Evaluation of LLMs-based Hidden States as Author Representations for Psychological Human-Centered NLP Tasks

Nikita Soni, Pranav Chitale, Khushboo Singh,
Niranjan Balasubramanian, H. Andrew Schwartz
Stony Brook University
{nisoni, pchitale, khusingh, niranjan, has}@cs.stonybrook.edu

Abstract

Like most of NLP, models for human-centered NLP tasks—tasks attempting to assess author-level information—predominantly use representations derived from hidden states of Transformer-based LLMs. However, what component of the LM is used for the representation varies widely. Moreover, there is a need for Human Language Models (HuLMs) that implicitly model the author and provide a user[1]-level hidden state. Here, we systematically evaluate different ways of representing documents and users using different LM and HuLM architectures to predict task outcomes as both dynamically changing states and averaged trait-like user-level attributes of *valence*, *arousal*, *empathy*, and *distress*. We find that representing documents as an average of the token hidden states performs the best generally. Further, while a user-level hidden state itself is rarely the best representation, we find its inclusion in the model strengthens token or document embeddings used to derive document- and user-level representations resulting in best performances.

1 Introduction

Human-centered NLP focuses on the humans generating language (Flek, 2020; Hovy and Yang, 2021; Soni et al., 2024b) Many human-centered NLP tasks focus on assessing human-attributes of a user based on their language (Lynn et al., 2017), utilizing a representation of the author, often within a specific time period or set of documents (Matero et al., 2022; Giorgi et al., 2024).

Like most sub-fields in NLP, human-centered NLP heavily relies on traditional LLMs (pre-trained on the language modeling task of either next word prediction or the missing word prediction) that do not directly contain a representation for the person. For example, no particular layer, token, or output is accepted as the best way to represent a person. More recently, *human language*

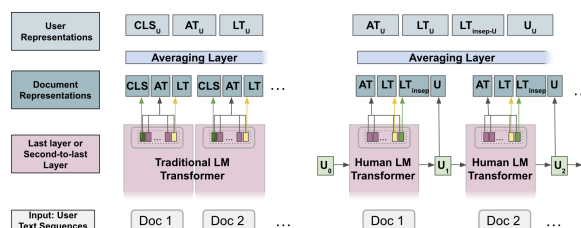[1]'Author', 'user' and 'person' used interchangeably.

Figure 1: Evaluating representations derived from different hidden states in both traditional Transformer-based LMs and human language models (HuLMs). Document representations as hidden states from: CLS, special token in auto-encoders; AT, averaged tokens in a document; LT, last token in a document for autoregressive models; LT_{insep}, special last token for HuLMs; and U, user state for HuLMs. User representation as averaged corresponding document representation types for all documents written by a user.

models (HuLMs)—those that directly model the language in the context of its "generator", i.e., the author (Soni et al., 2024c)—have been proposed but still lack evaluations for capturing human-level factors.

Here, we systematically compare and evaluate the commonly used ways to represent **words, documents, and users** in terms of hidden states (i.e., vector embeddings) from traditional LMs and HuLMs (Figure 1) for their ability to capture user-level information. We use the three levels of representations (word, document, user) to predict three levels of task outcomes: **document-level, wave-level, and user-level**. These outcomes are selected based on the prediction properties of: a) changing user states – as an effect of different situations/times and particular states when an author writes the document, b) changing user states over a range of (closely spaced) periods of time — as an effect of differences in user's states grouped together in one period of time (referring to it as a wave) and specifically separated from the next period of time (i.e., next wave), and c) sta-

ble trait-like user-level attributes. We select four well-established outcomes representative of psychological human-centered NLP tasks spanning fundamental human attributes: the affective circumplex (Russell, 1980; Mohammad, 2016) (valence and arousal), empathy, and distress (Batson et al., 1987; Litvak et al., 2016))

Transformer-based LM (Vaswani et al., 2017) architectures provide us with many choices for document representation: first token (i.e., CLS) (Devlin et al., 2019; Liu et al., 2019) in case of autoencoders, the last token in case of autoregressive models (Radford et al., 2019), or the average of all the tokens in a document. Since traditional LMs do not have a notion of a user, prior works have used a hierarchical user representation by averaging all document embeddings written by a particular user (Nagao and Katsurai, 2024) (Figure 1). While that is a posthoc way to represent a user, HuLMs provide us with user representations – a relatively more natural way for the same, by processing language in its multi-level structure i.e., words are a part of a document, and documents are generated by people (Soni et al., 2022). We select representative models for these different types of LMs currently available: autoencoders (BERT-like), autoregressive (GPT2-like), and HuLMs (HaRT-like (Soni et al., 2022)). HuLMs are trained on user-aware LM pre-training tasks that predict the next word conditioned on the contextual words *as well as a user state*, thereby processing text in the context of their historical language by tracking *user states* through recurrently updated vectors. HuLMs output word embeddings like traditional LMs (however, these are user-state-informed), and also output user states presenting additional opportunities for producing rich user representations from such user states. We hypothesized the user representations derived from user states—as the natural output of HuLMs—will result in better assessing the human-attributes considered at the three different levels of predictions.

Contributions. (1) We depict some of the different ways to represent words, documents, and users using traditional LMs and HuLMs hidden states and user states to predict a user's frequently changing states (i.e., document-level), their changing states over a specific period of time (i.e., wave), and their stable trait-like attributes (i.e., user-level). (2) We systematically evaluate the multiple representation types over broadly 2 levels of analysis of

user attributes: document- and wave-level dynamically changing over time (states) and somewhat stable user-level (trait-like) on an average. Traits are defined as averages of multiple state measurements (i.e. a trait defines the mean and sometimes the variance of one's distribution of states (Fleeson, 2001)). (3) We draw 2 major findings: (a) averaging the token embeddings as the basis for document representations (states) and hierarchical user representations (traits) perform the best on an average, and (b) contrary to our hypothesis, user states themselves are not the best representations, but consistently strengthen user representations derived from token and document embeddings where a user-state was included.

2   Related Work

Human-centered NLP tasks are often grounded in the person (Ganesan et al., 2024) and require user representations for downstream applications, such as assessing mental health (Ganesan et al., 2022; Varadarajan et al., 2024), estimating demographic attributes (Benton et al., 2016), assessing personality (Schwartz et al., 2013; Soni et al., 2024a), stance detection (Matero et al., 2021), and personalized recommendations (Qi et al., 2021; Wu et al., 2023). Commonly, a hierarchical approach has been used to average document representations across all text sequences from a particular user since the pre-LLMs era (He et al., 2019; Lynn et al., 2020). As Transformer-based LLMs become ubiquitous in human-centered NLP tasks, the hierarchical user representation approach continues to prevail (Lynn et al., 2020; Nagao and Katsurai, 2024). While this allows us to derive a user representation, it does not necessarily capture the author's context, as traditional LLMs process documents written by the same individual independently or, conversely, treating all language as if written by an average, universal person (Soni et al., 2024c).

A large body of past work explored generating user embeddings (Ning et al., 2024) and making LLMs personalizable (Zhang et al., 2024). Recently introduced Transformer-based Human Language Models (HuLMs) (Soni et al., 2022) offer a notion of the user by processing language in its multi-level structure, where words are part of a document, and documents are generated by people. This provides a relatively natural way of representing a user. Traditional LMs and HuLMs present a multitude of choices regarding which layer, token, or output to select for the best representation of a

person and their document- and token-level representations. To this end, we perform a systematic study to identify the best representations.

3 Datasets and Tasks

We select 2 datasets that allow for the multiple levels of analysis broadly categorized into (a) dynamic states (attributes that change with time and utterances), and (b) user-level (trait-like attributes on an average for a person). These datasets provide tasks of predicting fundamental human-attributes like affective circumplex grid (Russell, 1980) (valence and arousal), and empathy and distress (Batson et al., 1987).

Valence and Arousal. We use a subset of the DS4UD (Data Science for Unhealthy Drinking) dataset (Nilsson et al., 2024), consisting of affective circumplex grid self-reported questionnaires and essays written by US service industry workers on how they were feeling 3 times daily (minimum once) for 14-day periods (we call it a "wave"), across 6 data collection phases from 2021 to early 2024. We use self-reported outcomes derived from the affective circumplex grid: *Valence* (0-4, highly negative to highly positive affect), and *Arousal* (0-2, low to high energy). Our selection criteria include users who participated in at least 2 waves and wrote at least 2 essays ($\geq$ 10 words each), resulting in 120 users overall, 406 user instances across 6 waves (a user can have 2 to 6 wave instances, depending on the number of waves they participated in), and 10090 essays across all users. We compute outcomes labels across waves by averaging the labels corresponding to each essay for a user in a particular wave. Further, we average the labels corresponding to each wave for a user to get the user-level outcomes labels, following classical test theory (Rust and Golombok, 2014).

Empathy and Distress. We use a subset of the WASSA 2024 Shared Task on Empathy and Personality Detection in Interactions, specifically a part of Track 3 (Giorgi et al., 2024). The task is to predict the empathic concern and personal distress scores (Batson et al., 1987) for each essay written by crowd workers in response to news articles, with real-valued labels ranging from 1 to 7. We include train and dev data from WASSA 2024 and WASSA 2023 (Barriere et al., 2023), the years with user-level data, resulting in 180 users and 1837 essays. User-level trait-like outcome labels are obtained by averaging empathy and distress scores across each user's essays.

4 Methods and Experiments

Models and Types of Representations. In this study, we will evaluate different types of representations from 4 models, of which 2 are autoregressive: 1) HaRT (Soni et al., 2022), 2) GPT-2 small, and 2 are auto-encoders: 3) BERT, 4) RoBERTa.

HaRT enables us to use 4 types of **document representations**: a) Average of token embeddings (AT), b) Last token (LT), c) Last token as *insep* (LT_{insep}), d) User states (U) (Figure 1). Autoregressive GPT-2 can provide the first 2 of these 4 types, while auto-encoders (BERT, RoBERTa) normally use the CLS token embeddings (CLS) or the AT to represent a document. We employ the hierarchical averaging structure over the different types of document embeddings across all documents from a user, or in a specific time-period (i.e., wave) to retrieve different types of **user and wave representations** (refer Section 3), respectively.

HaRT Architecture and Variants. We sought to use an architecture that can handle multiple representations criteria: (1) a multi-level structure spanning words, documents, and user-level data, (2) last-token and special-token representations, and (3) applicability to traditional LM tasks. For our primary results we build on Human-aware Recurrent Transformer (HaRT) model that was pre-trained on the human language modeling task (Soni et al., 2022) (see Section 1). Originally, HaRT processes temporally ordered text sequences from a user, separated by a special *insep* token, chunked into blocks (up to 8 for training), each capped at 1024 tokens (referred to **HaRT_frozen**). HaRT maintains user state vectors, updated recurrently after processing each block. However, concatenated blocks of 1024 tokens didn't provide an easy way to extract user states at the end of each document for document- and wave-level prediction. To address this, we further pre-train HaRT on DS4UD essays data (refer Section 3) with a limit of one document per block (referred to **HaRT_ODPB**) and adjusting the maximum number of training blocks (to 55) to retain comparable user-level data, enabling user state extraction after each document. Additionally, in another variant (**HaRT_concat**), we continue to pre-train HaRT_frozen using the DS4UD essays data with the original block-settings to help tease out any spurious effects introduced bt altered block-settings in HaRT_ODPB. Following Soni et al., we derive user and wave rep-

Method	Dynamic States								Trait-like				
	Document-Level					Wave-Level			User-Level				
	Val	Aro	Emp	Dis	Avg	Val	Aro	Avg	Val	Aro	Emp	Dis	Avg
GPT-2													
LT	0.53	0.25	0.71	0.58	0.52	0.64	0.28	0.46	0.63	-0.14	0.37	0.42	0.32
AT	0.60	0.34	0.69	0.54	0.54*	0.77	**0.33**	**0.55***	0.75	0.18	0.52	**0.68**	0.53*
BERT													
CLS	0.58	0.30	0.70	0.66	0.56	0.74	0.28	0.51	0.71	0.09	0.45	0.42	0.42
AT	0.61	0.35	0.72	0.59	0.57*	0.75	0.26	0.50	0.73	0.14	0.45	0.39	0.43
RoBERTa													
CLS	0.62	0.35	**0.77**	0.63	0.59*	0.75	0.29	0.52	0.72	0.10	0.51	0.59	0.48
AT	0.63	**0.36**	0.72	0.63	0.58	0.75	**0.33**	0.54**	0.75	0.16	0.59	0.53	0.51
HaRT$_{\text{frozen}}$													
LT	0.52	0.22	0.70	0.57	0.50	0.69	0.21	0.45	0.73	0.18	0.42	0.60	0.48
LT$_{\text{insep}}$	0.55	0.23	0.76	0.67	0.55	0.69	0.22	0.46	0.74	0.23	0.45	0.62	0.51
AT	0.64	0.33	0.75	0.68	0.60*	0.79	0.27	0.53*	0.73	**0.35**	**0.67**	0.66	**0.60*†**
U	-	-	-	-	-	-	-	-	0.59	-0.01	0.46	0.49	0.38
HaRT$_{\text{concat}}$													
LT	0.53	0.22	0.72	0.62	0.52	0.70	0.16	0.43	0.74	0.18	0.42	**0.68**	0.50
LT$_{\text{insep}}$	0.57	0.26	0.76	0.56	0.54	0.75	0.30	0.53	**0.77**	0.15	0.61	0.55	0.52
AT	**0.65**	0.34	0.75	**0.70**	**0.61*†**	**0.80**	0.27	0.54	**0.77**	0.34	0.62	0.65	0.59*
U	-	-	-	-	-	-	-	-	0.65	0.06	0.59	0.49	0.45
HaRT$_{\text{ODPB}}$													
LT	0.57	0.27	0.68	0.56	0.52	0.72	0.17	0.44	0.72	0.24	0.41	0.55	0.48
LT$_{\text{insep}}$	0.64	0.34	0.74	0.66	0.59	0.79	0.27	0.53	0.72	0.09	0.40	0.50	0.43
AT	0.63	**0.36**	0.73	0.63	0.59*	**0.80**	0.28	0.54	0.75	0.12	0.58	0.51	0.49**
U	0.52	0.15	0.46	0.40	0.38	0.65	0.01	0.33	0.66	0.04	0.50	0.47	0.42

Table 1: Linear model predictions for 4 well-established psychological outcomes: valence, arousal, empathy, and distress, correlated (Pearson r) with the labels associated with the datasets in Section 3. Averaged token embeddings (AT) generally outperform LT, CLS, LT$_{\text{insep}}$, and U across respective models. Token and document embeddings informed by the user state perform best for averaged document- and user-level outcomes. No statistically significant difference found between HaRT$_{\text{concat}}$ and GPT-2 for averaged wave-level outcomes. Bold indicates best in column and * indicates paired t-test statistical significance with $p < 0.001$ and ** with $p < 0.05$ for each model's AT results against its respective second best result. † represents statistical significance ($p < 0.05$) across models for each of document-, wave-, and user-level analysis between best overall best (HaRT variants) and best traditional LM.

resentations by averaging user states across blocks in each HaRT variant (see Section 3).

Experiments. We use averaged token embeddings, special token embeddings, and user states embeddings (where available) from HaRT variants, GPT-2, BERT, and RoBERTa[2] to predict valence, arousal, empathy, and distress as both dynamic states and trait-like stable outcomes based on a person's text sequences (Table 1). We use DLATK (Schwartz et al., 2017) to apply ridge regression (aplha values: 1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04, 1.e+05, 1.e+06) over the different embeddings representations, learning only the ridge regression weights without any task-specific fine-tuning model parameters.

To evaluate vector representations, we measure Pearson r correlations between the linear model predictions and dataset labels (Section 3). We run 10- straitfied group-based cross-validation folds, ensuring no user overlap between the train and test splits in any fold, and report the averaged Pearson r. The experiments took ~5 hours on A6000 GPU

using approx ~24GB memory.

5 Results and Discussion

Overall, averaged across outcomes, we find HuLMs (HaRT variants) to perform the best for predicting both the changing states and trait-like outcomes at the document- and user-level (refer Table 1). We do not find any statistically significant difference in the performance of GPT-2 and HaRT$_{\text{concat}}$ for wave-level predictions.

Averaged Token Embeddings (AT). Using the averaged token embeddings helps more often than the last token or CLS embeddings for all types of models considered: autoregresive, auto-encoders, and HuLMs. Intuitively, AT embeddings perform better because they may be better able to capture the nuanced, context-dependent details needed to represent both stable traits and changing states— key aspects of human-level attributes—from each token of a document. Whereas, when we use a single special token's embeddings as a document representation, we may lose important nuanced information distributed across the tokens in the document in the process of compressing information into one token's embeddings. However, this is only

theoretical and we do not have evidence to prove the reasoning in the current study and encourage future studies to explore further.

Inclusion of User States. The user states by themselves are rarely the best embeddings but consistently the LT and AT-based embeddings were strong when a user-state was included. We see under the HuLMs (HaRT variants) in Table 1, representations from U to perform below par, however user-level representations derived from AT (and in some cases last tokens LT and LT_{insep}), that are processed in the context of the user states, boost the performance. While using AT is the current best choice, this also opens up avenues for future works to improve the natural representations of user states capturing the author's context and thus further improve the word's meaning embedded in tokens.

Last Token as a Special Token (LT_{insep}). Interestingly, we find representations derived from LT_{insep} embeddings from HuLMs (HaRT variants) to perform better than those from LT embeddings (Table 1) that suggests *insep* is able to capture the average of the tokens in a documents from a user. Theoretically, LT_{insep}, the special token for HuLMs, can be seen as analogous to the CLS token for auto-encoders.

Outcomes Predictions Analysis. Arousal prediction remains a challenging task across all 3 different levels of predictions (document, wave, and user). However, RoBERTa predicts it best at the document- and wave-level, while $HaRT_{frozen}$ is the best in predicting it as an average trait for the author (i.e., user-level).

Valence is better predicted at the user- and wave-level across models, suggesting it is harder to capture the frequently changing affect over quick succession of time (i.e., within the same day or consecutive days, as associated with each document written by an author); as opposed to capturing over a period of time. $HaRT_{concat}$ performs the best across document-, wave-, and user- levels of predictions.

Empathy is much better predicted as a frequently changing human state (i.e, at the document-level) versus as an average trait for the author. However, $HaRT_{frozen}$ is the best in predicting it as an average trait for the author (i.e., user-level).

Distress prediction showed mixed results with $HaRT_{concat}$ doing better for both document- and user-level predictions, while GPT-2 show same performance as $HaRT_{concat}$ for user-level but lower at the document-level. BERT-like models were sub-par at user-level.

Later Layers of Transformer-based Architectures. We performed initial evaluation by fetching hidden states from last layer and second-to-last layer, and comparing the different types of embeddings over the document-, wave-, and user-level tasks of predicting the human attributes (valence, arousal, empathy, and distress) (Appendix Table 2). We find a dominating trend with last layer hidden states for special tokens—LT and CLS tokens, whereas averaged token embeddings show similar results using both last and second-to-last layers with the latter having better results for auto-encoders at the document-level. We report the results from the last layer for representations using special tokens and second-to-last layer for AT representations to use a stronger baseline auto-encoder based results in Table 1.

6 Conclusion

We provided the first systematic evaluation of different types of word-, document-, and user-representations across three levels of analysis for tasks predicting both dynamically changing states and averaged trait-like user-level attributes (valence, arousal, empathy, and distress). We used traditional LMs (BERT, RoBERTA, and GPT-2) as well as human language models (versions of HaRT) and represented documents using the CLS token, the last token, the last token as *insep*, U (user state from a *HuLM*), and the averaged tokens hidden states for respective models. We hierarchically average the corresponding document representations written by a particular user to derive their user representation. Generally, document-, wave- (user over a specific time period), and user- representations based on averaged token hidden states outperform CLS, LT, LT_{insep}, and U for respective models. Although the user states themselves do not prove to be a powerful representation for predicting user-level outcomes, they improve the averaged user-level performance when using user representations derived from user-state-informed tokens hidden states making HuLMs (HaRT variants) as the best performing model. Overall, our study and findings provide a standardized comparison and identify the best representations to select when applying LMs for psychological human-centered NLP tasks.

Ethical Considerations

Models that incorporate the author's context and/or use data involving private and sensitive information must adopt a responsible use and release strategy. All models carry some human-level attribute information, and it is important for ethical applications (both for prosocial mental health use cases and for privacy preserving use cases) that we understand where these attributes are best represented. However, user states, user representations, or document/word representations are not mapped to any "one" identifiable user. These models are not trained to identify "individual users" or are not fed particular user-attributes, but instead use anonymized user identifiers while our evaluations compare how well they pick up on user attributes. While such models are essential for advancing research in human-centered NLP and understanding language in the context of the author, they may also pose unintended nefarious risks. For these reasons, the DS4UD data used in this study, which involves private user data, cannot be made publicly available due to ethical and privacy concerns. The affective grid circumplex tasks presented here were reviewed and approved or exempted by an academic institutional review board (IRB) as per the original dataset owners.

Limitations

The purpose of our study is to evaluate and standardize some of the commonly used Transformer-based approaches for representing users, their documents, and tokens in human-centered NLP tasks. We use some of the relevant and widely used models, exploring a few different types of representations. While our study focuses on specific models and methods, we encourage the NLP community to investigate lesser-explored approaches for modeling human context and deriving representations. We also acknowledge limitations inherent to these models and representations, such as context size and the extent of a user's historical language used as human context, which are beyond the scope for this study.

Traditional LMs and HuLMs have many different ways of deriving these representations but no study suggests the best route. And, we select representative models for these different types of LMs currently available: autoencoders (BERT-like), autoregressive (GPT2-like), and HuLMs (HaRT-like). At the same time, larger LLMs are mainly used for

prompting, and not by deriving embedding-based representations to use for human-centered tasks. We use GPT-2 and BERT-like models because these are comparable in the number of parameters with HaRT-like HuLMs. Currently, larger HuLMs are not available and we encourage future studies to explore this area of research and compare the performances.

Regarding datasets, some users and essays in the WASSA dataset were duplicates or skewed. We removed the duplicates and retained the remaining data for evaluation. Finally, models and datasets involving sensitive user information, such as the DS4UD dataset, require careful and responsible usage. Due to privacy concerns, the authors of these datasets cannot make them publicly available. The datasets for human-centered NLP tasks are often small-scale due to a variety of reasons including: (i) availability of user information, (ii) privacy issues, (iii) difficulty in collecting labeled datasets at the user level for a continued period of time. DS4UD dataset provides a rich historical language of authors (essays) collected for a continued period of time (e.g., DS4UD was collected over 3 years, from early 2021 until 2024), with associated multiple human-level attributes. Thus, this provides us with an opportunity to analyze the stable traits and changing human states of people. To the best of our knowledge, no studies provide rich longitudinal data at such scale. WASSA datasets are publicly available and provide rich textual data with associated labels for multiple documents written by a particular author, thus enabling the prediction of different empathetic (or distress) states of a person. To sum up, these datasets provide complex scenarios encapsulating changing human states and averaged traits of a person. More importantly, DS4UD also provides the temporally changing human-level attributes with self-reported labels instead of inferred attributes or annotated labels. Furthermore, these datasets are essay-like language which reduces the limitations of social-media datasets that are short and not necessarily continuous over time.

Acknowledgments

not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, any other government organization, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

C. Daniel Batson, Jim Fultz, and Patricia A. Schoenrade. 1987. Distress and empathy: two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55 1:19–39.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of Twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Berlin, Germany. Association for Computational Linguistics.

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

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

Lucie Flek. 2020. Returning the N to NLP: Towards Contextually Personalized Classification Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Adithya V Ganesan, Siddharth Mangalik, Vasudha Varadarajan, Nikita Soni, Swanie Juhng, João Sedoc, H Andrew Schwartz, Salvatore Giorgi, and Ryan L Boyd. 2024. From text to context: Contextualizing language with humans, groups, and communities for socially aware nlp. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 26–33.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H Andrew Schwartz. 2022. Wwbp-sqt-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 369–379.

Yun He, Jianling Wang, Wei Niu, and James Caverlee. 2019. A hierarchical self-attentive model for recommending user-generated item lists. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1481–1490, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. 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. *arXiv preprint arXiv:1907.11692*.

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, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Matero, Albert Hung, and H. Andrew Schwartz. 2022. Evaluating contextual embeddings and their extraction layers for depression assessment. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social*

Media Analysis, pages 89–94, Dublin, Ireland. Association for Computational Linguistics.

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

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier.

Hiroyoshi Nagao and Marie Katsurai. 2024. Researcher representations based on aggregating embeddings of publication titles: A case study in a Japanese academic database. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 277–282, Bangkok, Thailand. Association for Computational Linguistics.

August Håkan Nilsson, Hansen Andrew Schwartz, Richard N Rosenthal, James R McKay, Huy Vu, Young-Min Cho, Syeda Mahwish, Adithya V Ganesan, and Lyle Ungar. 2024. Language-based ema assessments help understand problematic alcohol consumption. *PLoS ONE*.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-llm: Efficient llm contextualization with user embeddings.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical user interest modeling for personalized news recommendation. 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)*, pages 5446–5456, Online. Association for Computational Linguistics.

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

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

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):e73791. Publisher: Public Library of Science.

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.

Nikita Soni, Niranjan Balasubramanian, H Schwartz, and Dirk Hovy. 2024a. Comparing pre-trained human language models: Is it better with human context as groups, individual traits, or both? In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328.

Nikita Soni, Lucie Flek, Ashish Sharma, Diyi Yang, Sara Hooker, and H Andrew Schwartz. 2024b. Proceedings of the 1st human-centered large language modeling workshop. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*.

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

Nikita Soni, H Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024c. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8623–8638.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, Lucie Flek, H. Andrew Schwartz, Charles Welch, and Ryan Boyd. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291, St. Julians, Malta. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra,

Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024. Personalization of large language models: A survey.

Method	Dynamic States											
	Document-Level								**Wave-Level**			
	Val		**Aro**		**Emp**		**Dis**		**Val**		**Aro**	
	L	**SL**	**L**	**SL**	**L**	**SL**	**L**	**SL**	**L**	**SL**	**L**	**SL**
GPT-2												
LT	**0.53**	0.48	**0.25**	0.18	0.71	0.71	0.58	0.58	**0.64**	0.61	**0.28**	0.15
AT	0.60	0.60	0.34	0.34	**0.70**	0.69	**0.58**	0.54	0.76	**0.77**	0.29	**0.33**
BERT												
CLS	0.58	**0.59**	**0.30**	0.29	0.70	**0.71**	**0.66**	0.60	**0.74**	0.70	0.28	**0.29**
AT	0.61	0.61	0.35	0.35	0.71	**0.72**	0.59	0.59	**0.77**	0.75	**0.29**	0.26
RoBERTa												
CLS	**0.62**	0.59	**0.35**	0.30	**0.77**	0.69	**0.63**	0.42	**0.75**	0.71	0.29	**0.30**
AT	0.63	0.63	0.36	0.36	0.71	**0.72**	0.60	**0.63**	**0.79**	0.75	0.33	0.33
HaRT$_{frozen}$												
LT	**0.52**	0.51	**0.22**	0.20	0.70	0.70	**0.57**	0.56	**0.69**	0.68	**0.21**	0.16
LT$_{insep}$	0.55	**0.56**	**0.23**	0.21	**0.76**	0.74	**0.67**	0.52	0.69	0.69	**0.22**	0.20
AT	0.63	**0.64**	0.33	0.33	**0.77**	0.75	**0.70**	0.68	0.79	0.79	0.27	0.27
U	-		-		-		-		-		-	
HaRT$_{concat}$												
LT	**0.53**	0.52	**0.22**	0.18	**0.72**	0.70	**0.62**	0.60	0.70	0.70	**0.16**	0.13
LT$_{insep}$	0.57	**0.59**	0.26	0.26	**0.76**	0.46	**0.56**	0.48	**0.75**	0.74	**0.30**	0.28
AT	0.64	**0.65**	0.34	0.34	**0.77**	0.75	**0.72**	0.70	0.80	0.80	**0.30**	0.27
U	-		-		-		-		-		-	
HaRT$_{ODPB}$												
LT	**0.57**	0.53	**0.27**	0.21	**0.68**	0.63	**0.56**	0.52	**0.72**	0.69	**0.17**	0.13
LT$_{insep}$	0.64	**0.65**	**0.34**	0.33	**0.74**	0.73	**0.66**	0.65	0.79	**0.80**	**0.27**	0.25
AT	0.63	0.63	0.36	0.36	**0.74**	0.73	**0.64**	0.63	0.80	0.80	**0.29**	0.28
U	0.52		0.15		0.46		0.40		0.65		0.01	

Table 2: We evaluate the performance differences in hidden states from the last layer ("L") and second-to-last layer ("SL") for all models. We find second to last layer to perform better for AT hidden states in case of auto-encoders for document-level tasks, and last layer to work generally well for the other types of representations across models. AT embeddings from L and SL show similar results for most other outcomes. Bold indicates better performance between L and SL for each model/representation/outcome.

Method	**Trait-like**							
	User-Level							
	Val		**Aro**		**Emp**		**Dis**	
	L	**SL**	**L**	**SL**	**L**	**SL**	**L**	**SL**
GPT-2								
LT	**0.63**	0.52	-0.14	**-0.01**	**0.37**	0.36	0.42	**0.47**
AT	0.75	0.75	**0.19**	0.18	0.52	0.52	0.64	**0.68**
BERT								
CLS	**0.71**	0.68	**0.09**	0.05	0.45	**0.48**	0.42	0.42
AT	**0.75**	0.73	**0.17**	0.14	0.44	**0.45**	0.37	**0.39**
RoBERTa								
CLS	**0.72**	0.61	0.10	**0.13**	**0.51**	0.48	0.59	**0.66**
AT	0.75	0.75	**0.20**	0.16	**0.62**	0.59	**0.58**	0.53
HaRT$_{frozen}$								
LT	**0.73**	0.72	**0.18**	0.13	0.42	**0.43**	**0.60**	0.59
LT$_{insep}$	**0.74**	0.73	**0.23**	0.20	**0.45**	0.44	**0.62**	0.47
AT	0.72	**0.73**	0.35	0.35	**0.68**	0.67	0.64	**0.66**
U	0.59		-0.01		0.46		0.49	
HaRT$_{concat}$								
LT	**0.74**	0.71	**0.18**	0.12	0.42	**0.43**	**0.68**	0.63
LT$_{insep}$	**0.77**	0.75	**0.15**	0.13	0.61	**0.65**	0.55	**0.59**
AT	0.76	**0.77**	**0.35**	0.34	**0.64**	0.62	0.63	**0.65**
U	0.65		0.06		0.59		0.49	
HaRT$_{ODPB}$								
LT	**0.72**	0.68	**0.24**	0.23	0.41	**0.42**	0.55	**0.56**
LT$_{insep}$	**0.72**	0.71	0.09	0.09	**0.40**	0.39	**0.50**	0.47
AT	0.75	0.75	**0.18**	0.12	0.57	**0.58**	**0.52**	0.51
U	0.66		0.04		0.50		0.47	

Table 3: We evaluate the performance differences in hidden states from the last layer ("L") and second-to-last layer ("SL") for all models. AT embeddings from L and SL show mixed and similar results for most outcomes at user-level. Last layer works generally well for the other types of representations across models. Bold indicates better performance between L and SL for each model/representation/outcome.