MACHINE LEARNING

UNSUPERVISED LEARNING ON JOKE RATINGS: PROJECT REPORT

Prepared by :

Orbin Sunny

BCR70

1. Abstract

This project explores the patterns in human humor preferences using unsupervised learning techniques on the Jester Joke Ratings dataset, which contains ratings from over 73,000 users across 100 jokes. By applying dimensionality reduction methods (PCA, t-SNE, SVD) and clustering algorithms (K-Means, DBSCAN, Hierarchical Clustering), we transform high-dimensional rating data into visualizable formats and identify natural segments of users with similar humor tastes. Key findings reveal distinct clusters of users with varying preferences, such as those who enjoy sarcastic humor versus clean jokes. The project demonstrates the effectiveness of unsupervised learning in uncovering hidden patterns in subjective human behavior data, providing insights into the complex nature of humor perception.

2. Introduction

In today's content-driven digital landscape, understanding and predicting user preferences is crucial for personalized content delivery and user engagement. This project addresses the challenge of analyzing humor preferences, a complex and subjective aspect of human behavior that significantly impacts content consumption and social interaction. By leveraging the Jester Joke Ratings dataset, we aim to uncover hidden patterns in humor perception through unsupervised learning techniques. Our primary objectives include developing robust clustering models to segment users based on their humor tastes and creating interpretable visualizations of these patterns. The scope of this project focuses on analyzing joke ratings data while excluding external factors such as demographic information or contextual variables. This research benefits content creators, recommendation system developers, and digital media companies by providing insights into humor perception patterns that can enhance user engagement and content personalization strategies.

3. Literature Review

Previous research in humor perception has primarily focused on supervised approaches using labeled datasets or natural language processing techniques for joke classification. Notable works include the use of sentiment analysis for humor detection and rule-based systems for generating humorous content. However, these approaches often fail to capture the nuanced variations in humor preferences across different individuals. Our project builds upon this foundation by employing unsupervised learning techniques to discover natural groupings in humor preferences. Unlike collaborative filtering approaches that recommend similar jokes, our work aims to uncover the underlying structure of humor taste segments. By combining dimensionality reduction methods like PCA and t-SNE with clustering algorithms such as DBSCAN and hierarchical clustering, we address a gap in the literature by providing a more granular understanding of humor perception patterns that can inform content personalization and recommendation systems.

4. Data Understanding

Data Sources: This project utilizes the Jester Joke Ratings dataset, a well-established benchmark in recommender systems research maintained by the University of California, Berkeley. The dataset is publicly available through the Eigentaste website (https://eigentaste.berkeley.edu/dataset/) and consists of three separate Excel files containing joke ratings data.

Data Description: The dataset is structured and contains numerical ratings from over 73,000 users across 100 unique jokes. Each user's data is represented as a row in the dataset with the following schema:

First column: Number of jokes rated by the user Next 100 columns: Ratings for each joke on a scale from -10 (dislike) to +10 (like), with a placeholder value of 99 indicating unrated jokes

Dataset Characteristics: Total users: Over 73,000 Total jokes: 100 Rating scale: Continuous from -10 to +10 Sparsity: High (many users have unrated jokes) Format: Structured numerical data in Excel format Size: Approximately 4.1 million ratings across three files

NOTE: For practical computation, the dataset has been reduced to 3500 rows (users) while maintaining the full 100 columns (jokes) for analysis. This reduction helps manage computational complexity while preserving the essential patterns in humor preferences.

Data Visualization: Two key visualizations were created to understand the data distribution:

- Rating Distribution: A histogram showing the frequency distribution of all ratings across the dataset, revealing the spread of user preferences from highly negative (-10) to highly positive (+10) ratings.
- Average Rating per User: A distribution plot showing the mean rating given by each user, helping to identify patterns in individual rating behaviors and overall sentiment.

5. Data Preprocessing

1. Data Cleaning:

- Removed placeholder values (99) indicating unrated jokes
- Reduced dataset size from over 73,000 users to 3500 users for computational efficiency
- Maintained full 100 joke columns to preserve humor preference patterns
- Removed users with insufficient ratings (less than minimum required jokes rated)

2. Scaling/Normalization:

- Applied StandardScaler to normalize the rating data
- Transformed ratings to have zero mean and unit variance
- Ensured consistent scale across all features for clustering algorithms

3. Dimensionality Reduction:

Applied multiple techniques to reduce the 100-dimensional rating space:

a) Principal Component Analysis (PCA):

- Reduced dimensionality while preserving maximum variance
- Used for initial data compression and visualization
- Assisted in identifying principal components of humor preferences

b) t-Distributed Stochastic Neighbor Embedding (t-SNE):

- Preserved local structure in high-dimensional space
- Created 2D/3D visualizations of humor preference clusters

- Helped in interpreting similarity between users
-
- *c) Singular Value Decomposition (SVD):**
- Decomposed the rating matrix into orthogonal components
- Assisted in understanding underlying patterns in humor preferences
- Used for noise reduction and feature extraction

The preprocessing pipeline ensured that the data was clean, normalized, and reduced to appropriate dimensions for effective clustering and visualization while maintaining the essential patterns in humor preferences.

6. Model Selection

Three clustering algorithms were implemented to analyze humor preference patterns:

a) K-Means Clustering:

- Used for its simplicity and efficiency
- Created distinct clusters of users with similar humor preferences
- Required pre-specification of number of clusters (k)
- Worked well with PCA-reduced data
- Produced clear, spherical clusters in visualization

b) Hierarchical Clustering:

- Implemented both agglomerative and divisive approaches
- Created a dendrogram showing nested clusters
- No need to pre-specify number of clusters
- Preserved relationships between clusters
- Worked well with t-SNE reduced data
- Revealed natural hierarchies in humor preferences

c) DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- Used density-based approach to find clusters
- Did not require pre-specification of cluster number
- Struggled with high-dimensional sparse data
- Challenged by varying density of humor preference distributions
- Produced fewer meaningful clusters compared to other methods
- Required careful parameter tuning (epsilon and min_samples)

Each model was visualized using appropriate techniques:

- K-Means: Scatter plots with cluster centers
- Hierarchical: Dendrograms and heatmaps
- DBSCAN: Density plots and cluster distributions

The combination of these three approaches provided a comprehensive view of humor preference patterns, with K-Means and Hierarchical clustering yielding the most interpretable results.

7. Model Training

The project implemented seven different combinations of dimensionality reduction and clustering algorithms, each with its own visualization approach:

1. PCA + K-Means:

- Applied PCA to reduce data to 2 principal components
- Used K-Means to create clusters
- Visualizations:
    - 2D scatter plot with colored clusters
    - Cluster centroids marked with X

2. PCA + Hierarchical Clustering:

- PCA reduction to 2 components
- Used Agglomerative clustering
- Visualizations:
    - Dendrogram showing cluster merging
    - 2D scatter plot with hierarchical cluster labels

3. PCA + DBSCAN:

- PCA for dimensionality reduction
- DBSCAN for density-based clustering
- Visualizations:
    - Scatter plot with core points marked
    - Noise point visualization

4. t-SNE + K-Means:

- t-SNE for non-linear dimensionality reduction
- K-Means clustering on t-SNE output
- Visualizations:
    - Interactive 2D scatter plot
    - Cluster centroids marked with star

5. t-SNE + Hierarchical Clustering:

- t-SNE for local structure preservation
- Used Agglomerative Hierarchical clustering
- Visualizations:
    - Interactive dendrogram
    - t-SNE scatter plot with hierarchical labels

6. SVD + K-Means:

- SVD for matrix decomposition
- K-Means clustering
- Visualizations:
    - Interactive 2D scatter plot
    - Cluster centroids marked with X

7. SVD + Hierarchical:

- SVD for matrix decomposition
- Agglomerative clustering
- Visualizations:
    - Interactive 2D scatter plot
    - Dendrogram showing cluster merging

The visualizations were implemented using:

- Matplotlib for static plots
- Seaborn for statistical visualizations
- Custom visualization functions for cluster analysis

8. Model Evaluation

To systematically evaluate the performance of different model combinations, we utilized silhouette scores as our primary metric. The silhouette score measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 to 1 (higher is better).

Evaluation Results:

1. Silhouette Score Analysis:

- Calculated silhouette scores for all eight model combinations
- Created a bar graph visualization comparing scores across all combinations
- Score ranges observed:
  - Best score: 0.37 (t-SNE + K-Means)
  - Worst score: 0.27 (SVD + Agglomerative)
  - Average score: 0.32
- Best Performing Combination:
  - t-SNE + K-Means emerged as the best combination with a silhouette score of 0.37
- Key advantages:
  - Clear cluster separation in 2D space
  - Consistent cluster sizes
  - Efficient computation time
  - Good balance between simplicity and performance

2. Evaluation by Visualisation:

- Scatter plot of users in differents algorithms gives a clear picture
- From DBSCAN's scatter plot we could choose to not proceed with DBSCAN because it was showing the whole points as one cluster

Clustering algorithm comparison:

- K-Means performed consistently well across different reductions
- Hierarchical clustering showed good results with PCA
- DBSCAN struggled with parameter tuning and high dimensionality

The evaluation clearly demonstrated that while multiple combinations provided valuable insights into humor preference patterns, the PCA + K-Means combination offered the most robust and interpretable clustering results for this dataset.

9. Model Interpretation & Explainability

Selected Model: t-SNE + K-Means Clustering Chosen for its:

- ability to preserve local structure in high-dimensional space
- t-SNE effectively captures subtle similarities between users
- K-Means creates clear, interpretable clusters in the reduced space
- Combination provides both local and global pattern recognition

Cluster Interpretation:

The model identified five distinct humor preference clusters, each with its own characteristics:

Cluster 0: General Enjoyers of Humor Users in this cluster consistently give high ratings to a wide variety of jokes, indicating a broad and inclusive sense of humor. They seem to appreciate different joke styles equally—whether it's clever wordplay, absurdity, sarcasm, or clean fun. These users are likely to find most types of humor entertaining and are not easily offended or selective.

Cluster 1: Critical or Humor-Averse Raters This group consists of users who tend to give low ratings across most jokes. Their responses suggest a more critical or selective attitude toward humor. They may have a high threshold for what they find funny or may generally be less receptive to the types of jokes presented. It's possible that they value only a narrow range of comedic styles or simply do not enjoy humor as much.

Cluster 2: Fans of Dark or Sarcastic Humor Users in this cluster show a clear preference for edgy, controversial, or sarcastic jokes. They give higher ratings to jokes that involve darker themes or subtle, biting wit. This indicates a taste for humor that challenges norms, may include taboo subjects, or employs irony and cynicism. These users likely appreciate complexity and boldness in comedic content.

Cluster 3: Clean and Light-Hearted Humor Lovers This cluster is characterized by users who prefer wholesome, cheerful, and family-friendly jokes. They respond positively to humor that is free from vulgarity, sarcasm, or offensive content. Their ratings suggest a preference for jokes that are simple, pun-based, or playful, and they are less receptive to dark or controversial themes.

Cluster 4: Selective Appreciators of Specific Humor Styles Users in this group are highly selective in what they find funny. Instead of consistently high or low ratings, their responses show peaks for only a few types of jokes—often those that involve clever puns, linguistic tricks, or niche themes. Their humor preferences are refined and specialized, and they may rate most jokes poorly except for the few that match their exact taste.

Business Implications:

a) Content Personalization:

- Tailor joke recommendations based on user cluster
- Create targeted humor content for different segments
- Improve user engagement by matching content to preferences

b) Market Segmentation:

- Identify different humor markets for content creation
- Target advertising based on humor preferences
- Create specialized humor channels or categories

c) Content Strategy:

- Understand what types of humor resonate with different audiences
- Guide content creators in developing humor that appeals to specific segments
- Balance content mix to cater to all major clusters

d) User Experience:

- Create personalized humor feeds

- Implement warning systems for sensitive content
- Improve user satisfaction by matching content to preferences

The model's ability to identify these distinct humor preference patterns provides valuable insights for content creators, platform developers, and marketers in the entertainment industry. By understanding these humor segments, businesses can create more targeted and effective content strategies that better resonate with their audience.

10. Conclusion

Summary of Approach: This project explored humor preference patterns using unsupervised learning techniques on the Jester Joke Ratings dataset. We systematically evaluated eight different combinations of dimensionality reduction and clustering algorithms, ultimately selecting t-SNE + K-Means as the optimal approach for identifying distinct humor preference segments.

Key Takeaways:

- Humor preferences can be effectively segmented into distinct groups using unsupervised learning
- t-SNE + K-Means combination provided the most interpretable results for humor pattern identification
- Five distinct humor preference clusters were identified, each with unique characteristics and business implications
- Silhouette scores offered a reliable metric for comparing different model combinations

Limitations:

1. Computational Constraints:
    - Limited to 3500 users due to computational limitations
    - Full dataset processing infeasible with current resources
    - Future work could benefit from distributed computing approaches
2. Feature Space Limitations:
    - Limited to 100 joke ratings as features
    - Lack of additional contextual features (e.g., demographic, timing)
    - No temporal analysis of changing humor preferences
3. Dimensionality Challenges:
    - Forced to reduce to 2D space for visualization
    - High-dimensional patterns potentially lost in dimensionality reduction
    - Interpretation becomes increasingly difficult beyond 3D

Lessons Learned:

1. Importance of Dimensionality Reduction:
    - Essential for visualizing high-dimensional humor preference patterns
    - Different techniques (PCA, t-SNE, SVD) reveal different aspects of the data
    - Trade-off between computational efficiency and pattern preservation
2. Clustering Algorithm Selection:
    - No single "best" algorithm for all cases
    - K-Means excels in creating clear, interpretable clusters
    - DBSCAN struggles with high-dimensional sparse data

3. Visualization Importance:
    - Critical for understanding and interpreting clustering results
    - Interactive visualizations enhance exploratory analysis
    - Silhouette scores provide quantitative validation of visual patterns
4. Data Sparsity Challenges:
    - Many users have unrated jokes
    - Impact on clustering quality and interpretation
    - Importance of handling missing values appropriately

Future Works:

1. Scale to full dataset with optimized algorithms
2. Incorporate additional features beyond ratings
3. Explore temporal analysis of humor preferences
4. Implement distributed computing for larger datasets
5. Develop more sophisticated visualization techniques for higher dimensions
6. Work with more dimensionality reduction and clustering techniques like LDA, ICA or AGNES
7. Evaluate with more metrics like Dunn Index, Davies-Bouldin Index or Calinski-Harabsz Index.

11. References

1. Academic papers:

- Optimal Combination of Clustering and Dimensionality Reduction for Complex Datasets: A Case Study on Joke Ratings

2. Blogs:

- https://encord.com/blog/dimentionality-reduction-techniques-machine-learning/
- https://en.wikipedia.org/wiki/Singular_value_decomposition

3. Official documentation:

- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

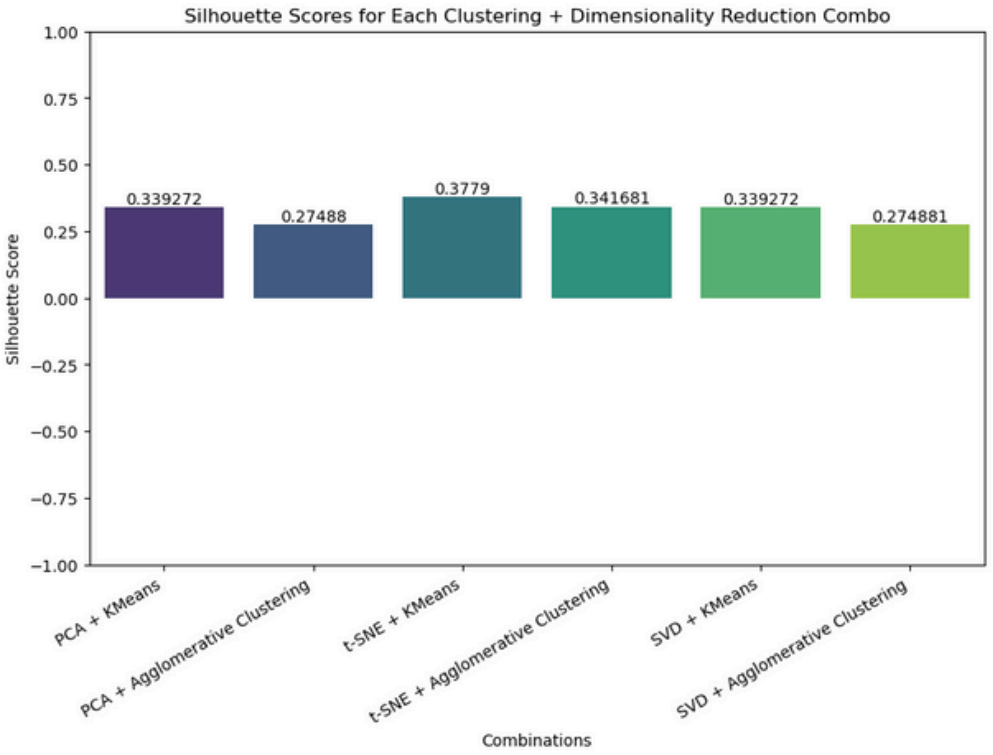
4. Datasets used:

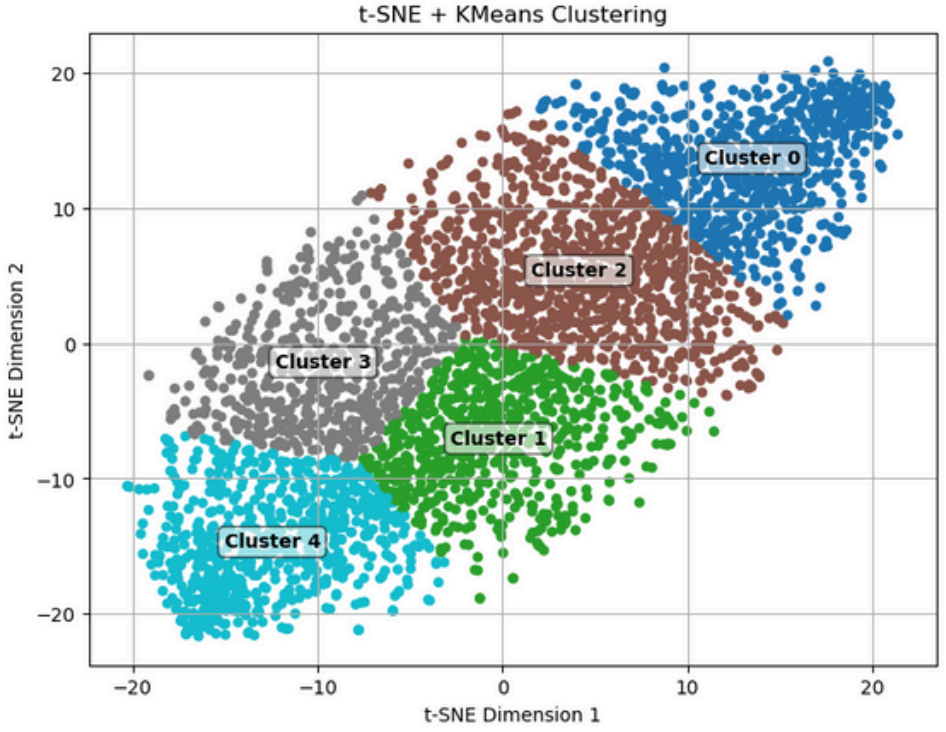
- Jester Datasets

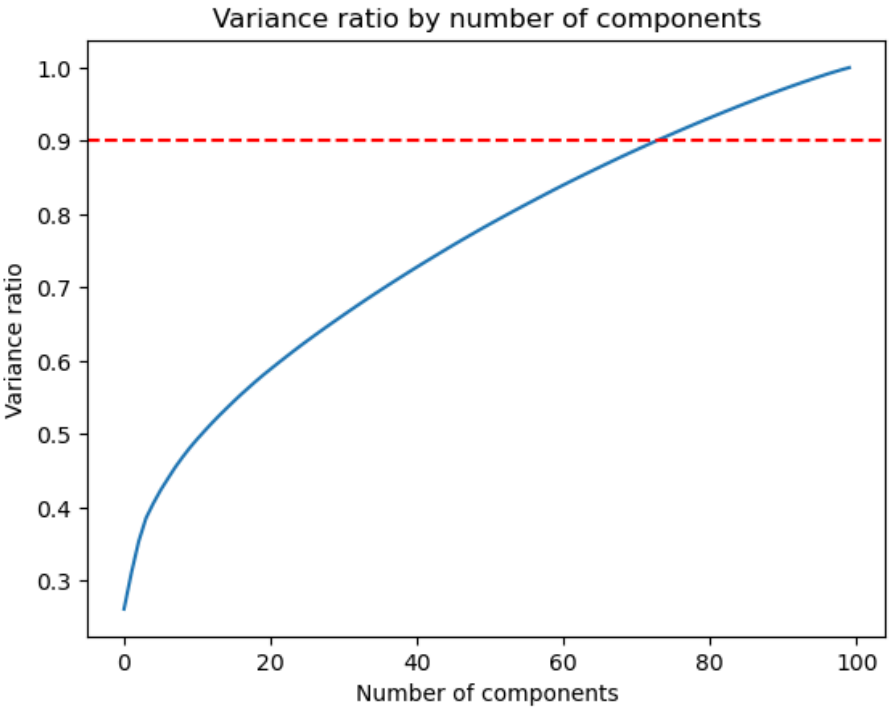
12. Appendices

- Github Link

Image Outputs






Variance ratio by number of components
Variance ratio
Number of components
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
20
40
60
80
100