Student Name: Darian Hamel

Instructor: Robert Guderian

Course: COMP 4560 Industrial Project

Discovering Relationships via Exploratory Analysis of Mood and Sleep using LLMs and Statistical Models (DREAMS (LLMs))

1. Introduction

This project explores whether AI techniques - specifically Large Language Models (LLMs), Neural Networks, and statistical models can uncover meaningful insights from user-generated wellness data. The goal is to forecast future mood states, explain mood fluctuations, and detect behavioral patterns using daily records of habits, sleep scores, and journal entries.

Approach:

1. LLM-Based Feature Extraction: Use pre-trained models to derive emotions and events from journal entries.
2. Time-Series Forecasting: Train neural network and statistical models to predict mood using structured features - mood score, sleep score, extracted events and emotions.

Core Question: Can behavioral data (sleep, habits, journal text) reliably forecast mood, and does added context (e.g., emotional extraction) improve accuracy?

2. Problem Statement

Objective: Predict next-day mood (0–4 scale) using:

* Journal Entries
* Sleep Scores
* Mood Scores

3. Data Overview

Dataset

* Training Dataset:
  + Synthetically generated personas with unique personalities using the Big 5 personality traits
    - 4 Personas
    - 1 primary persona with 100 entries for sleep and journal
    - 3 personas with 50 entries each for sleep and journals
* Real Dataset:
  + 3681 Entries
  + 377 Unique Users
  + Median entries 10
  + 2 Users with 10 and 3 entries respectively to test the model on

Preprocessing:

Extracting emotions and events utilizing pre-trained LLMs through transformer (a HuggingFace API for pre-trained models).

The emotional extraction was done using *j-hartmann/emotion-english-distilroberta-base*.

The event extraction was done using zero-shot classification using *facebook/bart-large-mnli*.

Tools: pandas, statsmodels, PyTorch, ChatGPT, Deepseek, pipeline, matplotlib, sklearn

Data Fields Table:

|  |  |  |
| --- | --- | --- |
| Field | Description | Example Value |
| Sleep Score | 0-4 Scale | 4 |
| Mood Score | 0-4 Scale (Target Variable) | 3 |
| Journal Entry | Concatenated daily text and synthesized for training | Had a great workout this morning. Feeling energized and ready to take on the week! |
| Event(s) | Extracted from journal entry | Exercising or moving your body |
| Emotion | Extracted from journal entry | joy |

4. Methodology

a. Emotion/Event Extraction (LLM):

* Input: Journal entry (e.g., "It rained softly this morning, like a secret the sky was whispering just to me. I took my time getting ready, savoring the slow pace for once.").
* Output: Structured features (e.g., {"emotion": "neutral", "event": "Spending time outdoors or in nature"}).

b. Time-Series Models:

* Feed Forward Neural Network: Trained on emotion, sleep, and past mood (window: t, t−1).
* Simple Feed Forward Neural Network: Trained on mood, sleep (window t, t-1).
* ARIMA/VAR: Baseline for linear trends.
* Random Forest: Trained on emotion, sleep, and past mood (window: t, t−1).

c. Feature Engineering:

* Mapped moods to Sentiment Classes
  + Mapped the 2 positive classes to 1 and similarly the 2 negative classes to 1
* Aggregated sequences of data from previous day and current day to predict next day mood.

Flow Diagram:

A screenshot of a diagram

AI-generated content may be incorrect.

5. Results

LLM prompted with the same data as the neural networks and even the random forest performed worse and was on par with random guessing.

Random forest had shown promise initially but could not get much better than 16% accuracy on 5 classes of emotions, 28% on 3 classes which is still worse than random guessing. The random forest also had all context, that is, yesterday’s data (emotion, events, mood and sleep) and today’s data.

The simple feed forward neural network with no emotion or event extraction had performed 29% on 5 classes, 54% on 3 classes. Not bad, but once you look at the results, you realize it was not learning – it was guessing positive almost entirely – and when the dataset has a positive bias of 58%, you can see that the model is not getting good results.

The feed forward neural network with the most context gave the best results, with optimistic results of upwards of 72% accuracy on 5 classes of moods and upwards of 77% accuracy on 3 classes of moods. This neural network performed 248% and 143% better than the neural network without the additional context against 5 and 3 mood classes respectively. It also performed 450% and 275% better than random forest between the 5 and 3 mood classes. These results show that utilizing neural networks, especially with more context will give us the best results.

Key Findings:

* Mapping mood to 3 classes improved prediction ability.
* Neural Networks performed the best when given the most context.
* Random Forest had most promise from simpler models before introducing neural networks but still had worse results than random guessing.

Visuals:

Neural Network with Event and Emotion Extraction

A graph with lines and dots

AI-generated content may be incorrect.

Seen Data (last 15 journals)

Non-Mapped Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 1 |
| Bad | 1.00 | 0.67 | 0.80 | 3 |
| Neutral | 1.00 | 1.00 | 1.00 | 3 |
| Good | 0.60 | 1.00 | 0.75 | 3 |
| Best | 0.60 | 0.60 | 0.60 | 5 |
| Accuracy |  |  | 0.73 | 15 |
| Macro Average | 0.64 | 0.65 | 0.63 | 15 |
| Weighted Average | 0.72 | 0.73 | 0.71 | 15 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 1.00 | 0.50 | 0.67 | 4 |
| Neutral | 1.00 | 1.00 | 1.00 | 3 |
| Positive | 0.80 | 1.00 | 0.89 | 8 |
| Accuracy |  |  | 0.87 | 15 |
| Macro Average | 0.93 | 0.83 | 0.85 | 15 |
| Weighted Average | 0.89 | 0.87 | 0.85 | 15 |

A graph with colorful lines and dots

AI-generated content may be incorrect.

Seen Data (All journals)

Non-Mapped Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 9 |
| Bad | 0.75 | 0.63 | 0.69 | 19 |
| Neutral | 0.91 | 0.62 | 0.74 | 16 |
| Good | 0.71 | 0.96 | 0.82 | 28 |
| Best | 0.65 | 0.83 | 0.73 | 24 |
| Accuracy |  |  | 0.72 | 96 |
| Macro Average | 0.60 | 0.61 | 0.59 | 96 |
| Weighted Average | 0.67 | 0.72 | 0.68 | 96 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 1.00 | 0.46 | 0.63 | 28 |
| Neutral | 0.90 | 0.56 | 0.69 | 16 |
| Positive | 0.71 | 1.00 | 0.83 | 52 |
| Accuracy |  |  | 0.77 | 96 |
| Macro Average | 0.87 | 0.68 | 0.72 | 96 |
| Weighted Average | 0.83 | 0.77 | 0.75 | 96 |

A graph with colorful lines and dots

AI-generated content may be incorrect.

Unseen Data

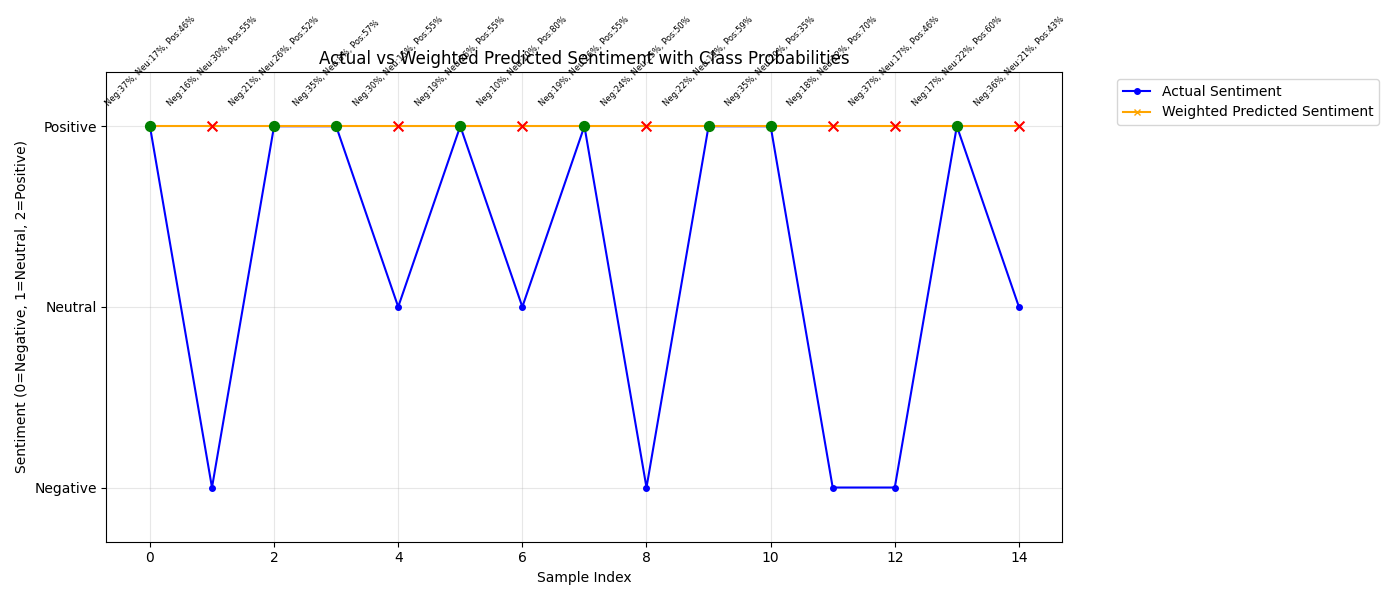
Non-Mapped Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 0 |
| Bad | 0.00 | 0.00 | 0.00 | 1 |
| Neutral | 0.00 | 0.00 | 0.00 | 5 |
| Good | 0.50 | 0.79 | 0.61 | 19 |
| Best | 1.00 | 0.22 | 0.36 | 23 |
| Accuracy |  |  | 0.42 | 48 |
| Macro Average | 0.30 | 0.20 | 0.19 | 48 |
| Weighted Average | 0.68 | 0.42 | 0.41 | 48 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 0.00 | 0.00 | 0.00 | 1 |
| Neutral | 0.00 | 0.00 | 0.00 | 5 |
| Positive | 0.86 | 0.90 | 0.88 | 42 |
| Accuracy |  |  | 0.79 | 48 |
| Macro Average | 0.29 | 0.30 | 0.29 | 48 |
| Weighted Average | 0.76 | 0.79 | 0.77 | 48 |

Simple Neural Network:



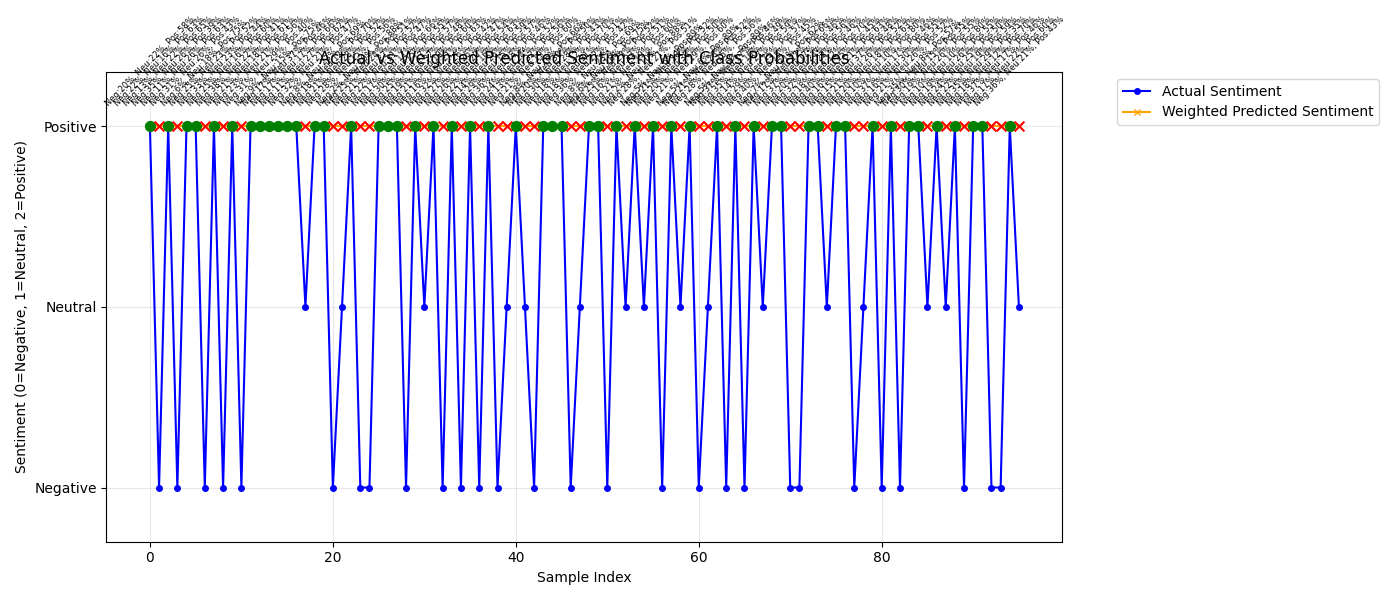
Seen Data (last 15 journals)

Non-Mapped Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 1 |
| Bad | 0.00 | 0.00 | 0.00 | 3 |
| Neutral | 0.00 | 0.00 | 0.00 | 3 |
| Good | 0.20 | 0.67 | 0.31 | 3 |
| Best | 1.00 | 0.20 | 0.33 | 5 |
| Accuracy |  |  | 0.20 | 15 |
| Macro Average | 0.24 | 0.17 | 0.13 | 15 |
| Weighted Average | 0.37 | 0.20 | 0.17 | 15 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 0.00 | 0.00 | 0.00 | 4 |
| Neutral | 0.00 | 0.00 | 0.00 | 3 |
| Positive | 0.53 | 1.00 | 0.70 | 8 |
| Accuracy |  |  | 0.53 | 15 |
| Macro Average | 0.18 | 0.33 | 0.23 | 15 |
| Weighted Average | 0.28 | 0.53 | 0.37 | 15 |



Seen Data (All journals)

Non-Mapped Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 9 |
| Bad | 0.12 | 0.11 | 0.11 | 19 |
| Neutral | 0.50 | 0.12 | 0.20 | 16 |
| Good | 0.28 | 0.68 | 0.40 | 28 |
| Best | 0.56 | 0.21 | 0.30 | 24 |
| Accuracy |  |  | 0.29 | 96 |
| Macro Average | 0.29 | 0.22 | 0.20 | 96 |
| Weighted Average | 0.33 | 0.29 | 0.25 | 96 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 0.00 | 0.00 | 0.00 | 28 |
| Neutral | 0.00 | 0.00 | 0.00 | 16 |
| Positive | 0.54 | 1.00 | 0.70 | 52 |
| Accuracy |  |  | 0.54 | 96 |
| Macro Average | 0.18 | 0.33 | 0.23 | 96 |
| Weighted Average | 0.29 | 0.54 | 0.38 | 96 |

A graph with a number of dots and lines

AI-generated content may be incorrect.

Unseen Data

Non-Mapped Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Worst | 0.00 | 0.00 | 0.00 | 0 |
| Bad | 0.00 | 0.00 | 0.00 | 1 |
| Neutral | 0.00 | 0.00 | 0.00 | 5 |
| Good | 0.40 | 1.00 | 0.57 | 19 |
| Best | 0.00 | 0.00 | 0.00 | 23 |
| Accuracy |  |  | 0.40 | 48 |
| Macro Average | 0.08 | 0.20 | 0.11 | 48 |
| Weighted Average | 0.16 | 0.40 | 0.22 | 48 |

Mapped Sentiment Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Negative | 0.00 | 0.00 | 0.00 | 1 |
| Neutral | 0.00 | 0.00 | 0.00 | 5 |
| Positive | 0.88 | 1.00 | 0.93 | 42 |
| Accuracy |  |  | 0.88 | 48 |
| Macro Average | 0.29 | 0.33 | 0.31 | 48 |
| Weighted Average | 0.77 | 0.88 | 0.82 | 48 |

6. Challenges & Limitations

* Limited real data to get real world feedback if the model truly works.
* Moods are subjective – it’s easier to predict when nothing out of the ordinary happens but hard to determine if a traumatic event (or extremely positive event) will happen the next day to change the mood. Further, moods fluctuate day to day but also within the day – making predictions in the real world likely harder.

7. Conclusion and Future Work

Utilizing LLMs for event and emotion extraction improve the results dramatically – but at a time cost – it could take upwards of 20 minutes to extract all the users’ events and emotions, and the dataset is still small. Mapping the emotion from 5 classes (Worst, Bad, Neutral, Good, Best) to 3 (Negative, Neutral, Positive) showed improvement in all models – this was expected, and a better fit for showing mood trends.

Further, utilizing neural networks showed the best performance and is the direction to go especially once the dataset increases as we could increase model complexity with less risk of overfitting.

Future Directions:

* Personalization: User-specific fine-tuning - that is, once data is large enough for a true generalized model – take the model and tune it to each individual.
* Expanded Context: Incorporate habit and to-do tracking for added context to improve predictions.
* Model Complexity: Once there is enough data, increase the complexity of the model by swapping to a LSTM (Long Short-Term Memory) neural network which *should* improve predictions.