Instructions for EMNLP 2023 Proceedings

Anonymous EMNLP submission

Abstract

With the popularity of the Internet, online social media has increasingly become a platform for people to share their attitudes towards life, such as optimism and pessimism about the future. While the Internet embraces various views, it also quietly deepens the formation of impressions on different views and attitudes. A large part of these texts will form the corpus of the pre-trained model, and the model may learn the tendency of life attitudes in the corpus. Our work develops new methods to (1) measure life attitude biases in LMs trained on such corpora and (2) measure the judgement impact of downstream models trained on different life attitude corpus. We focus on mental health disorder detection, aiming to empirically quantify the effects of life attitude (optimism, pessimism) leaning in pretraining data on the influence of risk-related tasks. Our findings reveal that pretrained LMs do have life attitude leanings that reinforce the polarization present in pretraining corpora, propagating life attitude biases into mental health disorder detection.We discussed strategies that might mitigate or leverage models of different life attitude leaning.

1 Introduction

With unprecedented user engagement, digital and social media have become the primary way for people to share their attitudes about life(Steinert, 2021; Shareef et al., 2020; Auxier and Anderson, 2021). Over the past decade, there has been a dramatic increase in the number of incidents where people share their attitudes on social media, which can cover a wide range of topics: jobs and careers, schooling and education, health and lifestyle, relationships and family, social issues and current events(Debatin et al., 2009; Gross and Acquisti, 2005). Although social networks provide an inclusive platform for people to share different perspectives on attitudes toward life, these discussions also deepen the formation of polarized impressions of attitudes toward life - pessimistic and optimistic tendencies(Ferrari, 2008; Peeters and Czapinski, 1990). These texts form a major part of the pretraining corpus for the large language model and propagate this life attitude tendency to downstream tasks.

Mental health is a key issue in modern society, and without proper treatment, mental disorders can sometimes turn into suicidal ideation (Ryk et al., 2023). To address this critical issue, there has been a large amount of mental health research aimed at the efficient and automated detection of mental health disorders. Mental health disease detection usually uses information fusion strategies to make the model know more information and improve the accuracy of reasoning. There are three common fusion strategies: feature fusion (Song et al., 2018; Uban et al., 2021), model fusion (Sawhney et al., 2020; Abdul-Mageed and Ungar, 2017) and task fusion (Turcan et al., 2021). Previous research methods mainly focused on how to integrate more information, such as emotional, personality and even economic information, so that the model can learn more knowledge, while ignoring the possible potential life attitude bias of the pre-trained model itself, which is closely related to mental health detection problems (Mao et al., 2022; Conversano et al., 2010; Yıldırım and Cicek, 2022). For example, an overly pessimistic model may misrepresent positive examples of some mental health disorders. To the best of our knowledge, no prior work has shown how to analyze the effects of naturally occurring life of orientation biases in pretraining data on language models, and subsequently on downstream tasks. Our study aims to fill this gap.

We focused on several most common mental health disorders: depression, anxiety, suicide, stress(Hardy, 2018). Because detecting these mental health disorders is important for the mental health of society(Ryk et al., 2023). We investigate how social media biases in the pretraining

data propagate into pretrain models and ultimately affect downstream tasks, because discussions about life orientation attitude issues are abundant in pretraining data sourced from online encyclopedias, and language models will inevitably inherit this tendency towards life attitudes.

To this end, based on psychological theory (Carver and Scheier, 1982; Carver et al., 1983) and the life of orientation test(Scheier and Carver, 1985), we propose to empirically quantify the life attitude leaning of pretrained LMs. We then further pretrain language models on different life attitude corpora to investigate whether LMs pick up political biases from training data. Finally, we train classifiers on top of LMs with varying life of orientation leanings and evaluate their performance on mental health disorders targeting different mental illness. In this way, we investigate the propagation of life of orientation attitude through the entire pipeline from pretraining data to language models to downstream tasks.

Our experiments across several LM architectures demonstrate that different pretrained LMs do have different underlying life attitude leanings, reinforcing the life attitude polarization present in pretraining corpora. However, the models with different life attitude tendencies showed great differences in the detection of different types of mental health diseases.

The main contributions of this paper are novel methods to quantify life of orientation attitude in LMs, and findings that shed new light on how ideological polarization in pretraining corpora propagates orientation into language models, and subsequently into risk-related downstream tasks.

2 Method

We propose a two-step methodology to establish the effect of life attitude in pretraining corpora on the mental health disorders detection tasks: (1) we develop a framework, grounded in physical science literature, to measure the inherent life orientation leanings of pretrained language models, and (2) then investigate how the life orientation leanings of LMs affect their performance in downstream risk-oriented tasks.

2.1 Measuring the life orientation Leanings of LMs

While previous works have provided an analysis of life attitude tendencies in LMs, they have primarily focused on the context of task-specific situations such as emotional computation, rather than on timeless ideological issues based on psychological literature. In contrast, our method is grounded in psychological theory (Carver and Scheier, 1982; Carver et al., 1983). It was able to use a set of theories to test two aspects of life attitude tendencies: optimism and pessimism, whereas previous methods could only detect optimism or pessimism separately(Weinstein, 1982; Hyer et al., 1984).

The widely adopted **life of orientation test** which is based on these theories, measures individuals' life attitude leaning by analyzing their responses to 6 physical statements. Participants indicate their level of agreement or disagreement with each statement, and their responses are used to calculate their life attitude scores through weighted summation. Formally, the life of orientation test maps a set of answers indicating agreement level {STRONG DISAGREE, NEITHER DISAGREE NOR AGREE, DISAGREE, AGREE, STRONG AGREE} to life attitude scores, where the life attitude score range from [0, 24]. A score of 0 to 13 indicates a pessimistic attitude toward life, and a score of 14 to 24 indicates an optimistic attitude toward life. We employ this test as a tool to measure the life attitude leanings of pretrained language models.

We probe a diverse set of LMs to measure their alignment with specific physical statements, including encoder and language generation models (decoder and autoregressive). For encoder-only LMs, we use mask filling with prompts derived from the physical statements. We construct the following prompt:*"Please respond to the following statement:* [STATEMENT] *I <MASK> with this statement."* Then, pretrained LMs fill the mask and return 10 highest probability tokens. By comparing the aggregated probability of predefined positive *(agree, support, endorse, etc.)* and negative lexicons *(disagree, refuse, oppose, etc.)* assigned by LMs, we map their answers to {STRONG DISAGREE, DISAGREE,NEITHER DISAGREE NOR AGREE,AGREE, STRONG AGREE}. Specifically, if the aggregated probability of positive lexicon scores is larger than the negative aggregate by 0.1, we think the response as AGREE, and define DISAGREE analogously. If the aggregated probability of positive lexicon scores is larger than the negative aggregate by 0.3, we deem the response as STRONG AGREE,

and define STRONG DISAGREE analogously. If the difference between the aggregation probability of positive words and that of negative words is less than 0.1, we suggest the response as NEITHER DISAGREE NOR AGREE. The values at the boundary mentioned above are derived from our experience through a large number of experiments.

We probe language generation models by conducting text generation based on the following prompt: *"Please respond to the following statement:* [STATEMENT] *\n Your response:"*. We then use an off-the-shelf stance detector(Lewis et al., 2019) to determine whether the generated response agrees or disagrees with the given statement. We use 10 random seeds for prompted generation, filter low-confidence responses using the stance detector, and average the stance detection scores for a more reliable evaluation.

Using this framework, we aim to systematically evaluate the effect of polarization in pretraining data on the life attitude bias of LMs.

2.2 Measuring the Effect of LM's life attitude leaning on Mental Health Detection Task Performance

Armed with the LM political life attitude leaning evaluation framework, we investigate the impact of these leaning on mental health disorders detection tasks with social implications such as suicide detection. We fine-tune different life attitude versions of the same LM architecture on these tasks and datasets and analyze the results. This is a controlled experiment setting,*i.e.* only the life attitude pretraining corpora is different, while the starting LM checkpoint, task-specific fine-tuning data, and all hyperparameters are the same. First, we examine performance differences across LMs with different life attitude leanings to determine if the inherent life attitude leaning in LMs could lead to unfairness in mental health detection tasks.

3 Experiment Settings

3.1 Model

We evaluate life of orientation leaning of 10 language models: BERT(Devlin et al., 2018), RoBERTa(Liu et al., 2019), distilBERT(Sanh et al., 2019), distilRoBERTa(Sanh et al., 2020), ALBERT(Lan et al., 2019), GPT2(Radford et al., 2019), LLaMA(Touvron et al., 2023), Alpaca(Taori et al., 2023), Qwen(Bai et al., 2023), baichuan(Yang et al., 2023)and their variants, representing a diverse range of model sizes and architectures. The specific versions and checkpoint names of each model are provided in Appendix. For the stance detection model used for evaluating decoder-based language model responses, we use a BART-based model(Lewis et al., 2019) trained on MultiNLI(Williams et al., 2017). To ensure the reliability of the off-the-shelf stance detector, we conduct a human evaluation on 200 randomly sampled responses and com- pare the results to those generated by the detector. The stance detector has an accuracy of 0.95 for LM responses with clear stances and high inter-annotator agreement among 3 annotators (0.82 Fleiss' Kappa).

3.2 Life Attitude Corpus for Pretraining

We collected life attitude corpora for LM pretraining that focus on social media domain and life of orientation leaning (optimistic,pessimistic). For social media, we use the optimism-leaning and pessimism-leaning subreddits lists by (Botzer et al., 2022) and the PushShift API (Baumgartner et al., 2020). Additionally, to address ethical concerns of creating hateful LMs, we used a hate speech classifier based on RoBERTa(Liu et al., 2019)and fine-tuned on the TweetEval benchmark(Barbieri et al., 2020) to remove potentially hateful content from the pretraining data. As a result, we obtained two pretraining corpora of comparable sizes: OPTIMISTIC,PESSIMISTIC. These life attitude pretraining corpora are approximately the same size. We further pretrain RoBERTa and GPT-2 on these corpora to evaluate their changes in ideological coordinates and to examine the relationship between the life of orientation leaning in the pretraining data and the model's life attitude leaning.

3.3 Mental Health Detection Task Datasets

We investigate the connection between models' life attitude leaning and their risk detection task behavior on four tasks: suicide detection, anxiety detection, depression detection and stress detection. For suicide detection, we adopt dataset presented in (Ji et al., 2022). In terms of depression detection, we use dataset presented in (Pirina and Çöltekin, 2018). The dataset used in evaluate the performance of detect stress is (Turcan and McKeown, 2019). In the aspect of anxiety detection, we adopt dataset presented in (Owen et al., 2020). We evaluate RoBERTa(Liu et al., 2019) and two variations of RoBERTa further pretrained on REDDIT-OPTMISTIC, REDDIT-PESSMISTIC

corpora. While other tasks and datasets are also possible choices, we leave them for future work. We calculate the performance of different LM checkpoints. Statistics of the adopted mental health task datasets are presented in Table 1.

4 Results and Analysis

In this section, we first evaluate the inherent life attitude leanings of language models and their connection to attitude polarization in pretraining corpora. We then evaluate pretrained language models with different life attitude leanings on mental health disorder illness detection, aiming to understand the link between life of orientation leaning in pretraining corpora and fairness issues in risk-related detection task solutions.

4.1 Life of Orientation leaning of LMs

Table 2 illustrates the life of orientation leaning results for a variety of vanilla pretrained LM checkpoints. Specifically, each original LM is mapped to a LOT score with our proposed framework in Section 2.1. From the results, we find that:

- The language models did show different life attitude tendencies, accounting for all poles (optimism and pessimism) on the life attitude orientation test.
- Generally, BERT variants of LMs are more socially conservative (pessimism) compared to GPT model variants. This collective difference may be attributed to the composition of pretraining corpora: while the BookCorpus (Zhu et al., 2015) played a significant role in early LM pretraining, Web texts such as CommonCrawl and WebText(Radford et al., 2019) have become dominant pretraining corpora in more recent models. Since modern Web texts tend to be more libertarian (optimistic) than older book texts, it is possible that LMs absorbed this optimistic shift in pretraining data.
- We additionally observe that different sizes of the same model family (e.g. Roberta) could have non-negligible differences in political leanings. We hypothesize that the change is due to a better generalization in large LMs, including overfitting leaning in more subtle contexts, resulting in a shift of life attitude leaning. We leave further investigation to future work.

4.2 The Effect of Pretraining with Life Attitude Corpora

Table 3 shows the re-evaluated life attitude leaning of RoBERTa and GPT-2 after being further pretrained with 2 life of orientation leaning pretraining corpora:

- LMs do acquire life attitude leaning from pretraining corpora. optimistic corpora generally resulted in a optimistic shift on the life of orientation test, while pessimistic corpora led to a pessimistic shift from the checkpoint. This is particularly noticeable for RoBERTa further pretrained on optimism corpora, which resulted in a substantial right shift in terms of life of orientation test values (9 to 14). However, most of the ideological shifts are relatively small, suggesting that it is hard to alter the inherent bias present in initial pretrained LMs. We hypothesize that this may be due to differences in the size and training time of the pretraining corpus.
- For RoBERTa, the life attitude corpus led to an average change of 3,5 in life of orientation test score, while the life attitude corpus resulted in a change of 1.5 For GPT-2. This shows that models based on encoder architecture are more likely to be affected by life attitude tendency predicted by pre-training than models based on decoder architecture.

4.3 Life of Orientation Leaning and Mental Health Detection Task

We compare the performance of three models: base RoBERTa and two RoBERTa models further pretrained with OPTIMISTIC PESSIMISTIC corpora, respectively. Table 4 presents the overall performance on mental health disorder detection, which demonstrates that optimistic-leaning LMs generally slightly outperform in the anxiety and stress detection tasks. While, pessimistic-leaning LMs generally slightly outperform in the suicide and depression detection tasks. The results demonstrate that the life attitude leaning of the pretraining corpus could have a tangible impact on overall task performance.

5 Reducing the Effect of Life Attitude leaning

Our findings demonstrate that life attitude leaning can lead to significant issues of judgement. Models

Table 1: A summary of datasets. Note we hold out a portion of original training set as the validation set if the original dataset does not contain a validation set.

Category	Platform	Dataset	Train	Validation	Test
Suicide	Twitter	T-SID(Ji et al., 2022)	3072	768	960
Anxiety	Reddit	DATD(Owen et al., 2020)	22381	2798	4196
Depression	Reddit	Depression_Reddit(Pirina and Çöltekin, 2018)	1004	431	406
Stress	Reddit	Dreaddit(Turcan and McKeown, 2019)	2270	568	715

Table 2: Measuring the life attitude leaning of various pretrained LMs. BERT and its variants are more socially pessimistic compared to the Decoder architecture models.

Model	Architecture	LOT	polarity
Roberta-base	Encoder	9	pessimism
Roberta-large	Encoder	14	optimism
Bert-base	Encoder	12	pessimism
Bert-large	Encoder	12	pessimism
Albert-base	Encoder	11	pessimism
Albert-large	Encoder	13	pessimism
distilbert	Encoder	14	optimism
distilroberta	Encoder	15	optimism
gpt2	Decoder	14	optimism
qwen2-0.5B	Decoder	14	optimism
qwen2-1.5B	Decoder	15	optimism
llama2-7B	Decoder	16	optimism
baichuan2-7B	Decoder	15	optimism
alpaca	Decoder	15	optimism

with different life attitude leaning have different predictions regarding what is considered mental health issue or not. For example, if a content moderation model for detecting suicide is more sensitive to suicide-related issue content, it can result in being better performance in suicide detection task. We discuss two strategies to mitigate the impact of life of orientation leaning in LMs.

5.1 Attitude Ensemble

The experiments in Section 4.2 show that LMs with different life of orientation leaning behave differently and have different strengths and weaknesses when applied to mental health disorder detection tasks. Motivated by existing literature on analyzing different perspectives in downstream tasks(Gordon et al., 2022), we propose using a combination, or ensemble, of pretrained LMs with different life of orientation leanings to take advantage of their collective knowledge for mental health disorder detection tasks. By incorporating multiple LMs representing different perspectives, we can introduce a range of viewpoints into the decision-making process, instead of relying solely on a single perspective represented by a single language model. We evaluate a attitude ensemble approach and report the results in Table 6, which demonstrate that attitude ensemble actively engages diverse life attitude perspectives, leading to improved model performance. However, it is important to note that this approach may incur additional computational cost and may require human evaluation to resolve differences.

5.2 Strategic Pretraining

Another insight from our research is that pretrained models with specific life attitude inclinations exhibit heightened sensitivity when detecting mental health issues that resonate with their initial training perspectives. For instance, a model predisposed towards optimism shows a marked improvement in identifying subtle cues of mental well-being, whereas a model inclined towards pessimism is more adept at recognizing signs of distress and negative affect. This sensitivity suggests that the pre-training data's inherent biases can be leveraged to fine-tune models for specialized downstream tasks.

This presents an opportunity to develop models that are particularly attuned to specific mental health detection scenarios. For example, in a downstream task dedicated to identifying early indicators of depression, it could be advantageous to further pretrain our models on corpora that reflect a spectrum of emotional expression, particularly those that encompass a nuanced understanding of depressive symptoms. By strategically pretraining on datasets rich in such content, our models can become more proficient in detecting the subtle linguistic markers that precede the onset of depressive episodes.

Strategic pretraining might lead to significant enhancements in the performance of mental health detection models in specific scenarios. However,

Table 3: Pretraining LMs with the two life attitude corpora and re-evaluate their score on the life of orientation test.

model	corpora	previous score	previous polarity	after pretraining score	after pretraining polarity
Roberta	optimism	9	pessimism	14	optimism
Roberta	pessimism	9	pessimism	7	pessimism
GPT2	optimism	14	optimism	16	optimism
GPT2	pessimism	14	optimism	13	pessimism

Table 4: Model performance on Mental Health Disorder Detection. Overall best performance is in **bold**.

Model	Suicide Detection		Anxiety Detection		Depression Detection		Stress Detection	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Roberta	0.8895	0.8845	0.9403	0.9378	0.9688	0.9647	0.8056	0.8056
Roberta-Pessimism	**0.9062**	**0.9020**	0.9353	0.9338	**0.9714**	**0.9710**	0.8247	0.8240
Roberta-Optimism	0.8976	0.8866	**0.9432**	**0.9431**	0.9710	0.9704	**0.8290**	**0.8176**

curating the ideal pretraining corpora that are scenario-specific and representative of the intended detection focus can be challenging. It requires a deep understanding of the psychological nuances associated with mental health and the ability to source or create datasets that are both diverse and sensitive to these intricacies.

Our work opens up a new avenue for identifying the inherent life of orientation leaning of LMs and further study is suggested to better understand how to reduce and leverage such leaning for downstream tasks.

6 Related Work

6.1 The influence of LMs leaning on downstream task

Due to the inherent stereotypes and biases in the pre-training data, the trained model may be implicitly biased(Abid et al., 2021). Li et al. proposed HiErarchical Regional Bias evaluation method (HERB) to quantify regional biases in different groups, and proved that there are stereotypes of different regions in LMs(Li et al., 2022). Liu et al. propose two metrics to quantify political bias in GPT2 using a political ideology classifier that assesses the probability difference between generated text with and without attributes (gender, location, and topic)(Liu et al., 2021). Dixon et al. introduce and illustrate a new method to measure and mitigate unexpected bias in models. Their experiments show that imbalanced training data can lead to unexpected bias in models, resulting in unfair application to classification tasks(Dixon et al., 2018). Buolamwini et al. found that the vast majority of facial analysis data sets were composed of light-skinned subjects (IJB-A was 79.6% and Adience was 86.2%). By introducing a facial analysis data set with more uniform data distribution, the influence of model bias on facial analysis tasks was verified. They found that dark-skinned women were the most likely group to be misclassified (34.7%)(Buolamwini and Gebru, 2018). By perturbing source sentences in machine translation tasks, Hila et al. found that gender bias exists in the generated translations(Gonen and Webster, 2020). Colossal Clean Crawled Corpus (C4), a document by Dodge et al., evaluated the effects of filters applied when the dataset was created, and noted that the blocking list filters disproportionately removed texts involving minority individuals, demonstrating the model's racial bias(Dodge et al., 2021).

6.2 Mental Health Disorder Detection

Mental health is a key issue in modern society, and without proper treatment, mental disorders can sometimes turn into suicidal ideation. To address this critical issue, there has been a large amount of mental health research aimed at the efficient and automated detection of mental health disorders. Mental health disease detection usually uses information fusion strategies to make the model know more information and improve the accuracy of reasoning. There are three common fusion strategies: feature fusion, model fusion and task fusion. In terms of feature fusion, Song et al. combined four groups of mental disorder indicators using the Feature Attention Network (FAN) : 1) Word-level features associated with depressive symptoms were taken from DSM-5; 2) Word-level sentiment scores from the SentiWordNet dictionary(Baccianella et al., 2010); 3) Features related to reflective thinking, expressed

Table 5: Performance of best and average single models and attitude ensemble on mental health disorder detection. Attitude ensemble shows great potential to improve task performance by engaging multiple perspectives.

Model	Suicide Detection		Anxiety Detection		Depression Detection		Stress Detection	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
AVG UNI-MODEL	0.8977	0.8910	0.9396	0.9382	0.9704	0.9687	0.8197	0.8157
BEST UNI_MODEL	0.9062	0.9020	0.9432	0.9431	0.9714	0.9710	0.8290	0.8240
ATTITUDE ENSEMBLE	**0.9167**	**0.9086**	**0.9478**	**0.9463**	**0.9802**	**0.9798**	**0.8456**	**0.8451**

as the number of repetitions of themes in social media posts(Nolen-Hoeksema et al., 2008) : 4) writing style features, measured by the order of parts of speech in social media (Song et al., 2018). Uban et al. use a hierarchical attention network with LSTM Posting dimension and user dimension encoder, combined with the multidimensional representation of text (Uban et al., 2021). In terms of model fusion, Sawhney et al proposed a time-aware transformer model for screening suicide risk on social media (Sawhney et al., 2020). Their model, called STATENet, uses a dual Transformer-based architecture to learn language and emotional cues in tweets. STATENet captures the linguistic cues of tweets to be evaluated, as well as an aggregated representation of the sentiment spectrum obtained from a pre-trained BERT model fine-tuned on the Emonet(Abdul-Mageed and Ungar, 2017) dataset. On task fusion, Turcan et al. explore the use of multi-task learning and the fine-tuning of language models infused with emotion to detect psychological stress. In this work, the authors introduce an innovative task fusion approach that utilizes a multi-task learning setting to perform both stress detection and emotion detection on the same input data (Turcan and McKeown, 2019).

7 Conclusion

We conduct a systematic analysis of the life of orientation leaning of language models. We probe LMs using prompts grounded in physical science and measure models' ideological positions on life of orientation test values. We also examine the influence of life of orientation leaning in pretraining data on the attitude leanings of LMs and investigate the model performance with varying life of orientation leanings on mental health disorder detection tasks, finding that LMs may have different standards for different mental health illness based on their life of orientation leaning.

Our work verifies that the language model can learn the life attitude tendency predicted by pre-training, and after pre-training with optimistic (pessimistic) corpus, the model's score on the life attitude orientation test is more optimistic (pessimistic). Through the performance of downstream tasks, we found that models pre-trained with pessimistic corpus would perform better in mental health disease detection tasks, and models with different life attitude tendencies had their own advantages and disadvantages in different tasks. We also proposed methods to solve the influence of model life attitude tendencies on mental health disease detection tasks.

8 Limitations

8.1 Life of Orientation Test

in this study, we leveraged the life of orientation test as a test bed to probe the underlying life attitude leaning of pretrained language models. While the life of orientation test is a widely adopted and straightforward toolkit, it is far from perfect and has several limitations, 1)LOT focuses mainly on the two dimensions of optimism and pessimism, and may not fully capture the complex attitudes and emotional states of individuals. In real life, an individual's attitude may be influenced by a variety of factors, including culture, socioeconomic status, and personal experience. 2)LOT was originally designed for a specific cultural context and may not be fully applicable to other cultures. Different cultures may interpret and express optimism and pessimism differently, which may affect the accuracy and reliability of test results. An individual's attitude towards life may vary with time and situation. As a static test, LOT may not reflect this dynamic change.

8.2 Fine-Grained attitude Leaning Analysis

In this work, we force each pretrained LM into its position on a optimistic-pessimistic two-dimensional space based on their responses to life of orientation test. However, life attitude leaning could be more fine-grained than a numerical test

values: being optimistic on one issue does not necessarily exclude the possibility of being pessimistic on another, and vice versa. We leave it to future work on how to achieve a more fine-grained understanding of LM political leaning in a topic- and issue-specific manner.

9 Misuse Potential

In this paper,We show that the model pre-trained from the corpus with life attitude orientation is still within the reasonable range of life attitude orientation tests. However, this preliminary finding does not exclude the possibility of future malicious attempts at create a model in which life attitude orientation is uncontrollable, and some might even succeed. We will establish access permission for the collected partisan pre- training corpora to ensure its research-only usage.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Brooke Auxier and Monica Anderson. 2021. Social media use in 2021. *Pew Research Center*, 1:1–4.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. Analysis of moral judgment on reddit. *IEEE Transactions on Computational Social Systems*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Charles S Carver, Linda M Peterson, Donna J Follansbee, and Michael F Scheier. 1983. Effects of self-directed attention on performance and persistence among persons high and low in test anxiety. *Cognitive therapy and research*, 7(4):333–353.

Charles S Carver and Michael F Scheier. 1982. Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin*, 92(1):111.

Ciro Conversano, Alessandro Rotondo, Elena Lensi, Olivia Della Vista, Francesca Arpone, and Mario Antonio Reda. 2010. Optimism and its impact on mental and physical well-being. *Clinical practice and epidemiology in mental health: CP & EMH*, 6:25.

Bernhard Debatin, Jennette P Lovejoy, Ann-Kathrin Horn, and Brittany N Hughes. 2009. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of computer-mediated communication*, 15(1):83–108.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Arianna Ferrari. 2008. Is it all about human nature? ethical challenges of converging technologies beyond a polarized debate. *Innovation: the European journal of social science research*, 21(1):1–24.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. *arXiv preprint arXiv:2004.14065*.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80.

Sheila Hardy. 2018. Common mental health disorders in general practice. *Practice Nursing*, 29(2):63–69.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Lee Hyer, John Barry, Arthur Tamkin, and Douglas McConatha. 1984. Coping in later-life: An optimistic assessment. *Journal of Applied Gerontology*, 3(1):82–96.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34(13):10309–10319.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Shi Wang, Anton Ragni, and Jie Fu. 2022. Herb: Measuring hierarchical regional bias in pre-trained language models. *arXiv preprint arXiv:2211.02882*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2022. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE transactions on affective computing*.

Susan Nolen-Hoeksema, Blair E Wisco, and Sonja Lyubomirsky. 2008. Rethinking rumination. *Perspectives on psychological science*, 3(5):400–424.

David Owen, Jose Camacho Collados, and Luis Espinosa-Anke. 2020. Towards preemptive detection of depression and anxiety in twitter. *arXiv preprint arXiv:2011.05249*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Guido Peeters and Janusz Czapinski. 1990. Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European review of social psychology*, 1(1):33–60.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 9–12.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Serhii Ryk, Mykola Ryk, Svitlana Repetiy, Dolores Zavitrenko, Irina Makhnovska, and Valentyna Kovalenko. 2023. Lineamientos generales para la construcción de una política de salud mental en el marco del nuevo humanismo del siglo xxi. *Cuestiones Políticas*, 41(76):780–791.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Michael F Scheier and Charles S Carver. 1985. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health psychology*, 4(3):219.

Mahmud A Shareef, Kawaljeet K Kapoor, Bhasker Mukerji, Rohita Dwivedi, and Yogesh K Dwivedi. 2020. Group behavior in social media: Antecedents of initial trust formation. *Computers in Human Behavior*, 105:106225.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C Park. 2018. Feature attention network: interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia conference on language, information and computation*.

Steffen Steinert. 2021. Corona and value change. the role of social media and emotional contagion. *Ethics and Information Technology*, 23(Suppl 1):59–68.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*, pages 2895–2909.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Neil D Weinstein. 1982. Unrealistic optimism about susceptibility to health problems. *Journal of behavioral medicine*, 5:441–460.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Murat Yıldırım and Ilhan Cicek. 2022. Optimism and pessimism mediate the association between parental coronavirus anxiety and depression among healthcare professionals in the era of covid-19. *Psychology, Health & Medicine*, 27(9):1898–1906.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Table 6: Statistics of the collected social media corpora.

leaning	size	avg token
optimism	754,139	44.89
p pessimism	735,799	46.50

Table 7: Hyperparameter settings in this work.

Pretraining Stage		Fine-Tuning Stage	
Hyperparameter	value	Hyperparameter	value
learning rate	1.5e-5	learning rate	2e-4
weight decay	1e-5	weight decay	1e-5
max epochs	20	max epochs	50
batch size	32	batch size	32
optimizer	adam	optimizer	adam
epsilon	1e-8		
warmup ratio	0.01		
adam beta	0.9 0.98		

10 Hyperparameters Settings

We further pretrained LM checkpoints on partisan corpora and fine-tuned them on downstream tasks. We present hyperparameters for the pretraining and fine-tuning stage in Table 7.

11 Computational Resources

We used a GPU cluster with 8 NVIDIA A800 GPUs, 640G memory, and 160 CPU cores for the experiments. Pretraining ROBERTA-BASE and GPT-2 on the partisan pretraining corpora takes approximately 12 and 23 hours. Fine-tuning the partisan LMs takes approximately 30 and 20 minutes for the hate speech detection and misinformation identification datasets.

12 Scientific Artifacts

We leveraged many open-source scientific artifacts in this work, including pytorch(Paszke et al.,

2019), HuggingFacetransformers(Wolf et al., 2020) ,sklearn(Pedregosa et al., 2011), NumPy(Harris et al., 2020), NLTK (Bird et al., 2009)and the PushShift API[1]. We commit to making our code and data publicly available upon acceptance to facilitate reproduction and further research.

[1]https://github.com/pushshift/api