Discovering User Behavior Patterns and Discussion Topics in Reddit ChatGPT Communities Through Clustering and Text Mining

Mahliq Obie
CS 4412 - Data Mining
Kennesaw State University
jobie1@students.kennesaw.edu

***Abstract*—This project investigates user behavior patterns and thematic structures within Reddit discussions about ChatGPT using unsupervised learning techniques. By applying clustering algorithms (K-Means, Hierarchical, DBSCAN) and text mining methods (TF-IDF, Topic Modeling) to a dataset of approximately 50,000 Reddit comments from four subreddits, we aim to discover natural user segments, identify emergent discussion topics, and characterize community-specific engagement patterns. This discovery-focused approach seeks to uncover hidden structures in social media discourse surrounding artificial intelligence tools, providing insights into how different online communities discuss and engage with emerging AI technologies.**

I. Dataset Description

A. Dataset Overview

Name: Reddit ChatGPT Comments Dataset
Source: GitHub Repository - https://github.com/Armita84/ChatGPT-Dataset-Reddit
Size: Approximately 50,000 comments collected from 4 distinct subreddits
Data Collection Period: Comments gathered during the initial public release and adoption period of ChatGPT

B. Dataset Description

This dataset represents a comprehensive collection of user-generated comments discussing ChatGPT across multiple Reddit communities. Reddit is a social media platform organized into topic-specific communities called subreddits, where users engage in threaded discussions. The dataset captures organic conversations, opinions, questions, and experiences related to ChatGPT, providing a rich source for understanding public perception and engagement patterns with AI tools.

The dataset is particularly valuable for pattern discovery because it contains authentic, unfiltered user discourse from diverse communities with varying technical backgrounds, interests, and perspectives on AI technology. Unlike curated or synthetic datasets, this represents real-world social media interactions that exhibit natural clustering patterns based on user behavior and content characteristics.

C. Key Features and Attributes

Table I presents the dataset schema with detailed descriptions of each attribute.

TABLE I
Dataset Schema

Feature	Type	Description
comment_id	String	Unique identifier for each comment
comment_parent_id	String	ID of parent comment (for threading)
comment_body	Text	Full text content of the comment
subreddit	Categorical	Source subreddit community

Features for Analysis:

- **comment_body:** Primary text data for content analysis, topic modeling, and sentiment extraction. This field contains the actual discussion content ranging from technical questions to personal experiences with ChatGPT.
- **comment_parent_id:** Enables analysis of conversation threads, reply patterns, and discussion depth. Useful for understanding user engagement styles and community interaction dynamics.
- **subreddit:** Categorical feature representing the source community. Allows for comparative analysis across different user populations and community-specific pattern discovery.
- **Derived Features:** We will engineer additional features including comment length, word count, lexical diversity, time-based features (if timestamps available), and TF-IDF vectors for clustering.

D. Data Quality Considerations

- **Missing Values:** Some comments may have null parent IDs (top-level comments). These will be handled appropriately in preprocessing.
- **Text Noise:** Reddit comments contain informal language, URLs, emojis, and markdown formatting that require cleaning.
- **Comment Length Variability:** Comments range from single-word responses to lengthy paragraphs, requiring normalization strategies.
- **Deleted/Removed Content:** Some comments may have been removed by moderators or deleted by users, appearing as ”[deleted]” or ”[removed]” in the dataset.
- **Spam and Low-Quality Content:** May include bot-generated comments or spam that should be filtered during preprocessing.

II. Discovery Questions

This project focuses on *pattern discovery* rather than prediction. Our research questions aim to uncover hidden structures, natural groupings, and emergent themes within the data.

A. Question 1: User Segment Discovery

What are the natural user segments based on commenting behavior and content characteristics in ChatGPT discussions?

Why This is Interesting: Understanding user segments reveals distinct personas or user types within AI discussions—such as technical experts, casual users, skeptics, or enthusiasts. This discovery can inform community management, content moderation strategies, and targeted engagement approaches. Different user segments may have different needs, concerns, and interaction patterns that are not immediately obvious from surface-level analysis.

Discovery Focus: This question seeks to uncover latent groups without predetermined labels. We will discover how many natural clusters exist, what characterizes each cluster, and how cluster membership relates to engagement patterns.

B. Question 2: Topic and Theme Emergence

What topics and thematic discussions emerge from the ChatGPT comment corpus, and how do these topics distribute across different communities?

Why This is Valuable: Topic discovery reveals the actual concerns, interests, and discussion themes that dominate AI-related discourse. Rather than assuming what people discuss, we let the data reveal emergent topics—which might include technical troubleshooting, ethical concerns, creative applications, comparison with other AI tools, educational uses, or workplace implications. Understanding topic distribution across subreddits also reveals community-specific interests and priorities.

Discovery Focus: This question uses unsupervised topic modeling to identify latent themes without predefined categories, allowing the data to speak for itself about what matters most to users.

C. Question 3: Community-Specific Patterns

What distinct patterns characterize user engagement and discussion styles across different subreddit communities?

Why This Matters: Different online communities develop unique cultures, communication styles, and engagement norms. Discovering these community-specific patterns can reveal how context shapes AI discourse—for example, whether technical communities discuss ChatGPT differently than creative communities, or whether certain communities show more critical versus enthusiastic engagement.

Discovery Focus: This question explores whether clustering algorithms naturally separate comments by subreddit origin or if cross-community patterns emerge, suggesting universal themes in AI discussion that transcend community boundaries.

III. Planned Techniques

Our methodology employs techniques from multiple data mining categories to comprehensively analyze the Reddit ChatGPT dataset. Figure 1 illustrates our planned analysis pipeline.

A. Category 1: Clustering Techniques

1) K-Means Clustering: **Application:** Partition users/comments into K distinct segments based on behavioral and textual features.

Relation to Discovery Questions: Directly addresses Question 1 (user segments) and Question 3 (community patterns). K-Means will identify compact, spherical clusters in the feature space representing different user types or discussion styles.

Implementation Details:

- Determine optimal K using elbow method and silhouette analysis
- Apply to TF-IDF vectors and engineered behavioral features
- Evaluate cluster quality using within-cluster sum of squares (WCSS)

2) Hierarchical Clustering: **Application:** Build a dendrogram revealing hierarchical relationships between comments and potential nested community structures.

Relation to Discovery Questions: Provides complementary insights to K-Means for Questions 1 and 3. Reveals whether user segments have sub-segments or hierarchical relationships (e.g., ”technical users” might split into ”developers” and ”researchers”).

Implementation Details:

- Test both agglomerative (bottom-up) and divisive approaches
- Compare linkage methods: single, complete, average, and Ward’s
- Visualize dendrograms to identify natural cluster cutoff points

3) DBSCAN (Density-Based Clustering): **Application:** Identify dense regions of similar comments while detecting outliers and noise.

Relation to Discovery Questions: Complements K-Means by discovering clusters of arbitrary shape and identifying anomalous comments for all three questions. Particularly useful for detecting unusual discussion patterns or niche user segments.

Implementation Details:

- Tune epsilon (neighborhood radius) and minPts parameters
- Identify and characterize outlier comments
- Compare density-based clusters with K-Means partitions

B. Category 2: Text Mining and Topic Modeling

1) TF-IDF Vectorization: **Application:** Transform comment text into numerical feature vectors capturing term importance.

Relation to Discovery Questions: Foundational technique enabling all clustering and topic analysis. Converts unstructured text into features suitable for mathematical analysis while emphasizing distinctive terms that characterize different discussion themes.

Implementation Details:

- Experiment with n-gram ranges (unigrams, bigrams, trigrams)
- Apply appropriate text preprocessing (lowercasing, stopword removal, lemmatization)
- Tune max_features parameter to balance dimensionality and information retention

2) Latent Dirichlet Allocation (LDA): **Application:** Probabilistic topic modeling to discover latent themes in ChatGPT discussions.

Relation to Discovery Questions: Directly addresses Question 2 (topic emergence) by modeling each comment as a mixture of topics and each topic as a distribution over words.

Implementation Details:

- Determine optimal number of topics using coherence scores
- Extract top terms for each discovered topic
- Analyze topic distributions across subreddits and user clusters
- Visualize topic relationships using pyLDAvis

C. Supporting Technique: Dimensionality Reduction

1) Principal Component Analysis (PCA): **Application:** Reduce high-dimensional TF-IDF vectors to 2-3 dimensions for visualization and computational efficiency.

Relation to Discovery Questions: Enables visual exploration of all three discovery questions by projecting high-dimensional clusters into interpretable 2D/3D space. Helps identify whether user segments, topics, and communities form visually distinct groups.

Implementation Details:

- Determine number of components explaining 80-95% of variance
- Create scatter plots with cluster/subreddit coloring
- Use for initial exploratory data analysis before full clustering

D. Technique Integration

Our approach integrates clustering and text mining to provide comprehensive pattern discovery:

1) TF-IDF vectors serve as input features for all clustering algorithms
2) LDA topics become additional features for user segment clustering
3) Clustering results inform topic distribution analysis (which topics dominate which user segments)
4) PCA visualization reveals relationships between clusters, topics, and subreddits

IV. Preliminary Timeline

Table II outlines the project schedule across milestones M2, M3, and M4.

A. Anticipated Challenges

1) High-Dimensional Sparse Data: Text data transformed into TF-IDF vectors creates high-dimensional, sparse feature spaces that can challenge clustering algorithms. The curse of dimensionality may cause distance metrics to become less meaningful.

Mitigation Strategy: Apply dimensionality reduction (PCA, TruncatedSVD) before clustering. Experiment with different n-gram ranges and max_features parameters to balance dimensionality and information content. Consider using cosine similarity instead of Euclidean distance for text data.

2) Determining Optimal Cluster Numbers: Unlike supervised learning, clustering requires determining the ”right” number of clusters without ground truth labels.

Mitigation Strategy: Use multiple evaluation metrics (elbow method, silhouette scores, Davies-Bouldin index) and validate results through qualitative inspection of cluster contents. Accept that there may not be one ”correct” answer—different granularities reveal different patterns.

3) Interpreting Discovered Patterns: Clusters and topics discovered by unsupervised methods can be difficult to interpret and label meaningfully.

Mitigation Strategy: Examine top terms, representative examples, and feature distributions for each cluster/topic. Use visualization tools like word clouds and t-SNE plots. Consider domain knowledge about Reddit communities and AI discourse to aid interpretation.

4) Computational Resource Constraints: Processing 50,000 text documents with multiple clustering algorithms and parameter combinations may be computationally intensive.

Mitigation Strategy: Start with stratified samples for initial experimentation. Optimize code using vectorized operations and efficient libraries (scikit-learn, gensim). Consider using dimensionality reduction to speed up clustering. If needed, utilize cloud computing resources or university computing clusters.

5) Class Imbalance Across Subreddits: The four subreddits may have unequal comment counts, potentially biasing clustering results.

Mitigation Strategy: Perform both global clustering (all comments together) and subreddit-specific clustering to understand whether imbalance affects results. Consider stratified sampling or weighting schemes if imbalance proves problematic.

V. Expected Outcomes

Upon completion, this project will deliver:

- Identification and characterization of 3-7 distinct user segments in ChatGPT discussions
- Discovery of 5-10 major topics/themes in the discourse
- Comparative analysis of discussion patterns across subreddit communities

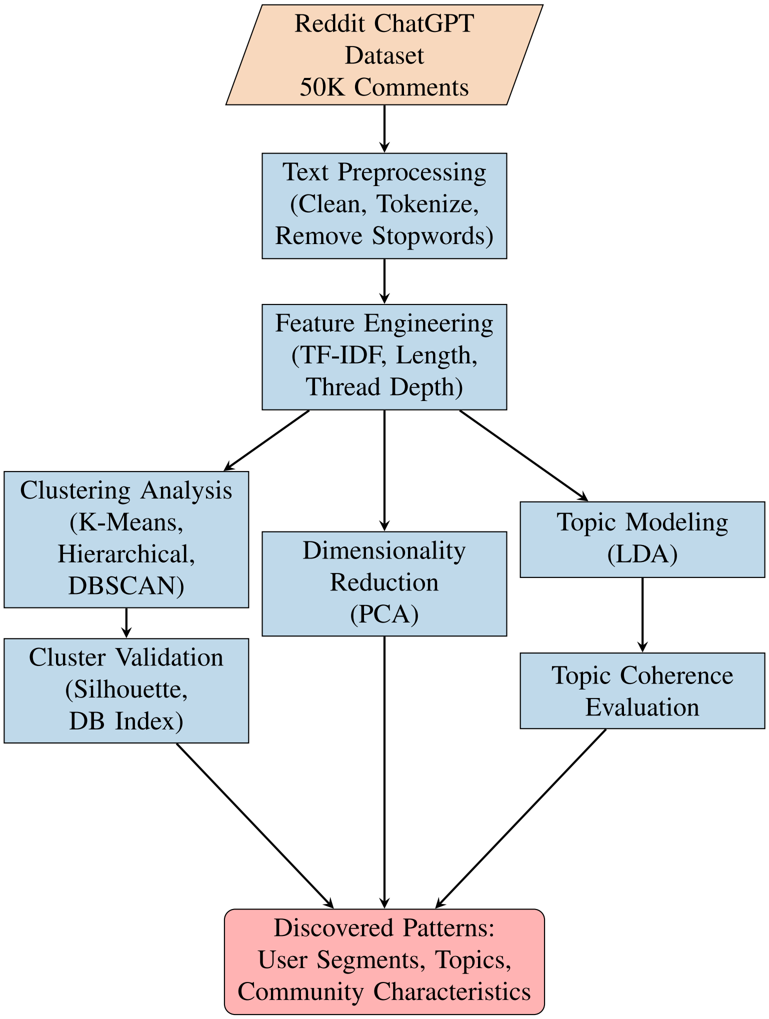

Fig. 1. Analysis Pipeline Overview

- Visualization dashboards illustrating clusters, topics, and their relationships
- Insights into how different user types engage with AI technology discussions
- Reproducible analysis pipeline documented in code and technical reports

These outcomes will contribute to understanding online discourse about AI tools, inform community management strategies, and demonstrate the power of unsupervised learning for pattern discovery in social media data.

VI. Conclusion

This project leverages unsupervised data mining techniques to discover hidden patterns in Reddit discussions about ChatGPT. By combining clustering algorithms (K-Means, Hierarchical, DBSCAN) with text mining methods (TF-IDF, LDA), we address three key discovery questions about user segments, emergent topics, and community-specific engagement patterns. The Reddit ChatGPT Comments dataset provides an authentic, substantial corpus for pattern discovery, enabling insights that cannot be obtained through supervised learning or manual analysis. Our systematic approach, comprehensive evaluation strategy, and awareness of potential challenges position this project for successful completion and meaningful discoveries about online AI discourse.

References

[1] Armita84, “ChatGPT Dataset - Reddit,” GitHub Repository, 2023. [Online]. Available: https://github.com/Armita84/ChatGPT-Dataset-Reddit

[2] M. J. Zaki and W. Meira Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed. Cambridge University Press, 2020.

[3] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 3rd ed. Cambridge University Press, 2020.

[4] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] D. Greene and P. Cunningham, “Practical solutions to the problem of diagonal dominance in kernel document clustering,” in *Proc. 23rd International Conference on Machine Learning (ICML’06)*, 2006, pp. 377–384.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

TABLE II
Project Timeline and Milestone Deliverables

Milestone	Timeline	Tasks and Deliverables
M2	Weeks 3-6	**Data Preprocessing & Exploratory Analysis** • Download and load dataset • Data cleaning: handle missing values, remove deleted comments • Text preprocessing: lowercasing, URL removal, stopword filtering, lemmatization • Exploratory Data Analysis (EDA): comment length distributions, subreddit statistics, word frequency analysis • Feature engineering: TF-IDF vectorization, derived behavioral features • Initial PCA visualization *Deliverable: EDA report with visualizations*
M3	Weeks 7-10	**Clustering & Topic Modeling** • K-Means clustering: parameter tuning, optimal K selection • Hierarchical clustering: dendrogram analysis, linkage comparison • DBSCAN: epsilon/minPts tuning, outlier analysis • LDA topic modeling: topic number selection, coherence evaluation • Cluster validation: silhouette scores, Davies-Bouldin index • Initial pattern interpretation *Deliverable: Clustering results with validation metrics*
M4	Weeks 11-15	**Analysis, Interpretation & Final Report** • Compare clustering algorithm results • Characterize discovered user segments (top terms, behaviors) • Analyze topic distribution across clusters and subreddits • Generate comprehensive visualizations (cluster plots, topic distributions) • Answer discovery questions with data-driven insights • Final report writing and presentation preparation *Deliverable: Final report, presentation, GitHub repository*