments [178, 65] and suicidal ideation prediction [173]. Our proposed approach, MIME, also utilizes an autoencoder but is distinguished in that it structures longitudinal features into mental health domains and age waves, enhancing interpretability in relation to psychopathology.

Clustering has shown value for discovering mental health co-occurrences, such as the link between depressive and anxiety disorders [70] or bipolar disorder and personality disorders [47]. Several studies in health intelligence have focused on cognitive or mental health [91] but only for point-in-time classification rather than representation generation, clustering, or forecasting. Works that focus on *risk prediction* have an inherent longitudinal aspect since it quantifies the probability of future condition development [119]. Psychological works that studied representation learning focused on the clinical vocabulary (i.e., semantics) of comorbid conditions [19]. Our adaptation of Agglomerative Clustering sets apart in that it is applied first to samples and then to features, which not only enables grouping of similar mental health trajectories but also alignment of highly co-occurring conditions, ultimately deriving longitudinal representations with predictive utility.

TABLE 4.1: Counts and percent prevalence of subjects diagnosed with each condition at each age wave

Domain	**Age 3**	**Age 6**	**Age 9**	**Age 12**	**Age 15**	**Age 18**
Depression	6 (1%)	19 (4%)	6 (1%)	26 (6%)	52 (12%)	93 (22%)
Anxiety	77 (18%)	58 (13%)	95 (22%)	101 (23%)	86 (20%)	100 (23%)
ADHD	9 (2%)	25 (6%)	54 (13%)	70 (16%)	60 (14%)	47 (11%)
DBD	41 (10%)	36 (8%)	16 (4%)	22 (5%)	15 (%3)	×
AUD/SUD	×	×	×	×	×	41 (10%)

4.3 Dataset

We use data from the Stony Brook Temperament Study [79] which consisted of diagnoses across core domains of mental health conditions (e.g., depression, anxiety, attention-deficit hyperactivity disorders (ADHD), disruptive behavior disorders (DBD), alcohol / substance use disorders (AUD/SUD)) at 3-year intervals (i.e., ages 3, 6, 9, 12, 15, 18) alongside subject demographics like biological sex. Along the waves of collected data, we note that differing diagnostic measurements are taken. At ages 3 and 6 one parent of each subject was interviewed to generate the responses used for diagnoses. At ages 9, 12, and 15 both the child and parent were interviewed, and at age 18 only the child was interviewed. The original dataset has 650 subjects, from which we applied an inclusion criterion to ensure high-fidelity longitudinal data, stipulating that subjects must have records for at least 80% of the measurements taken in time. This resulted in a total of 431 subjects. The subjects' condition counts per age wave and the prevalence of each condition per age are shown in Table 4.1. Diagnoses of alcohol use disorder (AUD) and substance use disorder (SUD) were not conducted for participants at ages 3 and 6, since the earliest onsets are known to be mid-teens [2]. Disruptive behavior disorder (DBD) is only given for input and AUD/SUD is only used for testing, resulting in 20 input variables – 4 conditions at 5 age waves. Likewise, the comorbidities of the subjects are presented in Table 4.2.

TABLE 4.2: Co-occurrences within conditions at age 18

	Depression	**Anxiety**	**ADHD**	**AUD/SUD**
Depression	24 (26%)	58 (58%)	17 (36%)	22 (54%)
Anxiety		31 (31%)	20 (43%)	16 (39%)
ADHD			19 (40%)	9 (22%)
AUD/SUD				13 (32%)
Total	93	100	47	41

4.4 Method

Our approach to building longitudinal mental health representations is guided by two over-arching goals – interpretability and utility.

Interpretability, the ability to understand how a machine learning model makes its predictions, allows us to apply models in safety-critical problem domains [5] such as clinical healthcare research with human subjects. *Utility*, the accuracy of a model at its specified task, aims to obtain a single overall representation that effectively conveys information about mental health trajectories for prediction and forecasting.

We present MIME and Agglomerative Clustering, which are modeled to accomplish both goals. MIME is an autoencoder-based architecture, designed to analyze and output representations per domain and age wave separately. A non-negativity weight constraint [26] was adopted to address the challenge of interpreting the ambiguity of neural network weights. Agglomerative Clustering is a more traditional and intuitive approach that groups subjects with similar mental health trajectories and aligns relevant features together.

Imputation Missing values were imputed using the mode of each feature. Practically, this is equivalent to setting the missing values to "no diagnosis" since the majority of participants are not diagnosed with any condition.

MIME The architecture of MIME is described in Figure 4.1. In the *encoding* stage, the model assembles the normalized values of features that belong to each age wave or domain,

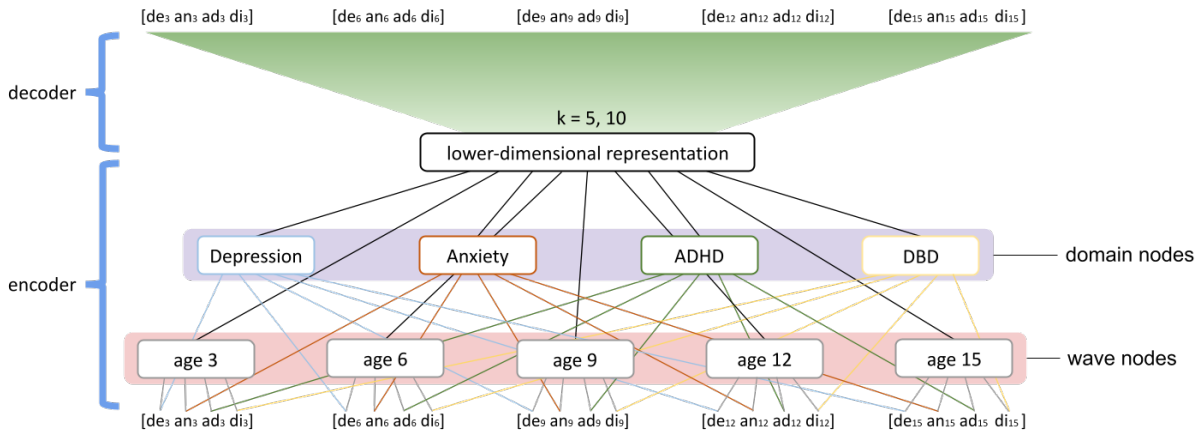


FIGURE 4.1: Architecture of the proposed *Multi-wave Integration for Multi-domain Encoding* (MIME). Wave nodes collect the patterns of each age wave, and domain nodes do the same for the corresponding conditions. The latent factor layer aggregates weights from the wave and domain nodes.

resulting in 9 unique factors – 5 waves and 4 domains. The values within each factor are normalized and then form a vector. The vectors pass through linear and activation layers and are reduced to a 1-D representation. Then the 9 values are concatenated to form an intermediate vector, and this vector again passes through the layers and is transformed into a k-dimensional representation. The *decoding* stage is the reverse of the encoding stage, where the goal is to minimize differences between the original input and reconstructed data. The 20 features are encoded into a lower-dimensional representation that effectively preserves essential information by organizing features into age waves and mental health domains.

Agglomerative Clustering We use Agglomerative Clustering with Euclidean distance as the metric and the *ward* linkage approach, which minimizes the variance within clusters during merging [122]. This method is applied in two stages – first on the samples, and then on the features. This is to bring together the samples with similar mental health trajectories and then put together the features that exhibit similar patterns across samples, hence facilitating interpretability.

Sample clustering is performed between the subjects by computing the Euclidean distance between the values of features, and we specify the number of the clusters so

that the unique mental health trajectories can be organized into a few patterns. The algorithm will merge the pairs of clusters that minimize this criterion. We then derive the mean of each value per feature within each cluster, which is visualized in the form of a heatmap.

Feature clustering is done between the features by computing the distance between their averaged values within each cluster derived from the first stage. Here we do not specify k, since the goal is to see which domains at which age waves are relevant to each other, not to arbitrarily group the features. This is displayed as a dendrogram that describes the iteration history of the clustering [90].

Outcome Control The difficulty of longitudinal forecasting tasks studying mental health conditions is partially attributable to their auto-correlation over time [127]. In other words, the past often functions as the best predictor of the future. Here, this phenomenon can be operationalized by incorporating knowledge about whether a subject was previously diagnosed with the given mental health condition. Thus, we take the *outcome control* approach [138] to integrate previous condition information with the learned representation so that subject history information is not lost among the many dimensions.

TABLE 4.3: Performance of forecasting conditions at age 18 measured in AUC, using representations derived from transforming the diagnoses from age 3 through age 15

	Baseline	**MIME**	**MIME-NN**	**AC**
Depression	.587	.664*	**.721***	.729*†
Anxiety	.659	**.729***	**.754***	.761*
ADHD	.820	**.892***	**.897***	.905*
AUD/SUD	.489	.453	**.571***†	.563*†

4.5 Evaluation

We tested the longitudinal representations transformed from the diagnoses at age 3 through age 15 via MIME and Agglomerative Clustering for their ability to forecast the onset of mental health conditions. We detect the risk of condition onsets in the future (age 18), which were omitted during the feature extraction. Repeating the diagnoses for the subjects at age 15 was used as a baseline set of predictions.

For MIME, each sample has its longitudinal disorder diagnoses formatted into a vector which is passed as input into the MIME architecture which outputs an intermediary k-dimensional representation.

For Agglomerative Clustering, the same input vector per sample is paired with the vector from each cluster, i.e., an array of values averaged across the samples in the cluster, to compute a Euclidean distance, also resulting in a k-dimensional representation per sample.

To focus on the value of the longitudinal representations themselves, we use a simple model – logistic regression with L2 regularization – to predict the presence of each diagnosis at age 18 using the representations derived from age 15 and younger. We apply a stratified 10-fold cross-validation and record the area under the receiver operating characteristics curve (AUC) averaged across the folds, following the practice of [61] for comparability on the same task. We exclude DBD at age 18 from testing since the number of positive cases is less than the number of cross-validation folds.

TABLE 4.4: Performance on forecasting mental conditions at age 18 using MIME-NN (MN) and Agglomerative Clustering (AC) given different numbers of dimensions (k)

	k=3		**k=5**		**k=7**		**k=10**	
	MN	**AC**	**MN**	**AC**	**MN**	**AC**	**MN**	**AC**
Depression	.694	.724	.702	.726	.715	.729	.721	.729
Anxiety	.740	.751	.723	.755	.762	.754	.754	.761
ADHD	.874	.910	.880	.893	.887	.896	.897	.905
AUD/SUD	.497	.585	.513	.583	.534	.586	.571	.563

Results Table 4.3 compares the AUC between the baseline and proposed models for forecasting diagnoses at age 18. MIME-NN is a variant of MIME with non-negativity constraints added to the architecture. The number of the latent dimensions (k) is equal to 10 for all models except the baseline. **Bold** results indicate models that do not perform significantly worse than Agglomerative Clustering ($p > .05$). * denotes significantly better than baseline, and † indicates significantly better than MIME ($p < .05$).

Both MIME and AC-based models achieve higher AUC than the baseline models ($p < 0.05$) for forecasting all conditions, except MIME on AUD/SUD. MIME-NN had a significantly greater AUC for AUD/SUD than MIME and in general, scored higher for all conditions. The non-negative formulation was originally motivated by enhancing interpretability but these results suggest it also improved the representations.

Amongst all domains and models, AUD/SUD is the most challenging to make forecasts on, likely due to data scarcity and imbalance. Since the earliest onset of AUD and SUD is during one's mid-teens, the diagnosis was not performed at ages 3 and 6. Additionally, none of the subjects were diagnosed with AUD/SUD at ages 9 and 12, and only one at age 15.

ADHD is the easiest to make forecasts on for the models. This may be because it is a relatively consistent condition across a lifetime, or a patient who has been diagnosed with ADHD is likely to be diagnosed with it again.

Also, our models achieve a similar level of accuracy for depression and anxiety with a previous work that used the earlier version of this dataset [61], where they achieved an AUC of .743 for forecasting depression and .777 for anxiety at age 15 when using the same type of model, i.e., L2-penalized logistic regression. Also, models are slightly better at forecasting anxiety than depression which is consistent with other longitudinal modeling works studying these conditions [96].

Next, we evaluated different dimensionalities for the representations from the two approaches as shown in Table 4.4. For MIME-NN, higher dimensionality leads to enhanced forecasting, except for anxiety. This could be caused by the domain already having a fairly

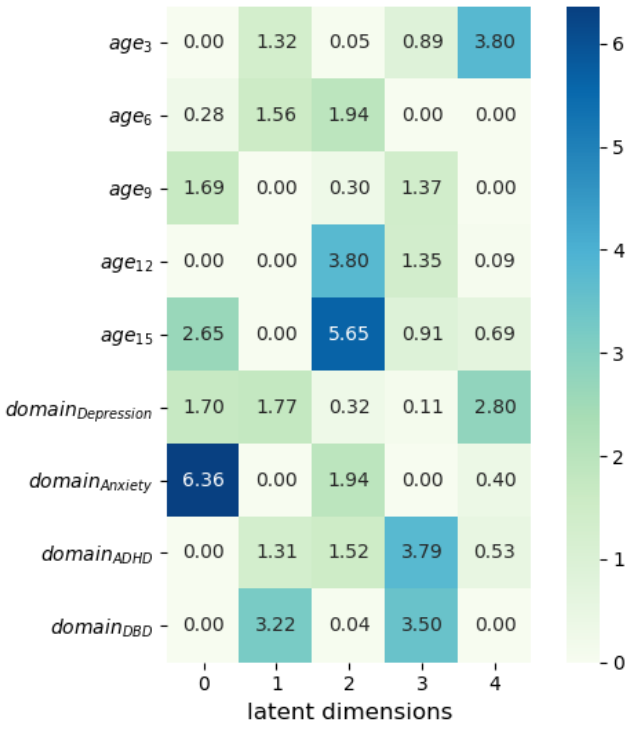


FIGURE 4.2: Coefficients between the lower-dimensional layer (X-axis) and individual factor layer (Y-axis) in the MIME-NN architecture. We interpret factors, i.e., domains and age waves, with high prevalence in each latent dimension as comprising mental health trajectories.

high prevalence at ages 15 and 18. The AUCs for Agglomerative Clustering fluctuated minimally as the dimensionality increased.

Interpretability Figure 4.2 displays the learned weights from the MIME-NN architecture between the nodes from the factor layer and final low-dimensional layer, displayed as black lines in Figure 4.1.

We can see that the dimensions that load on age- and domain-specific factors map onto relevant psychopathology. For example, Dimension 0 consists of high weights for depression and anxiety and emphasizes age 15. This agrees with the finding that the onset of Major Depressive Disorder, increases significantly during puberty [2]. This also agrees with Dimension 0 in the Agglomerative Clustering in Figure 4.3.

ADHD and DBD are loaded heavily on Dimensions 1 and 3. The former seems to indicate these conditions occur at earlier stages of life (i.e., ages 3 and 6), whereas the

latter spreads out across all ages except age 3.

Dimension 4 mainly loads on depression and age 3. Diagnoses at age 3 heavily rely on parent reports rather than self-reports. This may indicate suspected depression that has not manifested and becomes apparent when comparing the result from Agglomerative Clustering in Figure 4.3.

Finally, Dimension 2 maps anxiety and ADHD together, which is in coherence with previous findings about their co-occurrence [137, 133].

Figure 4.3 displays the five mental health trajectories obtained from the Agglomerative Clustering method. Dimension 0 contains samples that experienced anxiety throughout adolescence (i.e., ages 9, 12, 15) and depression at age 15. Dimension 1 shows a high ratio of ADHD and DBD throughout all age waves and a relatively high loading on anxiety across all age waves. Dimension 3 from MIME in Figure 4.2 exhibits a similar pattern. Dimension 2 consists of samples that experienced anxiety in childhood, i.e., ages 3, 6, and 9. Dimension 3 comprises samples that were relatively healthy throughout all age waves, showing a low ratio on all features. Dimension 4 shows prominence on anxiety and DBD at age 3 as well as depression at age 12. It is worth noting that anxiety at age 3 is loosely associated with that at other age waves as shown in Dimensions 2 and 4, as well as in the dendrogram on the left.

Finally, we performed an intruder detection task to estimate the interpretability of our approaches, using random clustering as a baseline. Four diagnoses are provided from each dimension, with three diagnoses among the most prominent ones and one of the least prominent ones, or the *intruder*. We then requested an expert annotator trained in psychopathology to discern the intruding condition for each set of four conditions. For random clustering, the diagnoses were given a random probability value. Out of 30 sets, the annotator correctly answered 25 sets (83%) for Agglomerative Clustering, 14 sets (47%) for MIME, and 8 sets (27%) for Random Clustering.

The primary distinction between the two approaches lies in their treatment of features. Agglomerative Clustering retains features in their original form, treating them

independently. By contrast, MIME integrates features from the same age waves or domains to determine their relative weights. This integration allows for a more streamlined representation of each factor, placing features on a comparable scale and facilitating an assessment of their importance.

4.6 Conclusion

In this work, we compared two interpretable modeling approaches, MIME and Agglomerative Clustering, for producing low-dimensional representations of longitudinal mental health diagnosis data, which to the best of our knowledge is the first attempt in the field. We leveraged the Temperament Study [79] Dataset comprising standardized and rigorously assessed mental health conditions. These representations were informative enough to forecast future conditions, with both proposed methods outperforming baselines at forecasting development of depression, anxiety, ADHD, and AUD/SUD, and demonstrating with greater interpretive coherence evidenced from an intrusion task. Furthermore, the latent dimensions corresponded with existing psychopathological findings. This work paves the way for clinically meaningful longitudinal modeling of mental health conditions by integrating both predictive utility and interpretability.

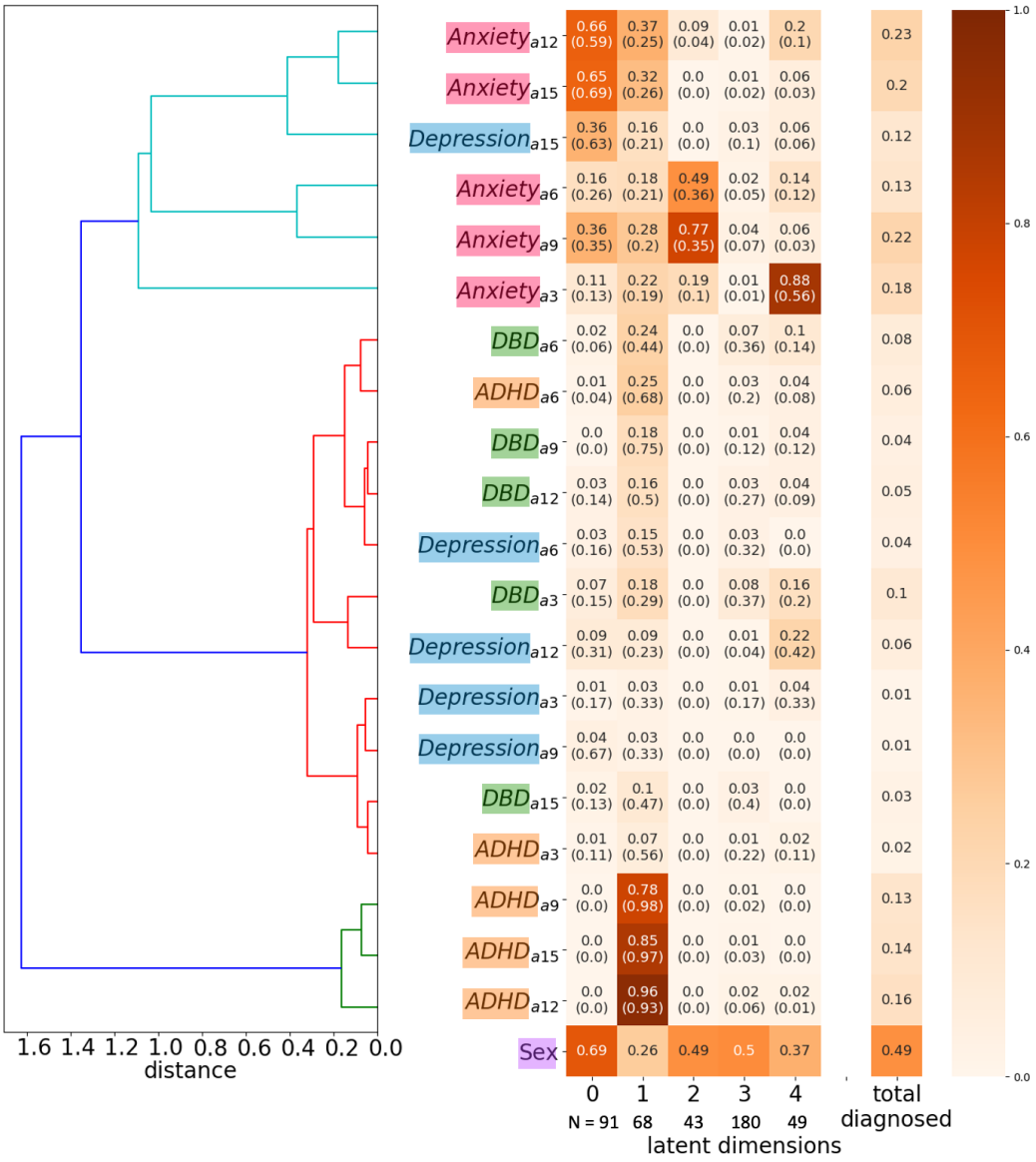


FIGURE 4.3: Mental health trajectories depicted by the Agglomerative Clustering approach. Each dimension, depicted as a column in the heatmap on the right, represents trajectories from childhood to early adulthood. Each cell contains two ratios: the top number shows the percentage of diagnosed samples in the cluster relative to all samples, and the bottom number shows the percentage relative to all diagnosed samples. The numbers on the far right, marked *total diagnosed*, are the ratios of the samples in each dimension that were diagnosed with each condition at each age wave.

Chapter 5

Proposed Work: Investigating the Association Between Social Norms and Anxiety

5.1 Introduction

Social norms are standards of acceptable behavior shared by social groups [27]. They function as a framework for regulating social interactions, increase the predictability of social behavior at scale [75], and hence contributing to the overall stability of society [17]. Norms exist on a spectrum, from widely accepted common sense (e.g., "Cover your mouth when you sneeze", "Be quiet when watching a movie in a theater") to subjective and culturally influenced rules (e.g., "Prioritize family over work", "Study hard and go to a prestigious university").

Members of the society may feel anxious when they perceive pressure coming from the norms and try to meet up to the expectations [41]. In an ideal setting, such induction of anxiety is natural and helpful in that it can help individuals navigate complex social landscapes, avoid rejection, and maintain cohesion within a group, therefore achieving social harmony and collective well-being [124].

In reality, however, social norms can sometimes be distorted or promoting harmful behaviors [3], at odds with one's own wishes or desires (i.e., cognitive dissonance) [8], or

even stigmatizing people [118]. Such downsides of social norms lead to excessive anxiety – the pressure to conform, fear of judgment, or internal confusion arising from conflicting personal values can create a persistent state of distress, negatively impacting one's mental health [174, 46] and daily functioning [42]. Researchers in the field of social psychology have studied specific social norms that induce anxiety, but there is a lack of literature from ML and NLP communities that comprehensively discuss various norms and their impact on anxiety in a data-driven manner.

In this work, we propose an approach that exhaustively explore the various social norms expressed in social media and their relations to anxiety. We specifically pay attention to the social expectations that an individual gets from the people that exercise influence over that person [73]. We will also evaluate the trends in the prevalence of different norms across different ethnicities, time, and geographic regions in the United States.

Our contributions include: (1) proposal of approaches for extraction and categorizations of social norms and norm drivers from social media; (2) analysis on the prevalence of various norms toward anxiety across different ethnicities, time, and geographic regions; and (3) examination on the impact of different norms on anxiety. We release norm extraction the social norms extraction tool to help facilitate future work in the area.

5.2 Related Work

Anxiety, a feeling of threat and turmoil stemming from anticipation of future events, is often induced by social structure and external pressures [147]. Researchers in the area of social psychology have studied the impact of various social norms on anxiety, including stigma on unemployment [158], workaholism culture [4], academic pressure [84], gender roles [95], marriage expectations [52], and beauty standards [32].

Studies in the fields of ML and NLP have also explored social norms in various directions, such as detection of social roles [13, 77] or stigma [161] from social media, and

identification integrating norms into [44] or measuring norms of language models [177]. Rai et al. [130] is the closest work to our proposal, which studied cultural difference in expression of shame and pride between United States and India by analyzing underlying social expectations from Bollywood and Hollywood movies. Our work is distinguished in that we seek to expand the cultural analysis on various ethnicities, examine their associations with anxiety, as well as perform longitudinal and geographical trends of social norms.

5.3 Dataset

We will utilize datasets from two sources. First is Reddit posts from subreddits that represent English-speaking ethnic groups, including `r/asianamerican`, `r/Hispanic`, `r/NativeAmerican`, `r/Blackpeople`, `r/blackladies`, `r/italianamerican`, `r/ABCDesis`, `r/KoreanAmerican`, `r/AsianParentStories`, `r/family`, and `r/firtstgenstudents`. After selecting the posts containing the norm phrases by the extraction method as described in the following section, the total number of is reduced down to 64,744 posts written by 47,256 users. We use this first dataset to define the categories of the norms and investigate the variance of prevalence of each norm by culture. We will also leverage the posts written by these users in subreddits other than the designated ethnic ones to estimate their level of anxiety using the anxiety prediction model proposed in Chapter 2. We will randomly select the same number of users who also wrote in the mentioned subreddits but never mentioned any social norms in order to see if the users' level of anxiety is associated with norm awareness.

Then we will also use County Tweet Lexical Bank (CTLB) [50] dataset, spanning from 2019 to 2023, to analyze the longitudinal and geographical trends of social norms within United States.

5.4 Method

Extraction of Norm Phrases and Drivers We first split each post into sentences and identify those containing regex patterns such as `expect / want / tell / force / allow me to VB` and `let me VB`, which we assume express social norms perceived by authors, or norm targets [87]. Using constituency parsing, we then extract the verb phrases following these patterns as the actual norms and the preceding subjects as norm drivers. For example, in the sentence *"He wants me to come downstairs and actually try to be a part of the family,"* the norm driver is *"he,"* and the norm itself is *"come downstairs and actually try to be a part of the family."*

Categorization of Norms and Norm Drivers To determine the categories of social norms, we apply LLooM [85], which is an LLM-based text analysis tool proposed as an alternative to topic modeling and unsupervised clustering. We would like to note that we have attempted running K-means clustering on the contextual embeddings of the extracted phrases, but the resulting clusters were not aligning with human judgment of semantic similarity. We find the categorization results by LLooM to be more intuitive and close to how humans would group social norms.

Annotation We will randomly select a portion of unique norm phrases from the Reddit dataset for annotation using LLooM, as it enables annotation on a 5-scale regarding relevance of each text to the derived norm categories. However, as LLM-based annotation is computationally expensive, we will run bootstrapped clustering to annotate the rest of the data based on the small portion of annotated samples.

We will also have to decide on the annotation methods, i.e., whether one norm phrase should only be labeled as one specific category of norm or can account for multiple categories, and whether the annotation scale should be binary or continuous (5-level).

Estimation of Anxiety To predict the level of anxiety of the Reddit users, we will use the language-based anxiety prediction model proposed in Chapter 2, by using the posts written by the included users in other subreddits than specified above. We will utilize the level of anxiety for users in the CTLB dataset already assessed in previous work [96].

5.5 Preliminary Results

We selected 3,000 unique norm phrases derived from the Reddit dataset to determine the categories of social norms. The resulting categories and the corresponding examples are shown in Table 5.3. As LLooM provides 5-scale annotation, i.e., 0., .25, .5, .75, 1., we only selected examples that are marked 1 to represent each category for now. Also, while there are only 5 categories, we noticed that because LLooM randomly selects a portion of the given texts for concept induction, the resulting categories can differ each time after execution. Therefore, we will execute multiple rounds of category inference using different sets of text to obtain as many categories of norms for comprehensive exploration.

In Table 5.2, the subreddits are sorted by the percentage of the posts that contain norm patterns out of all posts. The subreddits where the family is the topic by nature are ranked the highest.

The top 20 most frequently mentioned norm drivers from the Reddit dataset are presented in Table 5.5. We can see that personal pronouns and parents take up the majority of the norm drivers. We also asked ChatGPT to summarize the types of drivers among the ones that were mentioned more than 100 times, and the result is in Table 5.3. This step will be re-executed using LLooM as we also need them to be annotated in addition to categorization.

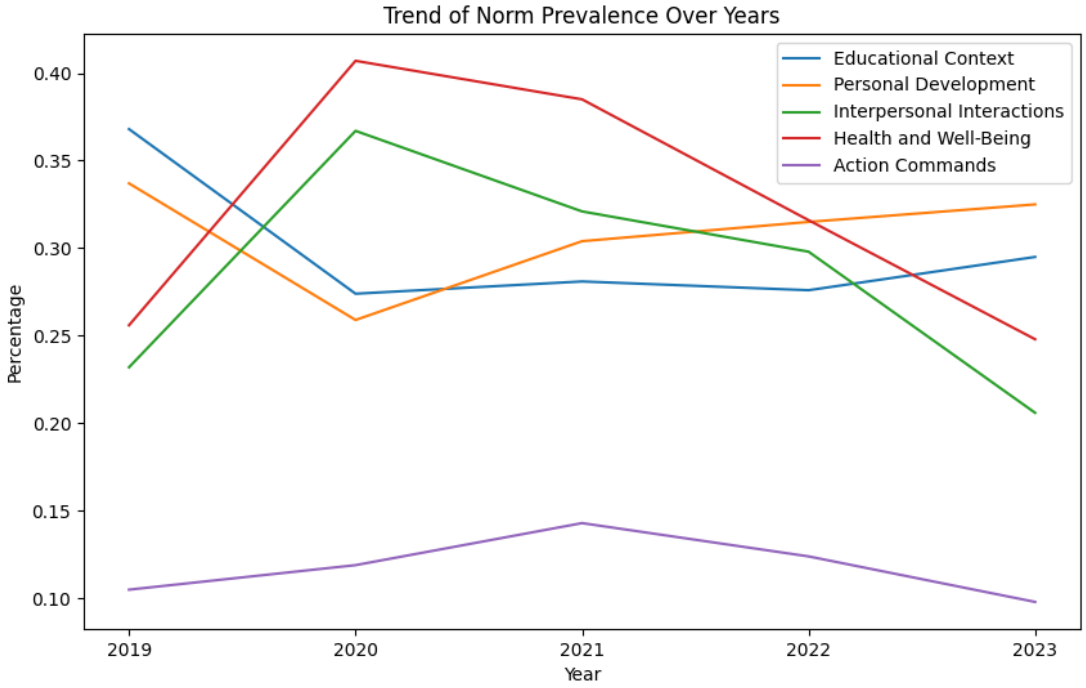


FIGURE 5.1: Prototype of line graph depicting prevalence of each social norm category derived from the CTLB dataset.

5.6 Expected Analysis

By combining annotation in LLooM and bootstrapped clustering, we will analyze the trends of social norms across various ethnic subreddits. For instance, we may find that norms on Educational Context may take up higher percentage out of all norm phrases in `r/AsianAmerican` and `r/ABCDesis` subreddits. We will validate whether the obtained categories are aligning with human judgment by performing intruder detection. We will also estimate the level of anxiety of the Reddit users by using the language-based anxiety prediction model applied on posts written by them outside the selected subreddits and investigate which norm categories are most or least associated with anxiety (Table 5.5). Furthermore, we will leverage Hadoop on CTLB data to enable analysis on longitudinal (Figure 5.1) and geographical trends of the social norms as well. For example, we may find an increase in the prevalence of norms on Health and Well-Being after 2020 due to the spread of COVID.

5.7 Conclusion

We propose an approach that extends our analysis on anxiety to societal dimensions by comprehensively exploring the various social norms expressed in Reddit and Twitter and their relations to anxiety. We were able to extract norm phrases and drivers using constituency parsing. Our initial result shows that five categories of social norms were derived from a smaller portion of the dataset. Based on the annotation on this segment of the dataset, we will perform bootstrapped clustering to facilitate resource-efficient labelling. We will also estimate the level of anxiety of the Reddit users by using the language-based anxiety prediction model applied on posts written by them outside the selected subreddits and investigate which norm categories are most or least associated with anxiety. Then, we aim to use Hadoop to investigate the trends in the prevalence of different norms across different ethnicities, time, and geographic regions in the United States. By integrating these analyses, our approach aims to provide a more comprehensive understanding of anxiety, bridging both individual experiences and broader societal influences, thereby contributing to the overarching goals of this thesis.

Category	Examples
Action Commands	• rant • explain • make
Personal Development	• take control in my life • become a varsity-level swimmer at Harvard who is also a senior engineer at Google • pay them, upgrade my car and buy a house with the salary from my first job
Health and Well-Being	• kill myself • figure out my weight loss and nutrition tracking for myself • go to the gym and go to clinic for wrestling
Educational Context	• get higher grades • go to the Ivy League (no scholarship money) • become a doctor
Interpersonal Interactions	• stop talking to them • be with her • do the thing my mom takes me to India for: to be a buffer between her and her family

TABLE 5.1: Categorization of norms and corresponding example phrases as determined by LLooM.

Subreddit	All Posts	Posts with Patterns	Percentage
r/AsianParentStories	35,204	9,340	26.53%
r/family	39,373	6,119	15.54%
r/KoreanAmerican	34	4	11.76%
r/firstgenstudents	53	3	5.66%
r/ABCDesis	20,957	1,020	4.87%
r/Hispanic	905	38	4.20%
r/asianamerican	6,885	256	3.72%
r/NativeAmerican	3,126	104	3.33%
r/italianamerican	361	5	1.39%

TABLE 5.2: Subreddits sorted by the percentage of posts containing norm patterns out of all posts.

Personal Pronouns	**Family Members**
• I / me / myself • You / you guys • He / him • She / her • They / them	• Parents • Siblings • Grandparents • Aunts/Uncles • Cousins • Step-parents • Spouse/Partner
Friends and Social Circles	**School and Education**
• Friends • Groups	• Teachers • School-related
Work and Authority Figures	**Objects / Non-People Entities**
• Bosses / Managers • Doctors / Healthcare • Law Enforcement and Government	• Technology • Pet • Concepts
Misc.	
• Religious / Spiritual • Filler Words / Expressions	

TABLE 5.3: Categories of norm drivers that are mentioned more than 100 times, as summarized by ChatGPT.

subject	count	percentage	subject	count	percentage
she	4,993	22.30%	her	176	0.79%
they	4,219	18.84%	parents	169	0.75%
he	2,449	10.94%	it	167	0.75%
my parents	1,592	7.11%	them	166	0.74%
my mom	1,245	5.56%	mom	142	0.63%
you	755	3.37%	my mum	134	0.60%
my dad	615	2.75%	my father	122	0.54%
my mother	281	1.25%	my family	116	0.52%
who	206	0.92%	him	91	0.41%
that	196	0.88%	my sister	77	0.34%

TABLE 5.4: Top 20 most frequent norm drivers.

Norm Category	Pearson r
Educational Context	.452
Personal Development	.431
Interpersonal Interactions	.367
Health and Well-Being	.209
Action Commands	.165

TABLE 5.5: Prototype table displaying correlations between each norm category and predicted anxiety.

Bibliography

[1] Samir Al-Stouhi and Chandan K Reddy. "Transfer learning for class imbalance problems with inadequate data". In: *Knowledge and information systems* 48.1 (2016), pp. 201–228.

[2] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Vol. 5. American psychiatric association Washington, DC, 2013.

[3] Catherine E Amiot, Sophie Sansfaçon, and Winnifred R Louis. "Investigating the motivations underlying harmful social behaviors and the motivational nature of social norms". In: *Journal of Applied Social Psychology* 43.10 (2013), pp. 2146–2157.

[4] Cecilie Schou Andreassen et al. "The relationships between workaholism and symptoms of psychiatric disorders: A large-scale cross-sectional study". In: *PloS one* 11.5 (2016), e0152978.

[5] Robert Andrews, Joachim Diederich, and Alan B Tickle. "Survey and critique of techniques for extracting rules from trained artificial neural networks". In: *Knowledge-based systems* 8.6 (1995), pp. 373–389.

[6] Bruce Arroll and Tony Kendrick. "Definition of Anxiety". In: *Primary Care Mental Health*. Ed. by Linda Gask et al. Cambridge University Press, 2018. Chap. 9.

[7] Jordan T Ash et al. "Deep batch active learning by diverse, uncertain gradient lower bounds". In: *arXiv preprint arXiv:1906.03671* (2019).

[8] Alessandro Balestrino and Cinzia Ciardi. "Social norms, cognitive dissonance and the timing of marriage". In: *The Journal of Socio-Economics* 37.6 (2008), pp. 2399–2410.

[9] Russell A Barkley. *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment*. Guilford Publications, 2014.

[10] Monica Ramirez Basco et al. "Methods to improve diagnostic accuracy in a community mental health setting". In: *American Journal of Psychiatry* 157.10 (2000), pp. 1599–1605.

[11] Krishna C Bathina et al. "Individuals with depression express more distorted thinking on social media". In: *Nature Human Behaviour* 5.4 (2021), pp. 458–466.

[12] Katja Beesdo, Susanne Knappe, and Daniel S Pine. "Anxiety and anxiety disorders in children and adolescents: developmental issues and implications for DSM-V". In: *Psychiatric Clinics* 32.3 (2009), pp. 483–524.

[13] Charley Beller et al. "I'm a Belieber: Social Roles via Self-identification and Conceptual Attributes". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 181–186. DOI: 10.3115/v1/P14-2030. URL: https://aclanthology.org/P14-2030/.

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

[15] Adrian Benton, Margaret Mitchell, and Dirk Hovy. "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*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 152–162. URL: https://aclanthology.org/E17-1015.

[16] Harsh Bhasin and R. K. Agrawal. "Multiple-Activation Parallel Convolution Network in Combination with t-SNE for the Classification of Mild Cognitive Impairment". In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. 2021, pp. 1–7. DOI: 10.1109/BIBE52308.2021.9635485.

[17] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. "Social norms". In: *The Stanford encyclopedia of philosophy* (2018).

[18] D Thomas Blair and Valerie A Ramones. "The undertreatment of anxiety: Overcoming the confusion and stigma". In: *Journal of psychosocial nursing and mental health services* 34.6 (1996), pp. 9–9.

[19] William J Bolton et al. "Co-morbidity Representation in Artificial Intelligence: Tapping into Unused Clinical Knowledge". In: *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*. Springer, 2024, pp. 173–196.

[20] Setiyo Budiyanto, Harry Candra Sihombing, and Fajar Rahayu IM. "Depression and anxiety detection through the Closed-Loop method using DASS-21". In: *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 17.4 (2019), pp. 2087–2097.

[21] Franziska Burger, Mark A. Neerincx, and Willem-Paul Brinkman. "Natural language processing for cognitive therapy: Extracting schemas from thought records". In: *PLOS ONE* 16 (Oct. 2021), pp. 1–24. DOI: 10.1371/journal.pone.0257832. URL: https://doi.org/10.1371/journal.pone.0257832.

[22] Christina Caron. "The Upside of Anxiety". In: *The New York Times* (2022). URL: https://www.nytimes.com/2022/01/19/well/mind/anxiety-benefits.html.

[23] Kimberly LH Carpenter et al. "Quantifying risk for anxiety disorders in preschool children: a machine learning approach". In: *PloS one* 11.11 (2016), e0165524.

[24] Simon Chapman, Wai Leng Wong, and Wayne Smith. "Self-exempting beliefs about smoking and health: differences between smokers and ex-smokers." In: *American journal of public health* 83.2 (1993), pp. 215–219.

[25] Jongwon Choi et al. "VaB-AL: Incorporating Class Imbalance and Difficulty With Variational Bayes for Active Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 6749–6758.

[26] Jan Chorowski and Jacek M. Zurada. "Learning Understandable Neural Networks With Nonnegative Weight Constraints". In: *IEEE Transactions on Neural Networks and Learning Systems* 26.1 (2015), pp. 62–69. DOI: 10.1109/TNNLS.2014.2310059.

[27] Adrienne Chung Adrienne Chung and Rajiv N Rimal Rajiv N Rimal. "Social norms: A review". In: *Review of Communication Research* 4 (2016), pp. 01–28.

[28] Gui Citovsky et al. "Batch Active Learning at Scale". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 11933–11944. URL: https://proceedings.neurips.cc/paper/2021/file/64254db8396e404d9223914a0bd355d2-Paper.pdf.

[29] Glen Coppersmith, Craig Harman, and Mark Dredze. "Measuring post traumatic stress disorder in Twitter". In: *Eighth international AAAI Conference on Weblogs and Social Media*. 2014.

[30] Glen Coppersmith et al. "Exploratory Analysis of Social Media Prior to a Suicide Attempt". In: *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, June 2016, pp. 106–117. DOI: 10.18653/v1/W16-0311. URL: https://aclanthology.org/W16-0311.

[31] Glen Coppersmith et al. "Natural language processing of social media as screening for suicide risk". In: *Biomedical informatics insights* 10 (2018), p. 1178222618792860.

[32] Antonios Dakanalis et al. "Internalization of sociocultural standards of beauty and disordered eating behaviours: the role of body surveillance, shame and social anxiety". In: *Journal of Psychopathology* 20 (2014), pp. 33–37.

[33] Anja Dalgaard-Nielsen. "Promoting exit from violent extremism: Themes and approaches". In: *Studies in Conflict & Terrorism* 36.2 (2013), pp. 99–115.

[34] Munmun De Choudhury et al. "Predicting depression via social media". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 7. 1. 2013, pp. 128–137.

[35] C. Lindsay DeVane et al. "Anxiety Disorders in the 21st Century: Status, Challenges, Opportunities, and Comorbidity With Depression". In: *AJMC* (2005).

[36] Jacob Devlin et al. "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)*. 2019, pp. 4171–4186.

[37] Shichao Du et al. "Social drivers of mental health: A US study using machine learning". In: *American journal of preventive medicine* 65.5 (2023), pp. 827–834.

[38] Melanie Ducoffe and Frederic Precioso. "Adversarial active learning for deep networks: a margin based approach". In: *arXiv preprint arXiv:1802.09841* (2018).

[39] Helen Link Egger and Adrian Angold. "Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology". In: *Journal of child psychology and psychiatry* 47.3-4 (2006), pp. 313–337.

[40] Liat Ein-Dor et al. "Active Learning for BERT: An Empirical Study". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7949–7962. DOI: 10.18653/v1/2020.emnlp-main.638. URL: https://aclanthology.org/2020.emnlp-main.638.

[41] Jon Elster. "Rationality, emotions, and social norms". In: *Synthese* (1994), pp. 21–49.

[42] Kirsten L Ferguson and Margaret R Rodway. "Cognitive behavioral treatment of perfectionism: Initial evaluation studies". In: *Research on Social Work Practice* 4.3 (1994), pp. 283–308.

[43] William P Fleisher and Laurence Y Katz. "Early onset major depressive disorder". In: *Paediatrics & Child Health* 6.7 (2001), pp. 444–448.

[44] Maxwell Forbes et al. "Social Chemistry 101: Learning to Reason about Social and Moral Norms". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 653–670. DOI: 10.18653/v1/2020.emnlp-main.48. URL: https://aclanthology.org/2020.emnlp-main.48/.

[45] Linton C Freeman. *Elementary applied statistics: for students in behavioral science*. New York: Wiley, 1965.

[46] Randy O Frost et al. "The dimensions of perfectionism". In: *Cognitive therapy and research* 14 (1990), pp. 449–468.

[47] Lorena de la Fuente-Tomas et al. "Classification of patients with bipolar disorder using k-means clustering". In: *PloS one* 14.1 (2019), e0210314.

[48] Adithya V Ganesan et al. "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. 2021, pp. 4515–4532.

[49] Salvatore Giorgi et al. "Lived experience matters: Automatic detection of stigma toward people who use substances on social media". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 18. 2024, pp. 474–487.

[50] Salvatore Giorgi et al. "The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1167–1172. DOI: 10.18653/v1/D18-1148. URL: https://aclanthology.org/D18-1148/.

[51] Daniel Gissin and Shai Shalev-Shwartz. "Discriminative active learning". In: *arXiv preprint arXiv:1907.06347* (2019).

[52] Tianhan Gui. "Coping with parental pressure to get married: Perspectives from Chinese "leftover women"". In: *Journal of Family Issues* 44.8 (2023), pp. 2118–2137.

[53] Sharath Chandra Guntuku et al. "What twitter profile and posted images reveal about depression and anxiety". In: *Proceedings of the international AAAI Conference on Web and Social Media*. Vol. 13. 2019, pp. 236–246.

[54] Chuan Guo et al. "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[55] Tom SF Haines and Tao Xiang. "Active rare class discovery and classification using dirichlet processes". In: *International Journal of Computer Vision* 106.3 (2014), pp. 315–331.

[56] Umme Marzia Haque, Enamul Kabir, and Rasheda Khanam. "Detection of child depression using machine learning methods". In: *PLoS one* 16.12 (2021), e0261131.

[57] Eddie Harmon-Jones and Cindy Harmon-Jones. "Cognitive dissonance theory after 50 years of development". In: *Zeitschrift für Sozialpsychologie* 38.1 (2007), pp. 7–16.

[58] Eddie Harmon-Jones and Judson Mills. "An introduction to cognitive dissonance theory and an overview of current perspectives on the theory." In: *Cognitive dissonance: Reexamining a pivotal theory in psychology* (2019).

[59] Eddie Harmon-Jones et al. "Left frontal cortical activation and spreading of alternatives: tests of the action-based model of dissonance." In: *Journal of personality and social psychology* 94.1 (2008), p. 1.

[60] Jason S Hartford et al. "Exemplar guided active learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13163–13173.

[61] Mariah T Hawes et al. "Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning". In: *Psychological Medicine* 53.13 (2023), pp. 6205–6211.

[62] Holly Hazlett-Stevens and Michelle G. Craske. "THE CATASTROPHIZING WORRY PROCESS IN GENERALIZED ANXIETY DISORDER: A PRELIMINARY INVESTIGATION OF AN ANALOG POPULATION". In: *Behavioural and Cognitive Psychotherapy* 31.4 (2003), 387–401. DOI: 10.1017/S1352465803004016.

[63] Timothy M. Hospedales, Shaogang Gong, and Tao Xiang. "Finding Rare Classes: Active Learning with Generative and Discriminative Models". In: *IEEE Transactions on Knowledge and Data Engineering* 25.2 (2013), pp. 374–386. DOI: 10.1109/TKDE.2011.231.

[64] Neil Houlsby et al. "Bayesian active learning for classification and preference learning". In: *arXiv preprint arXiv:1112.5745* (2011).

[65] Yufei Huang and Jianqiu Zhang. "Exploring factor structures using variational autoencoder in personality research". In: *Frontiers in psychology* 13 (2022), p. 863926.

[66] John A Johnson. "Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120". In: *Journal of Research in Personality* 51 (2014), pp. 78–89.

[67] Swanie Juhng et al. "Discourse-Level Representations can Improve Prediction of Degree of Anxiety". In: *Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

[68] Simona C Kaplan et al. "The Cognitive Distortions Questionnaire (CD-Quest): validation in a sample of adults with social anxiety disorder". In: *Cognitive therapy and research* 41.4 (2017), pp. 576–587.

[69] Siddharth Karamcheti et al. "Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 7265–7281. DOI: 10.18653/v1/2021.acl-long.564. URL: https://aclanthology.org/2021.acl-long.564/.

[70] Jun Kashihara et al. "Classifying patients with depressive and anxiety disorders according to symptom network structures: A Gaussian graphical mixture model-based clustering". In: *Plos one* 16.9 (2021), e0256902.

[71] Siegfried Kasper. "Anxiety disorders: under-diagnosed and insufficiently treated". In: *International Journal of Psychiatry in Clinical Practice* 10.sup1 (2006), pp. 3–9.

[72] Sean W Kelley and Claire M Gillan. "Using language in social media posts to study the network dynamics of depression longitudinally". In: *Nature Communications* 13.1 (2022), pp. 1–11.

[73] Theodore D Kemper. "Self-conceptions and the expectations of significant others". In: *The Sociological Quarterly* 7.3 (1966), pp. 323–343.

[74] Smith K Khare and U Rajendra Acharya. "An explainable and interpretable model for attention deficit hyperactivity disorder in children using EEG signals". In: *Computers in biology and medicine* 155 (2023), p. 106676.

[75] Sara B Kiesler. "Preference for predictability or unpredictability as a mediator of reactions to norm violations." In: *Journal of Personality and Social Psychology* 27.3 (1973), p. 354.

[76] Selina Kikkenborg Berg et al. "Anxiety predicts mortality in ICD patients: results from the cross-sectional national CopenHeartICD survey with register follow-up". In: *Pacing and Clinical Electrophysiology* 37.12 (2014), pp. 1641–1650.

[77] Sunghwan Mac Kim, Stephen Wan, and Cécile Paris. "Detecting Social Roles in Twitter". In: *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*. Ed. by Lun-Wei Ku, Jane Yung-jen Hsu, and Cheng-Te Li. Austin, TX, USA: Association for Computational Linguistics, Nov. 2016, pp. 34–40. DOI: 10.18653/v1/W16-6206. URL: https://aclanthology.org/W16-6206/.

[78] Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. "Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1152–1158. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.145.

[79] Daniel N Klein and Megan C Finsaas. "The Stony Brook Temperament Study: Early antecedents and pathways to emotional disorders". In: *Child Development Perspectives* 11.4 (2017), pp. 257–263.

[80] William M. Klykylo. "Comorbidity". In: *Encyclopedia of Psychotherapy*. Ed. by Michel Hersen and William Sledge. New York: Academic Press, 2002, pp. 475–479. ISBN: 978-0-12-343010-6. DOI: https://doi.org/10.1016/B0-12-343010-0/00053-2. URL: https://www.sciencedirect.com/science/article/pii/B0123430100000532.

[81] Suraj Kothawade et al. "SIMILAR: Submodular Information Measures Based Active Learning In Realistic Scenarios". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 18685–18697. URL: https://proceedings.neurips.cc/paper/2021/file/9af08cda54faea9adf40a201794183cf-Paper.pdf.

[82] Nikolaos Koutsouleris and Paolo Fusar-Poli. "From heterogeneity to precision: redefining diagnosis, prognosis, and treatment of mental disorders". In: *Biological Psychiatry* 96.7 (2024), pp. 508–510.

[83] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.

[84] Narasappa Kumaraswamy. "Academic stress, anxiety and depression among college students: A brief review". In: *International review of social sciences and humanities* 5.1 (2013), pp. 135–143.

[85] Michelle S Lam et al. "Concept induction: Analyzing unstructured text with high-level concepts using lloom". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–28. DOI: 10.1145/3613904.3642830. URL: https://doi.org/10.1145/3613904.3642830.

[86] Aziliz Le Glaz et al. "Machine learning and natural language processing in mental health: systematic review". In: *Journal of medical Internet research* 23.5 (2021), e15708.

[87] Sophie Legros and Beniamino Cislaghi. "Mapping the social-norms literature: An overview of reviews". In: *Perspectives on Psychological Science* 15.1 (2020), pp. 62–80.

[88] Shoushan Li et al. "Active Learning for Imbalanced Sentiment Classification". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 139–148. URL: https://aclanthology.org/D12-1013.

[89] Yichan Liang, Jianheng Li, and Jian Yin. "A new multi-choice reading comprehension dataset for curriculum learning". In: *Asian Conference on Machine Learning*. PMLR. 2019, pp. 742–757.

[90] Todd D Little. *The Oxford handbook of quantitative methods, volume 1: Foundations*. Oxford University Press, 2013.

[91] Houjun Liu et al. "A Transformer Approach for Cognitive Impairment Classification". In: *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*. Springer, 2024, pp. 93–102.

[92] Xin Liu et al. "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. Yokohama, Yokohama, Japan, 2021. ISBN: 9780999241165.

[93] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[94] Veronica Lynn et al. "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*. New Orleans, LA: Association for Computational Linguistics, June 2018, pp. 37–46. DOI: 10.18653/v1/W18-0604. URL: https://aclanthology.org/W18-0604.

[95] James R Mahalik et al. "Development of the conformity to masculine norms inventory." In: *Psychology of men & masculinity* 4.1 (2003), p. 3.

[96] Siddharth Mangalik et al. "Robust language-based mental health assessments in time and space through social media". In: *NPJ Digital Medicine* 7.1 (2024), p. 109.

[97] Md Maniruzzaman, Jungpil Shin, and Md Al Mehedi Hasan. "Predicting children with ADHD using behavioral activity: a machine learning analysis". In: *Applied Sciences* 12.5 (2022), p. 2737.

[98] William C Mann and Sandra A Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, 1987.

[99] Katerina Margatina et al. "Active Learning by Acquiring Contrastive Examples". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 650–663. DOI: 10.18653/v1/2021.emnlp-main.51. URL: https://aclanthology.org/2021.emnlp-main.51/.

[100] Marija Maric et al. "Distorted cognitive processing in youth: the structure of negative cognitive errors and their associations with anxiety". In: *Cognitive Therapy and Research* 35.1 (2011), pp. 11–20.

[101] Matthew Matero et al. "MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2959–2966. URL: https://aclanthology.org/2021.findings-emnlp.253.

[102] Matthew Matero et al. "Suicide risk assessment with multi-level dual-context language and BERT". In: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*. 2019, pp. 39–44.

[103] Sandra C Matz et al. "Predicting individual-level income from Facebook profiles". In: *PLOS ONE* 14.3 (2019), e0214369.

[104] April McGrath. "Dealing with dissonance: A review of cognitive dissonance reduction". In: *Social and Personality Psychology Compass* 11.12 (2017), e12362.

[105] Sean R McWhinney et al. “Principal component analysis as an efficient method for capturing multivariate brain signatures of complex disorders—ENIGMA study in people with bipolar disorders and obesity”. In: *Human Brain Mapping* 45.8 (2024), e26682.

[106] Abu Saleh Musa Miah et al. “Alzheimer’s disease detection using CNN based on effective dimensionality reduction approach”. In: *Intelligent Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)*. Springer. 2021, pp. 801–811.

[107] Jakov Milić et al. “High levels of depression and anxiety among Croatian medical and nursing students and the correlation between subjective happiness and personality traits”. In: *International Review of Psychiatry* 31.7-8 (2019), pp. 653–660.

[108] Muzafar Mehraj Misgar and MPS Bhatia. “Advancing ADHD diagnosis: using machine learning for unveiling ADHD patterns through dimensionality reduction on IoMT actigraphy signals”. In: *International Journal of Information Technology* (2024), pp. 1–13.

[109] Saif M. Mohammad and Peter D. Turney. “Crowdsourcing a Word-Emotion Association Lexicon”. In: *Computational Intelligence* 29.3 (2013), pp. 436–465.

[110] Elham Mohammadi, Hessam Amini, and Leila Kosseim. “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*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 34–38. DOI: 10.18653/v1/W19-3004. URL: https://aclanthology.org/W19-3004.

[111] Rei Monden et al. “Simultaneous decomposition of depression heterogeneity on the person-, symptom-and time-level: the use of three-mode principal component analysis”. In: *PloS one* 10.7 (2015), e0132765.

[112] Melissa Mulraney et al. “A systematic review of the persistence of childhood mental health problems into adulthood”. In: *Neuroscience & Biobehavioral Reviews* 129 (2021), pp. 182–205.

[113] Prateek Munjal et al. “Towards Robust and Reproducible Active Learning Using Neural Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 223–232.

[114] Elizabeth M. Muran and Robert W. Motta. “Cognitive distortions and irrational beliefs in post-traumatic stress, anxiety, and depressive disorders”. In: *Journal of Clinical Psychology* (1993). URL: https://psycnet.apa.org/doi/10.1002/1097-4679(199303)49:2%3C166::AID-JCLP2270490207%3E3.0.CO;2-6.

[115] Ludmila BS Nascimento et al. “Assessment of the Relationship Between Attribute Coding and the Interpretability of Machine Learning Models: An Analysis in the

Context of Children and Adolescents with Depression." In: *BIOSTEC (2)*. 2024, pp. 482–489.

[116] Yuval Netzer et al. "Reading Digits in Natural Images with Unsupervised Feature Learning". In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[117] Andrew Ng. "MLOps: from model-centric to data-centric AI". In: *Online unter https://www. deeplearning. ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centricAI. pdf [Zugriffam09. 09.2021] Search in* (2021).

[118] Ross MG Norman et al. "The role of perceived norms in the stigmatization of mental illness". In: *Social psychiatry and psychiatric epidemiology* 43 (2008), pp. 851–859.

[119] Mbithe Nzomo and Deshendran Moodley. "A Semantic Architecture for Continuous Health Monitoring, Risk Prediction, and Proactive Decision Making". In: *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*. Springer, 2024, pp. 265–281.

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

[121] Gregory Park et al. "Automatic personality assessment through social media language." In: *Journal of personality and social psychology* 108.6 (2015), p. 934.

[122] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[123] Christopher Peterson. "The meaning and measurement of explanatory style". In: *Psychological Inquiry* 2.1 (1991), pp. 1–10.

[124] Murray Petrie. *Institutions, social norms and well-being*. Tech. rep. New Zealand Treasury Working Paper, 2002.

[125] Emily Pitler et al. "Easily Identifiable Discourse Relations". In: *Coling 2008: Companion volume: Posters*. Manchester, UK: Coling 2008 Organizing Committee, Aug. 2008, pp. 87–90. URL: https://aclanthology.org/C08-2022.

[126] Livia Polanyi. "A formal model of the structure of discourse". In: *Journal of pragmatics* 12.5-6 (1988), pp. 601–638.

[127] Valery A Ponomarev et al. "Group independent component analysis (gICA) and current source density (CSD) in the study of EEG in ADHD adults". In: *Clinical Neurophysiology* 125.1 (2014), pp. 83–97.

[128] Rashmi Prasad et al. "The Penn Discourse TreeBank 2.0." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. 2008.

[129] John S Price. "Evolutionary aspects of anxiety disorders". In: *Dialogues in clinical neuroscience* 5.3 (2003), pp. 223–236.

[130] Sunny Rai et al. "Social Norms in Cinema: A Cross-Cultural Analysis of Shame, Pride and Prejudice". In: *arXiv preprint arXiv:2402.11333* (2024).

[131] Neil A Rector et al. "Examination of the trait facets of the five-factor model in discriminating specific mood and anxiety disorders". In: *Psychiatry Research* 199.2 (2012), pp. 131–139.

[132] Nils Reimers and Iryna Gurevych. "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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

[133] Frederick W Reimherr et al. "ADHD and anxiety: clinical significance and treatment implications". In: *Current Psychiatry Reports* 19 (2017), pp. 1–10.

[134] Ilaria Rocco et al. "Time of onset and/or diagnosis of ADHD in European children: a systematic review". In: *BMC psychiatry* 21 (2021), pp. 1–24.

[135] GA Roth. "Global Burden of Disease Collaborative Network. Global Burden of Disease Study through 2017 (GBD 2017) Results". In: *The Lancet* 392 (2018), pp. 1736–1788.

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

[137] David Beck Schatz and Anthony L Rostain. "ADHD with comorbid anxiety: a review of the current literature". In: *Journal of Attention disorders* 10.2 (2006), pp. 141–149.

[138] H. Andrew Schwartz et al. "DLATK: Differential Language Analysis ToolKit". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 55–60. DOI: 10.18653/v1/D17-2010. URL: https://aclanthology.org/D17-2010.

[139] H Andrew Schwartz et al. "Towards assessing changes in degree of depression through facebook". In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*. 2014, pp. 118–125.

[140] Ozan Sener and Silvio Savarese. "Active Learning for Convolutional Neural Networks: A Core-Set Approach". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=H1aIuk-RW.

[141] Burr Settles, Mark Craven, and Lewis Friedland. "Active learning with real annotation costs". In: *Proceedings of the NIPS workshop on cost-sensitive learning*. Vol. 1. Vancouver, CA: 2008.

[142] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[143] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. "Machine learning in mental health: a scoping review of methods and applications". In: *Psychological medicine* 49.9 (2019), pp. 1426–1448.

[144] 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*. Vancouver, BC: Association for Computational Linguistics, Aug. 2017, pp. 58–65. DOI: 10.18653/v1/W17-3107. URL: https://aclanthology.org/W17-3107.

[145] Yanyao Shen et al. "Deep active learning for named entity recognition". In: *arXiv preprint arXiv:1707.05928* (2017).

[146] Sagarika Shreevastava and Peter Foltz. "Detecting Cognitive Distortions from Patient-Therapist Interactions". In: *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Online: Association for Computational Linguistics, June 2021, pp. 151–158. DOI: 10.18653/v1/2021.clpsych-1.17. URL: https://aclanthology.org/2021.clpsych-1.17.

[147] Richa Shri. "Anxiety: causes and management". In: *The Journal of Behavioral Science* 5.1 (2010), pp. 100–118.

[148] T. Simms et al. "Detecting Cognitive Distortions Through Machine Learning Text Analytics". In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017, pp. 508–512. DOI: 10.1109/ICHI.2017.39.

[149] Karandeep Singh and Maria A Woodward. "The rigorous work of evaluating consistency and accuracy in electronic health record data". In: *JAMA ophthalmology* 139.8 (2021), pp. 894–895.

[150] Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. "Causal Explanation Analysis on Social Media". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[151] Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. "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)*. Abu Dhabi, United Arab Emirates

(Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 45–55. URL: https://aclanthology.org/2022.umios-1.5.

[152] Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. "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)*. Association for Computational Linguistics, 2022.

[153] Youngseo Son et al. "Recognizing Counterfactual Thinking in Social Media Texts". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 654–658. DOI: 10.18653/v1/P17-2103. URL: https://aclanthology.org/P17-2103.

[154] Youngseo Son et al. "World Trade Center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews". In: *Psychological Medicine* (2021).

[155] Charles Spearman. "The proof and measurement of association between two things". In: *The American Journal of Psychology* 100.3/4 (1987), pp. 441–471.

[156] Robert L Spitzer et al. "The structured clinical interview for DSM-III-R (SCID): I: history, rationale, and description". In: *Archives of general psychiatry* 49.8 (1992), pp. 624–629.

[157] Jeffrey P Staab et al. "Detection and diagnosis of psychiatric disorders in primary medical care settings". In: *Medical Clinics of North America* 85.3 (2001), pp. 579–596.

[158] Tobias Staiger et al. "Intersections of discrimination due to unemployment and mental health problems: the role of double stigma for job-and help-seeking behaviors". In: *Social psychiatry and psychiatric epidemiology* 53 (2018), pp. 1091–1098.

[159] David Stillwell and Michal Kosinski. "myPersonality project: Example of successful utilization of online social networks for large-scale social research". In: Jan. 2012.

[160] Kostas Stoitsas et al. "Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery". In: *Scientific Reports* 12.1 (2022), p. 16990.

[161] Nadiya Straton, Hyeju Jang, and Raymond Ng. "Stigma annotation scheme and stigmatized language detection in health-care discussions on social media". In: *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*. European Language Resources Association. 2020, pp. 1178–1190.

[162] Richard M. Suinn. "Anxiety and Cognitive Dissonance". In: *The Journal of General Psychology* 73.1 (1965). PMID: 14316953, pp. 113–116. DOI: 10.1080/00221309.

1965.9711258. eprint: https://doi.org/10.1080/00221309.1965.9711258. URL: https://doi.org/10.1080/00221309.1965.9711258.

[163] Maite Taboada and William C Mann. "Rhetorical structure theory: Looking back and moving ahead". In: *Discourse studies* 8.3 (2006), pp. 423–459.

[164] Bethany A Teachman. "Aging and negative affect: the rise and fall and rise of anxiety and depression symptoms." In: *Psychology and aging* 21.1 (2006), p. 201.

[165] Akim Tsvigun et al. "Towards Computationally Feasible Deep Active Learning". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1198–1218. DOI: 10.18653/v1/2022.findings-naacl.90. URL: https://aclanthology.org/2022.findings-naacl.90.

[166] Yevhen Tyshchenko. "Depression and anxiety detection from blog posts data". In: *Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia* (2018).

[167] Kaushik Vakadkar, Diya Purkayastha, and Deepa Krishnan. "Detection of autism spectrum disorder in children using machine learning techniques". In: *SN computer science* 2 (2021), pp. 1–9.

[168] Vasudha Varadarajan et al. "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, 2022. URL: https://aclanthology.org/2022.nlpcss-1.16.

[169] Vasudha Varadarajan et al. "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, 2023.

[170] Dan Wang and Yi Shang. "A new active labeling method for deep learning". In: *2014 International Joint Conference on Neural Networks (IJCNN)* (2014), pp. 112–119.

[171] Shufan Wang, Laure Thompson, and Mohit Iyyer. "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, 2021.

[172] Yizhong Wang, Sujian Li, and Jingfeng Yang. "Toward Fast and Accurate Neural Discourse Segmentation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 962–967. DOI: 10.18653/v1/D18-1116. URL: https://aclanthology.org/D18-1116/.

[173] Jun-Cheng Weng et al. "An autoencoder and machine learning model to predict suicidal ideation with brain structural imaging". In: *Journal of clinical medicine* 9.3 (2020), p. 658.

[174] Y Joel Wong et al. "Meta-analyses of the relationship between conformity to masculine norms and mental health-related outcomes." In: *Journal of counseling psychology* 64.1 (2017), p. 80.

[175] Xingjiao Wu et al. "A survey of human-in-the-loop for machine learning". In: *Future Generation Computer Systems* 135 (2022), pp. 364–381. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2022.05.014.

[176] Dani Yogatama et al. "Learning and evaluating general linguistic intelligence". In: *arXiv preprint arXiv:1901.11373* (2019).

[177] Ye Yuan et al. "Measuring Social Norms of Large Language Models". In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 650–699. DOI: 10.18653/v1/2024.findings-naacl.43. URL: https://aclanthology.org/2024.findings-naacl.43/.

[178] Effat Jalaeian Zaferani, Mohammad Teshnehlab, and Mansour Vali. "Automatic personality traits perception using asymmetric auto-encoder". In: *IEEE access* 9 (2021), pp. 68595–68608.

[179] Xueying Zhan et al. "A comparative survey of deep active learning". In: *arXiv preprint arXiv:2203.13450* (2022).

[180] Mike Zhang and Barbara Plank. "Cartography active learning". In: *arXiv preprint arXiv:2109.04282* (2021).

[181] Ayah Zirikly et al. "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*. 2019, pp. 24–33.