Can Natural Language Processing Effectively Track Negative Emotions in the Daily Lives of Adolescents?

Emotions play a fundamental role in human life, serving as essential cues that influence our actions and interactions with the environment (Barrett et al., 2007). Emotional states act as immediate alerts to potential benefits or dangers, driving us towards actions that align with our personal goals and away from potential threats (Frijda, 1988). Their dynamic nature, characterized by frequent shifts over time, significantly impacts mental health. Research shows that both the intensity of emotions and their temporal dynamics can contribute to the development of mood disorders (Fisher et al., n.d.; Houben et al., 2015; Houben & Kuppens, 2020; Kuppens & Verduyn, 2017; Schoevers et al., 2021). This is especially important for adolescents—a developmental period marked by high negative emotions and rapid mood fluctuations, all of which contribute to their elevated risk for mood disorders such as depression and anxiety (Bailen et al., 2019; Bennik et al., 2014; Hollenstein & Lanteigne, 2018). Given this profound impact, developing effective methods to measure daily emotional changes that are specifically suitable for adolescents is crucial. These measurements could identify both acute emotional distress and longer-term maladaptive patterns, serving two key purposes: enabling timely interventions during periods of high negative affect and facilitating targeted intervention to address maladaptive emotional patterns. Such methods could help predict increased risk of mood disorder onset, tailor treatment plans, and evaluate intervention efficacy, ultimately contributing to improved mental health outcomes.

Language offers a promising avenue for measuring emotion, as people convey their emotional states both explicitly and implicitly through their choice of words and their manner of speaking or writing (Davitz, 2013; Pennebaker et al., 2003). Language plays a dual role in the realm of emotions, serving not only as a primary tool for expressing and communicating

emotional states but also as an actively shaping how emotions are experienced (Jablonka et al., 2012; Lindquist, 2017). Therefore, the current study aims to explore whether text can be used to track within-person fluctuation in emotional states.

With the rise of digital communication, people, especially adolescents, now express themselves extensively through text-based interactions, including social media posts, messages, and online discussions. This shift has generated an unprecedented volume of language data, offering a unique opportunity to study adolescent emotions on a larger scale (Iliev et al., 2015). Advances in natural language processing (NLP), a multidisciplinary field combining computer science, artificial intelligence, and linguistics, have enabled researchers to analyze this data efficiently and meaningfully (Hirschberg & Manning, 2015; Manning, 1999). One key advantage of NLP over traditional, manual text analysis is its ability to handle large datasets—such as thousands of social media posts or digitized texts—quickly and efficiently. By leveraging these advancements, NLP methods can identify patterns in language use (e.g., sentiment, tone, and self-referential quality), offering real-time insights into emotional states without requiring participants to actively report their feelings (Carlier et al., 2022; Kahn et al., 2007; Sun et al., 2020). Importantly, text-based communication provides insights not only into the intensity of emotions but also into the context in which those emotions arise. This contextual information can reveal situational factors contributing to emotional experiences, offering a more nuanced understanding of emotional processes. While NLP has been used in behavioral science for decades, its increased accessibility and affordability have significantly improved, making it an increasingly popular tool (Feuerriegel et al., 2025).

In recent years, various text analysis tools have emerged (Boyd et al., 2022; Demszky et al., 2023; Eichstaedt et al., 2021; Neuendorf, 2017; Rathje et al., 2024; Tausczik & Pennebaker,

2010; van Loon, 2022), ranging from basic approaches like counting word frequencies to more advanced methods, such as leveraging large language models (LLMs), each with its own set of advantages and limitations. For example, closed-vocabulary programs such as LIWC (Pennebaker, 2001), use predefined dictionaries to categorize words. While highly interpretable, transparent, and efficient at summarizing concepts, these methods often neglect context, leading to potential misinterpretations (Eichstaedt et al., 2021). In contrast, open-vocabulary approaches, such as Latent Dirichlet Allocation (LDA), and word embedding methods, leverage data-driven techniques to examine a broader spectrum of words and topics (Blei et al., 2003; Griffiths et al., 2007; Sivakumar et al., 2020). These methods are better at capturing nuances, addressing ambiguous word meanings, and are less susceptible to misinterpretations. Limitations of open-vocabulary approaches include the need for more technical expertise, larger datasets, and careful consideration of parameter choices, as well as challenges in interpretability (Eichstaedt et al., 2021; van Loon, 2022). More recently, LLMs like GPT have shown great promise in accurately identifying various psychological constructs in text, overcoming many of the constraints of older methods (Rathje et al., 2024). These models can interpret the context of words and have achieved effective results across multiple languages with simple prompts. The main limitations of this approach include a lack of transparency in how inferences were generated and difficulties in reproducing results due to their probabilistic nature (Abdurahman et al., 2024). For example, although LLM-based approaches could predict whether a person experiences negative emotions based on social media posts, LLMs could not be used to identify which linguistic cues (such as the use of pronouns) predict this experience (Feuerriegel et al., 2025; Rathje et al., 2024). In the latter case, dictionary-based methods are preferrable. Ultimately, closed- and open-vocabulary approaches, along with advancements in LLMs, provide complementary strengths that

significantly enhance our ability to understand psychological states through language. By combining these methods, researchers can leverage their unique advantages while mitigating their limitations, offering a robust alternative to traditional and relatively cumbersome self-report measures assessing emotions.

Despite the increasing sophistication of these methods, significant gaps remain in the current literature. Most importantly, existing studies on text-based emotion prediction have focused on exploring between-person differences, primarily identifying which text features are associated with individuals experiencing high levels of negative emotions (e.g., Tackman et al., 2019). These approaches aggregate data across participants and aim to pinpoint common linguistic markers, such as the use of first-person pronouns or negative sentiment, that are indicative of heightened emotional distress and depression (Bathina et al., 2021; Funkhouser et al., 2024). However, this method overlooks both between-person differences in how linguistic markers relate to emotions and within-person fluctuations in emotional states. Even when using multilevel models to predict within-person fluctuations in emotions, models assumes that the same linguistic features signal high levels of negative emotions across different individuals (Funkhouser et al., 2024). Such assumptions fail to account for the fact that language is inherently idiosyncratic, with individuals expressing emotions in unique and context-dependent ways. Beck & Jackson, (2022) demonstrated the importance of idiographic approaches by showing that the psychological and situational antecedents that predicted future loneliness varied substantially across participants, with no two individuals showing the same pattern of predictive features. This highlights the need for personalized models that adapt to the distinctive ways people express emotions through text, enabling more accurate predictions tailored to each individual.

Moreover, previous research has typically studied language features in isolation, missing the potential benefits of combining different language-based methods (Carlier et al., 2022). This is especially important as features that may be significant for one individual may be not important for others. To address these limitations, this study integrates diverse language-based tools and leverages machine learning to develop personalized models capable of monitoring within-person fluctuations in emotional states.

The Present Study

The study's primary goal is to evaluate the effectiveness of various text analysis and machine learning techniques in tracking within-person fluctuations in negative affect (NA) over time. Since individuals express emotions in a variety of linguistic ways, this research aims to develop more personalized approaches to emotion tracking.

In addition to this main objective, we address two key questions:

1) Do idiographic (individual-level) models outperform nomothetic (group-level) models in predicting fluctuations? This question is motivated by the fact that individuals may express emotions through language in highly unique ways.
2) Is combining various NLP approaches beneficial for emotion prediction?

The research employs and compares three types of Natural Language Processing (NLP) approaches, which, as noted above, have complementary strengths and limitations:

a) Closed vocabulary (LIWC and VADER)

b) Open vocabulary (LDA)

c) Large Language Models (GPT)

By addressing these questions, the study aims to enhance the accuracy of emotion prediction models, potentially enabling closer monitoring of emotional fluctuations in daily life.

The ultimate goal is to use these improved emotion predictions to inform the delivery of scalable, real-time, and personalized interventions for alleviating high NA states and improving emotion regulation abilities.

Method

Participants

Participants were 98 English-speaking adolescents aged 12-18 (XX female, XX male; m_{age} = XX, *SD* = XX) recruited from the greater Boston area, derived from two larger studies that recruited typically developing (non-anhedonic) adolescents, as well as adolescents with elevated levels of anhedonia. Exclusion criteria included history or current diagnosis of any of the following DSM-5 psychiatric illnesses: major depressive disorder, schizophrenia spectrum or other psychotic disorder, bipolar disorder, substance or alcohol use disorder within the past 12 months or lifetime severe substance or alcohol use disorder, as well as current diagnosis of anorexia nervosa or bulimia nervosa for an additional information about sample inclusion and exclusion criteria see Murray et al., (2023).

Procedure

All procedures were approved by the Mass General Brigham IRB. Written informed consent was provided by participants who were 18 years of age, as well as from parents of participants who were under 18 years of age, along with the participant's written informed assent. At the baseline session participants were administered a semi-structured clinical interview, the Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS; Kaufman et al., 1997), either in-person or over Zoom. Participants also completed self-report measures and installed the MetricWire App on their smartphones to complete ecological momentary assessments (EMAs). EMA surveys were delivered 3-4 times per day whereby participants were

randomly signaled during two timeslots (4 pm to 6:30 pm and 6:30 pm to 9 pm) during a 5-day period (Thursday - Monday). A third survey was sent on weekends (11 am-4 pm). Of the participants, 39 completed the EMA for four consecutive weeks, while 59 completed it over four weeks on an every-other-week schedule.

Ecological Momentary Assessment (EMA): Participants were asked to rate on a 5-point Likert scale, ranging from 1, "Very slightly or not at all," to 5, "Extremely" the extent to which they were feeling several emotions immediately before they started the assessment. Participants' negative affect (NA) included responses for: "sad," "nervous," and "angry." Mean NA was measured by averaging the three NA variables. In addition, participants responded to the following open-ended questions: 1) What were you thinking about right before you started this survey? 2) Think about the most enjoyable or happy time since you completed the last survey (or if this is your first survey, then the last 24 hours). Very briefly, what happened (1-2 sentences is fine)? 3) Think about the most stressful or negative time since you completed the last survey (or if this is your first survey, then the last 24 hours). Very briefly, what happened (1-2 sentences is fine)? Participants with fewer than 30 observations were excluded from the analysis.

Language Measures

We used the following strategies to extract text features and generate quantitative summaries of the language data.

Close vocabulary: For close vocabulary, we extracted features using the Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker, 2010) and the Valence Aware Dictionary and sEntiment Reasoner (VADER; Hutto & Gilbert, 2014). LIWC is a computerized text analysis tool that categorizes words into over 90 linguistic and psychological dimensions based on an internal dictionary of approximately 6,400 words (Boyd & Schwartz, 2021). It

calculates the percentage of words that match each predefined category, offering insights into linguistic structures (e.g., pronouns), psychological constructs (e.g., affect), and broader language patterns (e.g., analytical thinking). The LIWC has been extensively validated across numerous studies and is widely used in psychological research to quantify language use patterns associated with various psychological states and traits. For this study, we used LIWC-22 (Boyd et al., 2022), the latest version of the software, the analyze participants' responses to the EMA open questions.

VADER is a simple rule-based model for general sentiment analysis optimized for social media. It provides four sentiment scores: negative, positive, neutral, and compound (an overall sentiment score from -1 to +1). VADER is particularly effective at handling sentiments expressed in short, informal text and accounts for factors like punctuation, capitalization, and modifiers (e.g., intensifiers like "very") that influence the intensity of the sentiment. Whereas LIWC excels in offering detailed psychological and linguistic insights across a wide range of texts, VADER is more attuned to detecting sentiment polarity and intensity in social media. Finally, text lengths were extracted as features for each question and included in the models.

Open vocabulary: Latent Dirichlet Allocation (LDA) is a probabilistic clustering method that groups words into topics based on their co-occurrence across a text corpus (Blei et al., 2003; Griffiths et al., 2007). Unlike predefined dictionaries, LDA generates topics directly from the data, identifying latent patterns in word usage. Each word is assigned to one or more topics, iterating until an optimal balance is reached. This results in a set of posterior probability distributions, which approximates the likelihood of each word occurring within each topic. This allows LDA to create semantically coherent clusters, overcoming word sense ambiguities by contextually assigning words to topics. We ran LDA on one- to three-word phrases, rather than

individual words alone, so that phrases such as ‘Summer camp’ are treated as a single term. We implemented Latent Dirichlet Allocation (LDA) using distinct approaches for nomothetic and idiographic analyses. For the nomothetic model, we extracted a uniform set of topics across the entire sample. For the idiographic model, we derived personalized topics unique to each individual.

The latent topics were extracted from the preprocessed text corpus using probabilistic modeling with the R package topicmodels (Grün & Hornik, 2011). For the whole-sample (nomothetic) LDA, before creating the model, we performed the tuning of the algorithm to define the number k of topics using *ldatuning*. To determine the optimal number of topics, we evaluated the quality of topic modeling using three metrics: Griffiths2004 (Griffiths2004), CaoJuan2009 (Cao et al., 2009), and Arun2010 (Arun et al., 2010). These metrics included both minimization criteria (CaoJuan2009 and Arun2010), which are optimized when minimized, and maximization criteria (Griffiths2004), which are optimized when maximized. Additionally, Deveaud2014 provided a reference point that decreases linearly with the number of topics. We selected six topics based on model fit metrics, as illustrated in Figure 1. The selection of six topics reflected a balance among these metrics. Table 1 represents each topic by its 15 most highly frequent words, providing a clear semantic foundation for further analysis.

Large Language Models (GPT). To generate emotion ratings using large language models (LLMs), we employed the Generative Pre-trained Transformer (GPT), an autoregressive AI language model developed by OpenAI. GPT is built on a transformer-based architecture, a neural network design that excels at processing sequential data by using self-attention mechanisms to capture contextual relationships across words. This allows GPT to generate human-like text by predicting the next word in a sequence based on the provided context.

Specifically, we utilized GPT-4, an advanced version pre-trained on a vast dataset (45TB), enabling it to generate coherent sentences and perform various tasks such as writing, answering questions, and engaging in conversations. In this study, we prompted GPT-4 to rate the extent to which a participant experienced one of the following emotions: Sadness, Anger, and Nervousness, on a scale of 1 to 5 (similar to the scale used in EMA), based on responses to three open-text questions. See the online supplement for the full prompt used to generate GPT responses.

Data Analysis

Model Specification

We compared both nomothetic (group-level) and idiographic (individual-level) models in their ability to predict variability in negative emotional states *within* individuals over time. We employed two machine learning approaches to predict negative affect: elastic net regression and random forest models. Elastic net regularization (ENR) is a popular regression technique that combines two types of penalties: ridge and lasso. This combination helps address issues related to multicollinearity by constraining the coefficients of correlated variables while also minimizing model overfitting. On the other hand, random forest (RF) is an ensemble learning method based on decision trees. Unlike ENR, random forest can capture complex nonlinear relationships and interactions between variables without having to specify them in advance. Given that these two methods use different strategies for selecting and weighting variables, comparing them can help to determine which one provides the most accurate predictions in a given context.

To assess model performance and generalizability while minimizing overfitting, we implemented a nested cross-validation (CV) procedure using the *nestedcv* package (Lewis et al., 2023). This method consists of two levels of cross-validation: 1. Outer loop: Evaluates the

overall performance of the model by repeatedly training and testing on different subsets of the data. 2. Inner loop: Conducts model selection and hyperparameter tuning to optimize performance. We used 10-fold cross-validation in both the inner and outer loops. In the outer loop, the dataset was divided into ten folds. The model was trained on nine folds and tested on the remaining fold, with this process repeated ten times so that each fold served as the test set once. This iterative process ensures that performance metrics are evaluated across multiple train-test splits, providing a robust estimate of generalizability. Within the inner loop, models were trained on each training fold, and hyperparameter tuning was conducted by selecting the parameter combination that maximized performance on a validation set. The best-performing model was then applied to the corresponding outer loop test set. By separating hyperparameter optimization from performance evaluation, this approach prevents data leakage and mitigates the risk of overfitting. Nested cross-validation offers several advantages over a simple train-test split. It allows for efficient use of all available data for both training and validation while reducing the bias introduced by a single, potentially unrepresentative data split. Additionally, by incorporating multiple train-test iterations, this approach yields a more robust and reliable estimate of model performance, ensuring that the findings generalize beyond the specific dataset used in training.

After iterating through all outer folds, we aggregated the predictions from the left-out test sets and compared them against true values to compute overall predictive performance. We assessed model accuracy using R^2 and root mean squared error (RMSE). We consider RMSE to be an intuitive metric for assessing predictive accuracy on a target variable measured on a 7-point Likert scale, as it accounts for larger errors more heavily and allows us to make statements such as, “on average, the model’s prediction deviates from momentary subjective stress by approximately 0.80 points on a 7-point scale.”

First, we ran nomothetic models that included features extracted from all three NLP approaches, including all subjects for group-level analysis. These models aimed to identify common predictors of negative emotional states at the group level. To assess how well group-level models performed for individual participants, we adopted a nomothetic-idiographic approach, where metrics derived from nomothetic models such as R^2 were calculated separately for each participant.

Finally, we employed fully idiographic models, building separate models for each participant to capture person-specific language-emotion associations. These models focus exclusively on within-person variability, enabling highly individualized predictions. By comparing the three abovementioned approaches, we aimed to understand their relative strengths in predicting negative affect. In these idiographic models, variables with low variance were removed based on standard frequency criteria, where the most common value for that variable could not exceed 95% of the total observations.

Feature importance

To evaluate feature importance, we used SHAP (SHapley Additive exPlanations) values, which quantify the contribution of each feature to the model's predictions in a consistent and interpretable manner by assessing how variations in a feature impact the model's output (Lundberg, 2017). SHAP values were calculated using the R packages *fastshap* and *ggbeeswarm*, which facilitates visualization and interpretation of feature contributions. This method allowed us to identify the relative importance of predictors in the model and their specific effects on the predicted outcomes.

Results

Demographic and clinical characteristics

Table 2 presents the demographic and clinical characteristics of the participants. Participants completed an average of XX EMA surveys (s.d. = XX). Average EMA compliance was XX% (s.d. = XX%).

Model performance

Model performance metrics for the nomothetic (one model combinging all observations from all participants), nomothetic-idiographic (same as above but calculating performance separately for each participant), and idiographic (building separate models for each individual) approaches using text features extracted from the three NLP approaches to predict negative affect are shown in Table 3.

The nomothetic models generally outperformed both the nomothetic-idiographic and idiographic models across all emotions. Specifically, when predicting negative affect, sadness, anger, and nervousness, the nomothetic RF model yielded R^2 values of 0.38, 0.33, 0.20 and 0.24, respectively, whereas the nomothetic ENR model yielded R^2 values of 0.17, 0.14, 0.13 and 0.11, respectively. The substantially higher performance of the random forest model can be attributed to its ability to account for participant ID (grouping by individuals), which Elastic net cannot. This inference is supported by the feature importance analysis shown in Figure 2, where participant ID emerges as the most important feature in the RF model. This finding points to substantial differences between individuals in their average levels of negative affect, emphasizing the importance of person-specific modeling approaches to better capture within-person fluctuations.

Building on this and aligning with the ultimate translational goal of using text to track *within*-person fluctuations in negative emotions (rather than between-person differences), we proceed to evaluate the model's predictive performance for each participant individually (nomothetic-idiographic approach). Performance decreased substantially, with mean R^2 values in the range of 0.06-0.11. However, there was considerable variability in predictive accuracy across participants. For example, when predicting negative affect, R^2 values ranged from 0.00 to 0.41. For anger, R^2 achieved values as high as 0.69, indicating that for some participants, text features explained little or no variability in negative affect, while for others, text features explained a significant portion of the variability.

The idiographic approach (individual models) showed similar average performance to the nomothetic–idiographic approach, with mean R^2 values ranging from 0.06 to 0.10. Figure 3 demonstrates examples of high vs. low accuracy person specific models. Figure 4 further shows that, despite this individual variation, the overall trend (represented by the dashed line) demonstrates a generally positive relationship between predicted and actual negative affect ratings. The comparison between the nomothetic–idiographic and idiographic approaches yielded mixed results. Differences emerged in their ability to achieve relatively high performance for specific individuals and in the proportion of significant associations across all participants. For negative affect, the idiographic approach exhibited a wider range of R^2 values across participants, implying that this model achieved higher performance for some participants. However, the nomothetic–idiographic approach slightly outperformed it in terms of the proportion of participants with significant associations (55–58% vs. 43–48%). A similar pattern was observed for sadness, suggesting that although person-specific models can yield higher

accuracy for certain participants, the nomothetic–idiographic approach is more robust for identifying significant associations across a larger number of individuals.

For anger, the nomothetic–idiographic approach not only showed greater variability in R^2 values but also identified more significant associations than the idiographic approach. In contrast, for nervousness, the fully idiographic approach demonstrated both greater variability—allowing some participants to achieve R^2 values as high as 0.71—and a larger number of significant associations overall.

Which Text Features Are Most Predictive in High- vs. Low-Performing Models?

Figure 5 displays the top 10 text features from the four best-performing subject-specific random forest models predicting negative affect. Critically, this visualization demonstrates significant variation in the features influencing each participant's model, suggesting differences in the relative importance of features derived from various NLP approaches (LIWC, VADER, LDA, and GPT) across individuals. Though feature importance varied across individuals, we also examined whether specific patterns emerged across high- and low-performing participant-specific models and whether certain NLP approaches provided more influential features in high-performing models compared to low-performing ones.

Participants were divided into groups based on model performance (R^2), with the top 25% (n=25) classified as "High R2" and the bottom 25% (n=25) as "Low R2." Feature importance scores were analyzed separately for RF and ENR models. Importantly, participants with high- and low-performing models showed no significant difference in number of observations (RF: $High_{mean}$ =72.84, Low_{mean}=60.64, t=1.69, p=0.09, CI=-2.31;26.71, ENR: $High_{mean}$ =62.72, Low_{mean}=71.36, t=-1.26, p=0.21, CI=-22.38;5.10) or in mean negative affect (RF: $High_{mean}$ =0.68, Low_{mean}=0.45, t=1.85, p=0.07, CI=-0.02;0.49, ENR: $High_{mean}$ =0.77, Low_{mean}=0.65,

t=0.72, p=0.47, CI=-0.20;0.43) but for RF high and low performing models differed significantly in variability (RF: $High_{mean}$ =0.61, Low_{mean}=0.40, t=3.26, p=0.002, CI=0.08;0.33, but not for ENR: $High_{mean}$ =0.59, Low_{mean}=0.50, t=1.29, p=0.20, CI=-0.05;0.22). Assessment number ("Time") emerged as the most important feature in both RF and ENR models, suggesting that time-related variability was a relatively strong predictor of NA among the “High R2” subjects. In addition, as shown in Figure 6, for the RF high-performing models (High R^2 group), the top 5 important variables included features from different approaches (LIWC, GPT, VADER, and LDA). Notably, 3 of the top 5 variables that differed the most between the High R^2 and Low R^2 groups were GPT features (e.g.. ….), suggesting that GPT-derived variables may have been particularly effective at enhancing model performance. In contrast, for high-performing ENR models, the most important predictive features were from LIWC (e.g., xxxx), with LIWC variables also showing the largest differences between high- and low-performing models.

Do Closed Vocabulary, Open Vocabulary, LLM, or Combined Approaches Perform Best in Predicting Model Performance?

Figure 7 displays the performance metrics for each NLP approach used in the idiographic models. When examining the idiographic predictive performance of each NLP approach separately, GPT generally showed the highest predictive power for negative affect, sadness, anger, and nervousness, with R^2 values around 0.10 for negative affect and sadness, and slightly lower for anger and nervousness. These results were comparable to those achieved by the idiographic models that combined all NLP approaches, except in the prediction of nervousness where GPT slightly outperformed the combined idiographic model. However, it is important to note that while GPT showed strong R^2 values, it also had a higher RMSE compared to the combined models, suggesting a higher average prediction error despite

accounting for more variance in the outcomes. LIWC+VADER and LDA demonstrated considerably lower predictive power, with the lowest average R^2 values and fewer significant associations across all three emotions and the general scale, especially for anger and nervousness. This highlights that the contextual understanding provided by GPT was more effective in capturing emotional nuances compared to the other methods when NLP approaches were tested individually. Overall, while GPT captures variability in negative affect (as reflected by R^2), its predictions are less precise in terms of exact values (as indicated by RMSE). Combined models, which balance capturing variability and minimizing prediction errors, offer a more robust and reliable approach to emotion prediction.

Discussion

In this study, we combined multiple Natural Language Processing (NLP) approaches to examine whether within-person fluctuations in emotion can be accurately tracked through text analysis. Recognizing the idiosyncratic nature of emotional communication, we compared idiographic models—tailored to individual patterns—with nomothetic models that capture common trends across groups. By leveraging advanced NLP and machine learning techniques to track moment-to-moment emotional changes through text analysis, our approach has the potential to enhance mental health monitoring, support clinical decision-making, and enable early detection of distress for timely, personalized interventions.

The results showed that, overall, nomothetic models showed high performance in continuously predicting negative emotions (R^2 range: 0.11-0.38). These findings align with previous studies demonstrating the utility of NLP approaches in detecting emotional states across participants (Akhtar et al., 2019; Tanana et al., 2021). However, while nomothetic models identify general trends across participants, they do not necessarily capture nuanced within-person fluctuations. When we calculated *performance metrics* for each participant individually

(nomothetic-idiographic approach), the mean models' performance declined significantly and showed high between-person variability, indicating that the nomothetic models' ability to track within-person changes varies considerably from one individual to another.

When we built separate models for each participant (idiographic models) and compared their performance to that of nomothetic-idiographic we found that the overall predictive performance was comparable between the two approaches. In addition, idiographic models revealed significant variability in the text features predicting emotional states, indicating that individuals express negative emotions in distinct linguistic ways. This raises an intriguing question: if individuals differ so markedly in their predictive features, how can a nomothetic model that captures only general trends perform just as well? One plausible explanation is a trade-off between statistical power and individual variation. While idiographic models capture unique, person-specific patterns, nomothetic models benefit from larger datasets that enable the estimation of multiple, weaker yet stable feature-emotion relationships. In other words, although individual differences exist, the robust common patterns identified by the nomothetic approach appear sufficient to achieve similar predictive accuracy.

Supporting this interpretation, idiographic models provided highly personalized predictions for some individuals (e.g., across all emotions, R^2 reached up to 0.84 for idiographic models vs. 0.69 for nomothetic-idiographic models). However, nomothetic models demonstrated greater consistency across participants, successfully tracking fluctuations in negative emotions for 58% of individuals, compared to 48% for idiographic models.

While we are not aware of studies directly comparing nomothetic and idiographic approaches in text analysis, previous research comparing their effectiveness in tracking mental states using passive sensor data from smartphones and actigraphy, as well as self-reported

ecological momentary assessment (EMA), has yielded similar results (e.g., Aalbers et al., 2023; Cheung et al., 2017; Rozet et al., 2019; for exception see Soyster et al., 2022 who found that nomothetic models outperform idiographic models). For example, Aalbers et al., 2023 used smartphone passive sensor data to predict stress levels. Their findings revealed that idiographic models demonstrated higher accuracy in tracking stress levels for some participants (Spearman's ρ rank-order correlation up to 1 in idiographic models vs. up to .65 in nomothetic models). However, nomothetic models significantly predicted stress for a larger proportion of participants (up to 23.2% for idiographic models vs. up to 55% for nomothetic models). Rozet et al. (2019) used a comparable method and initially found that nomothetic models performed better (i.e., were more accurate) than idiographic models. However, as more data accumulated, the performance of the idiographic model eventually equaled and then surpassed that of the nomothetic model, suggesting that idiographic models may be a better option when sufficient data is available.

When comparing specific NLP approaches, GPT outperformed other models in its ability to monitor fluctuations in emotional states. It achieved R^2 values comparable to a combined model and demonstrated significant associations with reported emotions for 57% of the participants. However, it is important to note that despite these strong R^2 values, GPT also produced higher RMSE scores compared to the combined models. This discrepancy implies that while GPT effectively captures the overall variance in emotional states, it may be less precise in predicting the exact numerical values of these states. Such findings highlight the trade-off between capturing broad patterns and maintaining fine-grained precision—a balance that future research should aim to optimize, with fine-tuning offering one promising avenue for enhancement.

These findings are consistent with recent work by Rathje et al., (2024), which reported high correlations (r = 0.66 to 0.75) between GPT-4 outputs and human ratings of emotion, whereas dictionary-based approaches showed much lower correlations (r = 0.22 to 0.30). The higher correlation between GPT and human ratings in Rathje et al. (2024) study is expected as both GPT and human raters evaluated the emotion directly expressed in text, whereas our study aimed to predict participants' self-reported emotional experiences—a task complicated by the fact that individuals do not always explicitly communicate their internal states. Thus, while GPT and human annotation in Rathje et al. represent agreement between two raters performing the same task, our approach required GPT to infer emotions that participants may not have explicitly expressed.

From an applied perspective, GPT's efficiency and minimal training data requirements make it particularly attractive for studies with limited datasets. However, the model's lack of transparency and inability to explicitly articulate the rationale behind its predictions remain critical limitations that researchers must carefully weigh (Feuerriegel et al., 2025). Given these considerations, we recommend that researchers thoughtfully evaluate GPT's role in their analytical pipeline, considering whether to employ it as a standalone tool or integrate it with complementary approaches. Future work should focus on developing methods that combine GPT's powerful inference capabilities with more transparent analytical approaches, potentially offering a more robust framework for emotion analysis in psychological research.

This study is the first to use different NLP approaches to directly compare nomothetic and idiographic models for tracking emotional fluctuations. Yet, findings of this study should be interpreted in light of several limitations. First, capturing the nuanced dynamics of emotional expression typically requires large amounts of data. In everyday interactions, subtle emotional

changes may only become apparent with extensive exposure to an individual's language use—much like knowing someone well enables you to discern small shifts in their mood. Therefore, future studies should strive to collect larger datasets, potentially sourced from daily-life communications, to enhance the ability to track and model these nuances.

Second, the overall predictive accuracy was modest, leaving significant room for improvement. In addition to increasing the quantity of data, future research could benefit from incorporating additional modalities (e.g., vocal features, facial expression, passive sensors) alongside text. Prior studies have shown that combining various modalities can enhance the performance of nomothetic models of emotion (see Gandhi et al., 2023 for a review), this multimodal approach may also improve idiographic models. Moreover, future work should explore whether individuals not only differ in the text features that predict their emotions but also in the types of modalities that best capture their emotional states. Relatedly, our analysis of text features importance across high- and low-performing person-specific models did not reveal a single pattern differentiating the groups. Instead, multiple linguistic factors contributed to model performance, with GPT-derived ratings emerging as robust predictors. One possible explanation is that individual differences, such as affective suppression, affect how emotions are conveyed in text—people with high suppression may mask their emotional states, complicating accurate predictions. Future studies should examine whether these participants require larger datasets to capture subtle nuances in their text, or if alternative modalities better detect changes in their emotional states.

Third, the current study compared idiographic and nomothetic approaches, demonstrating that each has its merits. Future studies could explore hybrid approaches that integrate both individual-specific and group-level information, balancing personalization with statistical power

to enhance predictive accuracy. Finally, the text data in our study consisted of responses to specific questions, which may limit the generalizability of our findings to other types of text. Future research should incorporate text from diverse sources—such as social media, conversational exchanges, and free-form writing—to determine whether these results extend to broader contexts.

In conclusion, our findings highlight the potential of combining NLP approaches to track within-person emotional fluctuations, demonstrating both the strengths and limitations of idiographic and nomothetic models. Nomothetic models effectively capture general trends, this one-size-fits-all approach may work well for some but not for others. Idiographic models offer a more nuanced understanding by identifying person-specific features that capture the unique context of an individual's emotional expression, although they may suffer from limited data. Our results further indicate that GPT shows promise in capturing fluctuations in emotional states, though further tuning is needed to enhance its precision in predicting specific ratings. Future research integrating hybrid modeling strategies—leveraging both group-level patterns and individual differences—could improve predictive accuracy and refine emotion-tracking methods. Expanding this work to incorporate diverse text sources and multimodal data streams may further advance the field. Ultimately, improving the ability to monitor emotions in real-time not only enhances our capacity to study emotional nuances at scale in daily life but also can help develop just-in-time interventions that go beyond identifying moments of distress to also consider the specific emotions and contextual factors in which they arise, enabling more personalized and effective interventions.

References

Aalbers, G., Hendrickson, A. T., Vanden Abeele, M. M., & Keijsers, L. (2023). Smartphone-Tracked Digital Markers of Momentary Subjective Stress in College Students: Idiographic Machine Learning Analysis. *JMIR mHealth and uHealth*, *11*, e37469.

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Akhtar, M. S., Ghosal, D., Ekbal, A., Bhattacharyya, P., & Kurohashi, S. (2019). All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. *IEEE Transactions on Affective Computing*, *13*(1), 285–297.

Arun, R., Suresh, V., Veni Madhavan, C., & Narasimha Murthy, M. (2010). *On finding the natural number of topics with latent dirichlet allocation: Some observations*. 391–402. https://doi.org/10.1007/978-3-642-13657-3_43

Bailen, N. H., Green, L. M., & Thompson, R. J. (2019). Understanding emotion in adolescents: A review of emotional frequency, intensity, instability, and clarity. *Emotion Review*, *11*(1), 63–73.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol.*, *58*(1), 373–403. https://doi.org/10.1146/annurev.psych.58.110405.085709

Bathina, K. C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, *5*(4), 458–466.

Beck, E. D., & Jackson, J. J. (2022). Personalized prediction of behaviors and experiences: An idiographic person–situation test. *Psychological Science*, *33*(10), 1767–1782. https://doi.org/doi.org/10.1177/09567976221093307

Bennik, E. C., Nederhof, E., Ormel, J., & Oldehinkel, A. J. (2014). Anhedonia and depressed mood in adolescence: Course, stability, and reciprocal relation in the TRAILS study. *European Child & Adolescent Psychiatry*, *23*, 579–586.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Carlier, C., Niemeijer, K., Mestdagh, M., Bauwens, M., Vanbrabant, P., Geurts, L., van Waterschoot, T., & Kuppens, P. (2022). In search of state and trait emotion markers in mobile-sensed language: Field study. *JMIR Mental Health*, *9*(2), e31724. https://doi.org/10.2196/31724

Cheung, Y. K., Hsueh, P.-Y. S., Qian, M., Yoon, S., Meli, L., Diaz, K. M., Schwartz, J. E., Kronish, I. M., & Davidson, K. W. (2017). Are nomothetic or ideographic approaches

superior in predicting daily exercise behaviors? *Methods of Information in Medicine*, *56*(06), 452–460.

Davitz, J. R. (2013). *The language of emotion*. Academic Press.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., & Johnson, M. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., & Buffone, A. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398–427. https://doi.org/10.1037/met0000349

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., & Mohammad, S. M. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 1–16. https://doi.org/10.1038/s44159-024-00392-z

Fisher, H., Fatimah, H., Pidvirny, K., Brown, H., Balkind, E., Pastro, B., & Webb, C. A. (n.d.). *Affect dynamics in adolescent depression: Are all equilibria worth returning to?*

Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, *43*(5), 349–358. https://doi.org/10.1037//0003-066x.43.5.349

Funkhouser, C. J., Trivedi, E., Li, L. Y., Helgren, F., Zhang, E., Sritharan, A., Cherner, R. A., Pagliaccio, D., Durham, K., & Kyler, M. (2024). Detecting adolescent depression through passive monitoring of linguistic markers in smartphone communication. *Journal of Child Psychology and Psychiatry*, *65*(7), 932–941. https://doi.org/10.1111/jcpp.13931

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*, 1–30. https://doi.org/10.18637/jss.v040.i13

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hollenstein, T., & Lanteigne, D. M. (2018). Emotion regulation dynamics in adolescence. In *Emotion regulation* (pp. 158–176). Routledge.

Houben, M., & Kuppens, P. (2020). Emotion dynamics and the association with depressive features and borderline personality disorder traits: Unique, specific, and prospective relationships. *Clinical Psychological Science*, *8*(2), 226–239.

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930.

Hutto, C., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. https://doi.org/10.1017/langcog.2014.30

Jablonka, E., Ginsburg, S., & Dor, D. (2012). The co-evolution of language and emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2152–2159.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263–286. https://doi.org/10.2307/20445377

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26.

Lindquist, K. A. (2017). The role of language in emotion: Existing evidence and future directions. *Current Opinion in Psychology*, *17*, 135–139.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Manning, C. (1999). *Foundations of statistical natural language processing*. The MIT Press.

Murray, L., Israel, E. S., Balkind, E. G., Pastro, B., Lovell-Smith, N., Lukas, S. E., Forbes, E. E., Pizzagalli, D. A., & Webb, C. A. (2023). Multi-modal assessment of reward functioning in adolescent anhedonia. *Psychological Medicine*, *53*(10), 4424–4433. https://doi.org/10.1017/ S0033291722001222

Neuendorf, K. A. (2017). *The content analysis guidebook*. sage.

Pennebaker, J. W. (2001). *Linguistic inquiry and word count: LIWC 2001*. Erlbaum.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.

Rozet, A., Kronish, I. M., Schwartz, J. E., & Davidson, K. W. (2019). Using machine learning to derive just-in-time and personalized predictors of stress: Observational study bridging the gap between nomothetic and ideographic approaches. *Journal of Medical Internet Research*, *21*(4), e12910.

Schoevers, R., Van Borkulo, C., Lamers, F., Servaas, M., Bastiaansen, J., Beekman, A., Van Hemert, A., Smit, J., Penninx, B., & Riese, H. (2021). Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychological Medicine*, *51*(11), 1906–1915. https://doi.org/10.1017/S0033291720000689

Sivakumar, S., Videla, L. S., Kumar, T. R., Nagaraj, J., Itnal, S., & Haritha, D. (2020). *Review on word2vec word embedding neural net*. 282–290.

Soyster, P. D., Ashlock, L., & Fisher, A. J. (2022). Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: A demonstration of parallel utility. *Psychology of Addictive Behaviors*, *36*(3), 296–306.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364. https://doi.org/10.1037/pspp0000244

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., Edwards, T. S., Pennebaker, J. W., & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-

language-task research synthesis. *Journal of Personality and Social Psychology*, *116*(5), 817–834. https://doi.org/10.1037/pspp0000187

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods*, 1–14.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

van Loon, A. (2022). Three families of automated text analysis. *Social Science Research*, *108*, 102798. https://doi.org/10.1016/j.ssresearch.2022.102798

Fig. 1

Estimation of the most preferable number of topics for LDA model

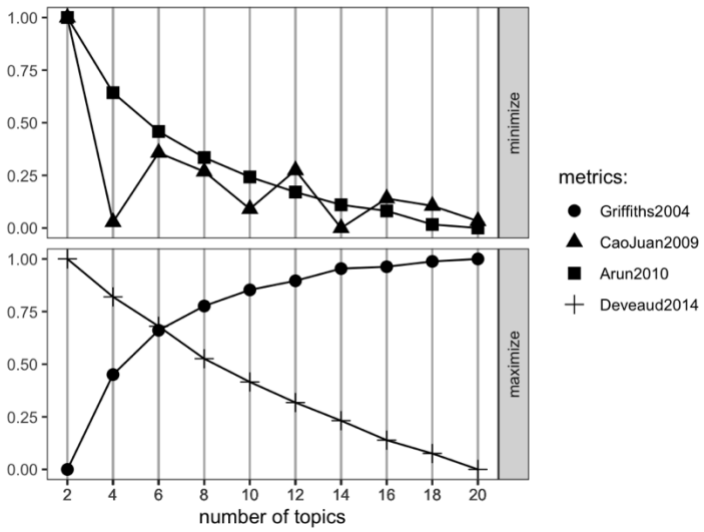


Table 1

The 15 most frequent words within the six topics extracted using LDA

Topic number	Topic interpretation	
1	**Social & Evening Activities:** focused on social interactions and nighttime activities with friends	friend, talk, sleep, last, night, last_night, see, time, morning, hang, wake, late, talk_friend, boyfriend, hang_friend,
2	**Academic Life:** centered on school-related activities and academic responsibilities	homework, school, think, class, finish, test, now, take, tomorrow, math, studi, right, finals, essay, stress
3	**Home activities**: capturing leisure activities, particularly around media consumption and family time	watch, nothing, eat, dinner, play, game, show, family, movie, tv, ate, video, eat_dinner, watch_tv, favorite, watch_movie
4	**Family Interactions:** reflecting family relationships and dynamics	mom, think, fun, sister, new, dad, brother, room, read, make, made, book, something, fight, clean
5	**Daily Activities:** representing routine daily activities like meals	go, went, home, walk, outside, lunch, food, drive, music, back, listen, shop, car, hurt, around, dog,
6	**Personal States & Obligations:** describing emotional states, needs, and work-related responsibilities	get, work, feel, want, today, day, need, good, think, just, felt, like, done, tired, sick,

Table 2

Comparison of Idiographic, Nomothetic, and Nomothetic–Idiographic Model Performance for

		Nomothetic			Nomothetic-idiographic			Idiographic		
		R^2	**R**	**RMSE**	**R^2**	**R^2 range**	**RMSE**	**R^2**	**R^2 range**	**RMSE**
Negative Affect	Random Forest	.38	.62	0.55	.10 (.10)	.00; .41 Sig.n= 53/97	0.52 (0.22)	.10 (.12)	.00; .47 Sig.n= 42/97	0.47 (0.20)
	Elastic net	.17	.41	0.64	.11 (.09)	.00; .39 Sig.n= 56/97	0.61 (0.27)	.10 (.12)	.00; .53 Sig.n= 47/97	0.48 (0.20)
Sad	Random Forest	.33	.58	0.75	.10 (.12)	.00; 0.53 Sig.n= 48/96	0.73 (0.27)	.10 (.13)	.00; .46 Sig.n= 38/94	0.68 (0.22)
	Elastic net	.14	.37	0.85	.09 (.09)	.00; .42 Sig.n= 51/96	0.82 (0.32)	.10 (.11)	.00; .51 Sig.n= 46/94	0.69 (0.23)
Angry	Random Forest	.20	.45	.69	.09 (.12)	.00; .69 Sig.n= 37/90	0.63 (0.33)	.06 (.07)	.00;.33 Sig.n= 27/85	0.64 (0.29)
	Elastic net	.13	.36	.72	.07 (.07)	.00; .45 Sig.n= 49/90	0.88 (0.32)	.07 (.08)	.00; .43 Sig.n= 39/85	0.66 (0.31)
Nervous	Random Forest	.24	.49	.85	.06 (.08)	.00; .44 Sig.n= 33/95	0.81 (0.29)	.08 (.12)	.00; 0.71 Sig.n= 33/94	0.75 (0.27)
	Elastic net	.11	.33	.92	.07 (.07)	.00; .35 Sig.n= 42/95	0.88 (0.33)	.09 (.12)	.00; 0.84 Sig.n= 43/94	0.76 (0.29)

Negative Affect, Sadness, Anger, and Nervousness Using Random Forest and Elastic Net

Figure 2

Feature importance of nomothetic RF and ENR models

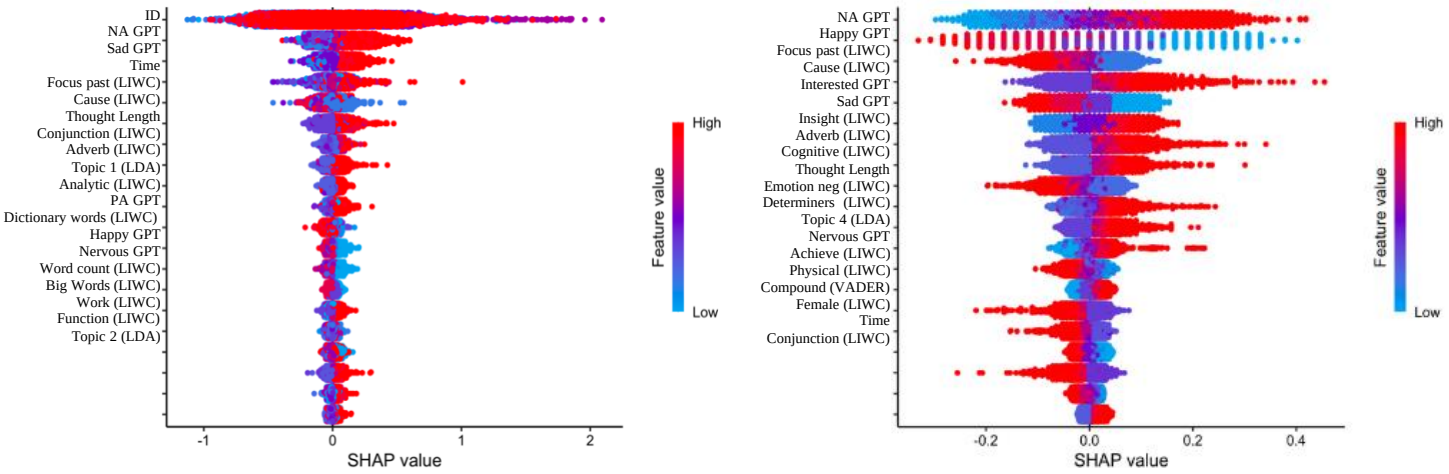


Figure 3

Examples of Person-Specific (Idiographic) Predictions of Negative Affect for High- and Low-Performance Models

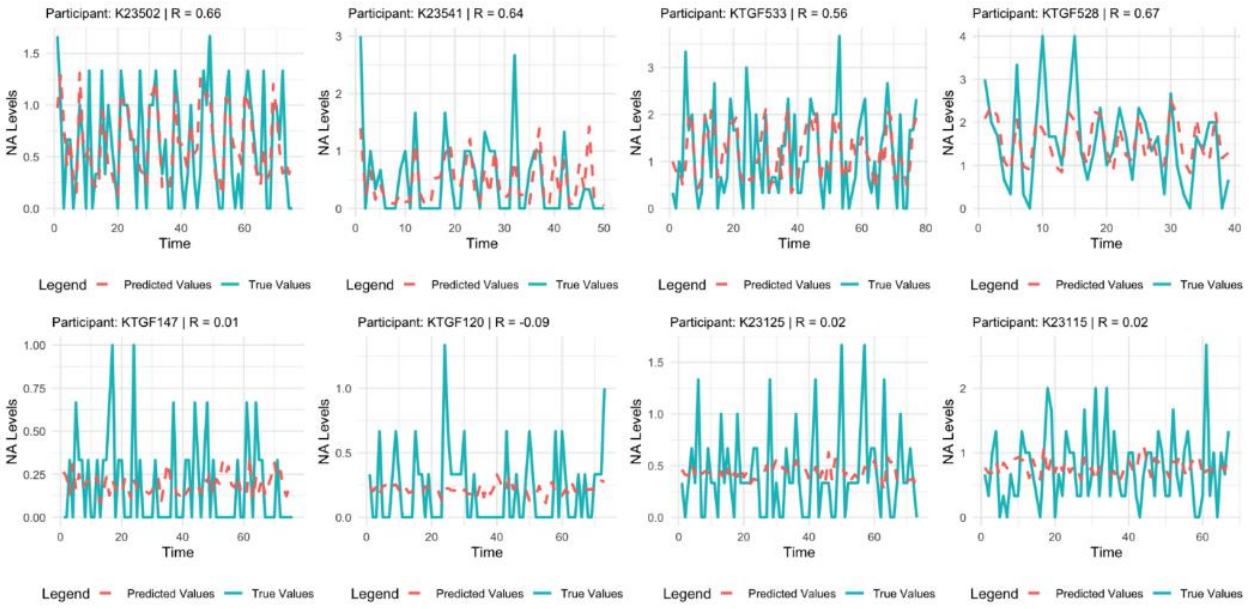


Figure 4

Person-Specific (Idiographic) Predictions of Negative Affect (a) Random Forest (b) Elastic Net

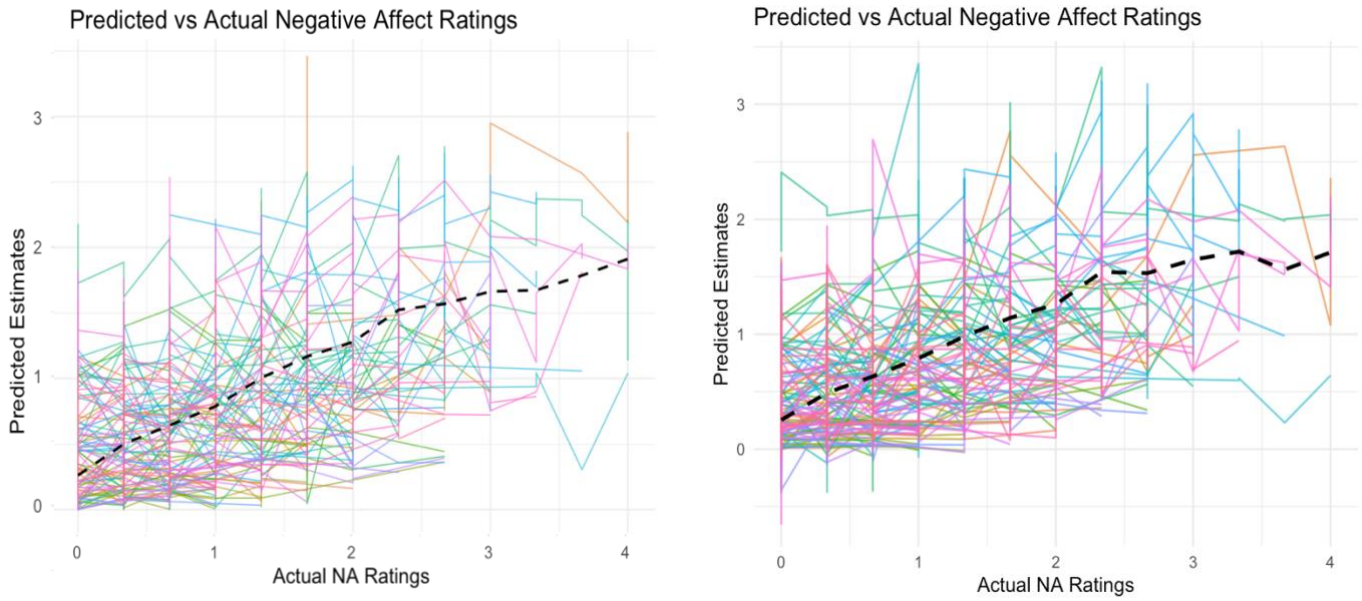


Figure 5

Top 10 Predictive Features Across Four Example Person-Specific Best-Performing Random Forest Models for Negative Affect Prediction

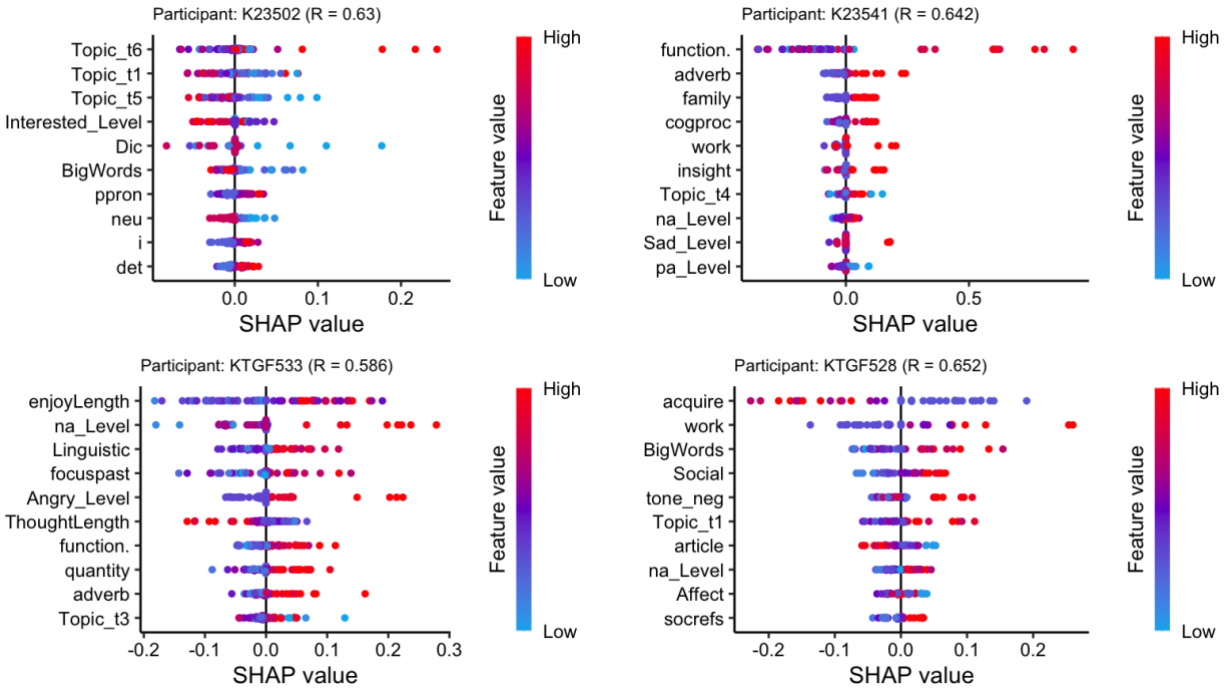


Figure 6

Group differences in feature importance predicting negative affect: random forest (RF) and elastic regularization (ENR)

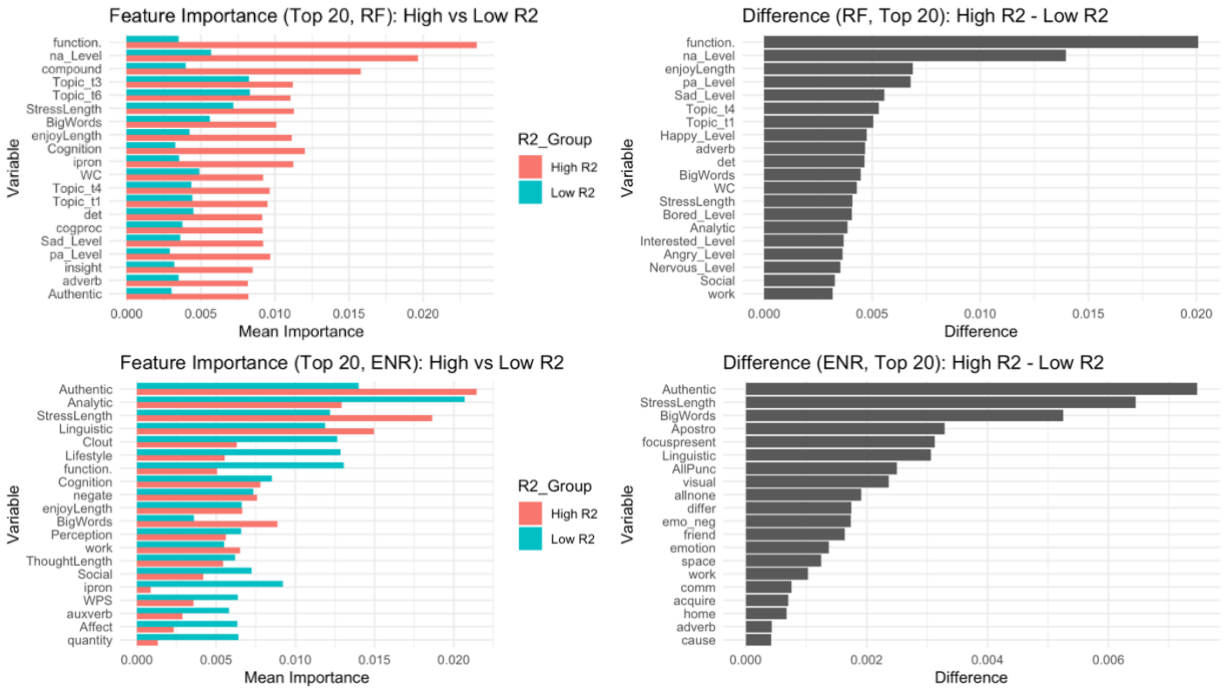


Figure 6

Group differences in feature importance predicting negative affect: random forest (RF) and elastic regularization (ENR)

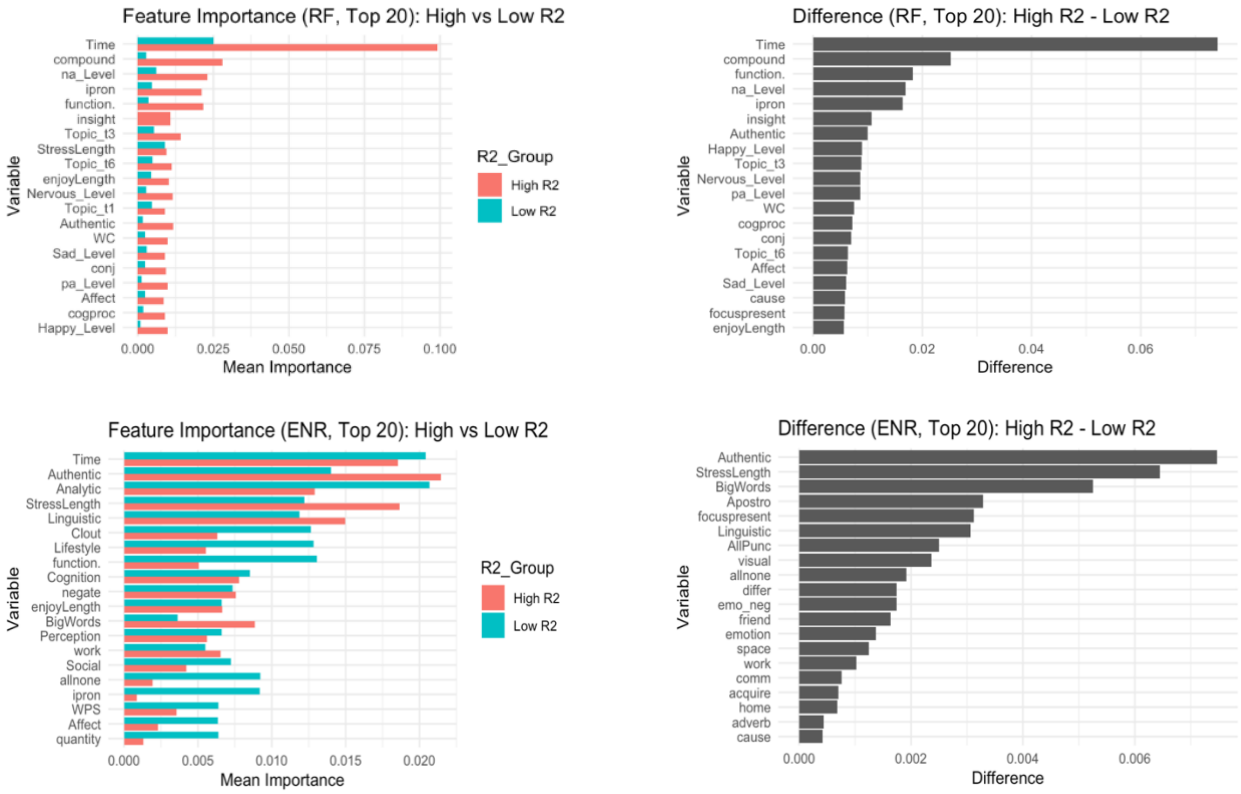


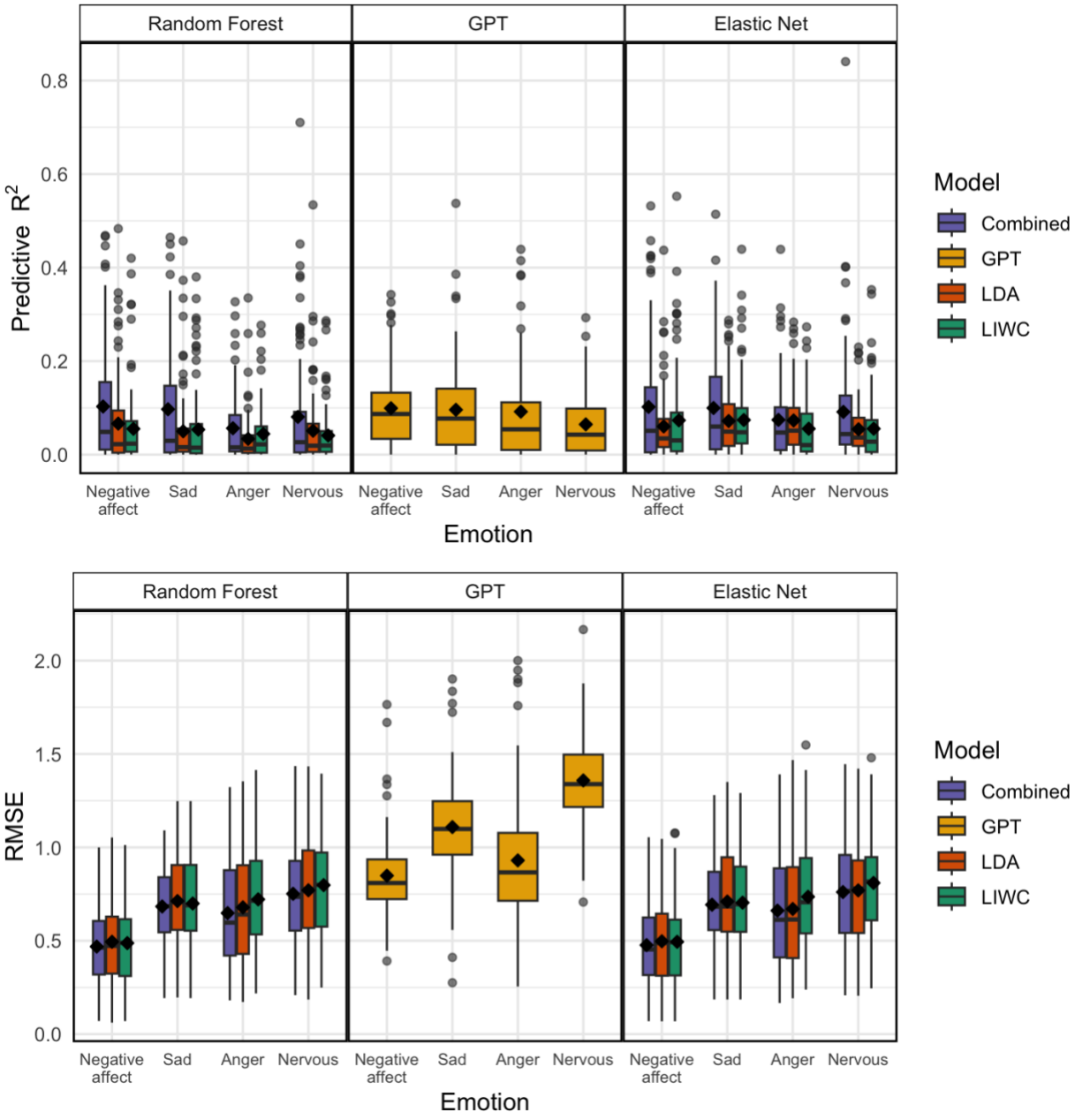
Table 3

Comparison of Idiographic Model Performance for Negative Affect, Sadness, Anger, and Nervousness Using LIWC+VADER, LDA, and GPT

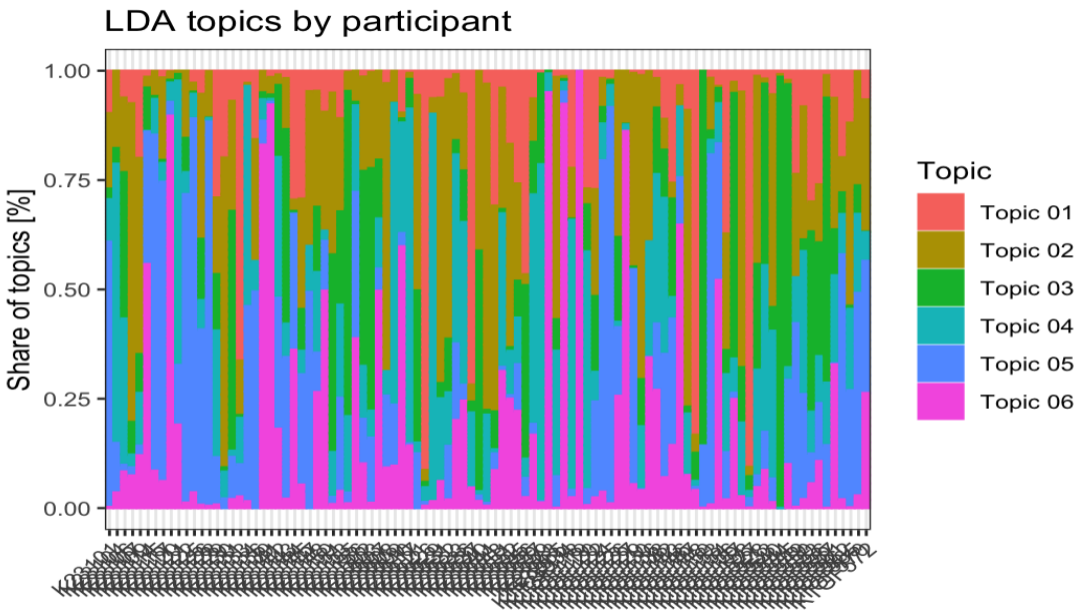
		LIWC+VADER			LDA			GPT		
		R^2	**R^2 range**	**RMSE**	**R^2**	**R^2 range**	**RMSE**	**R^2**	**R^2 range**	**RMSE**
Negative Affect	Random Forest	.05 (.08)	.00;.42 Sig.n= 31/97	0.49 (0.21)	.07 (.09)	.00;.48 Sig.n= 36/97	0.49 (0.21)	.10 (.08)	.00;.34 Sig.n= 55/97	0.85 (.20)
	Elastic net	.07 (.10)	.00;.55 Sig.n= 36/97	0.49 (0.22)	.06 (.07)	.00;.44 Sig.n= 35/97	0.50 (0.22)			
Sad	Random Forest	.05 (.08)	.00;.38 Sig.n= 24/95	0.70 (0.25)	.05 (.08)	.00;.46 Sig.n= 25/95	0.71 (0.26)	.10 (.10)	.00; .54 Sig.n= 55/96	1.11 (0.26)
	Elastic net	.07 (.08)	.00;.44 Sig.n= 42/95	0.70 (0.26)	.07 (.07)	.00;.29 Sig.n= 45/95	0.71 (0.27)			
Angry	Random Forest	.04 (.06)	.00;.28 Sig.n= 15/70	0.72 (0.29)	.03 (.05)	.00;.58 Sig.n= 12/85	0.68 (0.32)	.09 (.10)	.00;.44 Sig.n= 45/90	0.93 (0.34)
	Elastic net	.05 (.07)	.00;.27 Sig.n= 26/70	0.73 (0.31)	.07 (.07)	.00;.28 Sig.n= 43/84	0.67 (0.32)			
Nervous	Random Forest	0.04 (0.06)	.00;.29 Sig.n= 19/86	0.80 (0.26)	.05 (.08)	.00;.53 Sig.n= 29/94	0.77 (0.28)	.06 (.06)	.00;.29 Sig.n= 40/95	1.36 (0.22)
	Elastic net	.05 (.07)	.00;.35 Sig.n= 25/86	.81 (28)	.05 (.05)	00;.23 Sig.n= 36/94	0.76 (0.28)			

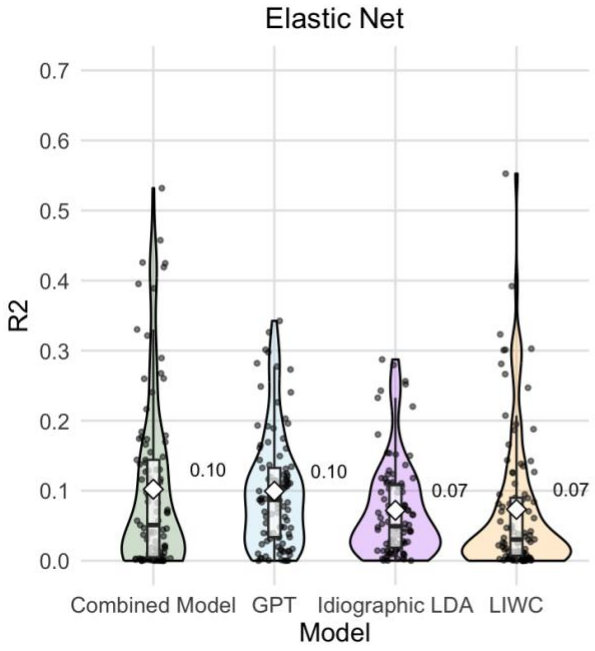
Figure 7

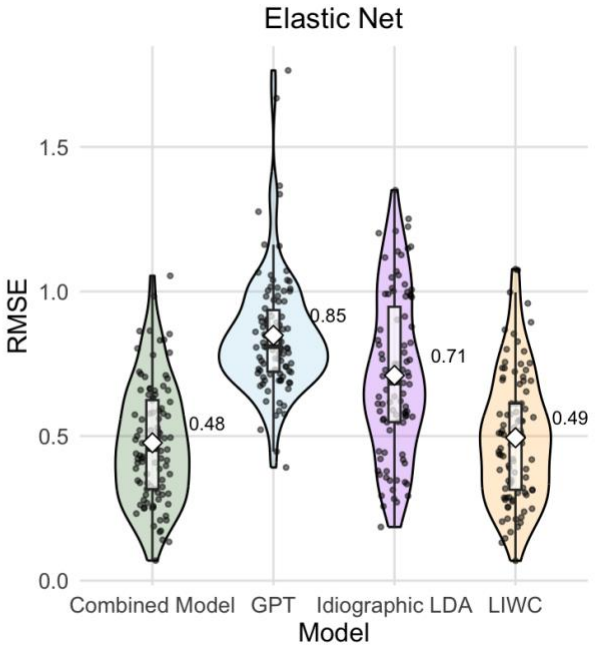
Predictive Performance of Idiographic Model Across NLP Approaches

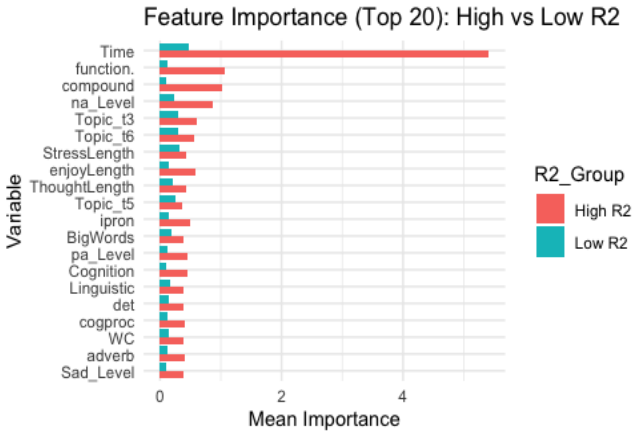


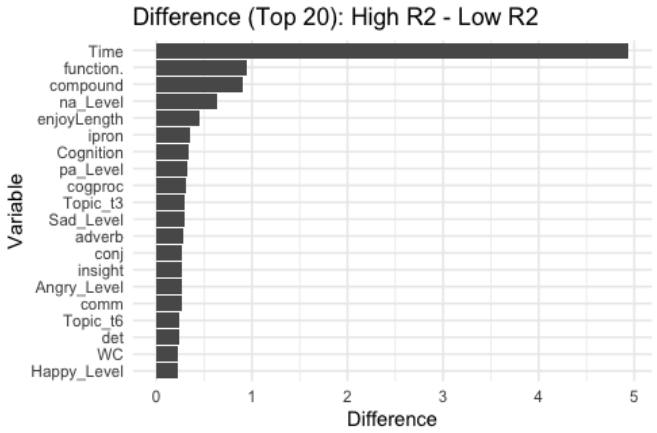
Supplement

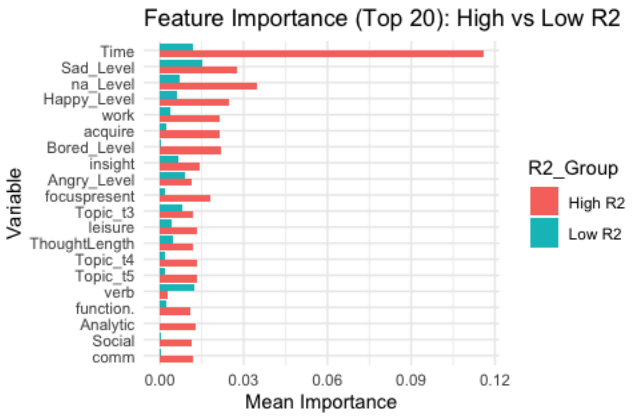

LDA topics by participant
Share of topics [%]
1.00
0.75
0.50
0.25
0.00
Topic
Topic 01
Topic 02
Topic 03
Topic 04
Topic 05
Topic 06

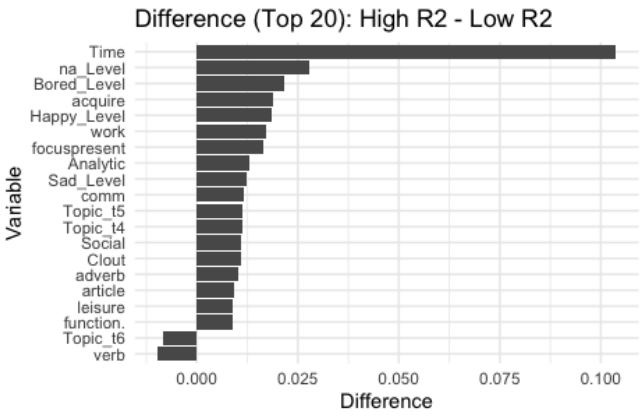

Elastic Net
R2
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.10
0.10
0.07
0.07
Combined Model
GPT
Idiographic LDA
LIWC
Model


Elastic Net
RMSE
1.5
1.0
0.5
0.0
0.48
0.85
0.71
0.49
Combined Model
GPT
Idiographic LDA
LIWC
Model


Feature Importance (Top 20): High vs Low R2
Time
function.
compound
na_Level
Topic_t3
Topic_t6
StressLength
enjoyLength
ThoughtLength
Topic_t5
ipron
BigWords
pa_Level
Cognition
Linguistic
det
cogproc
WC
adverb
Sad_Level
Variable
0
2
4
Mean Importance
R2_Group
High R2
Low R2


Difference (Top 20): High R2 - Low R2
Time
function.
compound
na_Level
enjoyLength
ipron
Cognition
pa_Level
cogproc
Topic_t3
Sad_Level
adverb
conj
insight
Angry_Level
comm
Topic_t6
det
WC
Happy_Level
Variable
0
1
2
3
4
5
Difference


Feature Importance (Top 20): High vs Low R2
Time
Sad_Level
na_Level
Happy_Level
work
acquire
Bored_Level
insight
Angry_Level
focuspresent
Topic_t3
leisure
ThoughtLength
Topic_t4
Topic_t5
verb
function.
Analytic
Social
comm
Variable
0.00
0.03
0.06
0.09
0.12
Mean Importance
R2_Group
High R2
Low R2


Difference (Top 20): High R2 - Low R2
Time
na_Level
Bored_Level
acquire
Happy_Level
work
focuspresent
Analytic
Sad_Level
comm
Topic_t5
Topic_t4
Social
Clout
adverb
article
leisure
function.
Topic_t6
verb
Variable
0.000
0.025
0.050
0.075
0.100
Difference

BERT (Bidirectional Encoder Representations from Transformers), in contrast, is a bidirectional model that takes into account both the left and right context of a sentence, making it more suited for tasks like sentiment analysis and natural language understanding (NLU). BERT's architecture enables it to understand the full context of a sentence by analyzing words from all directions simultaneously, which is particularly useful for tasks where understanding the meaning of a phrase within its broader context is essential. Both GPT-4 and BERT are pre-trained on large text datasets. However, while GPT-4 benefits from a significantly larger dataset, BERT's bidirectional architecture makes it better suited for tasks requiring a deep understanding of text context, such as sentiment analysis. In this study, BERT model, specifically trained for emotion classification ("bhadresh-savani/bert-base-go-emotion"), is used to tokenize the concatenated text inputs and generate emotion predictions.