Using Daily Language to Understand Drinking: Multi-Level Longitudinal Differential Language Analysis

Matthew Matero*[1], Huy Vu*[1], August Håkan Nilsson[2]
Syeda Mahwish[1], Young-Min Cho[3], James R. McKay[3]
Johannes Eichstaedt[4], Richard N. Rosenthal[1], Lyle Ungar[3], H. Andrew Schwartz[1]
[1] Stony Brook University [2] Oslo Metropolitan University
[3] University of Pennsylvania [4] Stanford University
{mmatero,has}@cs.stonybrook.edu

Abstract

Analyses for linking language with psychological factors or behaviors predominately treat linguistic features as a static set, working with a single document per person or aggregating across multiple documents into a single set of features. This limits language to mainly shed light on between-person differences rather than changes in behavior within-person. Here, we collected a novel dataset of daily surveys where participants were asked to describe their experienced well-being and report the number of alcoholic beverages they had within the past 24 hours. Through this data, we first build a multi-level forecasting model that can capture within-person change and leverage both the psychological features of the person and daily well-being responses. Then, we propose a longitudinal version of differential language analysis that finds patterns associated with drinking more (e.g. social events) and less (e.g. task-oriented), as well as distinguishing patterns of heavy drinks versus light drinkers.

1 Introduction

Language generated by people occurs at multiple levels of analysis, from tokens to documents to sequences of documents (Almodaresi et al., 2017). While past works have suggested modeling language hierarchically given the available history of a person's language (Acheampong et al., 2021; son; Lynn et al., 2017; Matero et al., 2021b; Soni et al., 2022), few techniques exist for language analyses geared toward eliciting language associated with psychological or behavioral changes (Tsakalidis et al., 2022). Where traditional techniques like differential language analysis (Schwartz et al., 2013) only reveal differences *between* people rather than changes *within* people around particular behaviors.

Typically, NLP-based approaches represent language from people as aggregations, such as of message or token embeddings over all time (Ganesan et al., 2021; Almodaresi et al., 2017; Matero et al., 2021a). While there have been some predictive-focused works that have experimented with forecasting based on language, they are either focused on psychological (latent) attributes (Halder et al., 2017; Matero and Schwartz, 2020) or focused on groups/communities of people rather than individuals (Matero et al., 2023), less has been done toward bringing out linguistic insights (e.g. differential language analysis (Schwartz et al., 2013)) leveraging the inherent multi-level longitudinal structure of human language. In this work, we present and evaluate (1) a longitudinal, multi-level approach to forecasting an individual's behavior rather than latent human attributes (e.g. emotions), namely daily consumption of alcoholic beverages, and (2) a longitudinal, multi-level differential language analysis to illuminate daily language patterns most commonly associated with heavier drinking both across different individuals and within one individual.

With roughly 10% of U.S. adults having an alcohol use disorder (NIH, 2023), research to understand an individual's alcohol consumption pattern and motivation is a pressing health concern. By modeling one's behavior over time we can more accurately predict future consumption or interpret their motivations for drinking alcohol through the use of longitudinal multi-level models. Such a model could be used to detect the risk of unhealthy drinking. These personalized models are naturally geared towards time-series forecasting, where the goal is to understand coming trends (Eichstaedt et al., 2018; Halder et al., 2017).

Our contributions include: (1) introduction of a sequential forecasting model that leverages language to accurately predict the number of alcoholic drinks a person will consume within a 24-hour window, (2) integration of user-level features (static across time) to build a multi-level sequential model for additional context in prediction, (3) empirical evaluation on dimensionality reduction of language

*Equal Contribution

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 133–144
March 21, 2024 ©2024 Association for Computational Linguistics

features cross-time concerning predictive power, and (4) insights into linguistic patterns that are longitudinally predictable of high or low daily drinking rates.

2 Related Work

Alcohol Consumption Psychological research has long demonstrated the complexities of alcohol consumption. On one hand, the general person drinks more on days when they feel more positive affect and not when they feel more negative effect (Dora et al., 2022), and general drinking level has a positive correlation to life satisfaction (Geiger and MacKerron, 2016; Massin and Kopp, 2014). On the other hand, this relationship is hump-shaped such that the happiest people are low to moderate drinkers and heavy drinkers are worse off with decreases in well-being (Geiger and MacKerron, 2016; Massin and Kopp, 2011).

Heavy alcohol consumption can lead to an Alcohol Use Disorder, a disorder that can cause morbidity (Carvalho et al., 2019) and decreased psychosocial functioning (Kendler et al., 2016). Predicting within-person alcohol consumption from scales that measure emotion such as positive affect have shown correlations between participant-aggregated affect and participant-aggregated number of drinks consumed of r = .10 and a non-significant relationship to negative affect (Dora et al., 2022). A likely reason for the positive relationship between drinking and positive affect is that most drinking occurs socially (Creswell et al., 2022) and spending time with others is strongly associated with reporting high levels of positive affect (Grimm et al., 2015; Killingsworth and Gilbert, 2010; Diener and Seligman, 2002).

Language and Drinking While there exists a few studies focused on predicting who is at risk for alcohol abuse from language, they use historical data to make a single prediction in time rather than predicting how behaviors may change. Both works of Jose et al. (2022) and Curtis et al. (2018) investigate the connection of historical social media language and their association with at-risk drinking. However, they both focus on different levels of analysis and outcomes with Jose et al. (2022) focusing on individual-level and the ability to predict one's risk-level for alcohol consumption (e.g. AUDIT-C) (Bush et al., 1998) and Curtis et al. (2018) leveraging county data with responses to the Behavioral Risk Factor Surveillance System (BRFSS); a U.S. health survey where someone may self-report their level of heavy drinking.

Longitudinal & Multi-level NLP is very familiar with sequence processing leveraging various techniques such as attention networks (Vaswani et al., 2017) and seq2seq modeling (Luong et al., 2015; Bahdanau et al., 2014). Even still, explicitly modeling the temporal dimension is largely under-utilized by most NLP models as words and sentences are often uttered at what can be assumed as the same point in time except for the case where language is considered to reflect a person (Soni et al., 2022; Matero and Schwartz, 2020). Sequential models designed explicitly for temporal modeling have been proposed but not widely adopted by the NLP community (Zhu et al., 2017; Che et al., 2018).

One could go one step further and adapt these sequential time-series models to account for the inherent hierarchical nature of language over time from a person through multi-level modeling. Multilevel modeling allows the model to operate on different levels of granularity and offers a natural way of framing the problem (Hox, 1998). Due to this natural hierarchy, in this case, defined by dynamic states and static traits cross-time (Su et al., 2019; Gana et al., 2019; Van der Werff et al., 2019), we can develop a model to account for this. Multilevel modeling is a common approach in psychology research, for example understanding substance cravings and personality (Parent-Lamarche et al., 2021; Alayan et al., 2019).

Lastly, we extend past works that explored the associations between social media language and drinking behavior by examining the association between topics of daily language, through self-reported experienced well-being responses, and alcohol consumption or risk. Differential language analysis (DLA) is commonly used to study topics of conversation and their ability to reliably predict certain outcomes (Schwartz et al., 2013; Eichstaedt et al., 2018; Schwartz et al., 2014; Kern et al., 2016). While the work of Jose et al. (2022) also investigated the relationship between specific social media topics and drinking risk, they focused on the between-person signals instead of within-person signals as in our approach. These within-person language signals are important for understanding what drives an individual to drink and are extracted via a fixed effects model that accounts for between-person heterogeneity (Hedges, 1994).

3 Data

We collected a novel dataset with the consent of study participants for a longitudinal investigation of drinking behavior. Upon enrollment, each participant also gave consent to access their Facebook posts and answer a “baseline” survey that asks various questions regarding mental health and well-being. Responses from the baseline survey include: measures of depression and anxiety (Johnson, 2014), AUDIT-C (Bush et al., 1998), and demographics.

Further, participants are asked to complete 14 days of ecological momentary assessments (EMA), short surveys expected to take a few minutes to complete on their phones. Each EMA contains a free response field called *affective essay*, where the participant describes their experienced well-being, emotions and daily experiences, as well as a question asking them how many alcoholic beverages they consumed in the past 24 hours[1]. Participants were selected to respond once or thrice daily (morning, afternoon, evening). The assignment was performed randomly (50/50) for which group a person was placed into.

The dataset samples from U.S. restaurant and hospitality workers (e.g., bartenders, servers, etc). Recruitment occurred between June 2020 and June 2021 from various sources such as organizations reaching out to their members via mailing lists or snowball sampling from social media. Sign-up and consent was handled via Qualtrix, where directions were given to download a companion app designed to be used for data collection.

Figure 1 illustrates the drinking behaviors from a random sample of 30 participants over the 14 days ordered by AUDIT-C score. The white empty cells indicate missing data points (no response) for that particular day. We observe that participants with higher AUDIT-C scores tend to drink more often and with a higher number of drinks.

Time-series Processing We split our time series into a train and test set based on out-of-sample time (e.g., forecasting) with a split such that each person’s last two days of responses are reserved for testing. When building our forecasting dataset, we filter participants for those that responded to at least three days of EMAs. This is done so that these users can still be used for testing, as they have at least one authentic response to use as input.

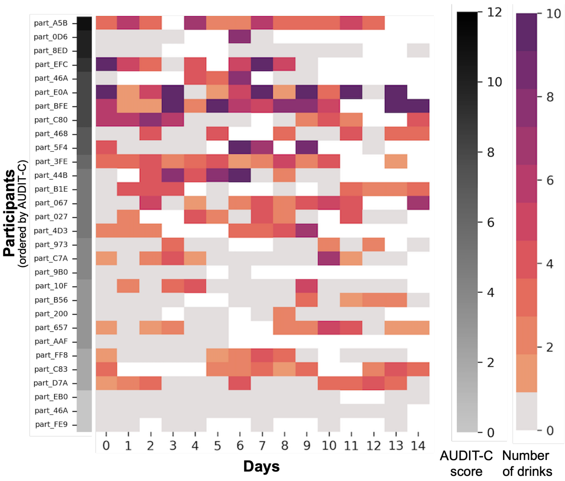

Figure 1: Overview of drinking behaviors data from a random sample of 30 participants ordered by AUDIT-C score. White cells indicate days to missing data where the participant did not respond to any EMA.

Additionally, we restrict to those users who were selected for three responses per day thus allowing our model to have more daily language to use as a signal for prediction. After applying this filter, we are left with 242 people, where 219 are kept for training and 23 are used as a held-out validation set for hyperparameter tuning.

For building our time-series features, we include an averaged RoBERTa embedding (Liu et al., 2019) of all *affective essays* of a given day, which is then dimensionality reduced to using pre-trained PCA models from Ganesan et al. (2021). At each time step, we concatenate these language features with the number of drinks and another small set of features representing a day-of-week marker defined as a 7-dimension one hot encoded feature space.

Lastly, to deal with participants who do not always remember to respond each day, we apply a simple imputation technique that fills missing gaps with the last available authentic response (Che et al., 2018).

4 Methods

Document Sequential Model We apply transformer networks (Vaswani et al., 2017) to our time-series as shown in Figure 2 describing our architecture. After the sequence is processed through the transformer network, the final representation is an average pooling over the output vectors for each time step. The average pooled representation is then run through a dense layer to predict the daily

[1] *affective essays* are 200 characters in length

number of drinks[2].

We also investigate multiple configurations of our models, namely multivariate and univariate forecasting. In the case of univariate, only past knowledge of drinking is used, such that a single variable represents each time-step. In multivariate, all available features per time-step are used as inputs.

Multi-level Sequential Model We incorporate both user-level variables and historic document-level social media language into our document-only sequential model. These features have been linked to both overall well-being and drinking behavior (Jose et al., 2022; De Choudhury et al., 2013). Thus, we include them as a separate module to perform a type of user-factor adaptation (Lynn et al., 2017).

The user-level features are as follows: degree of depression and anxiety, AUDIT-C, age, gender, and RoBERTa embeddings of the past two years of Facebook language that occurred before the start of the EMA period. The RoBERTa embeddings are reduced to 64 dimensions using the same pre-trained models from Ganesan et al. (2021). The models from Ganesan et al. (2021) are used as they have shown to be competitive on small data for human-level tasks and are pre-trained over a larger corpus.

These features are highlighted on the left side of Figure 2. They are concatenated with the average pooled representation of the document sequential transformer network and passed through a meta-learner, which is trained to perform the final prediction. The meta-learner used is a 2-layer feed-forward neural network with relu activation between the linear layers. The use of a small neural network as the meta-learner is motivated by allowing the model to adapt to the non-linear interactions between user-level and sequential features.

Alternative Models & Baselines We evaluated two heuristic baselines and two statistical baselines. These chosen heuristic baselines are often quite competitive in time-series applications, predicting the last observation again and an average of all past observations (Matero and Schwartz, 2020). In the case of our application, these are equivalent to predicting the last reported day's number of drinks and the average of all current and past days' drinks.

Our statistical baselines are a linear (ridge) autoregression and Gated Recurrent Unit (GRU) cell recurrent neural network (Chung et al., 2014). We train our GRU network using multi-head self-attention as introduced in Vaswani et al. (2017).

Language Association for Within-Person Drinkings Consumption To further understand the relationships between drinking behaviors and participants' language from *affective essays*, we analyze the associations between word usage and number of drinks quantitatively. We analyzed 4,939 *affective essays* from 489 participants. (some participants have missing data within the 14 days). Firstly, we employed Latent Dirichlet Allocation (LDA) (David M. Blei, 2003) topic modeling ($n = 200$, $\alpha = 2$) to identify the primary themes that emerged from the text to extract topic features for all essays. To identify the distinctive language used about drinking behavior, we applied differential language analysis (DLA) (Schwartz et al., 2017) to search for topic features that had the strongest positive or negative correlation with the number of drinks consumed on the previous day. To focus on the within-person signals, we applied fixed effects models, in which we mean-centered the input language features and output number of drinks with participant-wise averages across time. Consequently, this new multi-level differential language analysis shows insights into the language and behavior of participants changes compared to their daily language and average consumption. The reported correlations are beta coefficients from a standardized multi-level regression model where significance is validated via Benjimini-Hochberg correction (Benjamini and Hochberg, 1995).

Particularly for each participant i, with $X_{i,t}$ as the language topic features for the day t and $y_{i,t}$ as the number of drinks consumed 24 hours before the day t, consider the linear unobserved effects model:

$$y_{i,t} = X_{i,t}.\beta + \alpha_i + \epsilon_{i,t} \tag{1}$$

Where β is the parameter to be learned, α_i is the unobserved time-invariant individual drinking effect we aim to eliminate, and $\epsilon_{i,t}$ is the error term. Since α_i is not observable, it cannot be directly controlled for. To implement the fixed effects model, one can eliminate α_i by de-meaning X and y: $\ddot{X_{i,t}} = X_{i,t} - \bar{X}_i$ and $\ddot{y_{i,t}} = y_{i,t} - \bar{y}_i$, where t indexes the particular instance measurement for

[2] However, when mutl-level features are used, an FFNN is utilized.

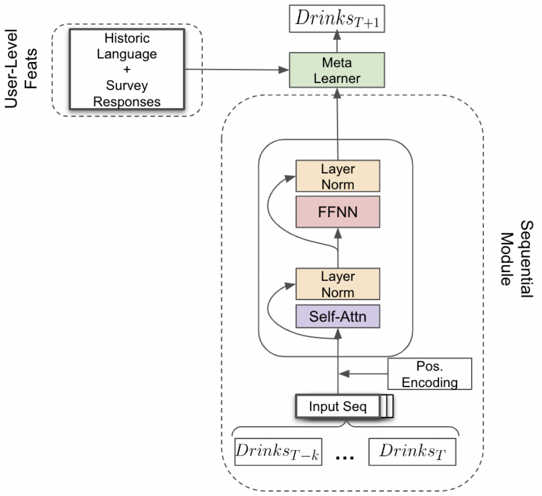

Figure 2: Architecture of our multi-level forecasting model with the contextual user-level module highlighted by a dashed box on the left-hand side. The sequential module is the document-level transformer that processes daily language and drinking data as per EMA responses. The avg-pooled sequential representation is concatenated with the user-level features and passed through a 2-layer FFNN called *meta-learner*.

participant i and the mean is over all instances of that user.

Since α_i is constant over time: $\ddot{\alpha_i} = \alpha_i - \bar{\alpha_i} = 0$ and the individual effect is eliminated. Thus equation (1) is transformed into equation (2) where the fixed effects estimator $\hat{\beta}_{FE}$ is then obtained by an OLS regression of $\ddot{y}$ and $\ddot{X}$.

$$\ddot{y_{i,t}} = \ddot{X_{i,t}}.\beta_{FE} + \ddot{\epsilon_{i,t}} \qquad (2)$$

Language Association for High and Low Risk Drinkers We partitioned the population into two groups based on their AUDIT-C scores for further investigation by gender. Males with scores greater than or equal to 5.5 and females with scores greater than or equal to 4.5 were deemed to belong to the high AUDIT-C category (Johnson et al., 2013), while the rest were placed in the low AUDIT-C category. The resulting sample comprised 234 high AUDIT-C participants (2,393 *affective essays*) and 241 low AUDIT-C participants (2,438 *affective essays*). For each category, we identified the top 30 topics correlated with the corresponding category. We then applied DLA algorithms to distinguish the language used to describe the drinking behaviors within each group.

Model (num days)	**MSE**	**MAE**	r
Heuristic Baselines			
Last Day	9.48	1.70	0.36
Average Drinks	5.02	1.44	0.58
Linear Models			
LinAR (5)	4.56	1.52	0.58
Deep Learning			
GRU (7)	4.62	1.44	0.59
TRNS (7)*	**4.22**	**1.33**	**0.62**

Table 1: Overall performance of our document sequential forecasting models. Models are trained using past drinking behavior, daily language features from EMA responses, and day-of-week markers. All models use the number of days found ideal during training, which was 7 for all except linear. **Bold** indicates best in column and * indicates statistical difference via paired t-test with $p < .05$ w.r.t GRU (7).

5 Results

Here, we showcase results using three separate metrics. First, we focus on mean squared error (MSE) as it is helpful to measure the impact of outliers where our models failed to predict as accurately and is also the metric we optimize for during training. Second, mean absolute error (MAE) shows errors within the same units (drinks per day). Lastly, Pearson r is used as a scale-invariant metric to show the relationship between model predictions and the actual trend.

For all tables shown, LinAR refers to a linear ridge (L2-normalized) autoregressive model, GRU is a gated recurrent neural network, and TRNS is our transformer based architecture.

Multivariate Forecasting We start by showing our best-performing multivariate sequential models compared to our baselines; shown in Table 1. We find that our heuristic baselines perform quite strongly, with the average number of drinks being the most competitive. In fact, we find that modeling the multivariate sequence using an autoregressive linear model fails to out-predict these baselines in 2 out of 3 metrics. However, both deep learning baselines offer improved performance, with both having a modest drop in MSE, showing their robustness to outliers. The transformer-based model performs better across all metrics, showcasing lower error and higher correlations. We believe this to be due to the superior modeling capabilities when it comes to modeling the complexities of changes in language over time.

Model (num days)	**MSE**	**MAE**	r
With Language			
TRNS (7)*	**4.22**	**1.33**	**0.62**
Without Language			
GRU (7)	5.48	1.51	0.46
LinAR (9)	4.49	1.35	0.59
TRNS (7)	4.29	1.43	0.62

Table 2: Comparison of predictive power using only past knowledge of number of drinks to forecast future number of drinks. All models use the number of days found ideal during training, which was 7 for all except linear. **Bold** indicates best in column and * indicates statistical difference via paired t-test with $p < .05$ w.r.t LinAR (9).

Model (num days)	**MSE**	**MAE**	r
Heuristic Baselines			
Last Day	9.48	1.70	0.36
Average Drinks	**5.02**	**1.44**	**0.58**
No Drinking History			
GRU (7)	7.21	1.80	0.28
TRNS (7)	5.85	1.77	0.42

Table 3: Evaluation of predictive power when the models do not have access to previous drinking behavior, a strong univariate signal, and instead are trained using only daily language and day-of-week flags. **Bold** indicates best in column.

Model (num dims)	**MSE**	**MAE**	r
TRNS (768)	4.91	1.45	0.54
TRNS (64)	4.39	1.34	0.60
TRNS (32)*	**4.22**	**1.33**	**0.62**
TRNS (16)	4.48	1.47	0.60

Table 4: Impact of number of language dimensions on predictive power. All models are trained with seven steps of history, which was found ideal. **Bold** indicates best in column and * indicates statistical difference via paired t-test with $p < 05$ w.r.t TRNS (768).

Univariate Forecasting We also compare our multivariate sequential models to the performance of univariate models in Table 2. None of the univariate models are capable of more accurate predictions than the best multivariate model, highlighting the importance language plays in detecting future behaviors. Interestingly, when shifting from multivariate to univariate, the GRU model fails to learn anything beyond the original average drinks baseline. On the other hand, the linear model sees quite a substantial performance improvement, implying that these models behave quite differently when limited to just a single feature dimension as input. Historically, linear univariate autoregressive models have been quite competitive with other sequential models such as RNNs (Matero and Schwartz, 2020; Sánchez Gavilanes, 2022; Menculini et al., 2021). At the same time, the modeling of language over time is likely too complex for such a model.

Covariates Only Next, in Table 3, we investigate the ability to forecast future drinking behaviors *without* knowledge of past drinking. For example, these models are trained using only a sequence of daily language as captured in the experienced well-being *affective essays* and the day-of-week markers. The transformer network is once again the best performing compared to the other statistical models, where we can get an absolute error close to that of knowing the number of drinks a person had the day before. This shows excellent utility for those running a study or clinicians already collecting language data from participants but do not have access to explicit drinking information. Only having a single open response field (experienced well-being) can predict future drinking almost as well as knowing how much a participant drank recently (past 24 hours).

Dimensionality Reduction We perform an additional sensitivity analysis over our models, where we explore the performance of the language features based on the number of dimensions. While previous studies have shown trends in the performance of dimensionality reduction sizes on human-level NLP tasks (Ganesan et al., 2021), they've not done so for tasks that span the temporal dimension or tasks specifically predicting beyond mental health or demographics. Thus, we show if these trends continue to hold in such a scenario in Table 4. We find that performance across all three metrics continues to increase as dimensions are reduced until only 16 language dimensions remain. This corroborates the findings of Ganesan et al. (2021), which suggests 32 dimensions for ideal results on a dataset of 200 people.

User-level Modeling In Table 5, we show the performance of using only the user-level features through the meta-learner as a stand-alone neural network (only using the user-module pipeline from Figure 2). We find that using only language gives a weak but reliable signal in terms of daily drinking. Alternatively, the baseline survey's psychological

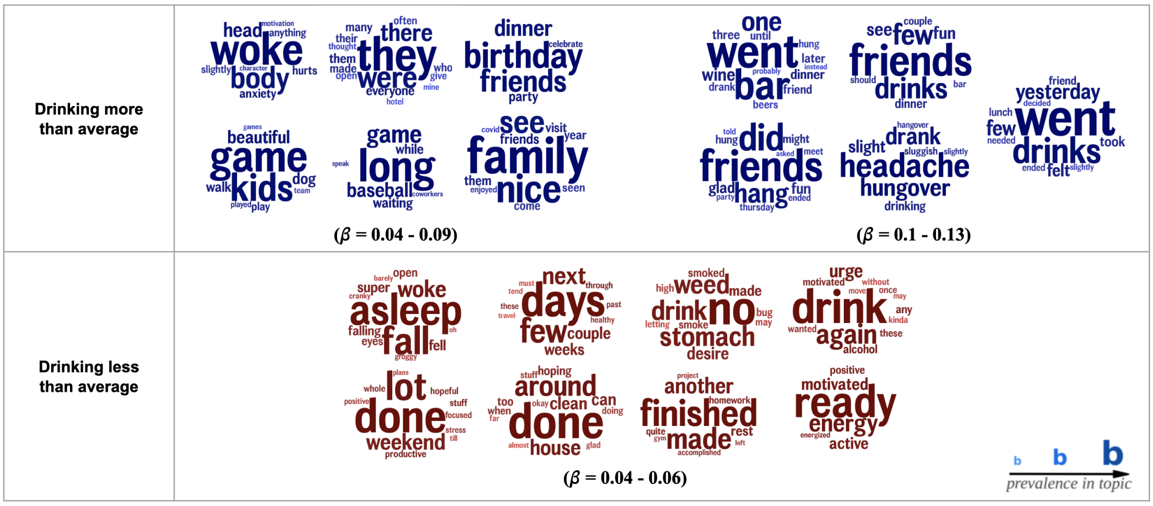

Figure 3: Worldcloud topics from the responses to *affective essays* associated with drinking more or less than average within participants. Association (β) is the coefficient from standardized multiple linear models ($p < 0.05$; Benjamini-Hochberg adjusted for false discovery rate, N=4,939 essays).

Model	MSE	MAE	r
Heuristic Baselines			
Last Day	9.48	1.70	0.36
Average Drinks	5.02	**1.44**	**0.58**
User-level			
Language	6.92	1.74	0.09
Survey	5.45	1.66	0.44
Lang+Survey*	**4.81**	1.50	0.51

Table 5: Performance of our user-level features as input into the meta-learner without using the document sequential (daily) module. Features use a user embedding representing past language used on social media and baseline survey responses. **Bold** indicates best in column and * indicates statistical difference via paired t-test with $p <$ 05 w.r.t Average Drinks.

and demographic features are quite competitive compared to the heuristic baselines. It is important to note that these survey features do *not* include any past information on drinking behaviors that the baselines have access to. When combining the language with the survey responses we see an increase in predictive power across all three metrics suggesting that the language features capture different covariance of drinking behaviors. While the user-level features do not outperform the heuristic baselines, they are still rather impressive as they are not leveraging the inputs of the sequential module and thus make the same prediction (static) for both testing days. Thus, there is likely a consistent personal factor for each individual that drives their drinking behaviors.

Model	MSE	MAE	r
Document Sequential			
TRNS	**4.22**	1.33	**0.62**
Multi-level Sequential			
TRNS*	**4.22**	**1.23**	**0.62**

Table 6: Performance of our multi-level model when incorporating contextual user-level information via the user module compared to using sequential data only. Both models use seven days of history, with the multi-level model also leveraging historic user-level features. **Bold** indicates best in column and * indicates statistical difference via paired t-test with $p < .05$ w.r.t Document Sequential TRNS.

Multi-level Sequential Forecasting Finally, in Table 6, we investigate the effect of using a multi-level forecasting model that leverages both the static user-level features and the dynamic time-series inputs. We see a small but significant increase in the ability to predict raw drinks per day (MAE) while maintaining the same level of MSE and Pearson r. This indicates that the feature spaces have overlapping covariance, but there are some aspects that are not accounted for in the sequential features. Especially concerning the absolute error in the raw number of drinks per day, in which most other approaches struggled to see large gains.

Language Association with Drinking Behaviors Figure 3 shows significant topics correlated positively (blue) and negatively (red) to drinking. Days when participants drink more than usual predomi-

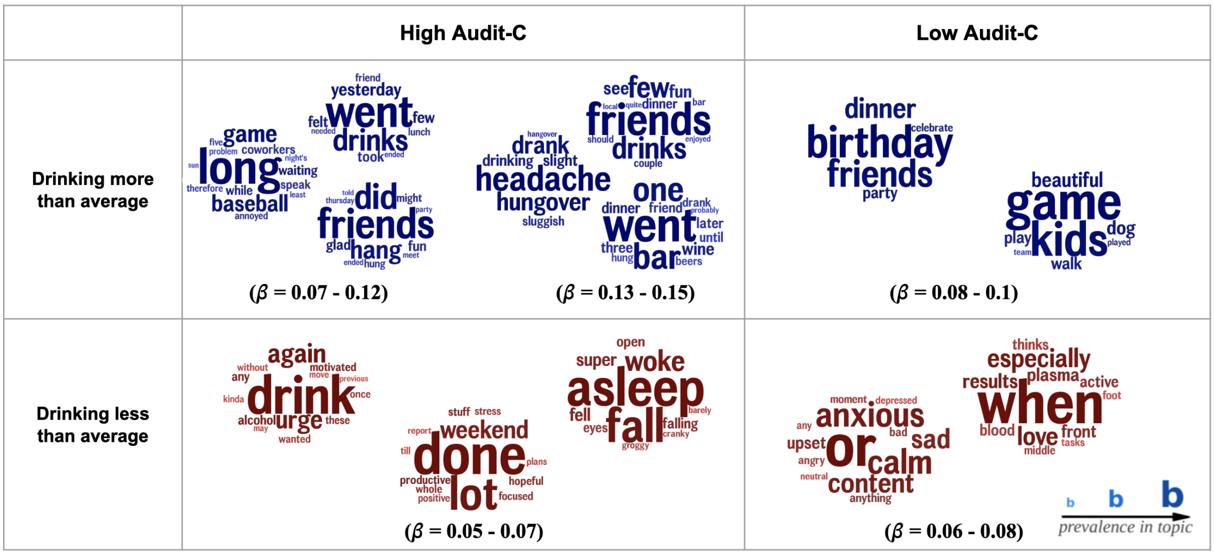

Figure 4: Worldcloud topics from the responses to *affective essays* associated with drinking more or less than average within participants, divided into groups of high AUDIT-C (N=2,393 essays) and low AUDIT-C (N=2,438 essays). Association (β) is the coefficient from standardized multiple linear models ($p < 0.05$; Benjamini-Hochberg adjusted for false discovery rate).

nantly relate to social experienced well-being language. For example, when participants drink more than usual, their language relates to friends, family, and social events (e.g., birthdays and dinners). Considering the positive relationship between spending time with others and positive affect (Grimm et al., 2015; Killingsworth and Gilbert, 2010; Diener and Seligman, 2002), and that positive affect rises on drinking days (Dora et al., 2022), the social language pattern related to drinking days is not surprising. Topics associated with drinking not related to social events include hangover-related language ("headache" and "woke, body, anxiety"). Conversely, the language associated with consuming less alcohol relates to accomplishment, energy, and urges to drink. Specific topics such as "energy, ready" and "fall asleep" seem contradictory. However, alcohol-consuming behavior is complex, and while the β values (0.05 - 0.15) are similar to the previous meta-analytic correlation between positive affect and drinking (Dora et al., 2022), the complexity of alcohol behavior (Geiger and MacKerron, 2016; Massin and Kopp, 2014) likely explain why language features divergent in meaning relate similarly to alcohol consumption. Further, no social language related to drinking less than normal, indicating that drinking can be *the* social platform for some individuals.

Language Analysis for High and Low AUDIT-C group The topics displayed in Figure 4 depict language that positively and negatively correlates with the number of drinks individuals consume, separated into high and low AUDIT-C. The high AUDIT-C group's motivations usually refer to social context, while the low AUDIT-C group refers to special occasions. For the low AUDIT-C group, the language significantly related to drinking was exclusively social, while for the high AUDIT-C group, the social aspects attenuated compared to the language pertaining to drinking, and the hangover language remained. When drinking less than usual, the high AUDIT-C group's language indicates the urge to drink and sleep, and the low AUDIT-C group mainly describes their common daily emotions.

Past research (Kornfield et al., 2018; Marengo et al., 2019; Moreno et al., 2016; van Swol et al., 2020; Jose et al., 2022) that has studied between-person signals across AUDIT-C scores find high AUDIT-C drinkers engage in discussions about alcohol consumption and profane language and low AUDIT-C drinkers often express an emphasis on religious beliefs. Here, we find that high AUDIT-C drinkers talk about alcohol consumption but do not use profane language, and low AUDIT-C drinkers do not mention religion. Our results provide an additional perspective on the complexities of drinking, where the language-based analyses demonstrate how divergent feelings and aspects can relate to drinking behaviors simultaneously.

6 Conclusion

Longitudinal, multi-level language analyses can be important for understanding human behavior, such as alcohol consumption and its motivations. In this work, we propose a multi-level longitudinal approach to analyze the language associations with drinking behaviors to find within-person signals. While much of previous work about language and drinking found characteristic differences *between people*, our approach yielded results that signal *day-to-day changes*, aligning with previous research on *within-person* changes in drinking associated with emotions and socializing. Our multi-level approach also yielded evidence for differing drinking motivations between people depending on their alcohol use disorder risk level, with lower AUDIT-C drinkers (those at lower risk) mentioning celebrations or special occasions more than those with higher risk.

7 Limitations

This study focuses on those who are potentially high-risk drinkers in the service industry, such as bartenders and restaurant workers in the United States. While participation was possible three times a day over 14 days, some participants dropped out after a few days or came in and out over the study. This lack of reports led to potentially noisy time series per participant, which had to be filled via interpolation techniques. All participants were also required to respond in English when crafting their experienced well-being *affective essay* responses and were filtered out if another language or spam was used.

Additionally, given that this dataset and task definition are novel, the size of the dataset used for forecasting could be considered small as it spans only 242 participants. While the data is longitudinal, with each participant having upwards of 14 days of data, the overall number of users motivates us to use techniques to avoid the curse of dimensionality (Ganesan et al., 2021).

Further, our multi-level model forecasts daily drinking consumption using focused language (*affective essays*), general public language (Facebook statuses), demographics (Age/Gender), and responses to psychological questionnaires (AUDIT-C, Depression, and Anxiety levels). The AUDIT-C is a shorthand questionnaire to get a rough estimate of one's level of alcoholism risk level. While there are more complete representations via the full AUDIT questionnaire, the structure of the study focused on short information-dense questionnaires as part of the initial participant baseline survey to capture many psychological outcomes.

8 Ethics Statement

This work aims to advance multi-disciplinary NLP-psychology *research* for understanding human behaviors associated with language. The models in this paper are not intended or validated for deployment in specific clinical settings and are not to be used for other commercial use cases, such as targeted marketing. The use cases this research is working towards are for developing more accurate and validated techniques for the benefit of society and human health. All participants in this research did so under informed consent without agreement to further share their non-anonymized individual data. The research was approved by an independent academic institutional review board (IRB).

This work is intended as a step toward an assistive tool, but it is not evaluated for such use at this point. Currently, we do not enable the use of our model(s) independently in practice to label a person's potential behaviors. Before our models are used by trained clinicians, they must demonstrate validity in a clinical setting for the target clinical population, with steps for evaluation reviewed by an ethical review board. Practice should follow clinical guidelines.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Nour Alayan, David Eddie, Lucille Eller, Marsha E Bates, and Dennis P Carmody. 2019. Substance craving changes in university students receiving heart rate variability biofeedback: A longitudinal multilevel modeling approach. *Addictive behaviors*, 97:35–41.

Fatemeh Almodaresi, Lyle Ungar, Vivek Kulkarni, Mohsen Zakeri, Salvatore Giorgi, and H Andrew

Schwartz. 2017. On the distribution of lexical features at multiple levels of analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 79–84.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

K. Bush, DR Kivlahan, MB McDonell, SD Fihn, and KA Bradley. 1998. The audit alcohol consumption questions (audit-c): an effective brief screening test for problem drinking. *Arch Intern Med*, 158(16):1789–1795.

Andre F Carvalho, Markus Heilig, Augusto Perez, Charlotte Probst, and Jürgen Rehm. 2019. Alcohol use disorders. *The Lancet*, 394(10200):781–792.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Kasey G Creswell, Yvonne M Terry-McElrath, and Megan E Patrick. 2022. Solitary alcohol use in adolescence predicts alcohol problems in adulthood: A 17-year longitudinal study in a large national sample of us high school students. *Drug and alcohol dependence*, 238:109552.

Brenda Curtis, Salvatore Giorgi, Anneke EK Buffone, Lyle H Ungar, Robert D Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H Andrew Schwartz. 2018. Can twitter be used to predict county excessive alcohol consumption rates? *PloS one*, 13(4):e0194290.

Michael I. Jordan David M. Blei, Andrew Y. Ng. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Ed Diener and Martin EP Seligman. 2002. Very happy people. *Psychological science*, 13(1):81–84.

Jonas Dora, Marilyn Piccirillo, Katherine T Foster, Kelly Arbeau, Stephen Armeli, Marc Auriacombe, Bruce D Bartholow, Adriene Beltz, Shari Blumenstock, Krysten Bold, et al. 2022. The daily association between affect and alcohol use: A meta-analysis of individual participant data.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

et al. Falcon, WA. 2019. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3.

Kamel Gana, Guillaume Broc, and Nathalie Bailly. 2019. Does the boredom proneness scale capture traitness of boredom? results from a six-year longitudinal trait-state-occasion model. *Personality and Individual Differences*, 139:247–253.

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. *arXiv preprint arXiv:2105.03484*.

Ben Baumberg Geiger and George MacKerron. 2016. Can alcohol make you happy? a subjective wellbeing approach. *Social Science & Medicine*, 156:184–191.

Carsten Grimm, Simon Kemp, and Paul E Jose. 2015. Orientations to happiness and the experience of everyday activities. *The Journal of Positive Psychology*, 10(3):207–218.

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.

Larry V Hedges. 1994. Fixed effects models. *The handbook of research synthesis*, 285:299.

Joop Hox. 1998. Multilevel modeling: When and why. In *Classification, data analysis, and data highways: proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation eV, University of Potsdam, March 12–14, 1997*, pages 147–154. Springer.

J Aaron Johnson, Anna Lee, Daniel Vinson, and J Paul Seale. 2013. Use of audit-based measures to identify unhealthy alcohol use and alcohol dependence in primary care: A validation study. *Alcoholism: Clinical and Experimental Research*, 37:E253–E259.

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

Rupa Jose, Matthew Matero, Garrick Sherman, Brenda Curtis, Salvatore Giorgi, Hansen Andrew Schwartz, and Lyle H Ungar. 2022. Using facebook language to predict and describe excessive alcohol use. *Alcoholism: Clinical and Experimental Research*, 46(5):836–847.

Kenneth S Kendler, Henrik Ohlsson, Jan Sundquist, and Kristina Sundquist. 2016. Alcohol use disorder and mortality across the lifespan: a longitudinal cohort and co-relative analysis. *JAMA psychiatry*, 73(6):575–581.

Margaret L Kern, Gregory Park, Johannes C Eichstaedt, H Andrew Schwartz, Maarten Sap, Luke K Smith, and Lyle H Ungar. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4):507–525.

Matthew A Killingsworth and Daniel T Gilbert. 2010. A wandering mind is an unhappy mind. *Science*, 330(6006):932–932.

Rachel Kornfield, Catalina L. Toma, Dhavan V. Shah, Troy J. Moon, and David H. Gustafson. 2018. What do you say before you relapse? how language use in a peer-to-peer online discussion forum predicts risky drinking among those in recovery. *Health Communication*, 33:1184–1193.

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*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

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.

Davide Marengo, Debora Azucar, Fabrizia Giannotta, Valerio Basile, and Michele Settanni. 2019. Exploring the association between problem drinking and language use on facebook in young adults. *Heliyon*, 5:e02523.

Sophie Massin and Pierre Kopp. 2011. Alcohol consumption and happiness: an empirical analysis using russian panel data. *Centre d'Economie de la Sorbonne*, pages 1–19.

Sophie Massin and Pierre Kopp. 2014. Is life satisfaction hump-shaped with alcohol consumption? evidence from russian panel data. *Addictive behaviors*, 39(4):803–810.

Matthew Matero, Salvatore Giorgi, Brenda Curtis, Lyle H. Ungar, and H. Andrew Schwartz. 2023. Opioid death projections with AI-based forecasts using social media language. *npj Digital Medicine*, 6(1):35.

Matthew Matero, Albert Hung, and H Andrew Schwartz. 2021a. Evaluating contextual embeddings and their extraction layers for depression assessment. *arXiv preprint arXiv:2112.13795*.

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.

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.

Lorenzo Menculini, Andrea Marini, Massimiliano Proietti, Alberto Garinei, Alessio Bozza, Cecilia Moretti, and Marcello Marconi. 2021. Comparing prophet and deep learning to arima in forecasting wholesale food prices. *Forecasting*, 3(3):644–662.

Megan A. Moreno, Alaina Arseniev-Koehler, Dana Litt, and Dimitri Christakis. 2016. Evaluating college students' displayed alcohol references on facebook and twitter. *Journal of Adolescent Health*, 58:527–532.

NIH. 2023. *Alcohol Use Disorder (AUD) in the United States: Age Groups and Demographic Characteristics*. NIH.

Annick Parent-Lamarche, Alain Marchand, and Sabine Saade. 2021. A multilevel analysis of the role personality play between work organization conditions and psychological distress. *BMC psychology*, 9(1):1–15.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ricardo Andrés Sánchez Gavilanes. 2022. *Univariate time series forecasting: comparing ARIMA & LSTM neural network to the random walk benchmark for exchange rates*. Ph.D. thesis, Instituto Superior de Economia e Gestão.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, pages 55–60.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H Andrew Schwartz. 2022. Human language modeling. *arXiv preprint arXiv:2205.05128*.

Rong Su, Gundula Stoll, and James Rounds. 2019. The nature of interests: Toward a unifying theory of trait-state interest dynamics. In *Vocational interests in the workplace*, pages 11–38. Routledge.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022. 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*, pages 184–198.

Lisa Van der Werff, Yseult Freeney, Charles E Lance, and Finian Buckley. 2019. A trait-state model of trust propensity: Evidence from two career transitions. *Frontiers in Psychology*, 10:2490.

Lyn M. van Swol, Chia T. Chang, Brianna Kerr, and Megan Moreno. 2020. Linguistic predictors of problematic drinking in alcohol-related facebook posts. *Journal of Health Communication*, 25:214–222.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2017. What to do next: Modeling user behaviors by time-lstm. In *IJCAI*, volume 17, pages 3602–3608.

A Appendix

A.1 EMA Question Details

The exact phrasings of the relevant EMA questions, number of drinks, and experienced well-being essays are as follows:

- How many standard drinks did you have in the past 24 hours?
- Using the box below, please describe in 2 to 3 sentences how you are currently feeling.

A description of ”standard drink” is given alongside the question describing the typical definitions in beer, malt liquor, wine, and distilled spirits. Such that the following are defined as a standard drink: (1) 12 fl oz of a 5% beer, (2) 8-9 fl oz of a 7% malt liquor, (3) 5 fl oz of 12% wine, and (4) 1.5 fl oz of a 40% spirit.

A.2 Implementation Details

All models were built using PyTorch (Paszke et al., 2019) and Lightning (Falcon, 2019) with hyperparameter tuning using Optuna (Akiba et al., 2019). Hyperparameters explored were learning rate between $5e-2$ and $5e-5$ and weight decay between 0.01 and 1.0. 10% of users were selected as a held-out validation set for hyperparameter tuning by random sampling. For these users, their last 2 days of drinking were only used for parameter tuning and thus were not included in the test set. However, their first k days of responses were included in training data using an out-of-sample time configuration (Matero and Schwartz, 2020). A random seed of 1337 was used for all training experiments.