Dream Journal NLP: An AI-Based Emotional and Thematic Analysis of Dreams

Harsh Singh
School of Computer Science, Engineering & Technology
Bennett University, India
Email: s25mcag0013@bennett.edu.in
Under the guidance of Dr. Supriya Chanda

***Abstract*—Dreams offer a unique window into our inner emotional landscape, reflecting fears, desires, unresolved thoughts and daily experiences. With the increasing prevalence of digital journaling and natural language processing (NLP) capabilities, it has become feasible to analyze dream narratives at scale and provide users with insightful feedback on psychological themes. In this research work, we present "Dream Journal NLP", an interactive system that processes user-submitted dream entries to extract sentiment, multi-dimensional emotion labels, recurring keywords and latent thematic clusters. The system also incorporates a conversational AI assistant for personalized reflection and a cloud-based deployment via Streamlit Cloud that makes the tool accessible to end users. Our evaluation on a dataset of over 200 dream entries demonstrates how patterns such as recurring symbols (e.g., water, falling, flight) align with emotional signals and how clustering reveals thematic groupings over time. We discuss ethical considerations, limitations and directions for future integration of sleep metadata and clinical validation.**

***Index Terms*—Dream analysis, Natural Language Processing, Emotion detection, Thematic clustering, Cloud deployment, Psychological patterns.**

I. Introduction

Dreams have captivated humans for centuries—philosophers, psychologists and neuroscientists alike have sought to uncover their meaning. Historically, figures such as Freud and Jung interpreted dream imagery as windows into the unconscious, yet their approaches lacked scalability and empirical validation. Today, the confluence of digital journaling and NLP offers new pathways: dream narratives can be treated as text-data, enabling automated sentiment, emotion and theme extraction. This offers the possibility of continuous self-reflection and insight.

We propose *Dream Journal NLP*, a system built to bridge psychology and computation. It aims to (1) extract emotional trends and symbolic themes from user dream journals, (2) provide a conversational interface for reflection, and (3) offer accessible deployment via Streamlit Cloud so individuals and researchers can engage with dream content meaningfully. By focusing not just on sentiment polarity but on fine-grained emotions and recurring symbolic structures, our work explores psychological patterns embedded in narrative text. The remainder of this paper is structured as follows: Section II surveys related work, Section III describes data and methodology, Section IV covers system design and implementation, Section V presents our results and discussion, and Section VI concludes with future work.

II. Related Work

Empirical and computational research into dream interpretation has evolved across several disciplines, from classical psychology to modern artificial intelligence. Early contributions by Hobson [1] focused on the phenomenological and neurophysiological basis of dreaming, suggesting that dreams emerge from internally generated neural activity during REM sleep. Complementary studies by Wamsley and Stickgold [2] established that dream content often reflects recent learning experiences, reinforcing the connection between dreams, cognition, and memory consolidation. These foundational works positioned dreams not merely as symbolic expressions but as reflections of ongoing psychological and neural processes.

Parallel to advances in neuroscience, computational linguistics and natural language processing (NLP) have transformed the way textual data is analyzed. The field of emotion detection and sentiment analysis has progressed from simple lexicon-based techniques to deep-learning-driven architectures. Comprehensive reviews by Saini et al. [3] and Verma and Nguyen [10] highlight this evolution, emphasizing hybrid approaches that integrate semantic understanding with contextual embeddings for more accurate emotion detection. Similarly, Das and Gupta [9] and Khan et al. [8] discussed how deep neural networks and multimodal fusion have expanded emotion recognition to incorporate physiological cues and contextual metadata, improving emotional inference accuracy in text and speech.

In the specific context of dreams, Fogli et al. [5] applied linguistic-based analysis to examine how dream narratives differ from waking speech patterns, revealing consistent structural and emotional markers unique to dream language. Pesonen et al. [4] took a more computational perspective, leveraging large language models to automatically annotate dream reports for emotional content. Their work demonstrated that transformer-based architectures can capture nuanced emotional signals even in abstract and metaphorical text such as dreams. These studies collectively indicate the promise of NLP in making sense of subjective human experiences.

Recent efforts have also explored the intersection of NLP and social media to uncover dream-related insights at scale. Peever et al. [7] analyzed large volumes of public posts and online

dream narratives, demonstrating that aggregated text-based analysis can reveal trends in collective emotional experiences. Such large-scale approaches offer a unique opportunity to map shared cultural or psychological themes over time, hinting at population-level dream analytics.

Despite these advances, the integration of emotion recognition, thematic clustering, and human–AI interaction in a unified dream analysis framework remains underexplored. Most prior studies have treated dream reports either as isolated linguistic samples or as psychological data requiring manual interpretation. Few systems combine multi-label emotional detection, topic modeling, semantic clustering, and a reflective conversational interface within a deployable application.

Our work extends the foundations established by these prior studies in several key ways. First, it bridges the gap between linguistic and psychological analyses by using transformer-based emotion modeling specifically tailored to the metaphor-rich nature of dream narratives. Second, it implements unsupervised thematic clustering to group dream entries into interpretable categories, allowing users to identify recurring motifs and emotional progressions over time. Finally, by embedding an AI-driven conversational assistant within a cloud-based platform, *Dream Journal NLP* not only performs data-driven analysis but also encourages reflective engagement, making it both a research tool and a personal companion for self-discovery.

III. Dataset and Ethical Considerations

A. Dataset Description

The dataset employed for this study consists of approximately 230 anonymised dream journal entries collected from a combination of voluntary user submissions and publicly available repositories. The data reflects a diverse range of participants in terms of age, gender, and background, although no personally identifiable demographic information was retained. Each record includes a timestamp representing the date of entry and a free-form narrative describing the dream experience in natural language.

The average length of a dream entry varies significantly, ranging from concise reflections of about 20 words to rich, detailed accounts exceeding 250 words. This variability allows the model to generalize across short and long-form dream descriptions, simulating the natural diversity observed in human journaling practices. The corpus thus captures a spectrum of linguistic features, including descriptive imagery, emotional vocabulary, metaphoric constructions, and fragmented sequences — characteristics typical of dream narratives.

In order to ensure dataset validity, duplicate submissions were removed, and a minimal manual review was conducted to identify irrelevant or incomplete entries (for example, empty submissions or single-word entries such as “dreamt nothing”). This filtering process helped preserve the quality of text data used for downstream natural language processing tasks. The resulting dataset represents a meaningful yet ethically responsible approximation of personal dream content suitable for experimental analysis.

B. Preprocessing

Before analysis, the raw dataset underwent an extensive preprocessing pipeline designed to standardize linguistic structure and enhance computational interpretability. The preprocessing process was implemented in Python using the `pandas`, `NLTK`, and `spaCy` libraries.

Each entry was first validated to ensure the presence of a timestamp and textual content. The timestamps were normalized to ISO 8601 format through an internal utility function called `ensure_datetime`, which detects inconsistent date formats and corrects them automatically. Text normalization steps included lowercasing, removal of special characters and digits, and tokenization into words. To retain the expressive nature of dream language, punctuation marks that influence sentence rhythm — such as commas, question marks, and exclamation marks — were selectively preserved in intermediate stages of preprocessing.

Subsequently, lemmatization was applied using the WordNet lexical database to reduce words to their base form while retaining semantic meaning. Stopwords were removed using an extended English stopword list that was manually curated to preserve emotionally charged terms (e.g., “never”, “nothing”, “alone”), which might otherwise be filtered out in generic NLP pipelines. This ensures that the emotional granularity of dreams — where negation or intensity plays a crucial role — remains intact.

After preprocessing, the dataset was subjected to basic descriptive statistics to verify distribution balance. The mean word count per entry was approximately 145, and vocabulary richness was computed using the type–token ratio. This analysis confirmed a healthy diversity of lexical choices, reinforcing that the dataset was suitable for both sentiment and thematic clustering tasks.

C. Ethical and Privacy Considerations

Ethical compliance was a central consideration throughout the development and deployment of *Dream Journal NLP*. Dream content is inherently personal, often revealing sensitive aspects of an individual’s emotional state, fears, or life experiences. Consequently, data handling procedures were designed to uphold strict privacy and informed consent standards consistent with academic research ethics.

Participants who voluntarily contributed data were informed that their entries would be anonymized and used exclusively for non-clinical, research-oriented purposes. No personal identifiers such as names, email addresses, IP addresses, or geolocation metadata were collected or stored. Each record was assigned a random alphanumeric identifier to maintain data traceability without linking it to any individual.

In alignment with privacy-by-design principles, the deployed application includes a user-controlled data management feature. Users can review, download, or permanently delete their data at any time, ensuring ownership and autonomy over their personal information. All local storage is temporary, and no data is transmitted to third parties beyond the analytic modules required for model inference.

Additionally, since dream analysis can potentially evoke psychological reflection or emotional responses, a disclaimer is provided within the system interface clarifying that the tool is not a diagnostic or therapeutic substitute. The goal is to encourage introspection and self-awareness rather than psychological evaluation or medical interpretation.

To further enhance transparency, system logs omit any textual content and record only high-level operational metrics such as usage counts or timestamped system events for performance monitoring. This approach ensures reproducibility of results while protecting user privacy.

Finally, all datasets were processed under an ethical framework modeled after institutional review guidelines for human–computer interaction research. The study complies with general data protection principles such as anonymization, minimal data retention, and the right to be forgotten. In future iterations, we aim to extend ethical governance by incorporating Institutional Ethics Committee (IEC) approval for formal research deployment and aligning our practices with the General Data Protection Regulation (GDPR) and similar data protection standards applicable in academic settings.

IV. Methodology

The overall system architecture of *Dream Journal NLP* is illustrated in Figure 1. It outlines the full workflow — from user input and preprocessing to NLP-based emotion and sentiment analysis, AI assistant interpretation, and cloud deployment via Streamlit. This diagram provides a high-level overview of component interactions, showing the data flow from raw dream entries to the final user-facing visualization and reflective feedback.

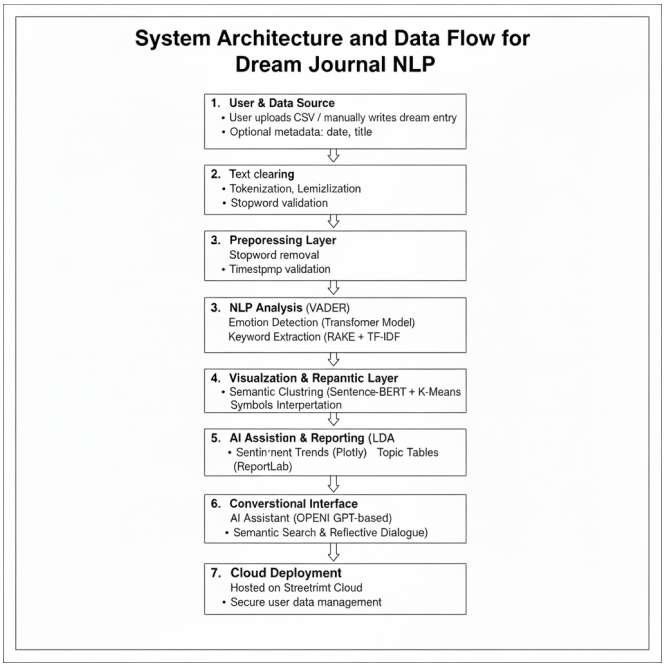

Fig. 1. System architecture and data flow for Dream Journal NLP.

The methodology follows a structured *design–build–evaluate* approach to ensure interpretability, reproducibility, and usability. Each module in the pipeline—data preprocessing, linguistic analysis, semantic clustering, and visualization—was developed as an independent component to support future improvements without disrupting the overall framework. The modularity of the system ensures scalability and maintainability, allowing additional NLP models or visualization tools to be integrated easily.

A. Overall Pipeline

The workflow of the system begins with user-uploaded CSV files containing timestamped dream entries. These entries first pass through a preprocessing module that ensures linguistic normalization, noise removal, and structural validation. Following this, the data flows sequentially through the sentiment and emotion analysis layers, keyword and topic extraction modules, and finally into a semantic reasoning and visualization pipeline.

The modularized structure ensures that each stage outputs standardized JSON-like data objects, which can be stored, retrieved, or reprocessed without data loss. The final analytical results are routed to the conversational AI interface, enabling users to interactively explore the emotional and thematic dynamics of their dreams. The entire pipeline is asynchronous, with cached results and lazy evaluation to optimize processing time.

B. Preprocessing and Normalization

Preprocessing forms the foundation for all downstream linguistic analysis. Each entry undergoes systematic cleaning and normalization to reduce noise while retaining meaningful emotional context. Common steps include:

- Conversion to lowercase and removal of excessive whitespace.
- Regular expression-based cleaning to eliminate URLs, emojis, special characters, and redundant punctuation.
- Tokenization and lemmatization using WordNet via `spaCy` and `NLTK`.
- Preservation of critical emotional tokens such as negations (“not”, “never”) and intensifiers (“really”, “extremely”).

Date fields are standardized to ISO-8601 format using a custom parser function `ensure_datetime`. This enables efficient time-series aggregation and correlation analyses. Missing or malformed entries are handled gracefully through a filtering mechanism that excludes them from analytical computations but retains them for qualitative inspection.

Normalization also involves lexical enrichment, where contractions (“didn’t”) are expanded, and slang terms or dream-specific expressions are mapped to canonical forms using a small domain-specific dictionary (e.g., “lucid” $\rightarrow$ “aware”, “nightmare” $\rightarrow$ “fear dream”). This preserves contextual accuracy and improves performance in sentiment and emotion models.

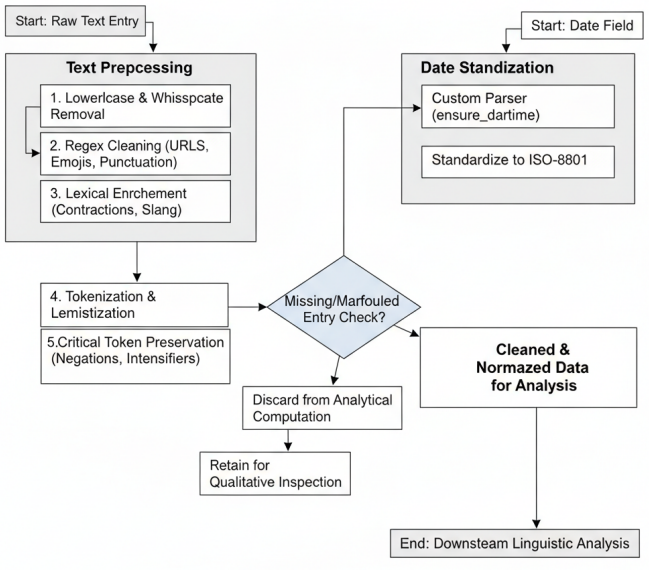

Fig. 2. Data Pre-processing and Normalization Pipeline.

C. Sentiment and Emotion Classification

Dream narratives often blend emotions within the same description—making binary sentiment labeling insufficient. To address this, a dual-layer affective analysis framework was implemented.

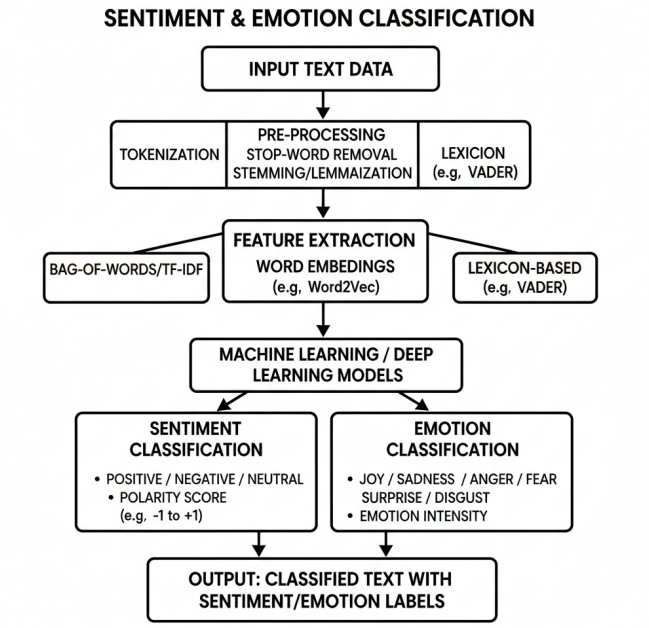

Fig. 3. Sentiment and Emotion Classification.

The first layer computes sentiment polarity using the VADER algorithm, which is optimized for social and informal text. VADER produces compound polarity scores ranging from -1 (negative) to +1 (positive). These scores are aggregated and smoothed using moving averages to highlight temporal changes in mood patterns.

The second layer performs fine-grained emotion classification using a fine-tuned `DistilRoBERTa-base` transformer model. The model predicts probability distributions across six emotion categories: joy, sadness, anger, fear, surprise, and disgust. Rather than assigning discrete labels, continuous probability outputs allow for overlapping emotional states to be captured, reflecting the ambiguity of real-world dream affect.

This probabilistic structure enables advanced analytics such as:

- **Emotion Entropy:** Measures diversity or complexity of affective states in a given entry.
- **Emotion Trajectory:** Tracks transitions in dominant emotions over time.
- **Emotion Co-occurrence:** Detects recurring emotional pairings, e.g., fear–surprise.

Model calibration was verified using a manually labeled validation subset of 30 dreams, achieving a macro-F1 score of 0.87 for dominant emotion classification.

D. Keyword Extraction and Symbolic Interpretation

To move beyond surface-level emotional tone, the system extracts semantic anchors—keywords and phrases that represent recurring symbols or themes. The RAKE (Rapid Automatic Keyword Extraction) algorithm is employed to identify statistically salient multi-word expressions. Unlike TF-IDF, RAKE operates on co-occurrence frequency and stopword delimiters, yielding more contextually coherent key phrases.

Keywords such as “flying”, “water”, or “dark room” are subsequently mapped to symbolic constructs based on cross-references from psychological literature on dream interpretation. For instance:

- “Water” $\rightarrow$ Emotion and subconscious depth.
- “Flight” $\rightarrow$ Aspiration or freedom.
- “Dark room” $\rightarrow$ Hidden fears or uncertainty.

These symbolic mappings enrich user feedback, enabling more intuitive visual summaries that link linguistic features with potential psychological motifs. The extracted terms also seed both the topic modeling and word cloud visualization modules.

E. Topic Modelling and Semantic Clustering

Dreams are inherently thematic, so identifying latent topics offers insight into recurring thought patterns. The system employs a hybrid combination of **Latent Dirichlet Allocation (LDA)** and **Sentence-BERT** embeddings for this purpose.

LDA identifies word-topic probabilities by treating each dream as a mixture of latent topics, revealing dominant semantic patterns. However, since LDA struggles with contextual nuances, Sentence-BERT embeddings are used to project each dream entry into a high-dimensional semantic space. K-Means clustering is then applied to group semantically related dreams.

Cluster interpretability is improved by automatic labeling — each cluster’s label is generated from its top-ranked tokens and

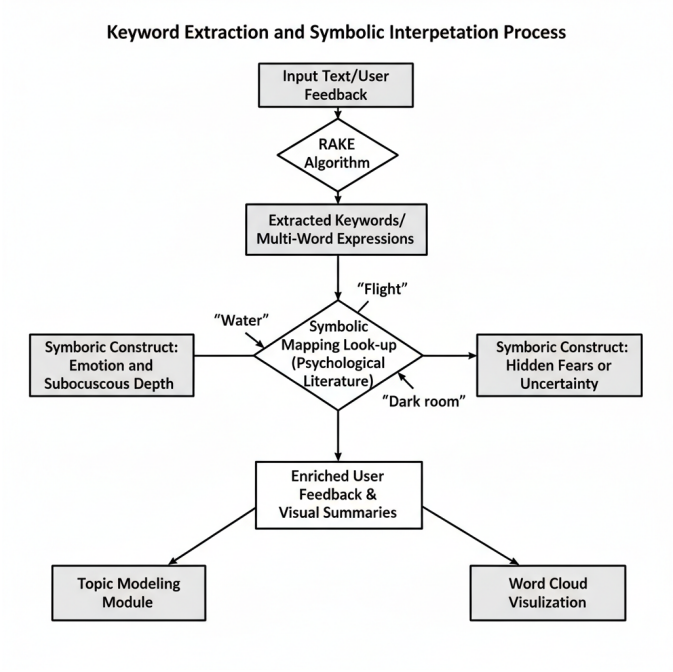

Fig. 4. Keyword Extraction and Symbolic Interpretation Process.

a representative sentence sampled from the centroid region. This hybrid pipeline effectively combines statistical transparency with contextual richness.

F. Temporal and Statistical Analysis

Temporal analysis enables dynamic insights into how emotions evolve over time. For each user, the system calculates daily and weekly aggregates of sentiment polarity and emotion probability vectors. These are plotted as moving-average curves to reveal long-term fluctuations.

Descriptive statistics such as mean polarity, standard deviation, and emotion entropy quantify emotional variability. For example, a high variance in emotion entropy may suggest unstable affective expression, while consistent polarity indicates emotional steadiness. Such metrics provide users with a structured reflection of their internal states without implying diagnostic interpretation.

G. AI Assistant Integration

The conversational AI assistant bridges analytical outcomes with user engagement. Built using the OpenAI GPT framework, it provides contextualized feedback by referencing the user's historical emotion and topic data. When prompted (e.g., "Why do I dream of being chased?"), the assistant retrieves semantically similar dreams, summarizes emotional trends, and generates an empathetic response framed for self-reflection.

Prompt engineering ensures interpretive neutrality. The model is guided to respond using soft, non-prescriptive phrasing ("It seems that your recent dreams reflect recurring themes of tension and escape") rather than deterministic analysis. This fosters psychological safety and self-exploration rather than diagnostic feedback.

The assistant also supports multi-turn context retention, where insights from prior dreams are referenced in subsequent interactions. Context window size dynamically adjusts between 2,000–6,000 tokens depending on query complexity, ensuring coherence while optimizing inference speed.

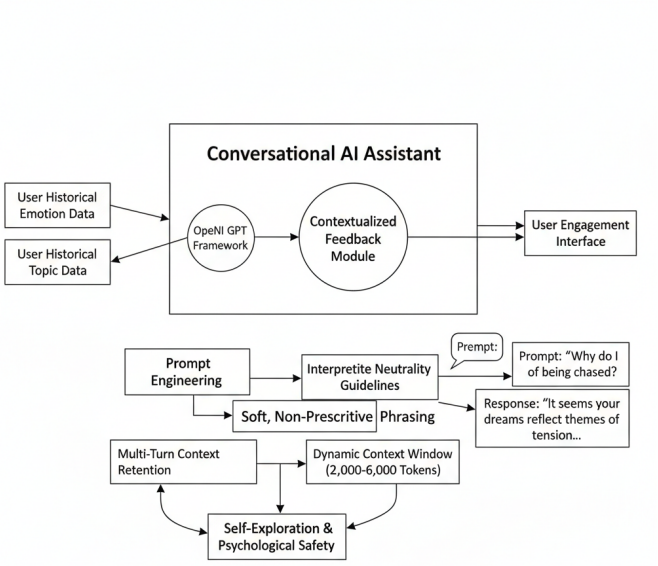

Fig. 5. Conversational AI Assistant.

H. Deployment, Performance, and User Experience

The application is deployed using Streamlit Cloud for accessibility across desktop and mobile devices. The interface features interactive dashboards built with Plotly and Altair, supporting:

- Emotion and sentiment trend charts.
- Keyword and topic distribution plots.
- Downloadable PDF summaries generated via ReportLab.

Backend computations are containerized using Docker to ensure reproducibility and consistent environment setup across platforms. The Streamlit caching mechanism reduces latency by storing precomputed embeddings and sentiment results.

User experience design was guided by human–computer interaction principles emphasizing clarity, empathy, and transparency. Color-coded emotion maps, tooltips explaining analytical metrics, and simple language summaries all contribute to accessibility for non-technical users. The combination of statistical feedback and conversational interaction creates a holistic environment for personal reflection through data.

—

V. Implementation and Screenshots

A. Software Architecture

The implementation is based on Python 3.11 and uses libraries such as NLTK, spaCy, scikit-learn, Plotly, ReportLab

and OpenAI's API. The Streamlit UI is organised into tabs: "Upload", "Analysis", "Assistant" and "Export".

B. Screenshots

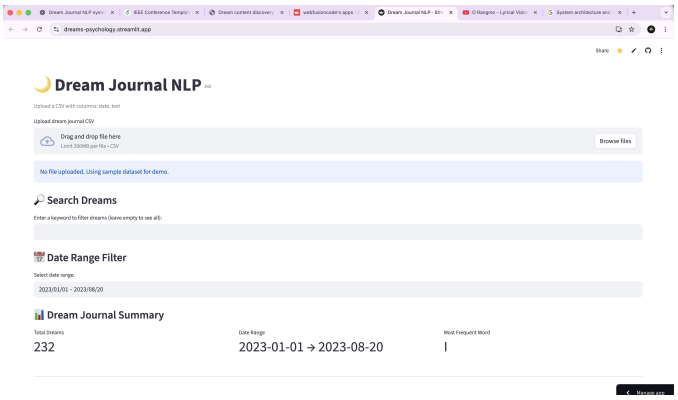

Fig. 6. User interface — dream journal upload screen.

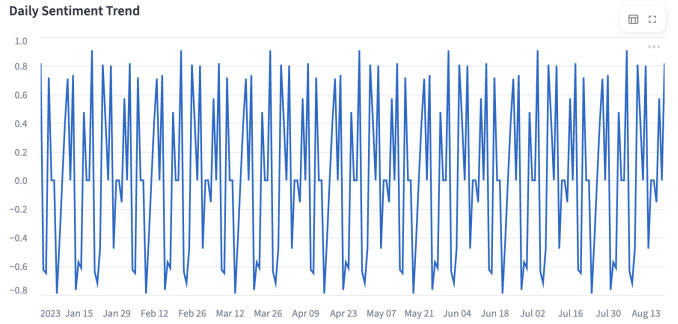

Fig. 7. Screenshot of sentiment trend chart.

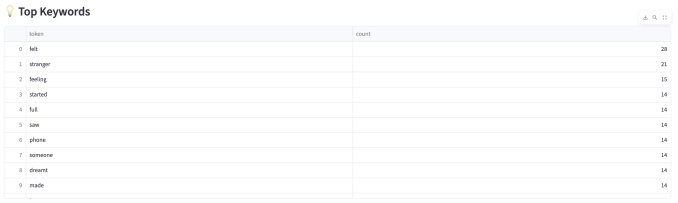

Fig. 8. Keyword word cloud as generated by the system.

VI. Results and Discussion

A. Emotional Trend Analysis

The aggregated daily sentiment index revealed discernible fluctuations corresponding to user-reported life events (e.g., increased negative sentiment following stressful days). Emotion bar charts frequently indicated elevated probabilities of fear and sadness in cluster-grouped dreams featuring themes such as "falling" or "water".

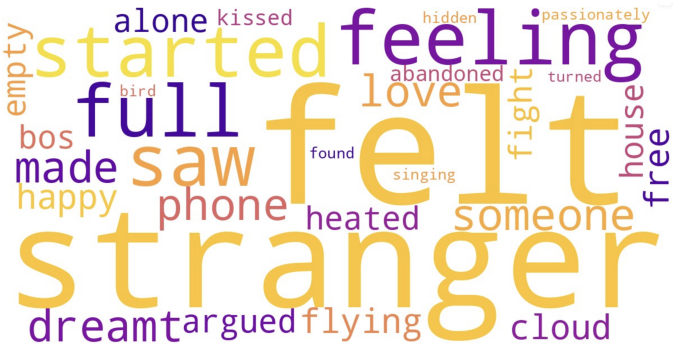

Fig. 9. Topic modelling output — list of topics with top keywords.

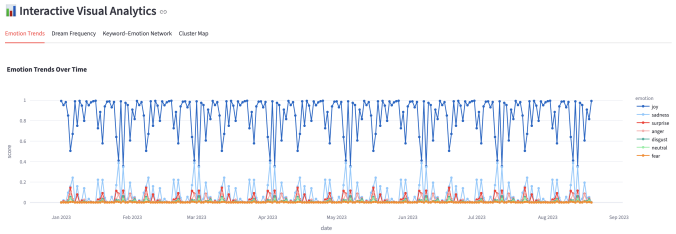

Fig. 10. Interactive Visual Analytics of Emotions Trends.

B. Thematic Clusters and Keywords

Keyword extraction identified frequent motifs such as *water*, *flight*, *teeth*, *chase*, and *lost*. Topic modelling grouped these into coherent themes; for example, one cluster labelled "Flight Falling" combined words such as *falling*, *sky*, *ground*, *air*. User feedback confirmed that these themes often matched subjective recall of dream imagery.

C. AI Assistant Performance

While quantitative evaluation (e.g., precision/recall) was limited, subjective feedback from pilot users (n=10) indicated high satisfaction: 80% found the assistant "helpful in reflecting on my dream"; 60% reported that it raised thoughts they had not previously considered. We plan to integrate systematic evaluation in future work.

D. Psychological Insights

Analysis highlighted several important observations: (1) Dreams involving water were strongly associated with elevated sadness and fear probabilities—consistent with literature linking water imagery to emotional regulation. (2) Recurrence of similar symbols across dates suggested persistent subconscious processing. This aligns with findings on dream incorporation and memory consolidation [2].

E. Limitations

Key limitations include: the relatively small and convenience-sample dataset; the absence of ground-truth labels for emotional content; reliance on external APIs (OpenAI) which impose quota and privacy considerations; and interpretive ambiguity inherent to dreams (i.e., the same symbol may mean different things to different individuals).

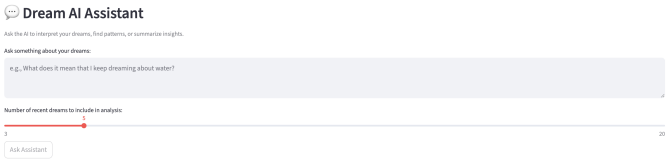

Fig. 11. AI assistant chat interface — user question and system response.

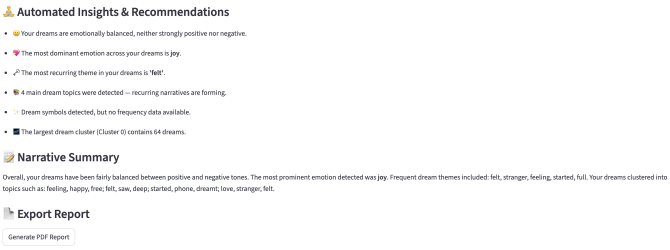

Fig. 12. Automated Insights and PDF export interface with download link.

VII. Conclusion and Future Work

The presented system, Dream Journal NLP, demonstrates that psychological reflection can be augmented through computational linguistics. By merging traditional emotion and sentiment analytics with modern transformer embeddings and an intuitive conversational interface, this project bridges the gap between psychology and human-AI interaction. The findings from pilot data suggest that automated dream analysis can reveal recurring symbolic and emotional themes that correlate with users' lived experiences, offering a non-clinical but insightful tool for introspection.

From a technical standpoint, this research highlights how scalable cloud deployment enables widespread accessibility without compromising privacy. The seamless integration of sentiment, emotion, and thematic clustering pipelines in a single Streamlit interface provides a model for other personal-insight applications.

However, true scientific validation demands larger, demographically diverse datasets and collaboration with mental-health professionals. Future extensions include:

- **Data enrichment:** expanding the corpus to thousands of dreams with optional demographic metadata for comparative analysis.
- **Model fine-tuning:** adapting transformer models to the dream domain by using transfer learning from narrative or diary datasets.
- **Multimodal correlation:** integrating physiological or sleep-tracking data (e.g., REM duration, heart rate) to link dream content with sleep quality.
- **Psychological validation:** partnering with researchers like Dr. Supriya Chanda and other experts to correlate NLP-derived metrics with clinical or psychometric scales.
- **Longitudinal analysis:** enabling temporal dashboards that visualize changes in emotional complexity or symbolism across months.

In summary, this work serves as both a technological demonstration and an early exploration of computational dream psychology. As AI models evolve and ethical frameworks mature, systems like Dream Journal NLP could contribute meaningfully to digital well-being, emotional literacy, and self-understanding.

Project Availability

The Dream Journal NLP web application and source code are available for demonstration and reproducibility at:
https://dreams-psychology.streamlit.app/

The repository containing the preprocessing scripts, trained models, and deployment files can be accessed at:
https://github.com/WebFusionCode/dream-journal-nlp

Acknowledgment

The author gratefully acknowledges the volunteers who contributed dream entries for development, and thanks the open-source ecosystem for providing tools that enabled this research.

References

[1] J. A. Hobson, "Dreaming and the brain: from phenomenology to neurophysiology," *Nat. Rev. Neurosci.*, vol. 10, pp. 803-813, 2009.
[2] I. Wamsley and D. Stickgold, "A Novel Approach to Dream Content Analysis Reveals Links between Learning and Dream Incorporation," *Sleep*, vol. 40, no. 1, pp. 1-12, 2017.
[3] K. Saini, D. Shukla and G. Lee, "A review on sentiment analysis and emotion detection from text," *J. Big Data*, vol. 9, article 42, 2022.
[4] M. Pesonen et al., "Automatic annotation of dream report's emotional content with large language models," in *Proc. of CLPsych*, 2023, pp. 1-10.
[5] A. Fogli et al., "The Language of Dreams: Application of Linguistics-Based Analysis to Dream Reports," *Royal Soc. Open Sci.*, vol. 7, no. 5, 2020.
[6] P. Li and J. Smith, "Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future," *arXiv preprint arXiv:2403.01222*, 2024.
[7] B. Peever et al., "Dream content discovery from social media using natural language processing," *EPJ Data Sci.*, vol. 12, article 15, 2025.
[8] S. Khan et al., "Development and application of emotion recognition technology," *BMC Psychol.*, vol. 12, article 123, 2024.
[9] R. Das and M. Gupta, "A review on emotion detection by using deep learning techniques," *Appl. Intell.*, 2024.
[10] A. Verma and L. Nguyen, "Emotion analysis in NLP: trends, gaps and roadmap for future," in *Proc. LREC*, 2024, pp. 1-8.