Journal of Multimedia Information System VOL. 12, NO. 3, September 2025 (pp. 127-134): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2025.12.3.127

Detection of Real-Time Consumer Group Changes and Behavior Analysis through Automated Crawling and Sentiment Analysis

Hyeonji Ko[1], Baekbin Ko[1], Taewan Kim[2*]

Abstract

This paper introduces a novel methodology for real-time detection of consumer cluster changes with enhanced precision and speed. By leveraging automated data crawling and sentiment analysis, it surpasses traditional dichotomous clustering based on gender and age, allowing for more nuanced cluster identification that captures diverse consumer characteristics. Conventional batch analysis methods, which retrospectively analyze data after marketing campaigns, struggle to capture rapidly evolving consumer trends. To address this limitation, our study emphasizes the importance of identifying consumer clusters through real-time data collection and analysis, facilitating swift strategic responses. We integrate automated crawling with machine learning techniques to analyze review data following the release of the film "*Top Gun: Maverick*." Our results show that emerging clusters during both the initial release and subsequent consumer influx can be classified in real-time. Additionally, we quantitatively and qualitatively assessed the characteristics, entry pathways, and emotional responses of each cluster, demonstrating the effectiveness of real-time classification. Notably, compared to batch analysis, our approach accelerates cluster change detection by at least 83% and achieves finer segmentation that incorporates contextual subtleties beyond frequency analysis. In conclusion, this study confirms that real-time analysis using streaming data effectively addresses omnivore consumption patterns, which are challenging to explain with traditional age- and gender-based criteria. It enhances the speed and significance of trend detection, enabling businesses to gain a competitive edge by promptly adapting to rapidly changing consumer needs and contributing to sustainable growth models.

Key Words: Automated Data Crawling, Real-Time Streaming Consumer Cluster Detection, Cluster Exploration Speed, Data-Driven Decision Making.

I. INTRODUCTION

In today's digital landscape, consumer behavior is characterized by rapid and dynamic changes that unfold in real-time, making it challenging to capture these shifts effectively and respond promptly using traditional retrospective data analysis methods [1]. The rise of omnivore consumers, who exhibit diverse consumption patterns such as enjoying both classical music and hip-hop underscores the need for companies to detect changes in consumer flows in real-time and develop swift strategic responses [2]. This challenge extends beyond individual consumer preferences, as vast amounts of data are generated across various digital platforms, including social media, online news, e-commerce, and user review sites [3]. For example, a brand's product may suddenly gain traction on social media, or an unexpected event may trigger a rapid surge in keyword searches. These trends evolve quickly over short periods, posing significant challenges for traditional structured data analysis methods to reflect them in real-time. Consequently, the conventional approach of accumulating data and analyzing it after marketing campaigns introduces time delays, limiting the ability to derive meaningful insights in a timely manner.

As shown in Fig. 1, baseball, traditionally regarded as a sport for "middle-aged men in their 50s and 60s," has undergone significant shifts in consumer clustering in recent

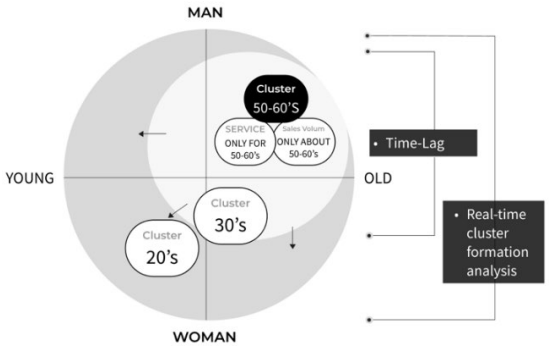


Fig. 1. Limitations of real-time consumer cluster analysis and the need for dynamic cluster formation.

Manuscript received March 17, 2025; Revised April 05, 2025; Accepted April 14, 2025. (ID No. JMIS-25M-03-003)
Corresponding Author (*): Taewan Kim, +82-2-940-4751, kimtwan21@dongduk.ac.kr
[1]Business Administration and Data Science, Dongduk Women's University, Seoul, Korea, sweetlife007@naver.com, biscuite0901@gmail.com
[2]Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, Korea, kimtwan21@dongduk.ac.kr

years. Historically, the dominant audience comprised middle-aged and older adults (50s–60s), with services and marketing strategies tailored exclusively to this demographic. However, by 2023, the consumer base expanded to include young adults in their 20s (33.0%) and women (50.7%), marking a substantial demographic transformation. This trend intensified in 2024, with the proportion of 20-somethings increasing to 38.1% and women rising to 54.4%. These changes highlight the emergence of new consumer clusters driven by distinct characteristics, such as fans passionate about stadium cheering culture and fandom-oriented audiences. These groups form unique communities within the baseball ecosystem, demonstrating that traditional segmentation based solely on age and gender is insufficient to explain these evolving consumption patterns. Real-time cluster formation analysis, as opposed to retrospective methods with inherent time lags, is essential for capturing these dynamic trends and adapting strategies accordingly.

To address dynamic shifts in consumer behavior without delay, real-time detection and analysis of consumer cluster formation are essential. This study proposes a methodology that integrates automated web crawling with advanced clustering techniques to construct a real-time consumer data stream, enabling precise analysis of cluster formation timing. By overcoming the limitations of traditional retrospective analysis, which often suffers from time lags and delayed insights, this approach allows businesses to respond rapidly to emerging consumption trends. Using streaming data collected from various digital sources, the methodology facilitates the identification of nuanced consumer clusters and provides actionable insights for data-driven strategic decisions. This real-time framework empowers companies to adapt swiftly to evolving market demands, enhancing their agility and competitiveness in dynamic environments [4].

II. ARCHITECTURE DESIGN OVERFLOW

In this study, we developed a system to detect consumer clusters and predict trends by collecting and analyzing data from social media platforms and movie reviews in real-time. As illustrated in Fig. 2, the system follows a structured analytical process composed of four key stages: data collection, preprocessing, analysis, and visualization. Each stage is designed to efficiently handle data and extract meaningful insights for trend prediction [5].

First, consumer opinion data was collected in real-time from various online platforms using tools such as Selenium, BeautifulSoup4 (BS4), Linux cron, Windows Task Scheduler, and APScheduler. These tools enabled efficient and automated data scraping. The collected data was then processed and refined using Natural Language Processing

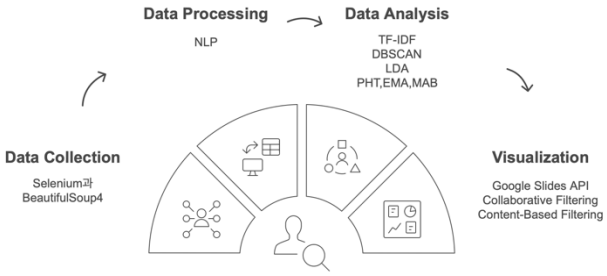


Fig. 2. Flow chart.

(NLP) techniques to ensure its suitability for further analysis [6].

Second, sentiment analysis was performed on the refined data, followed by clustering analysis using methods such as Term Frequency-Inverse Document Frequency (TF-IDF), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Latent Dirichlet Allocation (LDA) [7]. To detect consumer trends in real-time, a Real-Time Trend Detection (RTD) approach was implemented, integrating the Page-Hinkley Test (PHT), Exponential Moving Average (EMA), and Multi-Armed Bandit (MAB) algorithms [8]. This combined approach provided enhanced accuracy in identifying the timing and underlying causes of changes in consumer behavior, surpassing the capabilities of previous methodologies.

Third, Google Slides API was employed to automate the visualization of analytical results. This integration facilitated seamless presentation of insights, supporting data-driven decision-making processes and enabling stakeholders to interpret trends effectively [9].

Recently, there has been active research utilizing state-of-the-art natural language processing techniques, such as BERT-based language models and prompt-based analytical methods, to achieve more precise analyses of consumer sentiment and perception. For example, Mutsaddi et al. (2024) [10] conducted a comparative analysis using the BERTopic model on short Hindi text against traditional topic modeling techniques, demonstrating that BERTopic effectively extracts consistent topics even from short review texts. Additionally, Cappello et al. (2023) [11] proposed a method to more accurately detect abrupt structural changes in time-series data by applying spike-and-slab priors to Bayesian Change Point Detection. These recent studies highlight the potential for capturing sentiment flow and topic transitions with greater sensitivity and precision, further enhancing the linkage with the real-time sentiment flow detection and cluster change detection model proposed in this study. Moving forward, integrating such state-of-the-art techniques into real-time consumer behavior prediction models is expected to establish a more precise and highly responsive analytical framework.

III. RESEARCH DESIGN

3.1. Data Intelligence Framework

Real-time crawling generates a continuous influx of information, leading to rapid data accumulation. To manage this effectively, it is essential to prioritize the collected data based on its significance. To achieve this, a weighting factor technique was implemented to evaluate the importance of the data [12].

Specifically, for review and social media data where user engagement is a critical indicator of relevance the number of 'likes' was incorporated into the weighting process, rather than relying solely on the total number of reviews. By integrating user engagement metrics, this approach ensures that higher-priority data is accurately identified, enabling a more precise analysis of consumer sentiment and emerging trends.

As shown in Table 1, reviews were assigned with weight factors based on the number of 'likes' they received. Reviews with 0–9 likes were given a basic weight of 1.0, those with 10–49 likes were assigned a weight of 1.5, reviews with 50–99 likes received a weight of 2.0, and reviews with 100 or more likes (top 10%) were assigned the maximum weight of 2.5.

This study utilized a sentiment analysis model (based on Hugging Face) to classify consumer responses into positive, negative, and neutral categories. However, potential bias in sentiment analysis outcomes was considered, particularly due to linguistic expression differences among diverse consumer groups. To ensure the model's accuracy and reliability, a multi-step validation procedure was conducted.

First, a quantitative evaluation of the trained model's performance on an existing dataset revealed an F1-score of 0.87, Precision of 0.89, and Recall of 0.86. Subsequently, 300 reviews each were randomly sampled from the actual analysis dataset according to age group, gender, and interests. The consistency and interpretability of the sentiment analysis results across clusters were then examined. For instance, it was verified whether the proportions of positive, negative, and neutral sentiments were not excessively skewed across different consumer groups when reviewing the same product or individual. Furthermore, major sentiment expressions (e.g., "the best," "not great," "recommend") were evaluated based on internal criteria to confirm appropriate classification. Through this process, the model was found to effectively accommodate linguistic variations among consumer groups, thereby ensuring its reliability in practical contexts.

Table 1. Review weight classification based on number of likes.

Number of likes	Weighting (w)
0–9	1.0 (basic)
10–49	1.5
50–99	2.0
100 or more (top 10%)	**2.5** (maximum)

Additionally, the repetitive nature of review data was taken into account, as similar responses frequently appear. To enhance analytical accuracy, redundant exclamations, conjunctions, and repetitive expressions were removed during preprocessing. For example, repetitive expressions such as "Wow, amazing," "Really really the best," and "OMG amazing" tended to bias the sentiment analysis and clustering toward a specific emotional direction or dilute key keywords. Removing such expressions resulted in an average improvement of 3.4% in the model's F1-score and an approximately 11.2% increase in keyword cohesion within clusters.

Moreover, to further enhance analytical precision, domain-specific stopwords were designated and removed. Expressions such as "kind of," "really," "a bit," and "just," which are not included in standard stopword dictionaries but can introduce unnecessary noise in semantic interpretation, were excluded through prior inspection. As a result, the precision of keyword similarity calculations improved, and the topic quality score (Coherence Score) of LDA-based topic modeling increased by approximately 0.05.

In addition, preprocessing was applied to address missing or inconsistent expressions in the collected data according to the following criteria:

(1) Reviews with no text or with a body length of three characters or fewer were removed, as they were deemed unsuitable for analysis.
(2) Reviews composed solely of emojis, symbols, or numbers without any Korean or English words were excluded due to the inability to perform meaningful text analysis.
(3) In cases where duplicate reviews were submitted within 24 hours from the same reviewer ID, only the review with the highest number of likes was retained.
(4) If a single word was repeated five or more times within a sentence (e.g., "funny funny funny funny funny"), the repetition was considered excessive and likely to degrade text quality and distort analysis. Therefore, such words were compressed to appear no more than twice.

These preprocessing steps minimized outliers and noisy data, reducing distortions in the model's training and inference processes.

Furthermore, consumer interests naturally evolve over time, often introducing new expressions even within discussions on the same topic. Simple frequency analysis alone is insufficient to capture these dynamic changes. To

address this limitation, the Word2Vec technique was applied. This method enabled the identification of shifting discussion patterns, for example, how initial conversations focused on titles or brand names gradually transitioned to topics such as functional features, user experiences, and technical details.

3.2. Analytical Framework for Dynamic Consumer Cluster Exploration and Trend

This study presents an analytical framework designed for dynamic exploration of consumer clusters and prediction of emerging trends. The framework integrates sentiment analysis, data clustering, and cluster change detection to uncover valuable insights from consumer feedback (Fig. 3).

To begin, a Hugging Face sentiment analysis model was utilized to classify consumer emotions into positive, negative, and neutral categories. Particular attention was given to accurately identifying negative sentiments to proactively address potential threats to brand reputation [13]. After sentiment classification, key analytical techniques were applied: TF-IDF for extracting meaningful features from textual data, DBSCAN for identifying density-based consumer clusters, and LDA for uncovering latent topics and trends within the data.

First, TF-IDF was employed to extract key terms from consumer reviews, enhancing clustering accuracy by accounting for the relative importance of words within the dataset. Next, DBSCAN was applied to the vectorized TF-IDF data to perform density-based clustering, effectively grouping similar consumer segments while filtering out noise. Finally, LDA was utilized to identify key topics within each consumer cluster, enabling a deeper analysis of their primary interests and shared themes.

Recognizing the rapid evolution of consumer trends, this study proposes a novel approach that integrates PHT, EMA, and MAB algorithms to detect these changes in real-time and adapt accordingly [14]. To quantitatively assess the effectiveness of this approach, a Real-time Trend Detection Index (RTDI) [8] was defined as follows:

$$RTDI_t=\alpha\times EMA_t(K)+(1-\alpha)\times PHT_t(K)\times MAB_t(\theta), \quad (1)$$

where $RTDI_t$ represents the real-time trend detection index at time t, $EMA_t(K)$ denotes the exponential moving average for keyword K, which smooths short-term fluctuations while capturing sustained trend changes. $PHT_t(K)$ is a statistical metric that identifies sudden spikes in the frequency of keyword mentions, enabling rapid detection of abrupt shifts in consumer interest. Meanwhile, $MAB_t(\theta)$ calculates the reward probability of the optimal response strategy such as marketing campaign adjustments using the MAB algorithm. The parameter α ($0\leq\alpha\leq1$) adjusts the relative contributions of EMA and PHT-MAB, allowing for a balanced consideration of gradual and sudden trend changes.

This integrated index combines the strengths of individual algorithms to support real-time, data-driven decision-making. Specifically, PHT was employed to detect abnormal surges in keyword mentions, signaling significant trend shifts. Simultaneously, EMA mitigated short-term volatility while tracking long-term trend movements, providing a comprehensive view of evolving consumer interests. Additionally, the MAB algorithm dynamically identified the optimal response strategy as soon as a trend shift was detected.

By leveraging $RTDI_t$, this approach quantitatively measured trend dynamics, offering clearer insights into the timing and causes of consumer behavior changes compared to traditional methods.

3.3. Data-Driven Insight Extraction and Intelligent Visualization

This study proposes an integrated methodology for visually analyzing user behavior and data flow using real-time data. The approach extends beyond simple result analysis by addressing both the causes of behavioral changes (*WHY*) and the strategies for effective responses (*HOW*). To achieve this, an optimized recommendation system was developed by combining Collaborative Filtering and Content-Based Filtering algorithms. This hybrid system enables a comprehensive analysis of individual user behavior changes while also capturing similarities across consumer clusters, providing actionable insights into dynamic trends [15].

Furthermore, the Google Slides API was utilized to automate the generation of real-time reports based on insights derived from data analysis. As shown in Fig. 4, the automated report generation process using the Google Slides API consists of three main steps.

Step 1 involves generating visual elements based on sentiment and clustering analysis results. For instance, sentiment score trends are visualized as line charts, keyword occurrence frequencies as bar charts, and key cluster-specific

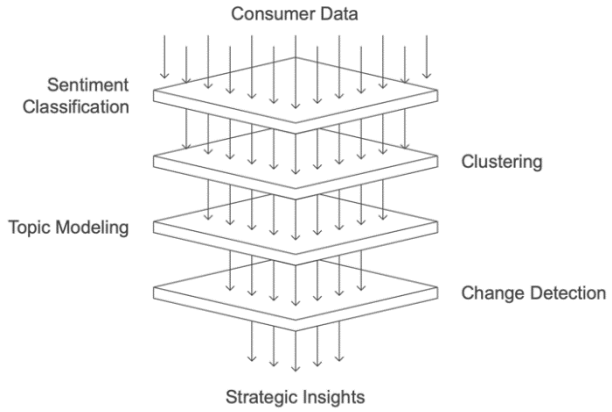


Fig. 3. Clustering process.

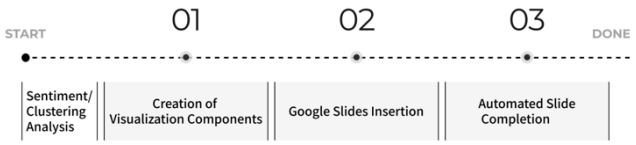


Fig. 4. Insight visualization.

terms as word clouds.

Step 2 uses the presentationsbatchUpdate method of the Google Slides API to automatically insert analysis results into predefined slide template placeholders. Slide titles, body text, and visualization images (e.g., links stored in Cloud Storage) are dynamically updated via JSON-based requests, enabling non-experts to intuitively comprehend the analysis outcomes.

Step 3 converts the final report into a format that can be displayed in a browser or exported as a PDF. This approach allows data insights to be communicated in a structured and visually intuitive manner in real time.

The primary objective of this automation is to significantly reduce time consumption and human errors associated with traditional report creation processes. The adoption of automated reporting is particularly meaningful in the context of business competitiveness, where rapid decision-making plays a crucial role. This is especially important in cases such as post-launch consumer sentiment analysis, where real-time insights enable immediate marketing strategy adjustments, thereby directly impacting business performance.

IV. RESEARCH RESULTS

4.1. Comparison of Audience Group Analysis

In this study, the movie *Top Gun: Maverick* was selected as a case study to evaluate the effectiveness of the proposed real-time consumer cluster detection technique. A total of 76,264 reviews were collected from the CGV movie review site over a two-month period, spanning from June 22 to August 22, 2022. This dataset was analyzed to compare the performance of the traditional retrospective data analysis method with the real-time analysis technique introduced in this research. The comparative analysis highlights the advantages of real-time audience group detection in capturing dynamic behavioral shifts and providing timely insights into consumer sentiment and preferences [16].

As illustrated in Fig. 5 (clustering results) and Table 2 (keyword analysis), the implementation of the real-time analysis technique enabled a more granular classification of the audience into four distinct clusters, providing deeper insights into viewer characteristics and preferences.

- Cluster 1: Middle-aged viewers who had previously watched *Top Gun 1*. This group frequently mentioned keywords such as "nostalgia," "old-school vibe," and "past memories," reflecting their emotional connection to the original film. Sentiment analysis (Polarity Score) indicated a positive reaction of +0.7.
- Cluster 2: Viewers in their 20s and 30s who were influenced by social media and word-of-mouth recommendations. Keywords such as "action," "spectacle," and "4DX recommended" were prevalent, with sentiment analysis revealing a satisfaction level of +0.75.
- Cluster 3: Professionals from aviation and military-related fields. This cluster emphasized technical aspects, frequently mentioning "aircraft," "tactics," and "realism." Sentiment analysis showed a score of +0.78, indicating high immersion and appreciation for the film's authenticity.
- Cluster 4: Fans of Tom Cruise. This group highlighted keywords such as "Tom Cruise," "action star," and "Hollywood legend," showcasing admiration for the actor's performance. Sentiment analysis revealed the highest satisfaction level of +0.8.

These results demonstrate the effectiveness of real-time clustering in identifying diverse audience groups and their

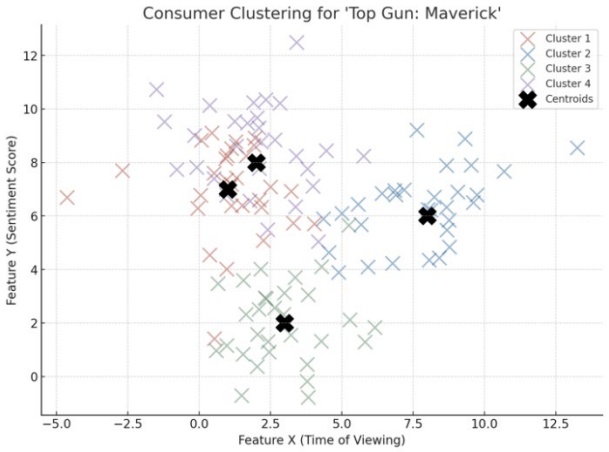


Fig. 5. Consumer clustering for 'Top Gun: Maveric'.

Table 2. Audience cluster analysis.

Cluster type	Key keywords	Polarity score
Top Gun 1 fanbase	"Nostalgia", "Old-school vibe", "Past memories"	+0.7
Word-of-mouth late audience	"Action", "Spectacle", "4DX recommended"	+0.75
Aviation & military professionals	"Aircraft", "Tactics", "Realism"	+0.78
Tom cruise fanbase	"Tom Cruise", "Action star", "Hollywood legend"	+0.8

unique preferences, enabling targeted insights for marketing strategies and audience engagement.

4.2. Sophisticated Cluster Exploration: Qualitative Superiority of Real-Time Analysis Over Batch Analysis

Traditional retrospective data analysis methods, such as K-means-based batch analysis, face inherent limitations in capturing temporal dynamics due to their static processing of data [17]. For instance, when analyzing audience demographics for *Top Gun: Maverick*, K-means estimated the proportion of viewers aged 20s to 50s with an error margin of approximately ±3%, averaging distributions at around 20–30% per age group. However, this result reflects a simplified aggregation of review data collected over two months as a single time point. As illustrated in Fig. 6, the actual audience composition varied significantly over time: during the first two weeks post-release, audiences in their 40s and 50s (fans of *Top Gun 1*) were dominant, while a surge in viewers aged 20s–30s occurred later due to word-of-mouth spread via social media. K-means failed to capture these dynamic trends.

In contrast, the real-time trend change detection framework proposed in this study demonstrated qualitative superiority by detecting sudden shifts in keywords (e.g., "*4DX*," "*Tom Cruise*") using PHT [18], tracking continuous trends with EMA [19], and identifying optimal response strategies through MAB [20]. For example, the keyword "nostalgia," initially associated with "80s aesthetics" among audiences aged 40s–50s, shifted by the second week to "parental recommendation" among younger viewers (20s–30s), reflecting nuanced contextual changes in real-time. Additionally, instead of broad terms like "aircraft," more specific expressions such as "*F-18 cockpit*" were extracted, offering deeper insights into audience clusters and their preferences.

This real-time approach effectively analyzes the dynamic flow of consumer behavior by accounting for both temporal variations and lexical precision, demonstrating its qualitative superiority over static batch analysis methods.

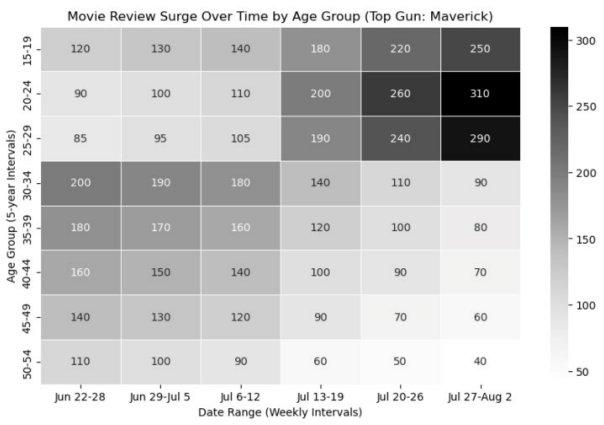


Fig. 6. Movie review surge over time by age group.

4.3. Rapid Trend Detection: Speed Advantage of Real-Time Analysis Over Batch Analysis

This study highlights the superior speed and accuracy of streaming analysis in detecting consumer trend changes compared to traditional batch analysis methods. Conventional post-hoc K-means clustering achieved an average classification accuracy of 70–75%. However, due to erroneous data handling during clustering, it failed to identify cluster transition points effectively, undermining reliability and limiting the utility of trend analysis [21].

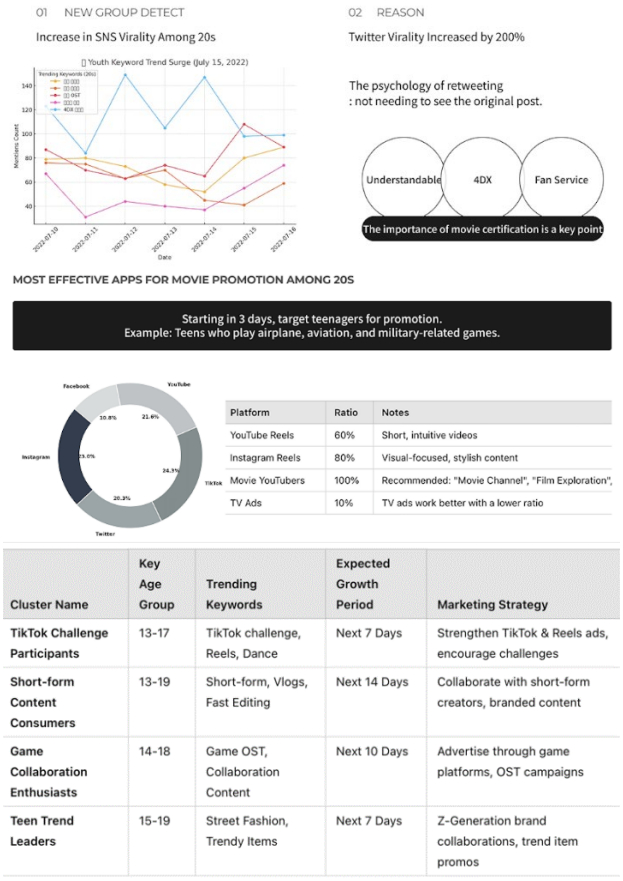
Additionally, the proposed method significantly improved the speed of cluster detection and detailed trend analysis compared to batch analysis. According to Table 3, the batch analysis method required an average of 35.83 hours (minimum 24 hours) to identify consumer clusters. Furthermore, an additional 14 days were required to collect and analyze detailed trend information.

The proposed real-time analysis framework, integrating PHT, EMA, and MAB algorithms, demonstrated significant improvements in both speed and detail for cluster detection and trend analysis. As shown in Table 3, the batch analysis method required an average of 35.83 hours (minimum 24 hours) to identify consumer clusters and an additional 14 days to collect and analyze detailed trend information. In contrast, the real-time framework detected cluster changes within an average of 3.56 hours (minimum 3 hours) and identified detailed trend shifts in just 3.6 hours. Furthermore, the previous anomaly detection model [22] proposed by Guha et al. (2016) required an average of 7 hours for anomaly detection, whereas the proposed framework completed this task in 3.6 hours—achieving a speed improvement of over 48.57%.

Overall, the real-time framework accelerated cluster discovery by approximately 10.06 times and reduced trend analysis time by approximately 93% compared to batch analysis methods. This dramatic improvement enables companies to respond to consumer trend shifts in real-time, enhancing agility in decision-making processes. Additionally, the accuracy of consumer trend change prediction reached 89.6%, representing a 14.6 percentage point improvement over previous studies. To further illustrate practical applications, an example of an automated reporting system utilizing the Google Slides API is provided in Fig.7.

Table 3. Exploration speed comparison.

Analysis method	Average cluster discovery time (hours)	Minimum cluster discovery time (hours)	Trend analysis time (hours)
Batch analysis	35.83	24.0	336.0
This study	3.56	3.0	3.6



Platform	Ratio	Notes
YouTube Reels	60%	Short, intuitive videos
Instagram Reels	80%	Visual-focused, stylish content
Movie YouTubers	100%	Recommended: "Movie Channel", "Film Exploration",
TV Ads	10%	TV ads work better with a lower ratio

Cluster Name	Key Age Group	Trending Keywords	Expected Growth Period	Marketing Strategy
TikTok Challenge Participants	13-17	TikTok challenge, Reels, Dance	Next 7 Days	Strengthen TikTok & Reels ads, encourage challenges
Short-form Content Consumers	13-19	Short-form, Vlogs, Fast Editing	Next 14 Days	Collaborate with short-form creators, branded content
Game Collaboration Enthusiasts	14-18	Game OST, Collaboration Content	Next 10 Days	Advertise through game platforms, OST campaigns
Teen Trend Leaders	15-19	Street Fashion, Trendy Items	Next 7 Days	Z-Generation brand collaborations, trend item promos

Fig. 7. Example of an Automated Report Based on Cluster Analysis.

V. CONCLUSION

Traditional retrospective data analysis methods struggled to quickly adapt to changes in consumer behavior. This study addressed these limitations by employing automated real-time data crawling and analysis to promptly identify cluster formation and propose strategic responses. Using *Top Gun: Maverick* as a case study, the approach demonstrated its effectiveness in incorporating cluster characteristics and behavioral patterns, highlighting the importance of detailed analysis for understanding diverse consumer groups.

The experimental results showed significant improvements in both the precision of cluster exploration and the speed of detection. By shifting from traditional segmentation methods based on demographics to behavior- and interest-based clustering, this study developed a more sophisticated consumer cluster model. The proposed real-time framework outperformed batch analysis, accelerating cluster detection by 10.06 times and reducing trend analysis time by 93%, proving its value in responding to rapidly changing consumer trends. Future research should explore the application of this methodology across various industries to further validate its effectiveness and enhance data-driven decision-making.

REFERENCES

[1] Y. Liu, "Word of mouth for movies: Its dynamics and impact on box office revenue," *Journal of Marketing*, vol. 70, no. 3, pp. 74-89, 2006.

[2] Y. Kim and J. Lee, "Cultural omnivorousness and social status in South Korea: An analysis of 2012 and 2018 Seoul survey data," *Korean Journal of Sociology*, vol. 57, no. 2, pp. 145-168, 2023.

[3] https://news.sbs.co.kr//news/endPage.do?newsId=N10 07673030&cooper=KAKAOTALK

[4] H. Choi and H. Varian, "Predicting the present with Google trends," *Economic Record*, vol. 88, no. s1, pp. 2-9, 2012.

[5] M. A. Russell, *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more,* Sebastopol, CA: O'Reilly Media, 2013.

[6] J. Y. Kim and D. S. Kim, "A study on the method for extracting the purpose-specific customized information from online product reviews based on text mining, "*The Journal of Society for e-Business Studies*, vol. 21, no. 2, pp. 151-170, 2016.

[7] M. Jang, S. Oh, and E. Kim, "Article analytic and summarizing algorithm by facilitating TF-IDF based on k-means," in *Proceedings of the 2018 Spring Conference of the Korea Information Processing Society*, 2018, pp. 271-274.

[8] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007, pp. 443-448.

[9] S. Few, *Show me the numbers: Designing Tables and Graphs to Enlighten,* Burlingame, CA; Analytics Press, 2012.

[10] A. Mutsaddi, A. Jamkhande, A. Thakre, and Y. Haribhakta, "BERTopic for topic modeling of Hindi short texts: A comparative study," *arXiv preprint, arXiv*:2501.03843, Jan. 2024.

[11] L. Cappello, O. H. Madrid Padilla, and J. A. Palacios, "Bayesian change point detection with spike-and-slab priors," *Journal of Computational and Graphical Statistics*, vol. 32, no. 4, pp. 1488-1500, 2023.

[12] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, Reilly Media, 2018.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[14] J. Gama, R. Sebastião, and P. P. Rodrigues, "On evaluating stream learning algorithms," *Machine Learning*, vol. 90, no. 3, pp. 317-346, 2013.

[15] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in Artificial Intelligence*, vol. 2009, pp. 1-19, 2009.
[16] CGV. (n.d.). Top Gun: Maverick–http://www.cgv.co.kr/movies/detailview/?midx=82120#commentReg
[17] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010.
[18] A. G. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Change-Point Detection*, Boca Raton, FL: Chapman & Hall/CRC, 2014.
[19] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice,* OTexts, 2018.
[20] J. Y. Audibert, S. Bubeck, and R. Munos, "Minimax policies for adversarial and stochastic bandits," in *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, 2009.
[21] C. Shim, *Optimal K-Means Algorithm Based on Automated Techniques for Real-Time Big Data Analysis (Final Research Report)*, Korea: Sunchon National University, 2020.
[22] S. Guha, N. Mishra, G. Roy, and O. Schrijvers, "Robust random cut forest based anomaly detection on streams," in *Proceedings of the 33rd International Conference on Machine Learning*, PMLR 48, 2016, pp. 2712-2721.

AUTHORS

Hyeonji Ko is a bachelor's student majoring in Business Administration and Data Science at Dongduk Women's University. She is particularly interested in data-driven decision-making and data marketing, exploring how data can enhance business strategies and consumer insights.

Baekbin Ko is a bachelor's student majoring in Business Administration and data Science at Dongduk Women's University. Her research interests include data analysis techniques and data-driven decision-making.

Taewan Kim received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Laboratory, SK Telecom, Seoul. In 2022, he joined as a faculty with the Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, where he is currently an Assistant Professor.
His research interests include computer vision and machine learning including continual and online learning.