SYMBIOSIS INSTITUTE OF TECHNOLOGY (SIT)

Constituent of SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

From Posts to Patterns

Sentiment–Engagement Analysis on Reddit

01

Introduction

- In today’s world, mental health challenges are more common than ever, yet they often remain hidden. Many individuals suffering from such challenges find it difficult to speak openly about what they feel. But even when they stay silent in real life, many turn to social media platforms to express their thoughts anonymously

- Among these platforms, Reddit is a goldmine where people freely discuss their feelings, experiences, and struggles under the comfort of anonymity. Every post carries a piece of someone’s emotional state. Hidden within these millions of posts lies a vast, untapped source of emotional data that can help us understand human wellbeing in the digital age.

What is Reddit and Why Reddit?

- A social news aggregation and discussion platform where users share posts, vote, and comment.
- Organized into subreddits (topic-specific communities), allowing focused discussions.
- Known for long-form content and in-depth conversations compared to other social platforms.

Quality of Data

- Reddit posts are usually longer and more descriptive, providing richer context for sentiment analysis.
- Twitter has a 280-character limit, making sentiment extraction harder.

Focused Communities

- Reddit allows targeting specific subreddits (e.g., r/MentalHealth), ensuring topic relevance.
- Twitter streams are broad and noisy, requiring heavy filtering.

Anonymity & Honest Sharing

- Reddit users often share personal experiences under anonymity, especially on mental health topics.
- Twitter tends to have public identities, which can lead to less candid posts.

Engagement Metrics

- Reddit provides upvotes, comments, and engagement ratios for deeper interaction analysis.
- Twitter's primary engagement is likes/retweets, which may not reflect discussion depth.

02 Literature Review

Year	Author/s	Title	Approach	Result
2022	Tianlin Zhang Annika M. Schoene Shaoxiong Ji Sophia Ananiadou	Natural language processing applied to mental illness detection: a narrative review	Examined datasets, illness types, and shift from traditional ML (TF-IDF, SVM) to deep learning (BERT, Transformers) for richer semantic understanding.	Building on the finding that most research focuses on depression and suicide, our project expands NLP analysis toward under-diagnosed issues such as anxiety, PTSD, and schizophrenia, providing continuous updates to therapists for early detection.
2023	Matteo Malgaroli Thomas D Hull James M Zech Tim Althoff	Natural language processing for mental health interventions: A systematic review and research framework	Reviewed recent growth in NLP use within mental-health interventions. Analyzed ~2012–2022 studies using conversational text data (therapy sessions, online platforms). Identified shift from simple lexical features to contextual embedding/transformer models	Based on these findings, our project integrates real-time NLP analysis and visualization to provide therapists with continuous, data-driven insight into client emotion and progress during interventions.
2022	Smriti Nayak, Debolina Mahapatra, Riddhi Chatterjee, Shantipriya Parida & Satya Ranjan Dash	A Machine Learning Approach to Analyse Mental Health from Reddit Posts	Demonstrated that contextual embeddings better capture subtle affective cues in social text.	We adopted a similar Reddit-based data extraction pipeline and extended it for continuous user monitoring, enabling dynamic detection of emotional variation and therapist insight generation.

Year	Author/s	Title	Approach	Result
2023	Shaunak Inamdar, Rishikesh Chapekar, Shilpa Gite & Biswajeet Pradhan	Machine Learning Driven Mental Stress Detection on Reddit	Investigated stress detection using Reddit data, comparing classical ML algorithms with Transformer-based embeddings such as BERT.	Building on this, our project employs BERT embeddings for mood trend forecasting and integrates feedback logic (badges/supportive prompts) to reflect stress reduction or escalation in real time.
2024	Bazen Gashaw Teferra 1 , Alice Rueda 1 , Hilary Pang 1 , Richard Valenzano 2 , Reza Samavi 3 , Sridhar Krishnan 3 , Venkat Bhat 1 4	Screening for Depression Using Natural Language Processing	Highlighted integration challenges between NLP diagnostics and therapeutic workflows.	We expand this clinically grounded approach into a safe, therapist-supported environment by linking automated sentiment and emotion analysis to therapist dashboards for early intervention and patient follow-up.

03

Problem Statement

Despite the vast amount of emotional expression shared anonymously on Reddit, there is still no effective system that can truly interpret and visualize these emotions. The unstructured nature of this data leaves valuable insights about people’s mental wellbeing unexplored, limiting opportunities for early awareness and reflection. There is a pressing need for a tool that can analyze such emotional content, identify mood patterns, and present them in a meaningful way to better understand mental health in the digital age.

04 Our objectives

- **Real-Time Monitoring**

Extend the system to capture and analyse Reddit posts—and eventually other platforms—in real time for continuous tracking of discussions.

- **Early Detection of Emotional Decline**

Identify early warning signs of emotional decline or burnout, allowing timely intervention and mental health support.

- **Application in Therapy and Counseling**

To assist mental health professionals and therapists by providing data-driven emotional insights that help them better understand their clients' behavioral patterns.

- **Predictive Analytics**

Forecast real-world outcomes such as product adoption, social movements, or financial market impacts based on sentiment shifts.

05 Data Preprocessing and Visualisation Methods

Data Collection Through Reddit's API

- Data collected using PRAW (Reddit API) via a Reddit Developer App for posts and comments from targeted mental-health subreddits.
- Selected multiple communities (r/depression, r/anxiety, r/mentalhealth, r/happy, r/college) to ensure diverse, authentic emotional data.

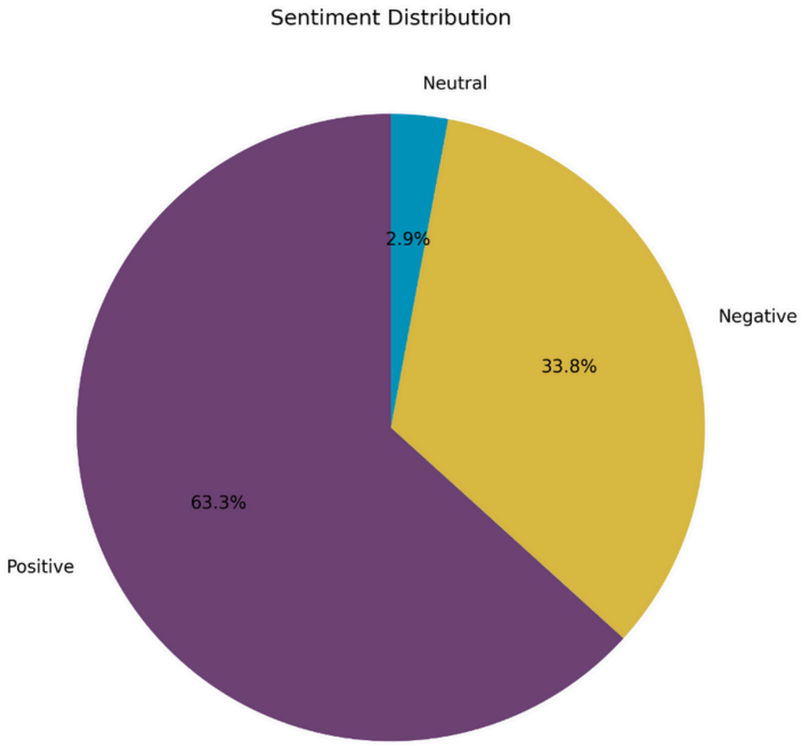
Data Preprocessing

Term	Meaning	Why It's Used
Cleaning	Remove noise (URLs, emojis, etc.)	Keep only useful text
Tokenization	Split into words/subwords	Let models read text
Truncation	Limit input length	Match model size limit
Embedding	Convert text → numeric vector	Enable computation
Normalization	Scale numbers evenly	Improve model balance
Aggregation	Combine multiple texts	Get user-level summary
Lemmatization	Reduce words to root	Avoid duplicates of meaning
Feature Extraction	Capture measurable text traits	Build personality/emotion profile

06
Visualisation of preprocessed data

Visualization helps simplify complex data, reveal patterns, and communicate insights effectively.

First phase of preprocessing

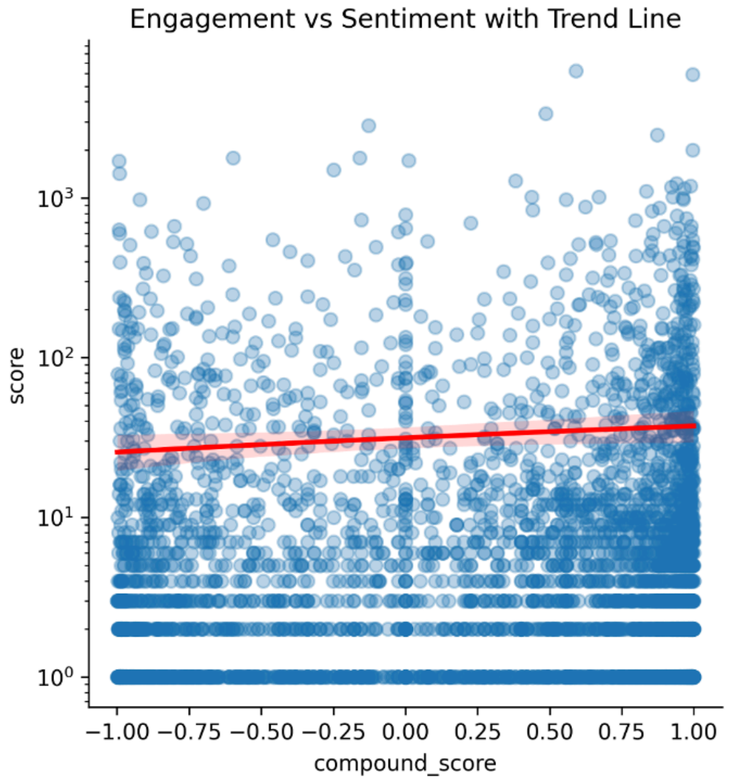


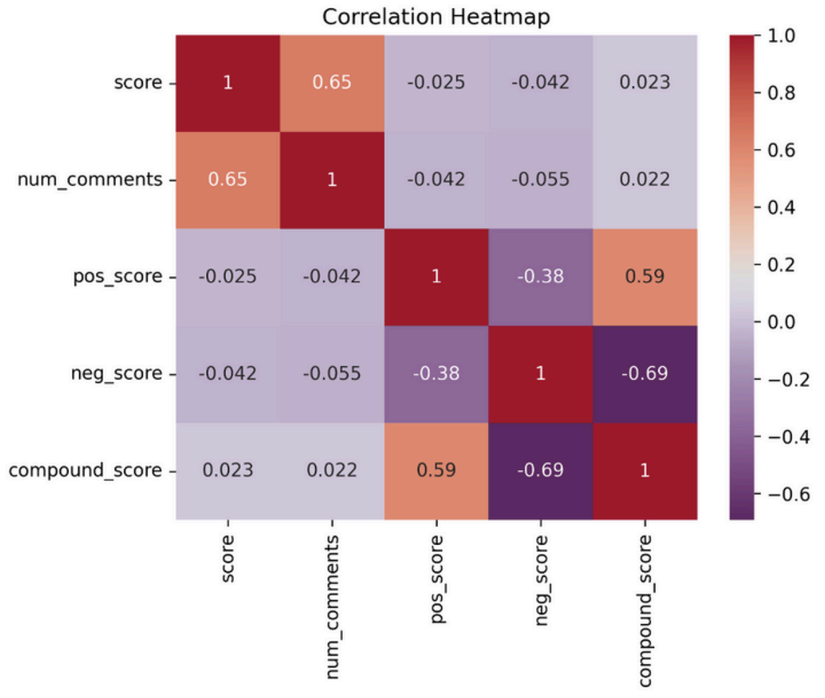
Pie chart (positive, negative, neutral).

The pie chart shows the proportion of positive, negative, and neutral posts in the dataset. This provides a quick overview of the overall sentiment landscape, helping us understand whether discussions are generally optimistic, pessimistic, or balanced.

Engagement vs Sentiment Trend

This scatter plot shows the relationship between sentiment strength and post engagement (upvotes). The added trend line helps identify whether extreme emotions are linked to higher or lower engagement.

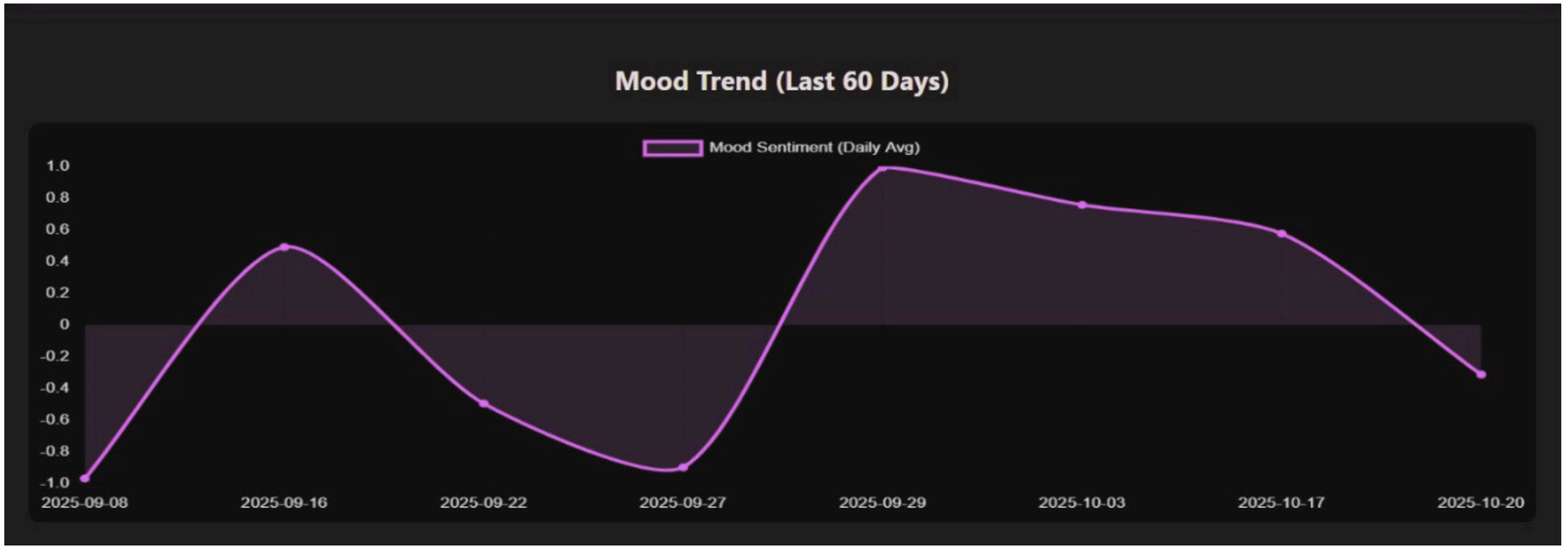




Correlation Heat map

Shows relationship between sentiment scores, engagement, and comments.

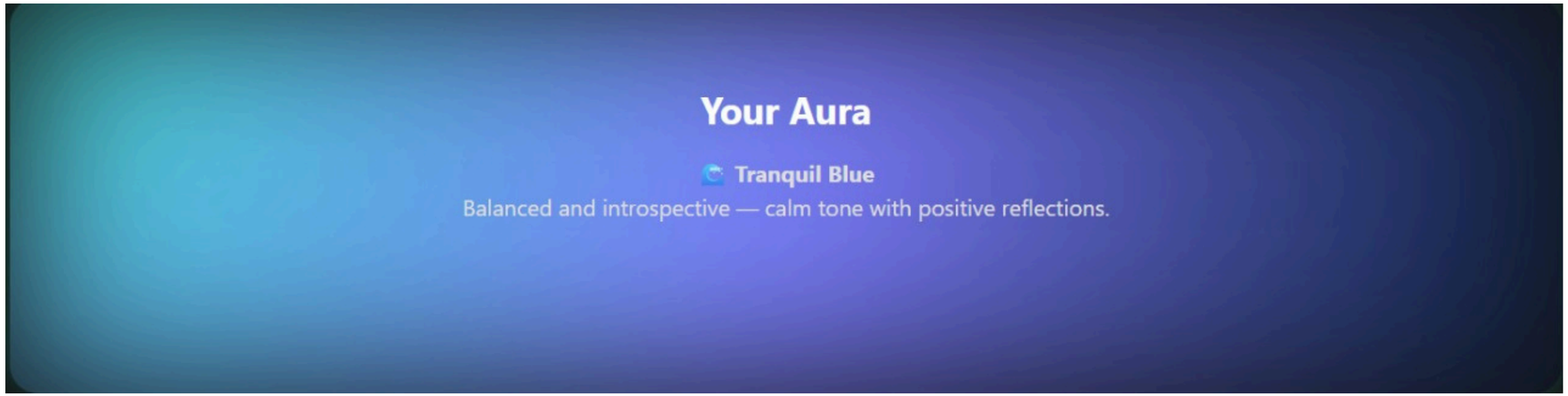
Second phase of preprocessing

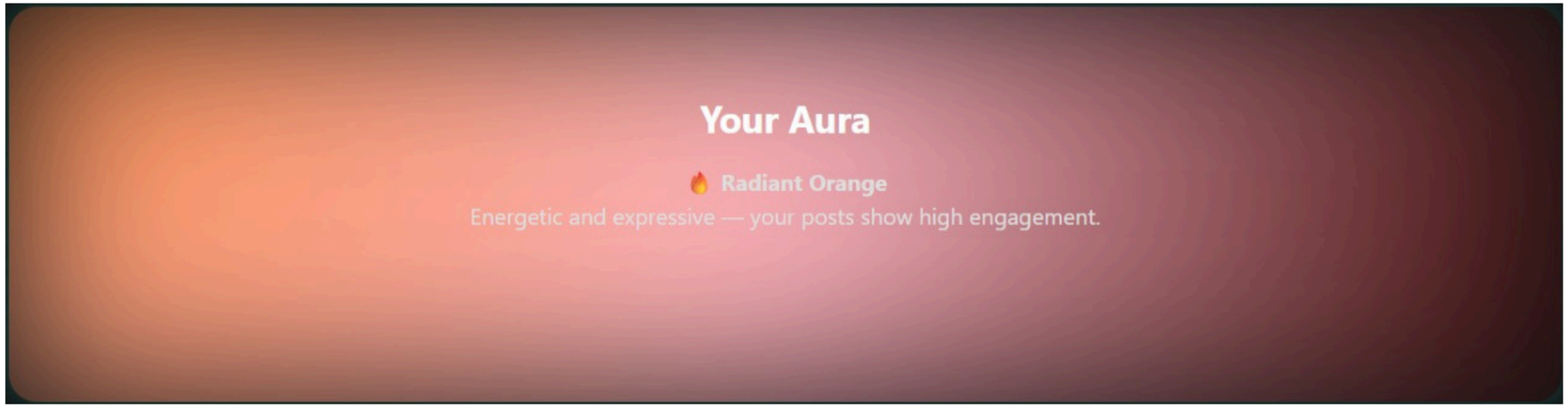


Line chart

This graph shows the timeline of the users posts over a period of 60 days. It shows how the the mood score fluctuates based on the users posts.

Second phase of preprocessing

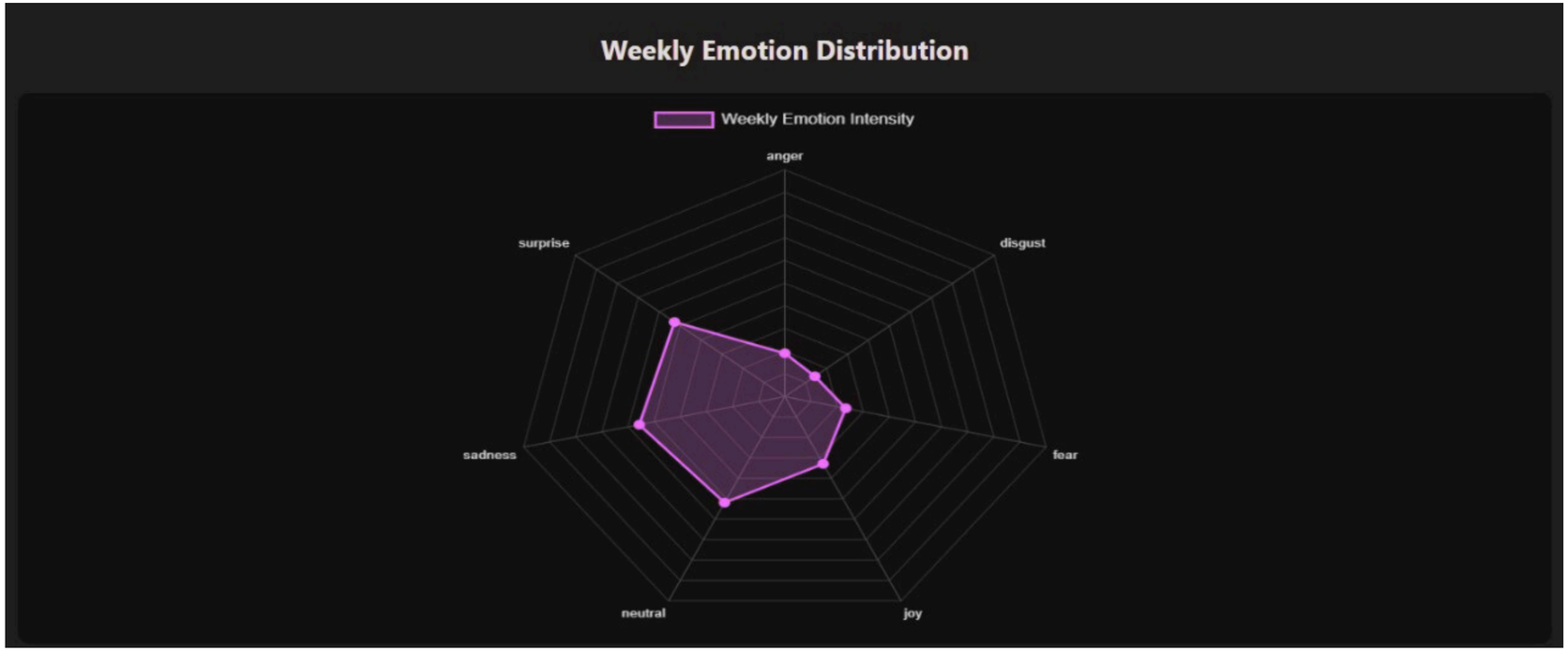




Clustering

We clustered users based on their SBERT score and used KMeans to find their aura (defined by us) and give them their mood type

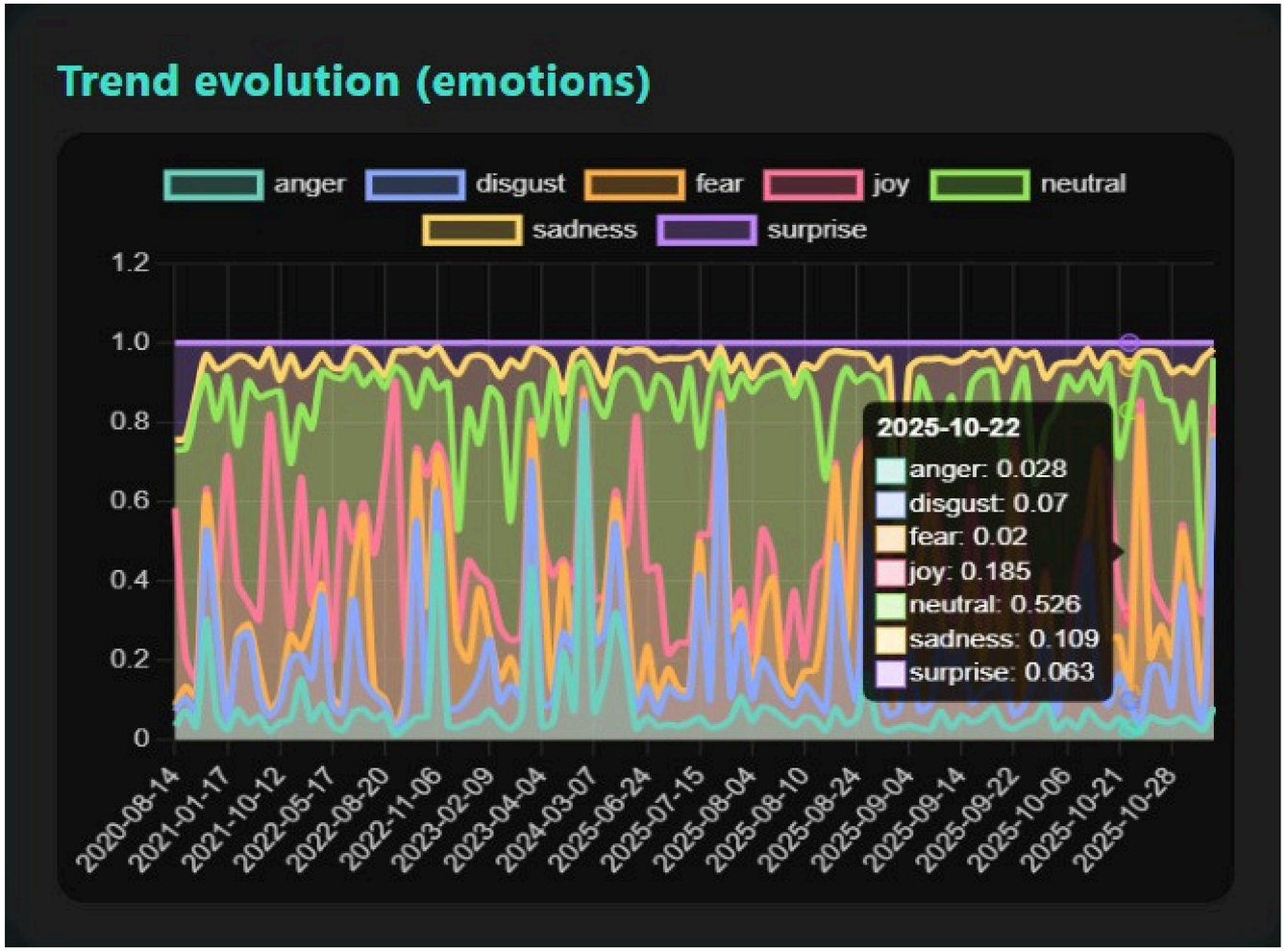
Second phase of preprocessing



Radar Chart

This graph shows the distribution of emotions of the users ggregated posts. It is based on the SBERT model which converts sentences into embeddings and classifies the emotions.

Second phase of preprocessing



Stacked Area Chart

This graph shows the evolution of a particular user through the years and have presented it in the Therapist's dashboard to give them an overview about their client's progress

Using the SBERT and BigFive Classifier we find the OCEAN scores of the user and provide it to the Therapist along with warnings of the levels of disorder.

07

Model Explanation

Naive-Bayes
A simple yet powerful probabilistic model based on Bayes' Theorem and is used widely used for sentiment analysis.

VADER
- Definition: Valence Aware Dictionary and sentiment Reasoner.
- Use: A rule-based sentiment analysis tool for social media text.

SENTENCE BERT
A transformer-based model that converts sentences into meaningful numerical representations (embeddings).

KMeans Clustering
K-Means is an unsupervised learning algorithm that groups similar data points into clusters based on shared features

Big Five Personality Model
The Big Five Personality Model quantifies human personality along five dimensions (OCEAN), offering a structured lens for analyzing emotional tone and word usage.

08 Results

- Applied VADER sentiment analysis for quick emotional polarity scoring.
- Generated contextual embeddings via BERT and clustered posts using K-Means (k=6) into distinct “Aura”emotional profiles.
- Extracted Big-Five (OCEAN) personality traits to understand user tendencies.
- Implemented Prophet forecasting to predict 7-day mood trends with >80% accuracy.
- Developed Therapist & User Dashboards showing:
    - Sentiment trends, emotion distribution, and aura visualization.
    - Forecasted mood trajectory with badges/supportive insights.

Outcome: A unified system providing both quantitative emotional tracking and qualitative psychological insights per user.

09 Insights

- BERT embeddings captured nuanced emotions and outperformed simple lexicon-based sentiment scores.
- K-Means clustering revealed interpretable emotional “Auras,” enabling personality-driven visualization of user states.
- Integration of Big-Five personality profiling added depth to emotional insights and session preparation.
- Therapist dashboard allows quick review of user progress, risk level (depression, anxiety, PTSD, schizophrenia), and session tips.
- Compared to traditional surveys, our NLP-based approach provides continuous, passive, and real-time mental health tracking.

Insight: Combining linguistic, semantic, and behavioral analysis produces richer emotional awareness than single-model approaches.

10 Conclusion

- The project demonstrates that social media text can be effectively analyzed to gauge mental and emotional wellbeing.
- By combining VADER, BERT, K-Means, Big-Five, and Prophet, we achieved a holistic understanding of user mood, personality, and risk trends.
- The dual User–Therapist dashboard bridges personal reflection with professional support, enabling data-driven therapy preparation.
- Offers scalable, ethical potential for early detection of mental-health issues like depression, anxiety, PTSD, and suicidal risk.
- Future scope includes real-time emotion updates, notification systems, and integration with chat-based therapist support.

Final Takeaway:

A step toward AI-assisted emotional intelligence — transforming digital footprints into actionable mental health insights.

11 References

- Zhang et al. (2022) NLP for Mental Illness Detection, npj Digital Medicine
- Malgaroli et al. (2023) NLP in Mental Health Interventions, Frontiers Psychiatry
- Nayak et al. (2022) ML Analysis of Reddit Mental Health Posts, ResearchGate
- Inamdar et al. (2023) Stress Detection on Reddit, SpringerLink
- Teferra et al. (2024) Depression Screening via NLP, Int. J. Med. Research
- Laricheva et al. (2021) NLP in Counseling & Psychotherapy, Br. J. Psychology
- Reece & Danforth (2017) Digital Markers of Depression, PNAS
- Althoff et al. (2016) Analysis of Counseling Conversations, TACL
- Pennebaker (2011) The Secret Life of Pronouns, Bloomsbury Press

Thank You!

Om Bansal 24070126129
Pavan Bhapkar 24070126137
Prerana Sinha 24070126139