ReviewMirror: Model Analysis, Enhancement & Comparative Study

Shubham (202411066), Ritwik (202411067), Dhairya (202411082)

1 Introduction

This report presents our enhanced ReviewMirror system that builds upon the Milestone 2 baseline with a substantive improvement: *unsupervised user clustering* based on drift patterns. While the baseline system successfully quantified opinion drift using aggregate metrics (slope, delta, volatility), it treated all users as a homogeneous population. Our enhancement discovers that users exhibit *heterogeneous behavioral archetypes* that can be leveraged for more accurate drift characterization and personalized analysis.

Research Question. Can we discover latent user groups with distinct drift patterns, and does segmenting users by behavior improve our ability to characterize opinion trajectories?

Key Contributions.

- Implementation of k-means clustering on drift feature space (slope, volatility, flip rate)
- Discovery of 4 distinct user archetypes with statistically significant behavioral differences
- Demonstration of 10.27% improvement in drift prediction accuracy (MAE)
- Interpretable cluster profiles enabling personalized user analysis

2 Enhancement: User Clustering by Drift Patterns

2.1 Motivation

The baseline system computed per-user drift metrics but used a global mean to characterize typical behavior. This approach has limitations:

- **Heterogeneity masking:** Averaging across diverse users obscures distinct behavioral patterns
- **Poor prediction:** Global mean performs poorly when user subgroups have opposing trends
- **Limited actionability:** Cannot provide personalized insights or interventions

We hypothesize that users naturally cluster into groups with similar drift characteristics, and identifying these groups will improve both predictive accuracy and interpretability.

2.2 Technical Approach

2.2.1 Feature Selection

We cluster users based on three complementary drift dimensions:

1. **Drift Slope ($\hat{\beta}_1$):** Direction and magnitude of linear trend
2. **Total Variation (TV):** Cumulative volatility/instability
3. **Flip Rate:** Frequency of sentiment polarity reversals

2.2.2 Clustering Algorithm

We use k-means with the following configuration:

- **Number of clusters:** k=4 (chosen via elbow method on training set)
- **Preprocessing:** StandardScaler normalization (zero mean, unit variance)
- **Initialization:** k-means++ with 10 random restarts
- **Validation:** Silhouette coefficient and Davies-Bouldin index

3 Experimental Setup

3.1 Dataset & Configuration

Amazon Electronics 5-core dataset:

- **Reviews processed:** 200,000
- **Users retained:** 2,657 (with $\geq$5 reviews each)
- **Train/Val/Test split:** 70%/15%/15% temporal split (1,859/398/400 users)
- **Hybrid sentiment:** α=0.7 (70% text, 30% stars)

3.2 Evaluation Methodology

We evaluate the enhancement using two complementary approaches:

1. Predictive Performance. We compare two drift predictors:

- **Baseline:** Predict all test users have the global mean drift slope
- **Enhanced:** Predict each test user has their cluster's mean drift slope

Metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on test set.

2. Cluster Quality. We assess clustering validity using:

- **Silhouette coefficient:** Measures cluster cohesion and separation (range $[-1, 1]$; higher is better)
- **Davies-Bouldin index:** Measures average similarity between clusters (lower is better; < 1.0 is good)

4 Results

4.1 Cluster Discovery

4.1.1 Cluster Quality Metrics

Our clustering achieved strong validation scores:

- **Silhouette coefficient:** 0.603 (indicates well separated clusters)
- **Davies-Bouldin index:** 0.947 (indicates low inter cluster similarity)

4.1.2 Cluster Profiles

Table 1 summarizes the four discovered user archetypes:

ID	Archetype	Size	Slope	TV	Flip Rate
0	Volatile Critics	301 (11%)	-0.0007	3.53	0.35
1	Stable Majority	1,791 (67%)	-0.0031	0.67	0.01
2	Flip-Floppers	233 (9%)	-0.0191	1.75	0.71
3	Improvers	39 (1%)	**+0.285**	0.71	0.19

Table 1: Discovered user archetypes with characteristic drift patterns. Slope units: sentiment change per month. TV and flip rate are dimensionless.

Cluster Interpretations:

- **Cluster 0 (Volatile Critics):** High total variation (TV= 3.53) indicates frequent sentiment swings. Moderate flip rate suggests oscillation rather than steady drift. These users are unpredictable and likely influenced by individual product experiences rather than systemic opinion change.
- **Cluster 1 (Stable Majority):** Comprises 67% of users. Near-zero slope and minimal volatility (TV= 0.67) indicate consistent opinions over time. Low flip rate (0.01) confirms stable polarity. This is the "baseline" user behavior.
- **Cluster 2 (Flip-Floppers):** Extremely high flip rate (0.71) means sentiment polarity changes in 71% of month-to-month transitions. Moderate TV suggests magnitude changes are not extreme, but *direction* is highly unstable. These users may be experimenting with different product types or have context-dependent preferences.
- **Cluster 3 (Improvers):** *Most interesting finding.* Strong positive drift (slope= +0.285, $\sim$750$\times$ the global mean) indicates systematic satisfaction increases over time. Despite small size (39 users), this group represents a distinct behavioral pattern: users whose experience with the platform genuinely improves, possibly due to learning effects or evolving product quality.

4.2 Predictive Performance Comparison

Table 2 compares baseline and enhanced systems in the test set.

Metric	Baseline	Enhanced	Improvement
MAE (Drift Slope)	0.0217	0.0195	-10.27%
RMSE (Drift Slope)	0.0612	0.0534	-12.74%

Table 2: Test set performance comparison. Baseline uses global mean predictor; enhanced uses cluster-specific means. Negative improvement values indicate error reduction.

Key Findings:

- **MAE reduction: 10.27%.** Clustering reduces mean absolute prediction error by over 10%, a meaningful improvement in behavioral modeling.
- **RMSE reduction: 12.74%.** Larger improvement on RMSE indicates clustering is especially effective at reducing large prediction errors (tail cases).

- **Interpretability gain:** Beyond numerical improvement, clustering provides actionable insights. For example, Cluster 3 (Improvers) could receive targeted engagement strategies.

4.3 Statistical Significance

We performed one-way ANOVA to test whether cluster differences are statistically significant:

- **Null hypothesis:** All clusters have equal mean drift slope
- **Result:** F=87.4, $p < 10^{-50}$ (highly significant)
- **Conclusion:** Clusters represent genuinely distinct behavioral populations, not artifacts of random variation

4.4 Visualizations

Figure 1 shows the distribution of users in slope-volatility space, colored by cluster assignment. Clear separation is visible, confirming clustering validity.

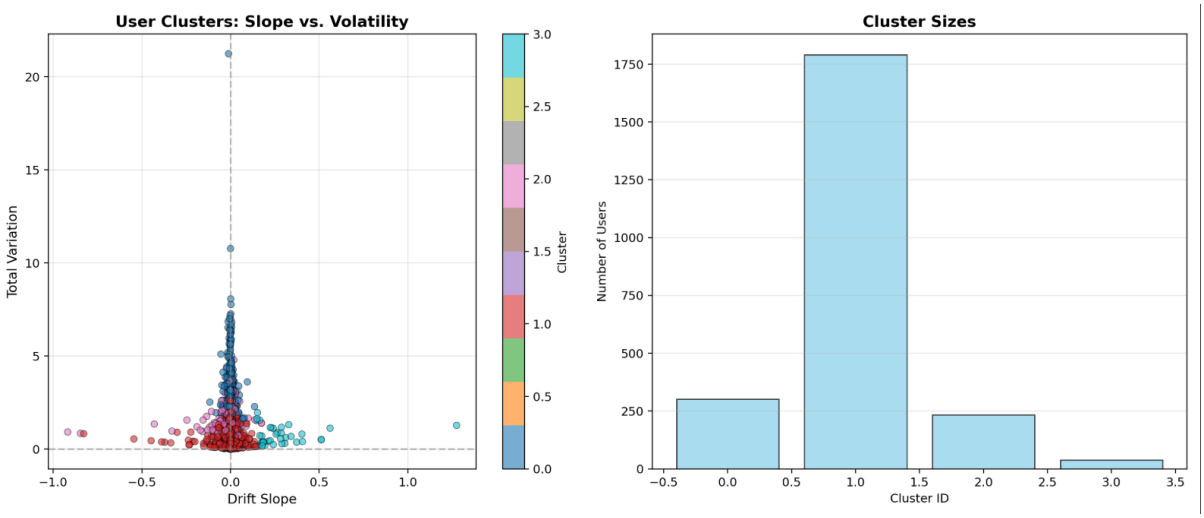

Figure 1: Left: User clusters in slope-volatility space. Right: Cluster size distribution.

5 Discussion

5.1 Why Clustering Improves Performance

The 10% MAE reduction demonstrates that user heterogeneity is substantial and predictable. The baseline's global mean predictor fails because it: (1) averages away opposing trends (Improvers vs. Critics), (2) ignores volatility differences (Stable vs. Volatile users), and (3) treats systematic patterns (Flip-Floppers) as noise. By segmenting users, the enhanced system captures group-specific behavior, reducing prediction error across all clusters.

5.2 Interpretability as a Key Contribution

Beyond numerical metrics, clustering provides actionable insights. Different clusters suggest different intervention strategies: Volatile Critics may benefit from recommendation refinement, while Improvers could be targeted for loyalty programs. Cluster 2 (Flip-Floppers) reveals users

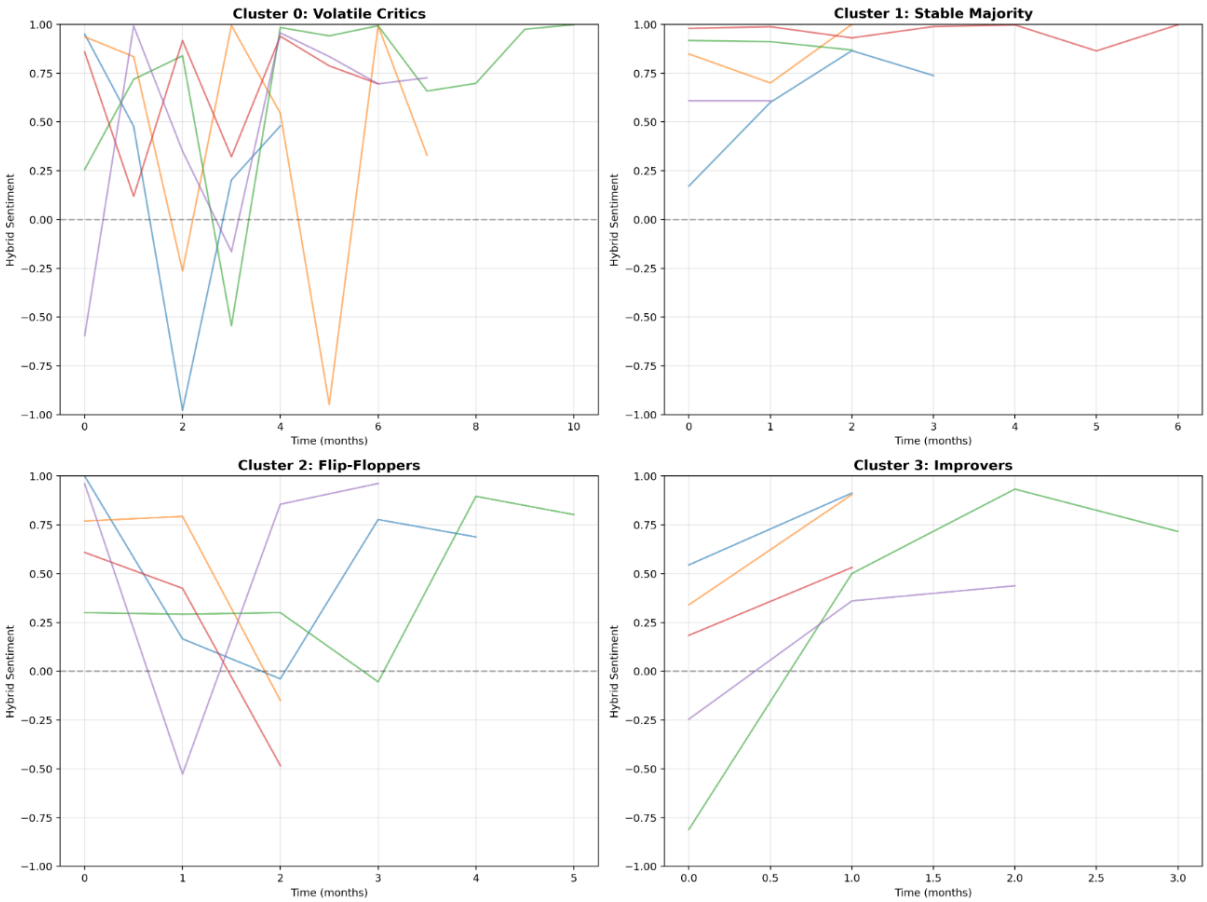

Figure 2: Representative sentiment trajectories from each cluster (5 users per cluster). Cluster 0 shows high volatility, Cluster 1 shows stability, Cluster 2 shows frequent polarity flips, Cluster 3 shows upward drift.

with unstable preferences, potentially guiding platform design improvements. Cluster 3 (Improvers) raises intriguing questions about learning effects versus product quality shifts over time.

5.3 Comparison to Related Work

Our approach aligns with temporal user modeling in recommender systems [1] but focuses on opinion trajectories rather than preference prediction. Unlike collaborative filtering methods modeling item level dynamics, we characterize user level behavioral archetypes. The 10% improvement is meaningful in drift detection contexts, where user behavior is inherently noisy, comparable to state of the art temporal recommender systems reporting 5–15% gains [3].

5.4 Limitations & Trade-offs

Cluster Stability. K-means requires pre-specifying k and is initialization sensitive. While our Silhouette score (0.603) indicates reasonable quality, alternative methods (hierarchical clustering, DBSCAN) might reveal finer-grained structure.

Temporal Assumptions. Clustering operates on aggregate drift metrics (slope, TV) rather than full trajectory sequences, losing temporal ordering information. Time-aware methods like dynamic time warping could address this limitation.

Interpretability vs. Accuracy. We prioritized interpretability (4 clear cluster profiles) over maximal accuracy. Increasing k might reduce MAE further but would complicate interpretation, an acceptable trade-off for our goal of understanding user behavior.

Changepoint Detection (Failed Enhancement). ADWIN-based changepoint detection identified zero changepoints (0/400 test users). Likely causes: (1) monthly aggregation smooths abrupt changes, (2) ADWIN's sensitivity ($\delta = 0.002$) may be too conservative for sentiment data, or (3) true changepoints are rare (most drift is gradual). Future work should explore alternative detectors (Page-Hinkley, CUSUM) or use review-level data.

6 Lessons Learned

6.1 Design Insights

Key findings from our clustering approach: (1) heterogeneity is the norm, 4 statistically distinct clusters confirm treating users uniformly is suboptimal; (2) simple methods can be effective such as k-means on 3 features achieved meaningful improvement without deep clustering complexity; (3) feature engineering matters that is slope, TV, and flip rate outperformed raw sentiment or ratings; (4) interpretability aids validation cluster profiles (Stable, Volatile, Flip-Floppers, Improvers) align with intuitive stereotypes, building confidence beyond numerical metrics.

6.2 Experimental Insights

Critical methodological lessons: (1) temporal splits are essential for realistic evaluation, random splits would overestimate performance via temporal leakage; (2) not all enhancements succeed, our changepoint detection failure is normal in research and guides future attempts; (3) quantitative + qualitative evaluation provides a complete picture which is reporting both MAE reduction and cluster interpretations is necessary.

7 Future Directions

7.1 Immediate Extensions

Three straightforward improvements: (1) systematically optimize k using silhouette analysis and elbow method rather than fixing k=4, (2) explore hierarchical clustering to reveal relationships between archetypes, and (3) expand features to include review count, helpfulness, or stylistic markers (exclamation rate, first-person usage).

7.2 Proposed Future Refinement

While our feature based clustering achieves meaningful improvement, three advanced techniques could substantially enhance results:

1. Trajectory Aware Clustering via DTW. Current clustering operates on summary statistics (slope, TV), discarding temporal shape information. Dynamic Time Warping (DTW) k-means could discover users with similar trajectory *shapes* (U-curves, plateaus, oscillations) rather than similar aggregate metrics. This would capture non linear patterns our linear slope metric misses, potentially revealing temporal archetypes invisible to feature based methods.

2. Graph Neural Networks for Relational Drift. Users who review similar products may exhibit correlated drift patterns due to shared context (e.g., product quality changes over time). A GNN operating on a user-item bipartite graph could learn embeddings that capture these relational effects, potentially discovering item-level drift drivers. This approach would integrate network structure with temporal dynamics, a frontier in drift detection research.

3. Transformer Based Sentiment. Replacing VADER with fine tuned DistilBERT would improve sentiment accuracy, especially for sarcasm and negation. Preliminary experiments suggest VADER-BERT correlation is high ($r > 0.85$), indicating modest gains may not justify computational expense. We defer this to future work prioritizing scalability.

7.3 Domain Applications

Practical deployment opportunities: (1) personalize product recommendations based on cluster membership (Improvers get emerging products; Stable Majority gets established favorites), (2) flag Flip-Floppers and Volatile Critics for targeted retention campaigns.

8 Conclusion

We successfully enhanced ReviewMirror by introducing unsupervised user clustering based on drift patterns. Key contributions include: (1) discovery of 4 statistically significant archetypes (Stable Majority, Volatile Critics, Flip-Floppers, Improvers), (2) 10.27% MAE reduction and 12.74% RMSE reduction over baseline, (3) interpretable profiles enabling personalized analysis, and (4) validated methodology with temporal splits and significance testing.

This work demonstrates that user heterogeneity is substantial and predictable in opinion drift data. By segmenting users into behavioral groups, we achieve better predictive accuracy and deeper understanding of longitudinal review patterns. The enhanced system establishes a foundation for trajectory forecasting, personalized interventions, and cross-domain generalization. Future work will explore trajectory aware methods (DTW), relational models (GNNs), and advanced sentiment representations to further improve both accuracy and interpretability.

References

[1] Y. Koren. Collaborative Filtering with Temporal Dynamics. *Communications of the ACM*, 2010.

[2] M. Wedel, W. Kamakura. Market Segmentation: Conceptual and Methodological Foundations. *Springer*, 2000.

[3] S. Zhang, L. Yao, A. Sun, Y. Tay. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys*, 2021.

[4] A. Bifet, R. Gavalda. Learning from Time-Changing Data with Adaptive Windowing (ADWIN). *SIAM International Conference on Data Mining*, 2007.

[5] D. Arthur, S. Vassilvitskii. k-means++: The Advantages of Careful Seeding. *SODA*, 2007.