WWBP-SQT-lite: Difference Embeddings and Multi-level Models for Moments of Change Identification in Mental Health Forums

Adithya V Ganesan[1], **Vasudha Varadarajan**[1], **Juhi Mittal**[2],
Shashanka Subrahamanya[3], **Matthew Matero**[1], **Nikita Soni**[1],
Sharath Chandra Guntuku[2], **Johannes C. Eichstaedt**[3] and **H. Andrew Schwartz**[1]
[1]Stony Brook University, [2]University of Pennsylvania, [3]Stanford University
{avirinchipur, vvaradarajan}@cs.stonybrook.edu

Abstract

Psychological states unfold dynamically; to understand and measure mental health at scale we need to detect and measure these changes from sequences of online posts. We evaluate two approaches to capturing psychological changes in text: the first relies on computing the difference between the embedding of a message with the one that precedes it, the second relies on a "human-aware" multi-level recurrent transformer (HaRT). The mood changes of timeline posts of users were annotated into three classes, 'ordinary,' 'switching' (positive to negative or vice versa) and 'escalations' (increasing in intensity). For classifying these mood changes, the difference-between-embeddings technique – applied to RoBERTa embeddings – showed the highest overall F1 score (0.61) across the three different classes on the test set. The technique particularly outperformed the HaRT transformer (and other baselines) in the detection of switches (F1 = .33) and escalations (F1 = .61). Consistent with the literature, the language use patterns associated with mental-health related constructs in prior work (including depression, stress, anger and anxiety) predicted both mood switches and escalations.

1 Introduction

Detecting shifts in mental health from language use could assist in identifying episodes of mental ill health and providing in-time treatment for conditions such as depression or anxiety. The accessibility and abundant usage of social media (Coppersmith, 2022) in comparison to traditional healthcare data (e.g. hospital visits) is enabling first steps toward unprecedented assessment and understanding of mental health, including detection of elevated risks (Choudhury et al., 2016; Zirikly et al., 2019; Guntuku et al., 2021). However, most language datasets for mental health classification are annotated statically such that a person has just one label across all of their language (Coppersmith et al., 2014; Lynn et al., 2018). Longitudinal language datasets can help analyze the mental state of a person over time (Halder et al., 2017; Matero and Schwartz, 2020; Son et al., 2021), but also open the door for many sequential, differencing, and time-series modeling techniques.

Here, we explore two types of modeling techniques that can capture changes over time: Human-aware Recurrent Transformers (Soni et al., 2022) and difference embeddings. These techniques were used as part of the WWBP-SQT-lite[1] system for the CLPsych 2022 shared tasks (Tsakalidis et al., 2022a): (Task A) modeling user state changes over time (Tsakalidis et al., 2022b), and (Task B) the suicide risk associated with the user (Shing et al., 2018), our **contributions** are as follows: (a) evaluation of Human-aware Recurrent Transformers (HaRT) and difference embeddings for Task A (b) exploring SoTA methods for predicting state escalations and switches, and (c) exploring theoretically related linguistic assessments.

2 Data

2.1 Task A

Task Data. The training data for task A contained $5,143$ Reddit posts comprising of titles and contents from 149 users spanning over 204 timelines. As described in Tsakalidis et al. 2022b, posts from each timeline were annotated with the Moment of Change (MoC) of the user's mood into three classes, namely, "Ordinary" (O), "In Switch" (IS) when the mood changes from positive to negative or vice versa, and "In Escalation" (IE) signifying mood progressions, i.e., changes from neutral or positive to more positive or negative to more negative. The posts were annotated in the context

[1] **SQT**: **S**eawolf, **Q**uaker, and **T**ree (the mascots of the schools composing our team); **lite**: due to constraints out of our control, we were restricted to just 4 days working with the data, covering only a portion of our planned human-level and temporal approaches.

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 251 - 258
July 15, 2022 ©2022 Association for Computational Linguistics

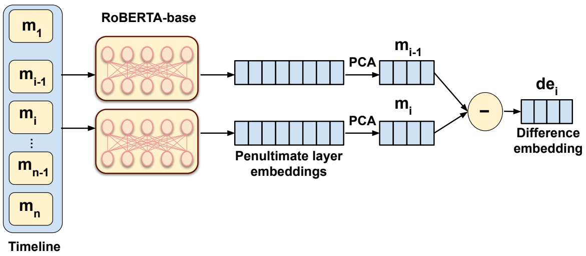

Figure 1: Difference embedding for the current message is obtained using a point-wise subtraction on the the current embedding and previous message's embedding.

of other posts from timelines as carried out in the CLEF eRisk 2020 dataset (Losada and Crestani, 2016; Losada et al., 2020). A small number of timelines in the CLPsych 2022 data was extracted from the CLEF eRisk 2020 dataset.

Internal Train & Validation sets. 119 (80%) randomly chosen users were sampled to form an internal train set, and the remaining 30 users were used for validation set. This resulted in 3,974 posts (78%) over 164 timelines in train set and 1,169 posts over 40 timelines in the validation set.

2.2 Task B

Task Data The goal of task B is to predict the Suicide risk level associated with the Reddit users into Low, Moderate or Severe. It utilizes the same data as Task A to the exclusion of 22 users which were annotated as "N/A". Thus a total of 127 users were present in task B, who collectively posted a total of 4,507 Reddit posts, averaging at around 35 posts per user. The risk level of the user was assigned as the maximum risk level annotation across all their posts. The suicide risk annotation followed the procedure described in Shing et al. 2018 and Zirikly et al. 2019.

Internal Train & Validation sets A random sample of 101 users (79.5%) were chosen for the internal train set – a total of 3,761 posts for training and the remaining 26 (20.5%) users for the validation set, resulting in 746 posts in the validation set.

2.3 Evaluation

For Task A, macro F1 and coverage-based (Arbeláez et al., 2011; Tsakalidis et al., 2022b) precision and recall scores were used to measure the performance of the models. The coverage based metrics are aimed at evaluating the model's ability to capture the regions of change. However, for task B, only macro F1 is used to evaluate model performance.

Dimension	β_O	β_{IE}	β_{IS}
Big 5 Traits			
Emotional Stability	.57‡	-.14‡	-.38‡
Extraversion	.18‡	-.05	-.12‡
Conscientiousness	.13‡	-.03	-.12†
Agreeableness	.13‡	-.01	-.12‡
Openness to Experience	-.04	.01	.03†
Anger	-.35‡	.09†	.24‡
Anxiety	-.48‡	.13‡	.31‡
Stress	-.58‡	.14‡	.39‡
Depression	-.59‡	.14‡	.39‡
Loneliness	-.82‡	.20‡	.56‡

Table 1: Association (standardized logistic regression coefficients, β) of theoretical features measures in language with the three classes of task A (ordinary, in-escalation, or in-switch). †: $p < .05$; ‡:$p < .001$

3 Methods

3.1 Task A

Beyond utilizing the best transformer based approaches, we also explore relevant theoretical features to understand the relationship between moment of change and psychological/demographic constructs. Furthermore, recent works (Sawhney et al., 2020, 2021) have shown the importance of joint modelling of such theoretical dimensions with the present-day neural approaches.

HypLex. To quantify the association of psychological and demographic constructs with the moments of change, 12 models trained on larger datasets were used to derive theoretical features which we call HypLex (short for Hypothesis-driven Lexica). These models include Cohen's stress (Guntuku et al., 2019a), depression, anger and anxiety (Schwartz et al., 2014; Son et al., 2021; Guntuku et al., 2019b), age and gender (Sap et al., 2014), loneliness expressions (Guntuku et al., 2019c), and the big 5 personality traits (Park et al., 2015). All these features were on a continuous scale.

HaRT. Recent works (Lynn et al., 2020; Matero et al., 2021b; Soni et al., 2022) have highlighted the importance of incorporating author context into the message representations through the use of history and multi-level modeling. We use the Human aware Recurrent Transformer model (Soni et al., 2022) which is built on GPT2 (Radford et al., 2019), to produce message representations that encode the latent representation of the author as well.

Method	Post-level Evaluation												Coverage-based	
	IS			IE			O			macro avg			macro avg	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Internal Validation														
HypLex	.00	.00	.00	.52	.34	.41	.84	.95	.89	.46	.43	.44	.27	.30
HaRT														
CLS	.16	.13	.14	.55	**.67**	**.60**	**.92**	.88	.90	.54	.56	.55	.39	.43
CLS+Last layer	.16	.15	.15	.56	.63	.59	.91	.89	.90	.55	.56	.56	.38	.44
PCA-Roba														
Curr	.09	.09	.09	.56	.46	.50	.90	.93	**.92**	.52	.49	.50	.36	.38
Diff	**.44**	**.34**	**.38**	.33	.03	.05	.81	**.98**	.88	.53	.45	.44	.39	.29
Curr+Prev	.42	.25	.31	.64	.53	.58	.89	.95	**.92**	**.65**	.57	.60	.43	.43
Curr+Diff	.38	.26	.31	.63	.46	.53	.88	.95	.91	.63	.56	.59	.42	.43
Curr+Prev+Diff	.42	**.34**	.37	**.65**	.52	.58	.89	.94	.91	**.65**	**.60**	**.62**	**.44**	**.46**
Test Set														
HaRT														
CLS	.23	.22	.22	.44	.46	.45	.85	.55	.85	.51	.51	.51	.34	.38
CLS+Last layer	.23	.20	.21	.43	.48	.45	.86	.84	.85	.50	.50	.50	.33	.37
PCA-Roba														
Curr+Prev+Diff*	**.42**	**.27**	**.33**	**.67**	**.56**	**.61**	**.86**	**.94**	**.90**	**.65**	**.59**	**.61**	**.46**	**.47**

Table 2: Results on internal validation and the test set for task A. IS, IE, and O refer to Switch, Escalation, and Ordinary classes respectively, and P, R and F1 refer to precision, recall, and F1 score. **Best** scores are highlighted. The variants of HaRT (Soni et al., 2022) refer to the fine tuning of the classification layer (CLS) and the last transformer layer (last Layer). The variants of PCA-Roba refer to the Current (Curr), Previous (Prev), and Difference (Diff) between the two on Roberta embeddings of text reduced using PCA. *The PCA-Roba (Curr+Prev+Diff) was turned in late due to technical difficulties.

We adapted HaRT in two ways. First, we try a frozen approach where we train using the message representation output from HaRT but only update weights of the classification layer. We call this approach **HaRT CLS**. Second, we allow a single transformer layer (the topmost layer) to also update its weights during fine-tuning, this variant is called **HaRT CLS+Last Layer**.

RoBERTa. Previous works have shown that contextual embeddings from large pre-trained language models can help improve downstream task performance (Matero et al., 2021a; Bao and Qiao, 2019). However, these models often output embeddings with a large number of dimensions, typically 768 or 1024, which can cause problems when training on small datasets (Li and Eisner, 2019; Bao and Qiao, 2019). Here, we leverage the dimensionality reduction approach proposed by V Ganesan et al. 2021, which suggests using RoBERTa embeddings (Liu et al., 2019) with PCA (Martinsson et al., 2011) to achieve the best performance in low data regime. Further, we incorporate techniques proposed in previous works on suicide risk-level assessment, such as modeling the title and message body of a post separately and concatenating them for a single representation (Matero et al., 2019).

To build our text representations, we extract separate transformer representations for title and body, from the second to last layer of RoBERTa. This allows us to keep highly relevant features, the individual words in the title, from getting underrepresented in the longer text from the body content. We then run our PCA reduction on each representation individually, down to 16 dimensions for title and 128 for the body, then concatenate them into a single representation of 144 dimensions. The number of reduced dimensions for title and body were chosen based on cross validation performance using 16, 64 and 128 dimensions. We observed no improvement in performance when increasing the dimensions for title from 16, but observed degradation in performance when decreasing the dimensions from 128 for the body.

Using dimension reduced RoBERTa (PCA-Roba) embeddings as a base, we build 5 separate models that each use different combinations of feature representations. (1) **Curr** uses only the current message as input features, (2) **Curr+Prev** uses both current and previous message representations concatenated, (3) **Diff** uses the difference in representation between the current and the previous messages as shown in figure 1, (4) **Curr+Diff** uses *diff* concatenated with only the *current*, and (5) **Curr+Prev+Diff** uses the *current*, *previous*, and *difference* representations all concatenated.

All feature representations are fit using a logistic regression model. To the exception of HaRT, experiments were performed using an open

Features	Cross Val F1(macro)	Internal Val F1(macro)
1gram	0.37	0.29
Roba	0.34	0.34
OpenVocab	0.39	**0.60** †
OpenVocab, HypLex	0.40	0.37
Roba, HypLex	0.36	0.35
OpenVocab+Roba, HypLex	**0.42**	0.37

Table 3: Results on the cross validation and internal validation set for task B. **Best scores** are highlighted. All the features were extracted for titles and message body separately. The OpenVocab consists of PCA reductions of the LDA Topics and 1-grams to 32 dimensions each. For Roba, we reduce 768 dimensions to 64 in case of contents and 16 in case of titles. HypLex is a set of 12 theoretical features as explained in §3.1.
† : the drastically high F-1 score is likely from chance due to the very low sample sizes afforded for the user-level task.

source python library for language analysis at scale, DLATK (Schwartz et al., 2017). The design of the library to support multiple levels of analysis for both linguistic and extra-linguistic features facilitated using it for both task A and task B, although the former maps an outcome to each message while the latter maps multiple messages to an outcome.

3.2 Task B

Open-Vocab Features. We explore three representations of a user's language for this task. First, **N grams** are extracted and normalized to obtain the frequencies, from the title and content for each user. The outliers are removed by retaining only the N grams that occur in at least 5% of users' posts. Next, we use the N grams to build **LDA Topics** which are generated using open-source data-driven word clusters Schwartz et al. (2013). These provide 2,000 topics trained on a corpus of 18 million Facebook posts. Each user is represented by the probability of usage for each topic across these 2,000 dimensions. The topic dimensions are then reduced down to 32 using PCA.

Additionally, we again used **PCA-Roba** as described in task A with the same dimension sizes, title/body split, and extraction layer. However, for this task we process all individual messages uttered by a user and average the message representations to build a user representation.

HypLex The HypLex (§3.1) models were run on the N gram counts of the user to obtain the theoretical HypLex features for task B.

We use both Open-Vocab and HypLex features as inputs for a logistic regression model. Internally we tested various combinations of features for this task, but only a single model was selected to be evaluated on the test set.

4 Results

4.1 Task A

As can be seen in Table 1, the mental-health-related hypothesis-driven lexica (HypLex)–including depression, anxiety, anger and loneliness–show high β associations (standardized logistic regression coefficients) with the outcome variables of task A. The 12 HypLex features alone produce a macro F1 of .44 on the internal validation set (Table 2) which demonstrates the power of these machine-learning-based language models learned on person-level survey responses. Throughout, mood 'switches' (from positive to negative and vice versa) where more easily predicted than mood 'escalations.' The absence of language signal related to negative affect (anger, anxiety, stress, depression, loneliness) predicted 'ordinary' mood states, as did the presence of language signal of the three personality traits typically associated with positive affect: extraversion, agreeableness and conscientiousness. Perhaps surprisingly, the language model for the low-arousal negative affect state of loneliness proved to be more predictive of both mood switches and escalations than the language models for high-arousal negative affect states (such as anger, stress, and anxiety).

Generally, the performance of auto-regressive transformer models are poorer than auto-encoder transformer models in classification tasks (V Ganesan et al., 2021; Zhou et al., 2020). However, the results on the internal validation set in Table 2 suggest that HaRT (CLS) performs better than RoBERTa embeddings (PCA-Roba Curr), primarily accounting for the importance of encoding history into text representations, especially for tasks spanning the temporal dimension. However, HaRT CLS+Last layer doesn't seem provide much improvement showing that fine tuning is not of much help. We would like to note that the hyperparameter values were chosen based on values reported in the paper due to the limited availability of time and computational resources.

It is evident from table 2 that the differencing approach of the PCA-Roba embeddings between the current and previous texts (PCA-Roba Diff) gives the best performance in capturing Switches on the internal validation set. However, the difference feature is very poor at capturing the gradual mood change (IE). It was found that 93% of the IE class was predicted as ordinary when using only the difference feature. This could potentially be because the difference in post embeddings for gradual mood changes are much more gradual and smaller, and may be mistaken for no mood changes - whereas the difference in switches have much more obvious and large differences in post embeddings.

The previous text representation in context to the current (PCA-Roba Curr+Prev) vastly improves the model to identify escalations besides improving the detection of switches. Overall the concatenation of Curr, Prev and Diff performs the best by bringing the best out of these individual features.

The strongest baseline from (Tsakalidis et al., 2022a) for task A utilizing tf-idf features trained with a logistic regression scores macro F1 of 0.49 on the test set. All our models perform significantly better than tf-idf features, particularly in capturing the switches by using transformer based embeddings and factoring previous message's representation for modelling the mood change.

4.2 Task B

The results on Table 3 suggests that using dimension-reduced RoBERTa (Roba) does not offer much advantage over dimension-reduced 1-gram features. This is likely due to the availability of small number of training samples where language models have shown to overfit (Bao and Qiao, 2019). The addition LDA Topics improves the performance of both 1gram model (OpenVocab) and the Roba HypLex model (OpenVocab+Roba, HypLex) showing the robustness of the topics trained on large external dataset in such low data regime.

The HypLex features too slightly improve the performance in cross validation. We get the best result in cross validation when we combine all the three – PCA-reduced RoBERTa embeddings + 1grams, PCA-reduced LDA Topics and HypLex features (OpenVocab+Roba, HypLex).

However, in the internal validation, we find that the best performing model was PCA-reduced OpenVocab. Since the performance in the cross validation was similar for all the listed models, we chose OpenVocab for the final predictions for test set on Task B, which scored an F1(macro) of 0.35.

5 Conclusion

We presented two approaches to detecting mental state changes in users through (a) a recurrent transformer model (HaRT) that encodes messages within context of previous ones and (b) a logistic regression model that relies on RoBERTA difference embeddings along with previous and current text representations to capture change in language over time. Compared to using other representation types, such as theoretically motivated (HypLex) or traditional open vocabulary features (N grams, Topics), both approaches saw improved model performance when predicting changes over time. Further, we found that theoretically relevant lexical scores had large associations with the change patterns. It showed emotional stability correlating with no change, and loneliness, depression, stress, anxiety and anger being associated with the mood change.

6 Ethical Consideration

We used publicly available data stripped of identifiable information which was collected in a non-intrusive manner for mental health research. Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20). Individuals of the study team who ran the analyses for this work are certified to conduct Human Subject Research and complied with the non-disclosure agreement signed with the dataset providers.

The findings of this work are intended for fellow researchers in Computational Linguistics and Psychology to improve technology for mental health assessments. Around 14 million adults in the United States face severe mental health issues (NIMH, 2022) and a very large part of this is marginalized communities that are underserved (Saraceno et al., 2007). However, given the prevalence of these communities in social media (Center, 2021), technology-enabled solutions can assist in detecting and providing assistance in a timely manner to a more diverse group of individuals. This work is a part of the growing body of mental health research aimed at applications for improving well-being. However, this shouldn't be deployed to use without collaboration of clinical practitioners.

Acknowledgements

The authors are grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for Task B, the American Association of Suicidology, to NORC for creating and administering the secure infrastructure and providing researcher support and to UKRI for providing funding to the CLPsych 2022 shared task organisers.

References

Pablo Arbeláez, Michael Maire, Charlotte Fowlkes, and Julien Malik. 2011. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:898–916.

Xingce Bao and Qianqian Qiao. 2019. Transfer learning from pre-trained BERT for pronoun resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 82–88, Florence, Italy. Association for Computational Linguistics.

Pew Research Center. 2021. Social media fact sheet.

Munmun De Choudhury, Emre Kıcıman, Mark Dredze, Glen A. Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen A. Coppersmith. 2022. Digital life data in the clinical whitespace. *Current Directions in Psychological Science*, 31:34 – 40.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, and Lyle H. Ungar. 2019a. Understanding and measuring psychological stress using social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):214–225.

Sharath Chandra Guntuku, Elissa V Klinger, Haley J McCalpin, Lyle H Ungar, David A Asch, and Raina M Merchant. 2021. Social media language of healthcare super-utilizers. *NPJ digital medicine*, 4(1):1–6.

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

Sharath Chandra Guntuku, Rachelle Schneider, Arthur Pelullo, Jami Young, Vivien Wong, Lyle Ungar, Daniel Polsky, Kevin G Volpp, and Raina Merchant. 2019c. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open*, 9(11).

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135, Copenhagen, Denmark. Association for Computational Linguistics.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2744–2754, Hong Kong, China. Association for Computational Linguistics.

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.

D. Losada and F. Crestani. 2016. A test collection for research on depression and language use. In *Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016*, pages 28–39, Evora, Portugal.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of eRisk 2020: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*. Springer International Publishing.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316, Online. Association for Computational Linguistics.

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

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. 2011. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30:47–68.

Matthew Matero, Albert Hung, and H. Andrew Schwartz. 2021a. Evaluating contextual embeddings and their extraction layers for depression assessment.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Matero and H. Andrew Schwartz. 2020. Autoregressive affective language forecasting: A self-supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

NIMH. 2022. Mental Illness Statistics.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Benedetto Saraceno, Mark van Ommeren, Rajaie Batniji, Alex Cohen, Oye Gureje, John Mahoney, Devi Sridhar, and Chris Underhill. 2007. Barriers to improvement of mental health services in low-income and middle-income countries. *The Lancet*, 370(9593):1164–1174.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn 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, Online. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16.

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

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Youngseo Son, Sean A. P. Clouston, Roman Kotov, Johannes C. Eichstaedt, Evelyn J. Bromet, Benjamin J. Luft, and H. Andrew Schwartz. 2021. World trade center responders in their own words: predicting ptsd symptom trajectories with ai-based language analyses of interviews. *Psychological Medicine*, page 1–9.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

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

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.