THE UNIVERSITY OF CHICAGO

A Multidimensional Analysis of Language Model Sentiment and Self-Reported Emotional States

By

Xiaochen Ding

March 2024

A paper submitted in partial fulfillment of the requirements for
the Master of Arts degree in the Master of Arts in
Computational Social Science

Faculty Advisor: Dr.Monica D. Rosenberg
Preceptor: Dr.Ali Sanaei and Dr.Sanja Miklin

Abstract

This thesis presents a multidimensional analysis of the relationship between sentiment as evaluated by a large language model (RoBERTa) and self-reported emotional states during mind-wandering episodes. The study utilized data from the Audiovisual Attention (AVA) study, which captured the spontaneous speech of participants. The RoBERTa model was employed to quantify the emotional valence of this speech, categorizing it into positive, negative, or neutral sentiments. In parallel, self-reported measures were taken, allowing for a comparison between computational sentiment analysis and participant self-assessment. Findings demonstrate that while both approaches provide valuable insights, they offer complementary information regarding the emotional valence of thoughts. The research highlights the prevalence of neutral sentiments in mind-wandering and the model's accuracy in classifying emotional tones, suggesting the potential of natural language processing tools in psychological research.

Keywords: Sentiment Analysis; Mind-Wandering; Computational Linguistics

1 Introduction

1.1 Mind Wandering

In our daily lives, we do not spend all the time focusing on specific tasks like solving a math problem. Sometimes our mind wanders freely between topics as varied as unfinished laundry and world peace. For example, when focusing on a primary task such as doing a math problem, our attention may shift away from this task toward more internal thoughts like remembering a classmate falling asleep in the previous math class or toward external stimuli not related to this problem, which is considered as mind wandering. This is related to spontaneous and self-generated thoughts, daydreams, or fantasies that occupy our minds during waking hours or when we are engaging in tasks that do not require full cognitive attention (Smallwood and Schooler, 2015). This mind process is not merely a lapse in attention, but a very important cognitive activity in human life, it reflects the brain's capacity to disengage from the external environment and explore internal thought processes (Christoff et al., 2016).

Researchers have demonstrated that mind wandering can occupy up to 50% of our waking hours varying across individuals (Killingsworth and Gilbert, 2010). It frequently occurs during monotonous or routine activities, for example, driving, reading, or attending lectures, our cognitive demand may be low during these tasks. Although it is common in our

daily lives, that does not mean it doesn't bring any consequences, for example, it may have emotional costs and may lead to unhappiness (Killingsworth and Gilbert, 2010). However, the narrative surrounding mind wandering is not solely negative. Scholars have begun to uncover the adaptive functions of mind wandering, suggesting that it plays a crucial role in future planning, creative problem-solving, and the consolidation of memories (Baird et al., 2012; Mrazek et al., 2013).

The dual nature of mind-wandering highlights the complexity of this cognitive process. Therefore it raises great interest and questions mind wandering. Given the dual impacts, accurately tracking and understanding mind wandering becomes crucial. This led to the exploration of two methodologies in cognitive research: Self-reports, which rely on participants' subjective interpretations and recollections of their mental state, In the context of this study, participants were asked to self-report (rated on a scale) on a series of nine questions immediately following mind wandering. And real-time verbal explanations, which involve participants articulating their thoughts as they occur during mind wandering. It potentially offers more immediate and objective insights into the wandering mind. For this thesis, I will focus on the comparative effectiveness and complementarity of these methods to address the question of how these two ways of tracking mind wandering give us overlapping or complementary information about thought content. I believe this inquiry is essential not only for advancing theoretical understanding but also for improving practical approaches to cognitive assessments and interventions.

1.2 Thought Valence: The Emotional Spectrum of Mind Wandering

To better understand the mind wandering's role in our daily lives, it is essential to explore the nature of the thoughts during different conditions. Specifically, the emotional valence of these thoughts, with the emotional value we assign to them, will significantly affect our mood, behavior, and cognitive function (Fredrickson, 2001). It can be generally categorized into positive, negative, and neutral states. This concept of 'thought valence' serves as a bridge between mind wandering and its psychological impacts. Positive thoughts, characterized by hope, optimism, or pleasure, can uplift and motivate us, as the study indicates that positive thoughts and emotions can also increase dopamine levels, and elevate our mood, which illustrates the beneficial aspects of mind wandering (Fredrickson, 2001). Conversely, negative thoughts, filled with worry, pessimism, or distress, might not only reflect our current emotional state but also potentially exacerbate negative moods, creating a reinforcing cycle of negativity.

The relationship between the valence of our thoughts—whether positive or negative—and our psychological resilience and susceptibility to mental health disorders is complex and potentially bidirectional. Instead of a one-way influence, it may be more accurate to describe

thought valence as being closely related to these psychological states. For instance, while a pattern of negatively valenced mind wandering is associated with a range of adverse outcomes, such as an increased risk of depression, anxiety, and other mood disorders (Marchetti et al., 2016), this relationship may also work in reverse. Individuals with lower psychological resilience or those more susceptible to mental health disorders might be more prone to experience negatively valenced mind wandering. Conversely, positive thought patterns, including those occurring during mind wandering, are correlated with enhanced well-being, offering a form of protection against stress (Marchetti et al., 2016; Erez and Isen, 2002). This bidirectional relationship underscores the importance of accurately assessing the emotional valence of mind wandering. In this study, I address a question: How similar are the measures of thought valence obtained from self-reports using a rating scale to those derived from analyses of free speech using sentiment analysis? By exploring this question, I aim to determine whether these methodologies provide overlapping or distinct insights into the emotional dimensions of mind wandering.

Moreover, the dynamic and often fleeting nature of mind wandering raises a unique challenge to its study: capturing and accessing the emotional valence of thoughts as they naturally occur. The more traditional self-report scale methods, such as inquiries, provide invaluable insights into participant's internal dynamics. But it would be subjective and may be oversimplified. In contrast, real-time verbal explanations of mind wandering offer a direct, in-the-moment account of thought valence, potentially providing a more objective and immediate depiction of the emotional landscape. However, how to accurately capture and measure the emotional valence of these expressions becomes a challenge.

By comparing self-report scales and expression tracking through verbal explanations, the research aims to discern the nuances of thought valence as experienced and reported by individuals, thereby offering a more comprehensive understanding of how mind wandering influences and is influenced by emotional states.

1.3 Measuring Thought Valence

Accurately measuring the valence of thoughts during mind wandering is crucial for understanding its psychological effects. Thought valence measurement has traditionally relied on self-report techniques. These tools are invaluable for capturing the subjective experience of mind wandering, as they allow individuals to report their thoughts and feelings directly. However, these methods come with inherent limitations. They depend significantly on their current emotional state, which can affect the accuracy of the reports (Kahneman et al., 2004).

In response to these limitations, advancements in computational linguistics have introduced sentiment analysis as a complementary approach. This method utilizes natural lan-

guage processing (NLP) to evaluate and categorize the emotional tone of verbal expressions collected in real-time during mind-wandering. Sentiment analysis employs sophisticated algorithms trained on large datasets to objectively assess whether spoken or written expressions reflect positive, negative, or neutral sentiments (Pennebaker et al., 2015). While this approach offers the potential to be more objective, it also faces challenges. Computational methods may not fully capture the nuanced emotional states that a person experiences, as they primarily analyze the emotional content of language, which may not always align perfectly with the speaker's internal emotional state.

By comparing the outcomes of self-reported assessments and computational sentiment analysis, this research aims to explore whether these methods provide overlapping or complementary insights. Such a comparison is crucial not only for validating each method's effectiveness but also for identifying potential synergies that could enhance our understanding of mind wandering. This approach aligns with my research question about how these two ways of tracking mind wandering, having people provide explicit ratings or verbally explaining their thoughts, give us overlapping or complementary information about thought content. This inquiry will help determine the extent to which these methods can be integrated or used selectively to study the emotional dimensions of mind wandering more effectively.

1.4 Study Overview

My study focuses on the exploration of the methodologies used to assess the emotional valence of mind-wandering, specifically comparing the effectiveness of self-reported rating-scale responses and verbal explanations of mind-wandering content. The juxtaposition of these methodologies seeks to uncover whether they provide overlapping or complementary insights into the emotional landscapes of mind-wandering episodes. I use a large language model, Twitter-based RoBERTa, for sentiment analysis, aiming to quantify the probability of positively and negatively valenced thoughts captured during episodes of mind wandering.

Drawing upon a large dataset from the Audiovisual Attention (AVA) study collected by the Rosenberg Lab, this research analyzes the transcribed spontaneous speech of participants during tasks that induced mind wandering. The sentiment of these transcriptions is quantitatively assessed using the Twitter-based RoBERTa model, which categorizes the emotional tone of text into positive, negative, or neutral sentiments. This sentiment analysis is complemented by self-reported measures of thought valence, allowing for a multidimensional comparison and correlation of self-report thought valence and computational language model analysis.

2 Data and Methods

The current study derived from the Rosenberg Lab's Audiovisual Attention (AVA) study and focuses on the relationship between participants' self-reported mind-wandering ratings and the large language model-generated sentiment probability based on participants' speech of ongoing thoughts. To achieve this goal, I utilize the recording of participants' articulated ongoing thoughts during the task designed in the AVA study. These recordings are transcribed into the text as a primary data source for examining the sentiment within spontaneous speech. Then I aim to use the pre-trained sentiment analysis model called Twitter-based roBERTa. I conduct sentiment analysis on these transcriptions. This approach enables us to generate probabilities indicating whether the content of the speech is predominantly negative or positive in sentiment, thereby facilitating a nuanced analysis of the relationship between self-reported mind-wandering dimensions and the emotional tone of speech.

2.1 Data

The AVA study aims to introduce an annotated resting-state fMRI paradigm to gain a deeper understanding of the connections between free-form cognition and brain activity (not analyzed here). During the study, there were two fMRI sessions; participants (N = 60) completed four 10-minute runs (32 rest periods total) of an annotated rest task (see below for details).

2.2 Participants

The current study includes data from 60 participants. They were recruited from the University of Chicago and the surrounding community (34 female and 26 male. Mean age of 22.86 years) All participants can speak fluent English and reported normal vision or corrected-to-normal vision. Study procedures were approved by the relevant Institutional Review Board of the University of Chicago and participants were compensated for their participation. Of the 60 total participants, three did not return for the second session and the testing computer broke during the second session for one participant.

2.3 Experimental procedure

In the AVA study, participants completed a series of tasks in an fMRI environment, coupled with post-scan evaluations. In the scanner, participants first watched a movie or listened to a podcast. They then performed the annotated rest task (Rest1), where they were given intervals of rest and prompted to articulate their spontaneous thoughts (see below for de-

tails). The subsequent in-scanner experience included an audiovisual task that involved engaging with combined audio and visual stimuli. Participants watched one of two movie clips' selected in random order for each session. This was followed by a second rest phase (Rest2), which mirrored the first, maintaining the same sequence of rest and thought articulation. Participants also completed one run per session of an audiovisual Continuous Performance Task (CPT)

2.4 Annotated rest task

In this task, participants had eight trials per run, and for each run, they had 30 seconds of rest followed by 10 seconds to describe ongoing thoughts verbally and then 35 seconds to answer a series of nine questions about their thoughts (Ho et al., 2020). The subsequent rest session mirrored the first, repeating the same sequence of rest thoughts.

The 9 questions (see Table 1) about their thoughts on the self-report rating task are shown in the table below. In my study, I focused on the self-reported score for Question 6: "My thoughts were... very negative/very positive" because I aim to explore the relationship between large language models, sentiment analysis, and self-reported emotional states.

Question	**Description**
Question 1	How aware/alert do you feel
Question 2	My thoughts were related to the external environment around me.
Question 3	My thoughts were about the future.
Question 4	My thoughts were about the past.
Question 5	My thoughts were about myself/about others
Question 6	My thoughts were very negative/very positive
Question 7	My thoughts were in the form of images.
Question 8	My thoughts were in the form of words.
Question 9	My thoughts were detailed and specific.

Table 1: Self-reported questions on thought content and character.

2.5 Self-Report Scale

Participants were asked to self-report their thoughts using a 9-point Likert scale ranging from 1 (not at all) to 9 (very much) across a series of nine questions immediately following mind wandering episodes. These questions were designed to assess various dimensions of thought content and emotional valence, including alertness, the relation of thoughts to the external environment, temporal focus (past, present, future), and emotional tone (very negative to very positive). For example, for the question "My thoughts were very negative/very positive," a score of 1 indicated very negative thoughts and a score of 9 indicated very pos-

itive thoughts. This scale allowed for a detailed capture of the participants' subjective mental states at different moments during the study.

2.6 Methods

2.6.1 Transcription

The description of ongoing thought is the core corpus for my sentiment analysis, which could serve as an indicator of the participants' present emotional state. Since the initial description is in audio format, two researchers assisted in transcribing the audio into text format. A third research assistant reviewed and compared these transcripts to identify the discrepancies and modify them to get the final version of a more refined and precise transcription. For the data cleaning phase, I systematically removed all transcripts marked with "[inaudible]" and "[silent]." There are several transcripts, including an inaudible part inside the body of the sentence (n=187, 9.904%). I maintained most of those sentences in my corpus because they still provide valuable information to be analyzed. 287 transcripts of 1,888 total (range = 16-32 transcripts per participant, mean = 31.47) are being excluded due to incomplete sentences and insufficient information. 1601 transcripts were included in the final analysis.

2.6.2 Sentiment analysis

To examine the relationship between participants' self-reported mental states and the sentiment derived from their speech on ongoing thoughts, I employ a pre-trained sentiment analysis model called "Twitter-roBERTa-base for Sentiment Analysis" (Barbieri et al., 2020). This model is trained with about 124M tweets from January 2018 to December 2021. The Twitter-roBERTa-base model is a derivative of the original RoBERTa-base model, which can accurately determine the sentiment of tweets, categorizing them into positive, negative, or neutral sentiments, further trained on Twitter data until 2022. This training approach follows Barbieri et al. (2020), utilizing the same settings for early stopping and learning rate adjustments, ensuring the model is well-adapted to the nuances of Twitter data, including the informal language and abbreviations commonly found on the platform. The reason why I selected this model is because of its proficiency in understanding and classifying the sentiment of text data, particularly the format of the informal and concise nature of language commonly found on social media platforms like Twitter, evidence can be found in a comprehensive evaluation of language models. The model was assessed using the TweetEval framework, which involves multiple tweet classification tasks such as sentiment analysis, hate speech detection, and irony detection. These tasks are vital for proving the model's ability to accurately interpret and classify Twitter data. The model was also compared

against other systems like SVM, FastText, and BERTweet, where it shows competitive or superior performance in various tasks. (Loureiro et al., 2022) This comparative evaluation highlights its robustness across different linguistic challenges found in social media data. Thereby demonstrating its effectiveness in processing the informal and concise language prevalent on social media. By using this state-of-the-art tool, I aim to generate reliable probabilities of the sentence transcript being negative, neutral, and positive. Thereby, I can further analyze the relationship between participants' self-reported states and the sentiments expressed in their spontaneous speech.

2.6.3 Clustering Analysis

To dissect and understand the clusters of transcripts with different probability scores on the 3 dimensions,' positive',' neutral', and 'negative'. I implemented the K-Means clustering algorithm, a method for assigning a dataset into distinct, non-overlapping subsets or clusters.

Determining the optimal number of clusters is critical for meaningful segmentation. Therefore, I employed the Elbow Method, a useful method for determining the number of clusters in a dataset. (Jain, 2010) The method involves plotting the explained variance as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. In this context, the sum of squared distances from each point to the centroid of its assigned cluster was plotted against a range of potential cluster counts. These clusters of transcripts may provide a classification for subsequent analyses and interpretations regarding the relationship between sentiment expression and self-reported emotional states.

2.6.4 Validation of sentiment scores and categories with human ratings

To test the reliability and accuracy of the sentiment analysis conducted through computational methods, it is imperative to validate these automated scores with human assessments. This validation process checks the congruence between the sentiment identified by the computational tools and the subjective human interpretation of the same text.

One hundred transcriptions derived from participants' verbal explanations were selected for validation. This subset consisted of randomly chosen segments to avoid any selection bias, ensuring that the samples accurately reflect the range of sentiments expressed across the study. And I manually classify the emotional tone of each selected transcription as positive, negative, or neutral. The human-assigned sentiment scores were then compared with the scores generated by the sentiment analysis model. This comparison was quantified using statistical measures and a confusion matrix, which provide insights into the model's performance against the human standard.

2.6.5 Principal Component Analysis

To interrogate the multidimensional nature of self-reported thought scores, Principal Component Analysis (PCA) was employed. PCA can help in filtering out noise from the data by focusing on the principal components that capture the most significant variance. Also, through the decomposition of the self-reported scores into principal components, PCA can reveal hidden patterns that might not be immediately apparent in the data. By concentrating on the principal components that hold the most variance, PCA aids in sifting through the noise to unveil the latent structure of the data.

The first step is preprocessing, standardizing the self-reported scores to ensure comparability across different measurements. Then apply PCA to this standardized data. This decision was driven by the goal of capturing a substantive amount of variance within the dataset without overcomplicating the model. The output of the PCA was a new data frame comprising the principal components, labeled Thought_PC1, Thought_PC2, Thought_PC3, and Thought_PC4. Additionally, the PCA loadings are also interpreted and analyzed to explore the pattern behind the sentiment components, which is useful for detecting the dominant and significant sentiment in this dataset.

In the interpretation phase, the loadings of the self-reported variables on the PCA components were analyzed. This involved determining which aspects of the thought scores—ranging from alertness to the emotional valence of thoughts—were most influential in each component. This analytical step is essential as it offers insight into which self-reported measures are most significant in terms of the variance they contribute to the participants' cognitive and emotional profiles.

2.6.6 Comparing thought ratings between sentiments

Upon analyzing the sentiment scores for verbal expressions, which were generated by a computational model, my next plan is to compare these two variables. This integration is key to unraveling potential patterns and relationships between these distinct yet interrelated measures of mind wandering.

Within each defined sentiment cluster, I computed the average thought rating for each question and compared scores across clusters. These averages yield mean scores that encapsulate the collective emotional valence ascribed to each sentiment, providing a quantifiable reflection of the subjective mental states expressed during the mind-wandering episodes.

By using this methodological framework, I will start the first step toward a more nuanced comprehension of the interplay between varying approaches to tracking mind wandering.

2.6.7 Comparing sentiment and self-report ratings and PCs

To elucidate the relationship between participants' self-reported mental states and the probability of positive, negative, and neutral sentiments derived from their current thoughts and speech, I conducted a comparative analysis combining the sentiment scores generated by RoBERTa and the self-reported positive emotion scores. To assess the relationship between sentiment expressions and self-reported positivity scores, I employed Spearman's rank correlation coefficient, a non-parametric measure that evaluates the strength and direction of association between two ranked variables. I chose this method because of the ordinal nature of my data. Each participant and session's correlations were calculated separately for each sentiment category against the behavior scores. This means the correlation is between the probabilities of having one sentiment and the behavior score, for example, if there is a negative correlation in the negative sentiment category, it indicates that the more likely the sentence is to be negative, the less likely the participant will give a high positive emotion rating. This may potentially uncover the patterns of correlation between sentiment and self-reported emotional states.

The next step was to examine the correlations between these sentiment probabilities and the principal components derived from PCA of the self-reported thought scores. This correlation analysis is trying to explore if there is a systematic relationship between the computational sentiment categorizations and the multidimensional thought content as defined by participants' self-reports.

I calculated the correlation coefficients between the three sentiment groups and the principal components that represent the condensed self-reported thought scores. Given the ordinal nature of both the sentiment categories and PCA components, Spearman's rank correlation coefficient was the chosen statistical method.

This method aims to uncover potential patterns, such as whether certain clusters characterized by a preponderance of positive sentiment are associated with particular aspects of the thought content identified by the PCA components.

3 Results

3.1 Sentiment Analysis

For the first step, I applied the pre-trained Twitter-roBERTa-base sentiment analysis model to 1,601 transcriptions of participants' verbalized thoughts of the moment, which categorize participants' sentiments into negative, neutral, and positive categories. The result revealed that there is a predominant transcription being classified as neutral sentiment, with approximately 66.8 % (n = 1203) of the processed transcription being classified as such. This result

suggests a tendency for participants to articulate thoughts that, accoring to the model, are not likely positive nor negative. The probabilities for the assigned three sentiment classifications ranged from 34.3% to 94.9%, indicating the model's variable confidence across different instances. The average confidence level for the sentiment determination was 66.3%, suggesting that there is a reasonable certainty in the classification made by the model. However, the distribution of the sentiment probabilities also captures the nuanced spectrum of emotional states in the participant's thought process, both positive and negative sentiments are non-negligible to be presented.

The distributions for negative, neutral, and positive sentiment scores each show unique patterns: Negative Sentiment Scores display a skewed distribution, with a high frequency of scores towards the lower end, indicating that fewer transcripts were strongly negative. Neutral Sentiment Scores show a broader distribution, with a peak around the mid-range scores. This suggests that a significant portion of the transcripts were classified with moderate neutrality. Positive sentiment scores also display a skewed distribution similar to negative scores, with most scores concentrated towards the lower end, indicating fewer transcripts were strongly positive.

The correlation between each sentient score highlights the interplay between different emotional valences expressed in the participants' verbalized thoughts, with each sentiment category inversely related to the others. For Negative and Neutral Sentiments, there is a moderate negative correlation (r=0.613, $p < .001$). A negative correlation (r=0.407, $p < .001$) is also observed here for Negative and Positive Sentiments, indicating an inverse relationship between negative and positive sentiments. Finally, for Neutral and Positive Sentiments, this pair also shows a negative correlation (r=0.473, $p < .001$), suggesting that transcripts classified with higher neutral scores tend to have lower positive scores.

3.1.1 Model Accuracy

Before further analysis, an accuracy test was performed. To evaluate the performance of the sentiment analysis model, I employed precision and recall metrics for each sentiment category (Negative, Neutral, Positive). Precision refers to the proportion of true positive identifications (correctly identified sentiment) over all positive identifications by the model (both correct and incorrect), while recall measures the proportion of true positive identifications over all actual positives (the amount of actual sentiment that was correctly identified).

Here are the precision and recall metrics for each category: Negative Sentiment: Precision = 75%, Recall = 70%; Positive Sentiment: Precision = 67%, Recall = 65%; Neutral Sentiment: Precision = 92%, Recall = 90%.

The overall accuracy, which reflects the proportion of all correctly predicted sentiment categories over all evaluations, was 84%. This overall accuracy metric demonstrates the

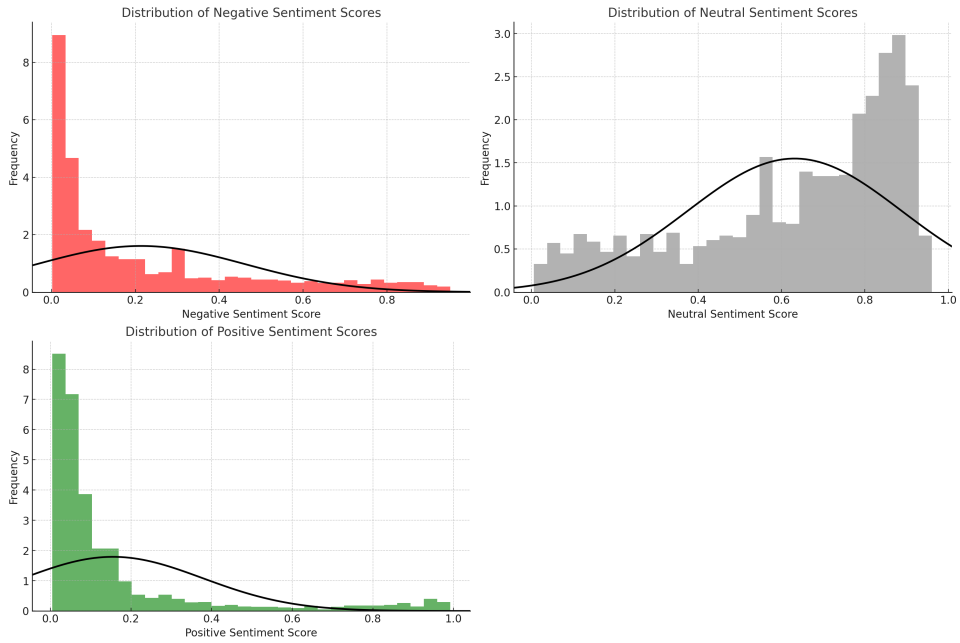

Figure 1: The histograms above illustrate the distribution of sentiment scores across all transcripts.

model's general effectiveness across all sentiment categories.

The confusion matrix illustrates the distribution of predictions across categories, highlighting the model's ability to correctly identify sentiments as negative, neutral, or positive. In sum, this roBERTa model provides high accuracy and reasonable precision and recall rates in this validation exercise, which enhances its credibility as a tool for sentiment analysis.

3.1.2 Sentiment example

To better explore the analysis and classification done by the model, I extracted several examples that are representative of each category. However, to protect the privacy of the participants, I summarize the example here.

High Probability of Neutral Sentiment: One transcript focused on logistical concerns regarding session schedules, showing the participant reflecting on the duration and end times. The model assessed the probability of neutral sentiment as 96.1%, indicating a high likelihood of neutrality, with comparatively lower probabilities for negative (1.1%) and positive sentiments (2.8%).

High Probability of Positive Sentiment: The participant's remarks about an upcoming event reflect significant enthusiasm, anticipation, and excitement. The model assigned a

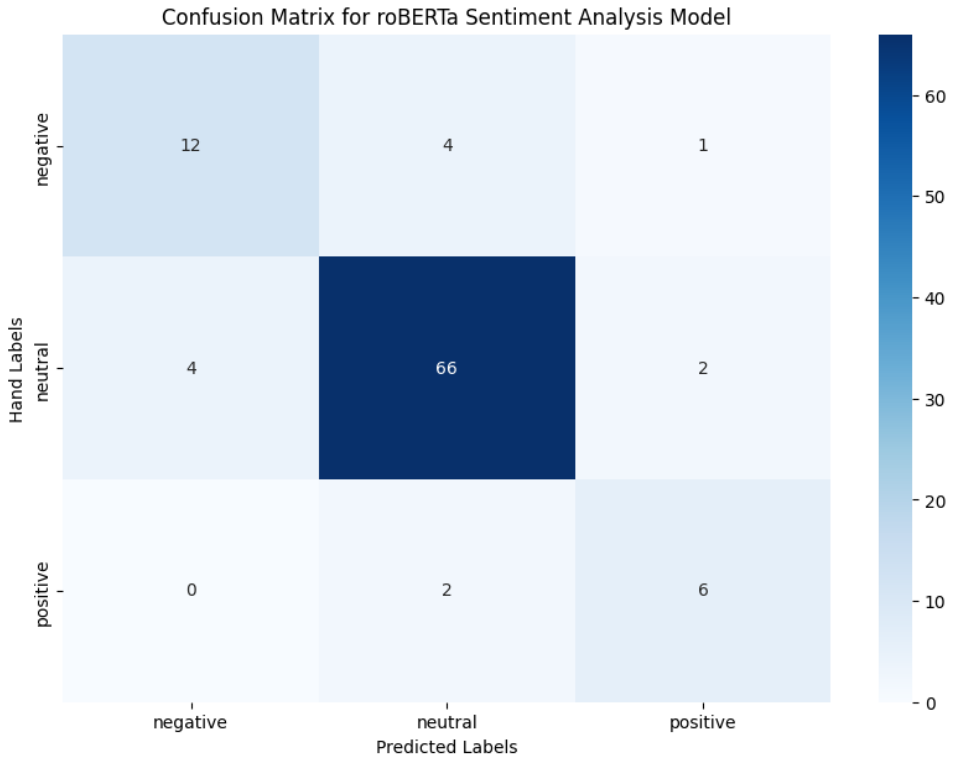

Figure 2: This figure presents a confusion matrix evaluating the performance of the RoBERTa sentiment analysis model. The matrix compares the predicted sentiment labels (Negative, Neutral, Positive) against the hand-labeled (true) sentiment categories. The cells of the matrix show the count of transcripts that fall into each category of the true label versus the predicted label intersection. The shades of blue in each cell represent the count of transcripts, with darker shades indicating higher numbers.

high probability of positive sentiment at 99.2%, suggesting a strong likelihood of positive sentiment, with very low probabilities for neutral (0.6%) and negative sentiments (0.2%)."

High Probability of Negative Sentiment: This case involves a participant expressing frustration over disorder and messiness, which they find particularly bothersome. The sentiment scores reveal a strong possible negative reaction (Negative: 95.1%), with very slight neutral (Neutral: 4.3%) and positive emotion (Positive: 0.6%) probability.

Overall, the model did a satisfactory job of classifying sentiment with limited context and a relatively short corpus. Which benefited from the training on Twitter posts which share similar features with the self-reported thoughts of the participants. However, the model is not perfect. There are some ambiguous thoughts or hidden sarcasms that represent the strong complex nature of human language and linguistic structure. The model may face challenges to such transcriptions and make mistakes.

Potential Model Mistakes (Ambiguous Transcripts)

In the study, some examples were noted where the emotional content of speech was nuanced and could potentially challenge the sentiment analysis model. One noticeable scenario involved an expression reflecting on the business of a week, with mixed feelings of excitement and weariness expressed. The sentiment scores associated with this example were closely distributed across negative, neutral, and positive categories, highlighting the complexity of human emotions.

Those ambiguous thoughts make it hard to determine the exact sentiment even from a human's perspective. Such instances underscore the importance of considering the context and the limitations of sentiment analysis models, especially when dealing with complex human emotions and thoughts. However, they are relatively rare cases (the percentage of ambiguous sentiment choices made by the RoBERTa model, based on the criterion that no sentiment category has a probability exceeding 50%, is approximately 2.38%) and I decided to allow them to remain in the corpus and further analysis.

3.2 Self-reported thoughts

The histograms exhibit a range of distributions across the different self-reported measures. Awake scores generally show a trend towards higher ratings, suggesting that most participants felt alert during the thought sampling process. External focus exhibits a skewed distribution, with a significant number leaning towards lower scores, indicating a tendency for thoughts to be less about the external environment. Conversely, ratings for future-oriented thoughts have a relatively even spread, implying no strong bias towards or against future-focused thinking among participants. Scores related to past thoughts and other-directed thinking suggest some variability among participants, with no single dominant

trend. Notably, the distribution of positive thought content ratings reveals a skew towards higher scores, suggesting that participants often reported positive thoughts.

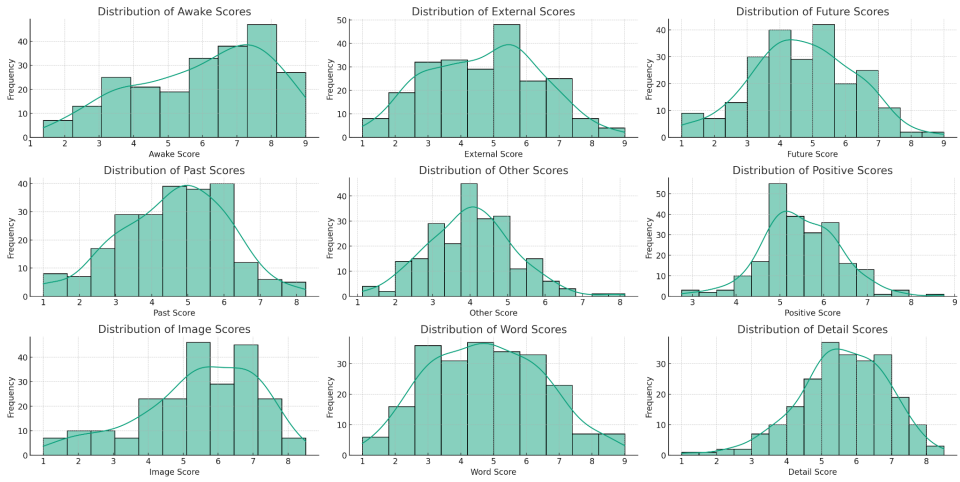

Figure 3: This figure presents a series of histograms that visualize the frequency distributions of self-reported scores across different thought content categories obtained from a study on mind wandering. Each histogram corresponds to a specific category, with scores ranging from 1 to 9.

3.3 Principal Component Analysis

In the process of examining the complex spectrum of mind wandering, Principal Component Analysis (PCA) was performed on the self-reported thought scores. The analysis revealed the principal components that encapsulate key cognitive and emotional patterns within the participants' experiences as quantified through their self-reports.

In order to determine the optimal component number, I decided to choose a number of components that capture a high percentage of the variance. I draw a scree plot that shows the cumulative explained variance ratio. Then I look for the point where the incremental benefit of adding another component starts to decline significantly. Based on the explained variance and the plot, we can see that the curve begins to flatten after the fourth component, which brings the cumulative explained variance to about 66.4%. After this point, each additional component seems to add progressively less to the model. Therefore, I chose 4 components (66.4% explained variance) as a balance between complexity and explanatory power.

The PCA unveiled that the first principal component (Thought_PC1) captures approximately 27.45% of the variance within the data, highlighting its pivotal role in encapsulating the primary facets of the thought content. This component is significantly characterized by

strong negative loadings on 'Image', 'Positive', and 'Other', denoting an inverse association with thoughts related to visual imagery, positive emotions, and other-directed thoughts.

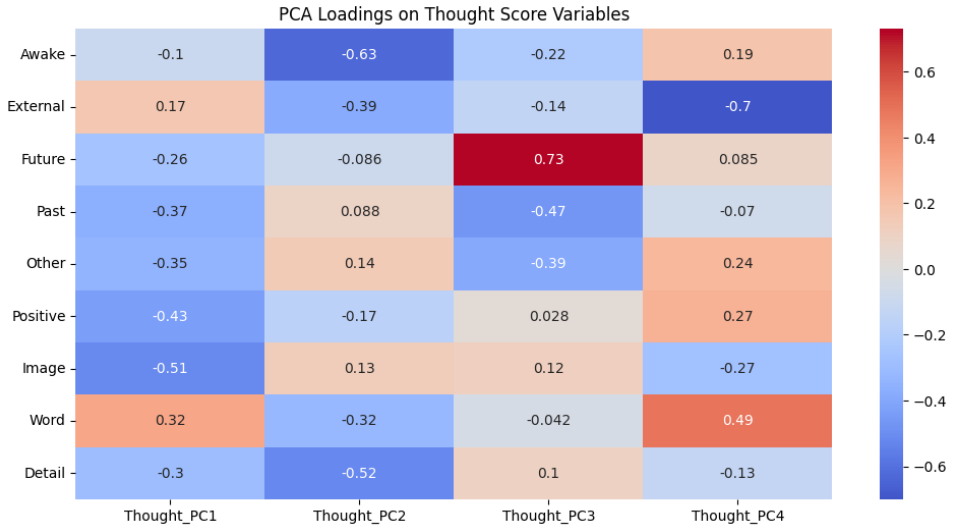

Figure 4: Heatmap Representation of PCA Loadings on Self-Reported Thoughts. The heatmap's color intensity mirrors the significance of each thought category's contribution across the four principal components (Thought_PC1, Thought_PC2, Thought_PC3, and Thought_PC4). This visualization is instrumental in interpreting the direction and magnitude of the influence exerted by each dimension on the components generated from the PCA.

The second principal component (Thought_PC2), which accounts for 15.38% of the variance, is distinguished by a substantial negative loading on 'Awake', revealing an inverse relationship with the participants' self-reported alertness during the thought process. This component also shows strong negative loadings on 'Detail', suggesting a contrast with the intricacy of thoughts, and to a lesser extent, 'External', pointing to a less external focus.

The third component (Thought_PC3), explaining 11.91% of the variance, is marked by a large positive loading for 'Future' thoughts, while 'Past' thoughts are strongly negatively correlated, highlighting a dimension that contrasts future-oriented thinking with retrospective or past-focused rumination.

The fourth component (Thought_PC4) accounts for 11.71% of the variance. It shows a pronounced negative loading on 'External', indicating that it captures an element of thought content that is less related to external stimuli. Conversely, there is a positive loading on 'Word', implying a connection to language or narrative thought processes.

Together, these four components elucidate 66.45% of the total variance, providing a comprehensive understanding of the thought dimensions examined. This structured anal-

ysis through PCA facilitates an in-depth exploration into the relationships between the sentiment scores derived from the Twitter-based RoBERTa model and the PCA components of thought content.

3.4 Comparing thought ratings between sentiment

The exploration of cognitive processes through sentiment analysis reveals a complex relationship between the emotional tone of thoughts and the cognitive content they encompass. By merging and analyzing data from transcripts classified by RoBERTa into positive, neutral, and negative sentiments, and contrasting these with thought ratings across nine distinct dimensions, some patterns emerge that reveal the intricate interplay of sentiment within our cognitive thought dimensions.

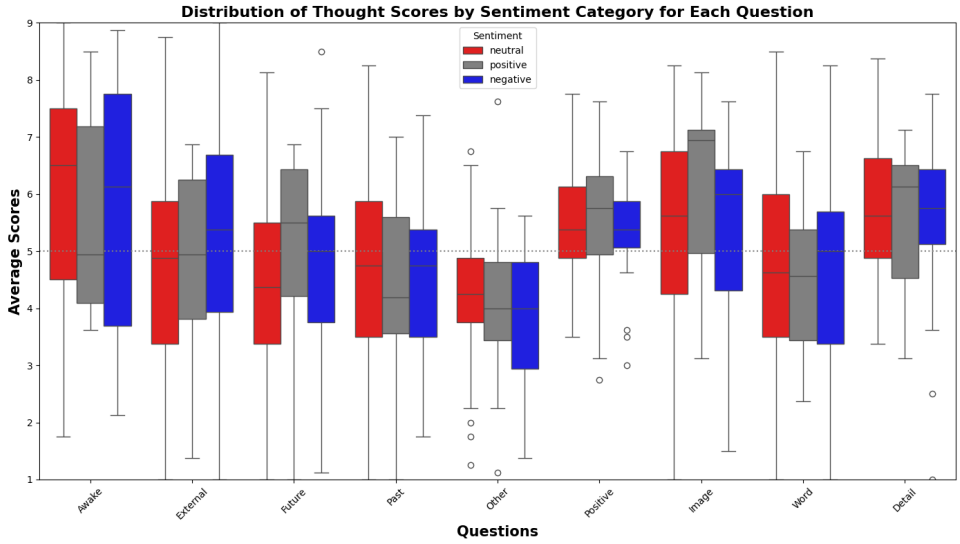

Figure 5: Distribution of Thought Scores by Sentiment Category for Each Question: This figure illustrates box plots that detail the spread and central tendency of thought scores for different sentiment categories across various questions. Each box plot represents the interquartile range of scores, the median and outliers. The horizontal dotted line at y=5 indicates the midpoint of the scoring scale, providing a reference for assessing score distribution relative to the scale's mid-value.

Upon merging, the data provided a platform to compare the average thought ratings against the backdrop of sentiment classifications. Notably, the ANOVA results reveal significant differences in thought ratings for the categories of "Future" ($F(2, \_) = 3.377$, $p = .034$), "Image" ($F(2, \_) = 3.618$, $p = .027$), and "Other" ($F(2, \_) = 3.085$, $p = .046$), suggesting that sentiment has a measurable impact on these specific domains of mind wan-

dering. For instance, thoughts about the "Future" and the presence of "Image" in thoughts demonstrate a variance that aligns with the emotional valence ascribed by the participants, indicating that positive emotional states might indeed foster a more vivid or expansive visual imagination.

Interestingly, while "Awake" did not reach statistical significance ($p = .073$), the trend suggests a possible relationship between emotional neutrality and heightened alertness or present-moment awareness.

Besides these findings, the data revealed that transcripts with a positive sentiment, as classified by real-time verbal explanations, corresponded to a higher self-report of positive valence. Inversely, transcripts with a negative sentiment corresponded to a lower self-report of emotional valence. This finding illustrates the alignment between subjective self-reports and objective real-time measures when capturing the positive aspects of thought content.

3.5 Clustering of Sentiment Scores

A clustering analysis was conducted to categorize the sentiment scores into discrete groups that represent unique sentiment profiles. The scatter plot visualizing the clustering results depicted the distribution of sentiment scores, with each data point colored according to its cluster assignment. Four clusters were identified, as shown in the plot:

Cluster 0 (Purple): This cluster predominantly consists of data points with high 'Positive' sentiment scores and low 'Negative' sentiment scores, indicating expressions of high probabilities of positive sentiment.

Cluster 1 (Blue): This cluster is characterized by the lowest positive sentiment scores and moderately high negative probability scores, indicating a predominance of negative sentiment expressions.

Cluster 2 (Green): Comprising data points with moderate positive and moderate negative scores, this cluster reflects a balanced sentiment expression with neither high probabilities of positive or negative sentiments.

Cluster 3 (Yellow): Representing sentiments with the lowest 'Positive' scores and ranging 'Negative' scores, this cluster captures the more neutral or ambiguous sentiment expressions.

The distribution and positioning of clusters on the scatter plot suggest the presence of a spectrum of sentiment, from highly positive to neutral or ambiguous states, as verbalized by the participants.

After clustering, frequency plots have been generated for each cluster to better visualize the most frequent terms within the transcriptions. These frequency plots provide a quantitative view of the sentiment thematic content within each cluster.

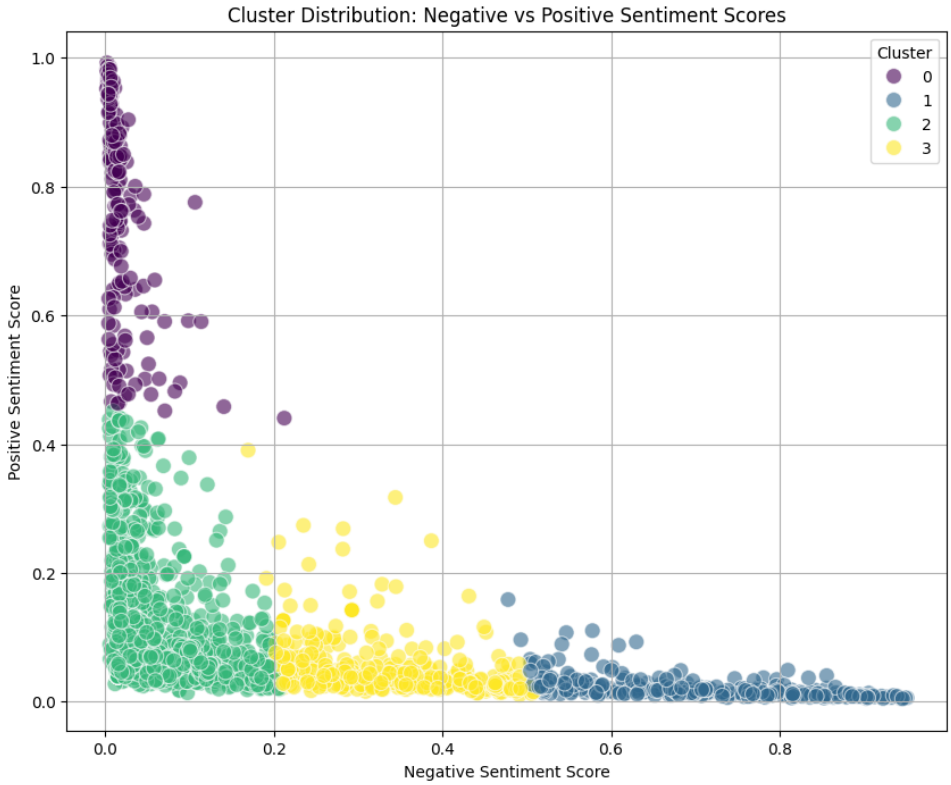

Figure 6: This scatter plot illustrates the clustering of sentiment scores based on negative versus positive sentiments. The x-axis represents the negative sentiment scores and the y-axis the positive sentiment scores, both ranging from 0 to 1. Data points are colored based on their cluster group: Cluster 0 (purple), Cluster 1 (blue), Cluster 2 (green), and Cluster 3 (yellow).

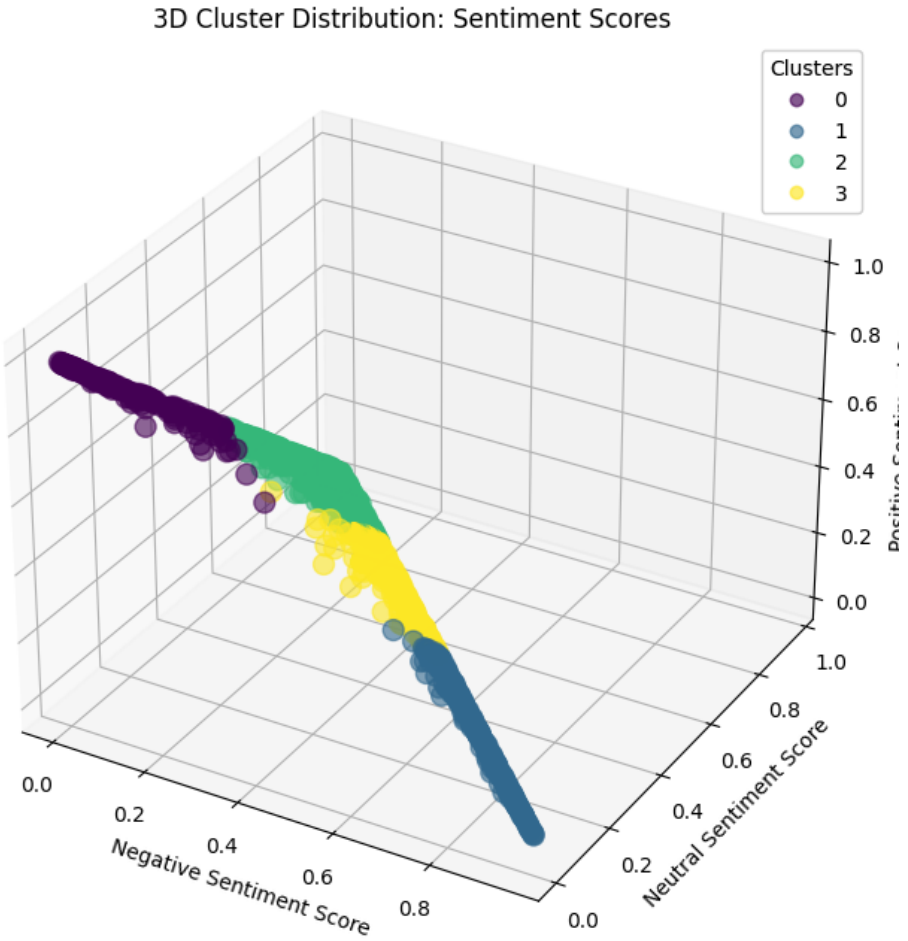

Figure 7: This 3D scatter plot presents a multi-dimensional view of sentiment clusters. Along with the negative and positive sentiment axes, there's a third axis for neutral sentiment scores, providing a comprehensive overview of how these scores interact. Each data point represents a sentiment score for a transcript, with the colors corresponding to the different clusters, the same as in the 2D plot.

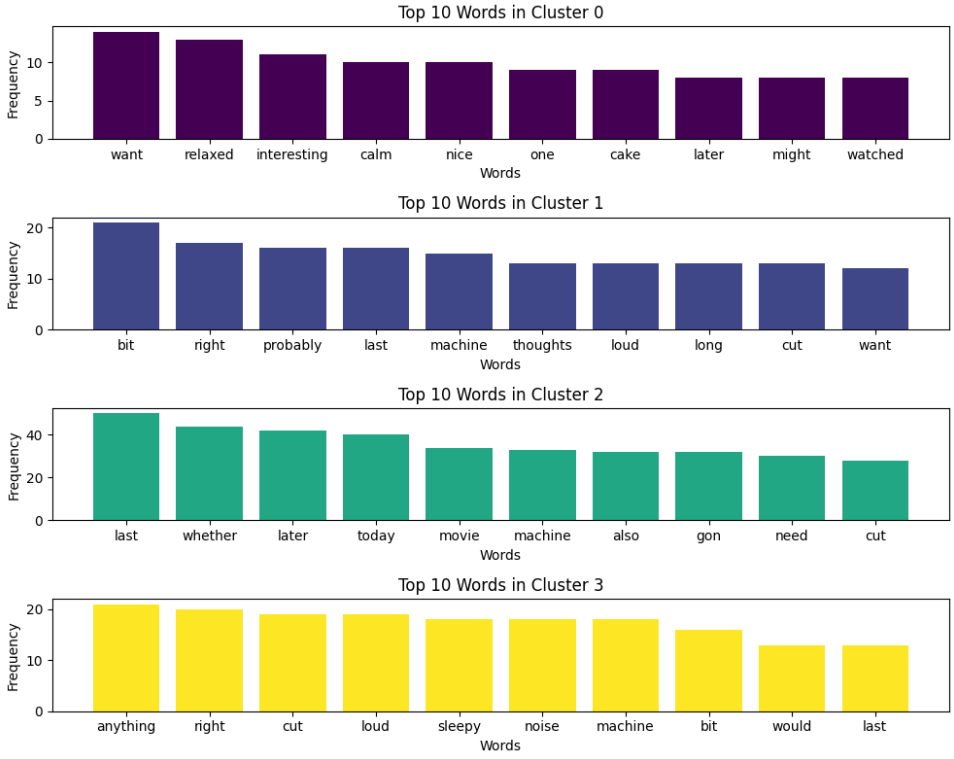

Figure 8: This figure is comprised of four horizontal bar charts, each depicting the frequency of the top 10 words within a distinct sentiment cluster identified in the sentiment analysis. The clusters are labeled from 0 to 3, each represented by a different color: Cluster 0 (purple), Cluster 1 (blue), Cluster 2 (green), and Cluster 3 (yellow).

For cluster 0, the predominant terms are indicative of a neutral sentiment, highlighting words such as “relaxed,” “interesting,” and “calm,” which suggest a predominantly composed or contemplative state among the participants. Such terms are reflective of the participant’s current state of mind and suggest a more positive expression of emotional states.

For cluster 1, the terms such as “bit,” “right,” and “probably” imply a level of uncertainty or measured thoughtfulness. The presence of words like “loud” and “machine” may be related to the relative negative focus on the external environment like the fMRI machine.

Cluster 2 features words such as ”last,” ”whether,” and ”today,” which may imply contemplation or consideration of recent or upcoming events, and ”movie” might reflect specific references to the naturalistic task. Which suggests a more neutral sentiment to this cluster.

For Cluster 3, words like ”anything,” ”right,” and ”loud” can be associated with a range of sentiments but may lean towards negative sentiments given the terms like ”loud” and ”sleepy,” which can be associated with discomfort or fatigue. This correlates with the cluster’s placement on the scatter plot, which shows lower positive sentiment scores.

The words enrich our understanding of the clustering result and the layers of textual data under sentiment classifications. We can see that the clusters identified are characterized not only by the sentiment polarity but also by the word selection and language used by the participants. This analysis confirms that negative sentiment is a significant differentiator in our dataset, while also highlighting that positive sentiment and thematic context play a crucial role in the sentiment structure revealed by clustering.

For the next step, by examining the relationship between these clusters and the PCA components of self-reported thought scores, I can investigate the potential correlations between computationally generated sentiment scores and the subjective dimensions of mind wandering captured through self-reports.

3.6 Correlation Between PCA Components of Thought Scores and RoBERTa Sentiment Scores

The comparative analysis of RoBERTa-generated sentiment probabilities and self-reported behavioral states reveals that there are some notable correlations, which indicate the interplay between participants’ expressed sentiment probabilities and their introspective evaluations. The Spearman’s rank correlation analysis showed that higher probabilities of negative sentiment classification by the RoBERTa model are correlated with lower positivity scores from “Question 6: My thoughts were. very negative/very positive ” ($r = -0.19$, $p < 0.001$). This suggests that a higher model certainty in negative sentiment classification tends to coincide with less positive self-reported emotional states. Similarly, a higher likelihood of being classified as positive sentiment is correlated with higher positivity scores ($r = 0.18$,

$p < 0.001$), indicating that participants' expressions more certain to be positive are aligned with their higher self-reported emotional positivity.

Moreover, the probability of a negative sentiment classification positively correlated with the participants' reported external focus level (r = 0.18, $p < 0.001$), suggesting that those more absorbed in external stimuli tend to express more negative sentiments. Conversely, transcripts classified with higher probabilities of positive sentiment tended to correlate with a more internally focused orientation (r = -0.14, $p < 0.001$). This analysis not only aligns the computational sentiment categorizations with self-reported cognitive assessments but also highlights nuanced interrelationships between the valence of spontaneous thoughts.

These results underscore the potential of using advanced sentiment analysis models to gain deeper insights into the emotions and thoughts that occur spontaneously, highlighting the value of integrating computational and psychological methods in research.

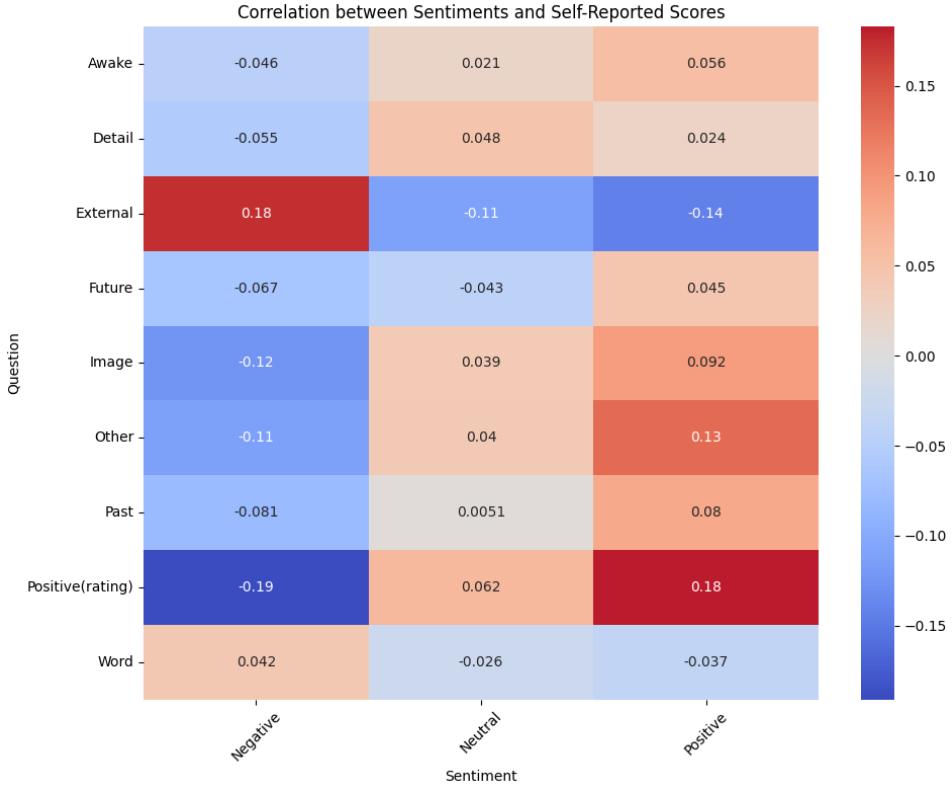

Figure 9: This heatmap presents a matrix of correlation coefficients, with rows representing different questions and columns representing sentiment categories. The cells within the heatmap contain the correlation coefficients, with the scale ranging from -0.15 to 0.15. Blue shades indicate negative correlations, while red shades indicate positive correlations. The intensity of the color corresponds to the strength of the correlation.

In furthering the analysis of the relationship between computational sentiment scores and self-reported thought content, a correlation analysis was conducted between the PCA components derived from the thought scores and the sentiment scores indicating the probability of a transcript being classified as positive, negative, or neutral generated by the RoBERTa model. This analysis aimed to uncover any significant linear relationships and offer insight into how the dimensions of self-reported thought content may align with the sentiment expressed in participants' speeches.

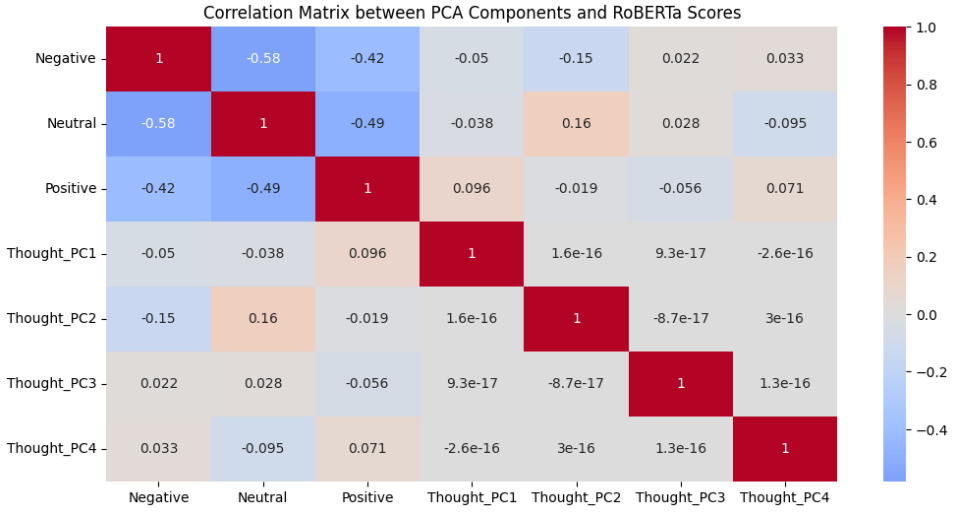

Figure 10: The heatmap shows a grid where each cell represents the correlation coefficient between different sentiment scores and PCA components, with the color intensity and shade corresponding to the strength and direction of the correlation. Red indicates a positive correlation, blue indicates a negative correlation, and white represents no correlation.

The heatmap of the correlation matrix delineates the correlation coefficients between the four PCA components of self-reported thought scores and the 'Negative', 'Neutral', and 'Positive' sentiment scores. The heatmap's color spectrum, which ranges from blue for negative correlations to red for positive correlations, visually quantifies the strength and direction of each correlation, with color intensity conveying the magnitude. For Thought_PC1, it manifested a marginal positive correlation with 'Positive' sentiment (r = 0.096) but held a negligible connection with both 'Negative' and 'Neutral' sentiment scores. This insinuates that Thought_PC1 while capturing a prominent share of the variance in thought content, is only slightly aligned with more positive expressions of sentiment. Thought_PC2 presented a weak inverse relationship with 'Negative' sentiment (r = -0.15) and a correspondingly weak direct relationship with 'Neutral' sentiment (r = 0.16), which suggests that Thought_PC2 could be inversely related to negatively valenced sentiments and positively associated with

a neutral state of sentiment. Thought_PC3 revealed no substantive correlation with any sentiment scores, signaling that the characteristics embodied by Thought_PC3 do not have a discernible linear correspondence with the sentiment scores analyzed. Thought_PC4 demonstrated a very weak inverse correlation with 'Neutral' sentiment (r = -0.095) and a similarly weak positive correlation with 'Positive' sentiment (r = 0.071), hinting at a minor propensity of Thought_PC4 to be inversely related to neutral sentiments and slightly inclined towards positive sentiments.

These findings suggest that while there are some associations between the sentiment scores and the principal components of thought scores, the relationships are generally weak. It indicates that the sentiment scores and the components of thought content, as derived from the PCA, might reflect different facets of the participants' experiences.

4 Discussion

4.1 Sentiment Analysis and PCA Findings

The utilization of the Twitter-based RoBERTa model on 1,821 verbalized thought transcriptions indicated that a significant majority, approximately 66.8%, were classified with a neutral sentiment. This prevalence of neutral sentiments suggests that during mind wandering, individuals might not engage in intense emotional introspection or that they may prefer to articulate emotionally balanced thoughts. However, the negative and positive sentiments are negligible.

The RoBERTa model achieved an 84% accuracy rate when compared with human ratings of a random subset of 100 transcripts, highlighting its effectiveness in correctly identifying the emotional tones embedded in the verbalized thoughts. This precision demonstrates the model's potential as a reliable tool for psychological research, providing a method to objectively measure emotional states that avoids some biases associated with self-reports.

PCA identified distinct dimensions of thought content indicated with self-report ratings, revealing the complex interplay between different types of thoughts and their emotional valences. Notably, the analysis showed that thoughts about the future and past are often emotionally charged and may influence an individual's emotional well-being.

4.2 Clustering of Sentiment Scores

Clustering analysis of sentiment scores revealed distinct clusters that correspond to different types of emotional experiences during mind wandering. These clusters can be used to identify patterns or typical profiles of emotional thought processes.

4.3 Thought Ratings Between Sentiments

The analysis that compared thought ratings between different sentiment classifications indicated that certain types of thoughts were more frequently associated with specific sentiments. For example, thoughts that were rated as more future-oriented or worry-related tended to have higher negative sentiment scores. This pattern suggests that the emotional valence of thoughts during mind wandering can vary significantly depending on the content and focus of the thoughts, highlighting the need to consider the qualitative aspects of thoughts in addition to their emotional tone.

4.4 Correlations Between Sentiment Scores and Self-Reported Emotional States

The significant correlation found between computationally derived sentiment scores and self-reported emotions directly addresses my research question concerning the emotional dynamics of mind wandering. This correlation demonstrates not just alignment but also the supplementary nature of the two methodologies in capturing the emotional valence of thoughts.

While computational methods like the RoBERTa model provide real-time, objective assessments of sentiment, they do not necessarily offer a less biased alternative; instead, they complement self-report methods by adding a layer of objective data that can validate and enrich the subjective reports provided by participants. The fact that computational sentiment scores align with self-reported emotions supports the idea that these methods can be integrated, each contributing its strengths to a more comprehensive understanding of emotional experiences during mind wandering.

Also, by utilizing both computational and self-report methods, researchers can achieve a more accurate and nuanced view of the emotional aspects of mind wandering. This dual approach allows for the validation of self-reported data through computational analysis

4.5 Limitations and Future Directions

The integration of NLP models such as RoBERTa into psychological research offers considerable advantages, including the ability to process large datasets efficiently and provide insights into complex patterns of emotional expression. However, this approach is not without its limitations, which must be acknowledged to fully understand the implications of my findings.

One significant concern involves the discrepancies that may arise between computationally derived sentiment scores and manually labeled data. For instance, if RoBERTa assigns lower positivity scores to statements that participants self-rated as 'very positive' compared

to those rated as 'positive', this raises questions about what is being measured: the relationship between self-reports and the textual content of speech or RoBERTa's capabilities in sentiment classification? Similar issues arise with negative sentiments. Such discrepancies could potentially indicate limitations in the NLP model's ability to discern nuanced emotional expressions compared to human judgments.

To mitigate these limitations mentioned, here are some potential directions for future research. For example. enhancing Model Training, future research could focus on training RoBERTa and similar models on more diverse datasets that include a wider range of emotional expressions and contexts. This could help improve the model's sensitivity and reduce bias. Also, we can use Hybrid Annotation Approaches, combining computational methods with manual annotations might provide a more robust framework for sentiment analysis. This hybrid approach can leverage the efficiency of NLP tools while ensuring the nuance and context sensitivity provided by human annotators. By addressing these limitations and exploring these future directions, research can better unlimit the capabilities of NLP in psychological research while minimizing potential drawbacks.

5 Conclusion

My thesis investigated the emotional dynamics of mind wandering, specifically examining how computational sentiment analysis via NLP and self-report methods can capture the emotional valence of thoughts. The research revealed that both methodologies, while distinct, provide complementary insights that deepen our understanding of emotional experiences during mind wandering.

The Twitter-based RoBERTa model, applied to 1,821 verbalized thought transcriptions, demonstrated a significant prevalence of neutral sentiments, suggesting that mind wandering frequently involves emotionally balanced thoughts rather than extreme emotional states. This finding is important as it challenges the assumption that mind wandering is predominantly characterized by emotionally charged thoughts. The accuracy of the RoBERTa model in classifying emotional tones, rated at 81.25%, underscores the potential of NLP tools as reliable aids in psychological research.

Moreover, the Principal Component Analysis (PCA) and clustering of sentiment scores further elaborated on the complexity of thought processes during mind wandering. PCA identified distinct dimensions of thought content, linking them with specific emotional valences, while clustering helped categorize different emotional experiences.

A crucial aspect of this research was the demonstration of a significant correlation between the sentiments derived computationally and those reported by participants. This finding not only supports the validity of using NLP in conjunction with self-reports to as-

sess emotional states but also highlights how these methods can be integrated to provide a more nuanced view of emotional valence.

However, the study also acknowledged limitations, particularly concerning the potential discrepancies between computationally derived sentiment scores and self-reports. These discrepancies underscore the need for further refinement of NLP tools to better capture the nuances of emotional expression.

Data and Code Availability Statement

The datasets generated and analyzed during this study contain information that could compromise the privacy of research participants and are therefore not publicly available. Data are held securely within the CAB Lab, and access is restricted to research team members.

The code used in the analysis can be shared upon request. The code is available under the MIT License. For further inquiries, you can contact the corresponding author.

References

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological science*, *23*(10), 1117–1122.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature reviews neuroscience*, *17*(11), 718–731.

Erez, A., & Isen, A. M. (2002). The influence of positive affect on the components of expectancy motivation. *Journal of Applied psychology*, *87*(6), 1055.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist*, *56*(3), 218.

Ho, N. S. P., Poerio, G., Konu, D., Turnbull, A., Sormaz, M., Leech, R., Bernhardt, B., Jefferies, E., & Smallwood, J. (2020). Facing up to the wandering mind: Patterns of off-task laboratory thought are associated with stronger neural recruitment of right fusiform cortex while processing facial stimuli. *Neuroimage*, *214*, 116765.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651–666.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*(5702), 1776–1780.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932–932.

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). Timelms: Diachronic language models from twitter. *CoRR*, *abs/2202.03829*. https://arxiv.org/abs/2202.03829

Marchetti, I., Koster, E. H., Klinger, E., & Alloy, L. B. (2016). Spontaneous thought and vulnerability to mood disorders: The dark side of the wandering mind. *Clinical psychological science*, *4*(5), 835–857.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and gre performance while reducing mind wandering. *Psychological science*, *24*(5), 776–781.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual review of psychology*, *66*, 487–518.