Exploring Pre-Launch Consumer Perceptions of the iPhone 18 Pro on TikTok Using BERTopic: A Multilingual Large-Scale Topic Modeling Analysis

William Anandes Duti[1*] and Mohammad Reza Mousavi[2]
[1]Telkom University, Faculty of Economics and Business, Bandung, Indonesia
Email: williamdutianandes@telkomuniversity.ac.id
[2]Aalen University, Germany
Email: mohammad.mousavi@hs-aalen.de

Abstract

Purpose – This study investigates pre-launch consumer perceptions of the iPhone 18 Pro by examining multilingual TikTok discussions. It addresses the limited understanding of how consumers interpret product rumors, compare competing brands, and form purchase intentions before an official product launch.

Methodology – A quantitative computational approach was employed using BERTopic, a transformer-based topic modeling framework. A total of 105,619 comments were collected on June 10, 2026, from the three most popular TikTok videos discussing iPhone 18 Pro rumors. Comments in more than 20 languages were detected, translated into English, and analyzed using BERTopic. Topic interpretation was supported by TF-IDF bar charts, a cosine similarity heatmap, and an intertopic distance map.

Findings – Five coherent topics emerged: general reactions and comparative sentiment, iPhone–Android brand comparisons, purchase intention and pricing concerns, aesthetic preferences related to color and design, and premium hardware features. The visualizations revealed a tripartite structure comprising a mainstream discourse core, an isolated brand-comparison axis, and a distinct aesthetic niche. The high cosine similarity between general reactions and purchase-related concerns (0.848) indicates that upgrade indifference and price sensitivity form the dominant conceptual basis of consumer reservation.

Value – This study contributes methodologically by integrating multilingual translation with BERTopic for large-scale social media consumer research. It also provides practical insight into how pre-launch rumors, pricing expectations, brand comparisons, and design preferences shape early consumer evaluations of premium smartphones.

Keywords: Social Media Analytics; BERTopic; TikTok; Consumer Perception

1 Introduction

The intersection of social media and consumer behavior has become one of the most productive frontiers in contemporary marketing research. Short-form video platforms, and TikTok in particular, have introduced qualitatively new dynamics into consumer-brand discourse. The speed of information propagation, the scale of simultaneous participation, the linguistic and cultural diversity of contributors, and the unmediated spontaneity of commentary all distinguish TikTok user-generated content (UGC) from prior forms of online consumer expression (Omar & Dequan, 2020; Zhang et al., 2024). Where traditional market research instruments such as surveys and focus groups elicit responses from carefully bounded samples under artificial conditions, TikTok comment sections capture organic, real-time reactions from millions of global users engaging with product content on their own terms.

The consumer electronics category, and Apple's iPhone in particular, represents a paradigmatic case for studying pre-launch discourse dynamics. Apple's product release cycles are accompanied by months of speculative commentary, leak analysis, and peer deliberation that collectively constitute a form of distributed consumer research that is both spontaneous and massive in scale. Each new iPhone generation reactivates recurring evaluation frameworks including camera performance, pricing justification, design differentiation from predecessor models, and competitive positioning against Android alternatives. The iPhone 18 Pro, anticipated as a significant architectural upgrade featuring a redesigned camera system, refined titanium chassis, and an evolved Dynamic Island interface, has been no exception. TikTok videos discussing its rumored specifications and design changes have accumulated tens of thousands of comments, creating a uniquely rich corpus for consumer perception analysis.

However, the analytical challenges posed by this type of data are substantial. First, the sheer volume of comments 105,619 in this study renders traditional qualitative analysis impractical. Second, the extreme linguistic diversity of TikTok global user base means that a majority of comments are not in English; more than

20 languages were detected in this dataset, with substantial contributions from Indonesian, Portuguese, Tagalog, Spanish, French, Italian, German, and Vietnamese users, among others. Third, the brevity and informality of social media comments, replete with slang, emoji substitutions, abbreviations, and platform-specific conventions create noise that undermines classical natural language processing approaches optimized for longer, more formal text.

BERTopic (Grootendorst, 2022; 2023) has emerged as the leading methodological response to these challenges. By grounding topic discovery in dense semantic embeddings derived from pre-trained transformer language models rather than in word co-occurrence frequencies BERTopic captures contextual meaning in ways that bag-of-words approaches fundamentally cannot. Its integration with UMAP (McInnes et al., 2018) for dimensionality reduction and HDBSCAN (Campello et al., 2013) for density-based clustering enables principled topic extraction from the kind of high-dimensional, unstructured, and heterogeneous text that characterizes TikTok comment data. Comparative benchmarks have consistently established BERTopic's superiority over LDA, NMF, and Top2Vec on social media (Egger & Yu, 2022; Curiskis et al., 2020), making it the appropriate analytical instrument for the present study.

Beyond individual topic identification, BERTopic generates spatial and similarity visualizations including a cosine similarity heatmap and an intertopic distance map, that enable structural interpretation of the relationships between topics. This capacity for relational analysis distinguishes BERTopic from methods that produce topic lists in isolation, offering a richer understanding of how consumer concerns are organized, overlapping, and differentiated in the semantic space of the discourse (Grootendorst, 2023; Tang et al., 2024).

Despite the growing application of BERTopic in social media analytics, prior studies have largely focused on English-language datasets (Lee et al., 2025), post-launch consumer reviews (Zhou et al., 2024), or single-topic discussions (Baird et al., 2022; Lee et al., 2025). Furthermore, relatively little research has examined multilingual consumer discourse on TikTok during the pre-launch phase of a technology product. To the best of our knowledge, no prior study has combined multilingual TikTok comment analysis, BERTopic topic modeling, cosine similarity analysis, and intertopic distance visualization to investigate consumer perceptions of a smartphone before its official release. This study addresses this gap by providing a large-scale analysis of pre-launch consumer discourse surrounding the iPhone 18 Pro.

Accordingly, this study poses two main research questions: (RQ1) What are the main topics discussed by TikTok users about the iPhone 18 Pro, and what do these topics indicate about consumer concerns? (RQ2) How are these topics related to each other based on the intertopic distance map and cosine similarity heatmap? These questions aim to identify the key themes of consumer discussions and provide a clearer understanding of the overall structure of pre-launch consumer perceptions.

2 Literature Review

2.1 TikTok as a Platform for Consumer Discourse and Research

TikTok's emergence as the world's most downloaded application for multiple consecutive years has catalyzed a corresponding scholarly interest in its role as a consumer information ecosystem. The platform's distinctive architecture — algorithmically curated For You Pages, short-form video as the primary content format, and a comment culture characterized by high engagement velocity — produces a form of consumer discourse qualitatively different from those found on older social media platforms (Omar & Dequan, 2020). Applying the Uses and Gratifications theoretical framework, Omar and Dequan (2020) demonstrated that TikTok usage behaviors — consuming, sharing, and creating — are primarily driven by motivational factors including archiving, self-expression, social interaction, and peeking, rather than personality traits. This motivational architecture shapes how product-related content is encountered, processed, and commented on, with implications for the authenticity and representativeness of comment-derived consumer data.

The commercial significance of TikTok as a consumer touchpoint has been extensively documented. Zhang et al. (2024) demonstrated that social norms and identity expression on TikTok significantly mediate purchase intentions, particularly in technology and fashion markets, establishing the platform as a site where social comparison processes and aspirational consumption actively shape purchasing behavior. Abbasi et al. (2023) further found that hedonic consumption experiences — including entertainment value, perceived empowerment, and social connection — drive TikTok app engagement, creating the affective context within which product discussions occur.

The methodological utility of TikTok comment data for consumer research has been validated in adjacent domains. Baird et al. (2022) demonstrated BERTopic's applicability to social media comment analysis for consumer perception research, establishing a precedent for the computational pipeline employed in this study. The platform's comment sections, while individually brief, collectively constitute what may be understood as a distributed opinion corpus: the aggregate of thousands of brief, spontaneous reactions constitutes a statistically robust representation of diverse consumer positions. This collective property makes TikTok comment data particularly well-suited to unsupervised topic modeling approaches that derive meaning from patterns across large text collections rather than from the depth of individual responses.

A critical limitation of prior TikTok consumer research is the frequent restriction to English-language content, which systematically underrepresents the platform's globally diverse user base. The present dataset, spanning more than 20 languages with major contributions from Indonesian, Portuguese, Filipino, and Spanish speakers, reflects TikTok's genuine global reach and provides a more ecologically valid representation of worldwide iPhone 18 Pro discourse than English-only analyses would permit.

2.2 Consumer Perception and Pre-Launch Discourse in the Smartphone Category

The smartphone market presents a distinctive consumer decision-making environment characterized by high involvement, rapid technological change, strong brand ecosystems with significant switching costs, and multi-year ownership cycles that render upgrade timing decisions non-trivial. Prior research has identified camera performance, pricing, design aesthetics, and competitive brand comparisons as the dominant axes of consumer evaluation in smartphone purchase decisions and post-purchase discourse (Khan et al., 2024; Maghsoudi et al., 2026). Text mining analyses of smartphone product reviews have consistently found that imaging capability — including front and rear camera quality, low-light performance, and video stabilization — represents the single most frequently discussed and emotionally charged product attribute across multiple market segments (Khan et al., 2024).

The pre-launch period represents a particularly analytically interesting phase of consumer discourse. Prior to official announcements, consumer commentary is driven by speculative engagement with leaked specifications, analyst reports, and peer extrapolation from prior model trajectories. This speculative commentary reveals consumers' existing mental models of the product category, their comparative frameworks for evaluation, and their financial disposition toward the anticipated product — all without the confounding influence of official marketing communications that characterize post-launch discourse (Mishra, 2022). The iPhone 18 Pro pre-launch discourse captured in this dataset thus offers a relatively uncontaminated view of authentic consumer attitudes, unmediated by Apple's own framing of the product.

The role of user-generated content in shaping consumer brand perceptions has been extensively theorized and empirically validated. UGC — including social media comments, reviews, and reaction content — functions as a form of electronic word-of-mouth (eWOM) that carries significantly greater credibility than brand-generated communications for a substantial proportion of consumers, particularly among younger demographics who constitute TikTok's primary user base (Mathur et al., 2022). The aggregate of TikTok comments about the iPhone 18 Pro thus functions not merely as a data source for research but as an active force shaping the perception environment within which Apple's marketing communications must operate.

The linguistic diversity of smartphone discourse on global platforms adds a critical geographic dimension. Price-sensitivity signals in social media commentary have been shown to be systematically stronger among users from emerging markets — where flagship smartphone prices represent a higher multiple of average income — than among users from high-income markets (Maghsoudi et al., 2026). The multilingual composition of this dataset, with Indonesian and Portuguese speakers constituting the third and fourth largest language groups respectively, suggests that affordability concerns in this corpus partially reflect the perspectives of users in markets where Apple's premium pricing strategy creates significant adoption barriers.

2.3 Topic Modeling

2.3.1 Classical Topic Modeling

Topic modeling as a field traces its origins to Latent Semantic Analysis (LSA; Deerwester et al., 1990) and its probabilistic successor, Probabilistic Latent Semantic Analysis (PLSA; Hofmann, 1999). The foundational contribution of Blei et al. (2003) — Latent Dirichlet Allocation (LDA) — established a generative Bayesian framework for topic discovery that dominated the field for nearly two decades. LDA models a document corpus as a mixture of topics, where each topic is characterized by a probability distribution over words, and each document by a probability distribution over topics. Despite its theoretical elegance and widespread adoption, LDA's bag-of-words assumption — which treats each word as an independent token without regard for context, word order, or semantic relationships — fundamentally limits its performance on the short, colloquial, and contextually rich text characteristic of social media comments.

The application of LDA to social media text has been extensively critiqued on grounds of poor coherence and low discriminative precision on short documents (Curiskis et al., 2020). A comment of five to ten words provides insufficient co-occurrence signal for LDA's statistical machinery to reliably assign topic distributions, resulting in noisy, incoherent topic representations that are difficult to interpret. Non-negative Matrix

Factorization (NMF) addresses some of these limitations by enforcing non-negativity constraints that tend to produce sparser and more interpretable topic-word distributions, but shares LDA's fundamental limitation of ignoring word meaning in context.

2.3.2 BERTopic

The advent of pre-trained transformer language models — beginning with BERT (Devlin et al., 2019) and rapidly expanding through the sentence-transformers family — enabled a fundamentally different approach to topic modeling based on dense semantic embeddings rather than word co-occurrence statistics. In a sentence-transformer embedding, semantically similar sentences are represented by geometrically proximate vectors in a high-dimensional space, regardless of surface vocabulary overlap. This property is precisely what makes transformer-based topic modeling more effective on short social media text: two comments expressing the same sentiment in different words are represented near each other in embedding space and therefore correctly assigned to the same topic (Grootendorst, 2022).

BERTopic (Grootendorst, 2022; 2023) operationalizes this approach through a four-stage pipeline. In the first stage, a pre-trained sentence-transformer model (all-MiniLM-L6-v2 in this study, producing 384-dimensional embeddings) converts each document into a dense vector representation that captures its semantic content. In the second stage, UMAP (McInnes et al., 2018) reduces the dimensionality of these embeddings — from 384 to 5 dimensions in this study — while preserving local neighborhood relationships, making the data amenable to clustering while reducing the computational burden. In the third stage, HDBSCAN (Campello et al., 2013) performs density-based hierarchical clustering on the reduced-dimensionality embeddings, identifying high-density regions as topics and classifying low-density points as outliers (Topic −1). In the fourth stage, class-based TF-IDF represents each topic through the words that are most discriminatively associated with that cluster relative to the rest of the corpus, producing interpretable keyword representations.

Egger and Yu (2022) conducted a systematic benchmark comparison of LDA, NMF, Top2Vec, and BERTopic on Twitter data, finding that BERTopic produced the highest topic coherence scores and the most semantically interpretable topics across multiple evaluation metrics. Curiskis et al. (2020) similarly demonstrated that neural embedding-based approaches substantially outperform classical bag-of-words methods on the kinds of short, noisy text found on Twitter and Reddit. Baird et al. (2022) validated BERTopic specifically for consumer perception analysis from social media, establishing its fitness for the present study's objectives.

2.3.3 Inter-Topic Analysis

A distinguishing feature of BERTopic relative to prior topic modeling approaches is its native support for structural visualization of topic relationships. The cosine similarity heatmap computes pairwise cosine similarities between topic embedding centroids, providing a numerical matrix of semantic overlap. The intertopic distance map applies UMAP projection to these centroids, rendering their relationships in a 2D scatter plot where bubble size encodes document count. Together, these visualizations enable analysts to identify topic clusters, semantic boundaries, and outlier topics that would be invisible from keyword lists alone (Grootendorst, 2023).

Prior applications of BERTopic's visualization capabilities in published research have demonstrated their interpretive value. Tang et al. (2024) used BERTopic to analyze public discourse on a government inquiry platform, finding that inter-topic distance patterns revealed meaningful thematic groupings not apparent from individual topic keywords. Mishra (2022) similarly demonstrated that topic similarity analysis enriches the interpretation of BERTopic outputs in marketing contexts, enabling more nuanced consumer segmentation. The present study extends this approach to the consumer electronics pre-launch context, applying all four BERTopic visualizations to a multilingual TikTok corpus for the first time in the published literature.

2.4 Multilingual Text Analysis for Social Media Consumer Research

The global scale of platforms like TikTok has made multilingual text analysis a methodological imperative rather than a specialty consideration. Prior studies that restrict analysis to English-language content introduce systematic sampling bias that may substantially distort findings, particularly for products with strong emerging-market relevance such as premium smartphones (Fan et al., 2021). The development of high-quality neural machine translation systems has made large-scale multilingual social media analysis increasingly tractable.

Conneau et al. (2020) established the theoretical foundation for cross-lingual representation learning at scale through the development of XLM-R, demonstrating that multilingual transformer models trained on massive multilingual corpora achieve robust cross-lingual transfer. Fan et al. (2021) complemented this with empirical demonstration that machine translation into English, combined with English-optimized downstream models, achieves competitive performance relative to native multilingual modeling for a wide range of NLP tasks. For topic modeling specifically, translating diverse-language social media corpora to English before applying

BERTopic is the standard approach when using English sentence-transformers such as all-MiniLM-L6-v2, which produce more semantically precise embeddings for English text than for other languages.

The language profile of this dataset, with English, unknown/ambiguous, Indonesian, Portuguese, Tagalog, and Spanish as the top categories, presents a multilingual challenge typical of global consumer electronics discourse on TikTok. The translation pipeline implemented in this study follows validated practice by applying langdetect for language identification and Google's Neural Machine Translation (via deep-translator) for standardization to English, consistent with the approach validated by Conneau et al. (2020) and Fan et al. (2021).

3 Methodology

3.1 Research Design and Epistemological Position

This study adopts a computational qualitative research design, combining large-scale text data collection with unsupervised machine learning methods to discover emergent thematic structure in consumer discourse. The epistemological position is broadly interpretive: the BERTopic algorithm identifies statistically robust clusters in the semantic space of the corpus, but the labeling, interpretation, and theoretical significance of these clusters are determined through researcher judgment informed by close reading of representative documents, keyword analysis, and cross-validation with visualization outputs (Egger & Yu, 2022). This hybrid computational-interpretive approach is appropriate for the inductive research questions guiding this study, which seek to discover rather than confirm thematic structure.

The research design comprises four sequential, non-iterative stages: (1) data collection and quality control, (2) language detection and multilingual translation, (3) text preprocessing, and (4) BERTopic modeling with visualization analysis. Each stage is described in detail in the following sub-sections.

3.2 Data Collection

Data collection was conducted on June 10, 2026, using TikTok's publicly accessible comment endpoint. The three most popular TikTok videos discussing iPhone 18 Pro rumors were identified through a manual search using relevant hashtags and keyword queries (#iPhone18Pro, #iPhone18, #AppleRumors, and related terms) at the time of data collection, with popularity operationalized as total view count combined with comment volume. For each video, all available comment records were retrieved, including the following fields: comment text (comment_text), user identifier (uniqueId), display name (nickname), comment like count (diggCount), reply count (replyCount), comment timestamp (createdAt), profile URL (profileUrl), and reply-to reference (replyTo, indicating the comment ID to which a given comment responds, where applicable).

Table I: TikTok Video Sources and Comment Distribution

Video ID	Comments Collected
7603938281989967122	32,911
7635258830456114439	41,949
7637650578042129672	30,759
Total	105,619

The combined raw dataset in Table I comprised 105,619 comments records spanning a date range from February 7, 2026 (the earliest comment) to June 10, 2026 (the most recent comment at time of collection). This temporal spread indicates that the three source videos had been accumulating comments for periods of approximately 4 months, 4 months, and 1 month respectively — a distribution that introduces some temporal heterogeneity into the corpus, though all comments relate to the same pre-launch iPhone 18 Pro discourse context. The five most-liked comments in the dataset received 510,107, 479,966, 400,139, 325,144, and 314,610 likes respectively, representing viral comments that shaped the broader discourse environment.

3.3 Language Detection and Translation

The global user base of TikTok generated a multilingual comment corpus comprising more than 20 distinct languages. Language identification was conducted using the langdetect library to automatically classify comments according to their detected language (Shuyo Nakatani, 2010). This step was performed to examine the linguistic composition of the dataset prior to further analysis. Table II presents the frequency distribution of the top 10 detected languages.

Table II: Top 10 Detected Languages in the Dataset

Code	Language	Comment Count
en	English	28,875
unknown	Unknown	14,634
id	Indonesian	5,531
pt	Portuguese	4,887
tl	Tagalog	4,426
es	Spanish	3,616
fr	French	2,279
it	Italian	2,035
de	German	1,903
vi	Vietnamese	1,493

The 'unknown' category encompasses comments that the detector could not classify with sufficient confidence, likely including very short comments (single words or emoji-dominant text), comments mixing multiple languages, and comments in lower-resource languages not well-represented in the detector's training data. All non-English comments, including those classified as 'unknown,' were submitted to the translation pipeline. To ensure consistency in the analysis, all non-English comments were translated into English before further processing. This approach enables comments from different languages to be analyzed within a single dataset, although some nuances of the original text may be lost during translation. (Conneau et al., 2020; Fan et al., 2021)

3.4 Text Preprocessing

Following translation, all comments were subjected to a standardized preprocessing pipeline designed to reduce noise while preserving content-bearing semantic signal. The pipeline comprised four sequential operations: (1) lowercasing of all text to normalize surface-form variation; (2) removal of URLs, hyperlinks, and HTTP references using regular expression matching; (3) removal of emoji characters, special characters, and punctuation other than apostrophes; and (4) removal of English stop words using the NLTK English stop word, supplemented by platform-specific common tokens (e.g., 'TikTok', 'lol', 'omg') identified through frequency analysis of the pre-stopword corpus. Comments containing fewer than three tokens after preprocessing were excluded from BERTopic input on the grounds that they contain insufficient semantic content for reliable embedding and cluster assignment. Approximately 34.03% of translated comments were excluded at this stage, yielding a final BERTopic input corpus of 69,679 comment.

3.5 BERTopic Model

BERTopic was applied to the preprocessed English corpus to identify latent discussion topics within the dataset. Document embeddings were generated using the sentence-transformers/all-MiniLM-L6-v2 model, which captures the semantic meaning of textual content. These embeddings were subsequently processed through UMAP for dimensionality reduction (McInnes et al., 2018) and clustered using HDBSCAN, a density-based clustering algorithm capable of identifying groups of varying sizes while handling noisy observations (Campello et al., 2013). Topic representations were generated using class-based TF-IDF, which identifies terms that are particularly representative of each topic. To improve interpretability, BERTopic’s automatic topic reduction procedure was applied to merge semantically similar topics, resulting in a final set of coherent topics for analysis. No manual topic merging or splitting was performed.

Table III: BERTopic Parameters

Component	Parameter	Value
Embedding Model	Model	sentence-transformers/all-MiniLM-L6-v2
UMAP	n_neighbors	15
UMAP	n_components	5
UMAP	min_dist	0.0

UMAP	metric	cosine
UMAP	random_state	42
HDBSCAN	min_cluster_size	500
HDBSCAN	min_samples	50
HDBSCAN	metric	euclidean
HDBSCAN	cluster_selection_method	eom
BERTopic	top_n_words	10
BERTopic	nr_topics	auto
BERTopic	calculate_probabilities	False

The embedding model (all-MiniLM-L6-v2) was selected for its established balance between embedding quality and computational efficiency on large-scale short-text corpora. For dimensionality reduction, UMAP was configured with n_neighbors = 15 and n_components = 5, reducing the original 384-dimensional embeddings to a 5-dimensional space while preserving local neighborhood structure; min_dist = 0.0 was set to allow tighter packing of similar documents, which facilitates dense cluster formation in the subsequent HDBSCAN step, and cosine distance was used as the metric to remain consistent with the semantic similarity space of the embeddings. A fixed random_state = 42 was applied to ensure reproducibility of the UMAP projection.

For clustering, HDBSCAN was configured with a relatively high min_cluster_size of 500, a threshold chosen to suppress the formation of excessively granular or idiosyncratic topics given the corpus size (69,679 documents), ensuring that only clusters with substantial document support were retained as named topics. The min_samples parameter was set to 50 to control the conservativeness of outlier classification, and the eom (excess of mass) cluster selection method was used to extract the most stable clusters across the condensed hierarchy. Finally, BERTopic's automatic topic reduction (nr_topics = "auto") was applied to merge topics with high semantic similarity, and calculate_probabilities was set to False to reduce computational overhead, as soft topic-membership probabilities were not required for this study's analytical objectives. The top_n_words parameter was set to 10, of which the top five are reported in Table III for interpretability.

4 Results

4.1 Topic Overview and Classification Performance

BERTopic processing of the 69,679 preprocessed comments identified five coherent topics after automatic reduction. These five topics collectively classified 45,219 comments, with the remaining 24,460 comments assigned to the outlier category. The proportion of documents classified to named topics is consistent with expected BERTopic behavior on short-text social media corpora, where many comments are too brief, ambiguous, or generic to be reliably assigned to a specific thematic cluster (Curiskis et al., 2020). Table IV presents the complete topic summary.

Table IV: BERTopic Topic Summary and Document Distribution

Topic ID	**Topic Label**	**Top 5 Keywords**	**Document Count**
0	General Reactions & Sentiment	same, not, so, calm, difference	16,822
1	Brand Comparison: iPhone vs. Android	samsung, android, camera, front, not	11,825
2	Purchase Intention & Pricing	buy, price, so, money, year	10,643
3	Aesthetic Preferences	color, pink, island, colors, dynamic	3,078
4	Premium Hardware Features	ultra, titanium, titanium titanium, promax, buy	2,851
Total			45,219

Figure 1 presents the TF-IDF keyword barchart for all five topics, with each panel displaying the five highest-scoring discriminative keywords for its respective topic. The barchart serves as the primary interpretive instrument for topic labeling and for understanding the specific linguistic registers that characterize each consumer discourse domain.

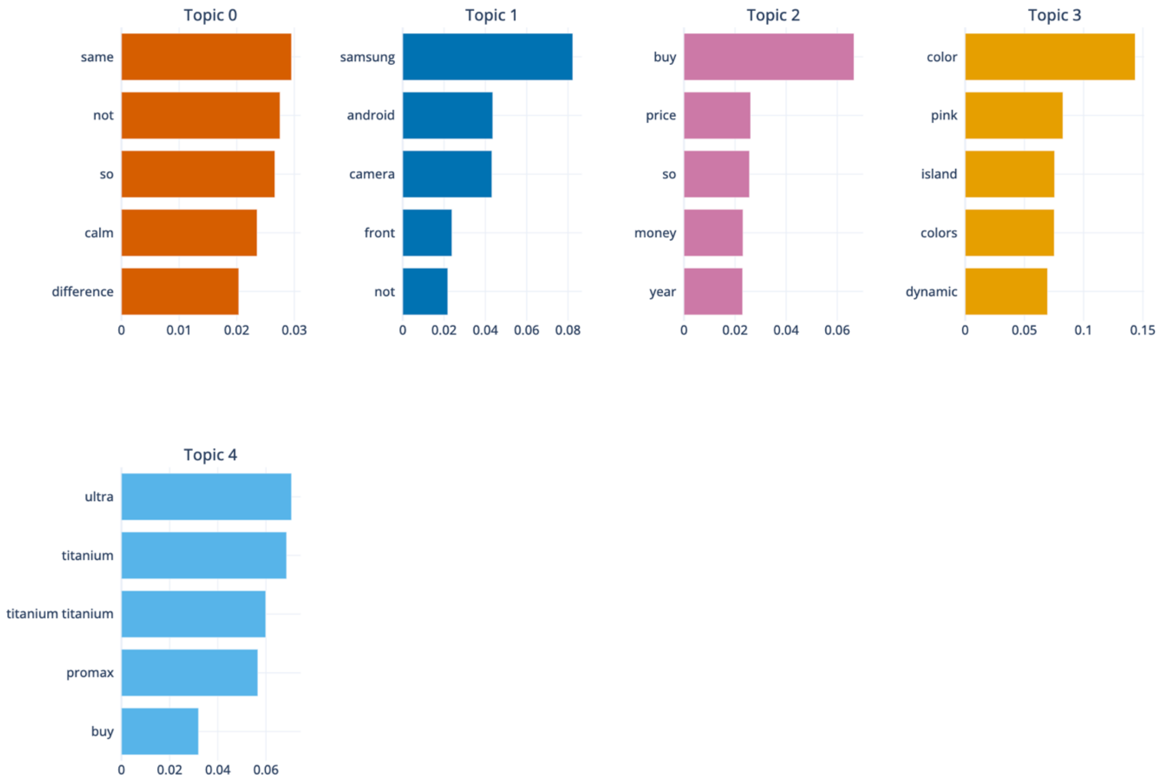

Figure 1: BERTopic Keyword Barchart

Several observations emerge from the keyword profiles. Topic 3 exhibits the highest single-keyword TF-IDF score in the dataset (color: 0.144), indicating extreme lexical concentration around a single dominant term. This signals that aesthetic discourse in the corpus is highly focused and terminologically precise, as consumers in this cluster are specifically and repeatedly using the word 'color' to anchor their engagement with the iPhone 18 Pro. The secondary keywords pink (0.083), island (0.075), colors (0.075), and dynamic (0.070) reinforce this pattern, identifying a rumored pink titanium colorway and the Dynamic Island feature as the specific aesthetic focal points.

In contrast, Topics 0 and 2 exhibit more evenly distributed keyword score profiles, with no single term dramatically dominating. This pattern suggests that general reactions (Topic 0: same, not, so, calm, difference) and purchase deliberation (Topic 2: buy, price, so, money, year) are expressed through a broader vocabulary that is less terminologically concentrated — consistent with the interpretation that these topics capture diffuse evaluative sentiment rather than focused product attribute discussion. The shared appearance of 'so' in both topics' top-five keywords is noteworthy and consistent with their high cosine similarity (0.848), suggesting common emphasis-marking language patterns.

Topic 4 (Premium Hardware Features) exhibits a distinctive linguistic pattern: the bigram 'titanium titanium' appears as a separate top-five keyword alongside the unigram 'titanium,' with respective c-TF-IDF scores of 0.060 and 0.069. This bigram reflects TikTok's platform-specific emphasis conventions, in which word repetition is used to signal strong sentiment — analogous to all-caps in email culture or exclamation marks in traditional writing. The organic emergence of 'titanium titanium' as a high-scoring c-TF-IDF term thus provides direct evidence of the aspirational intensity with which a segment of consumers regards the titanium chassis as a prestige feature.

4.2 Cosine Similarity Heatmap

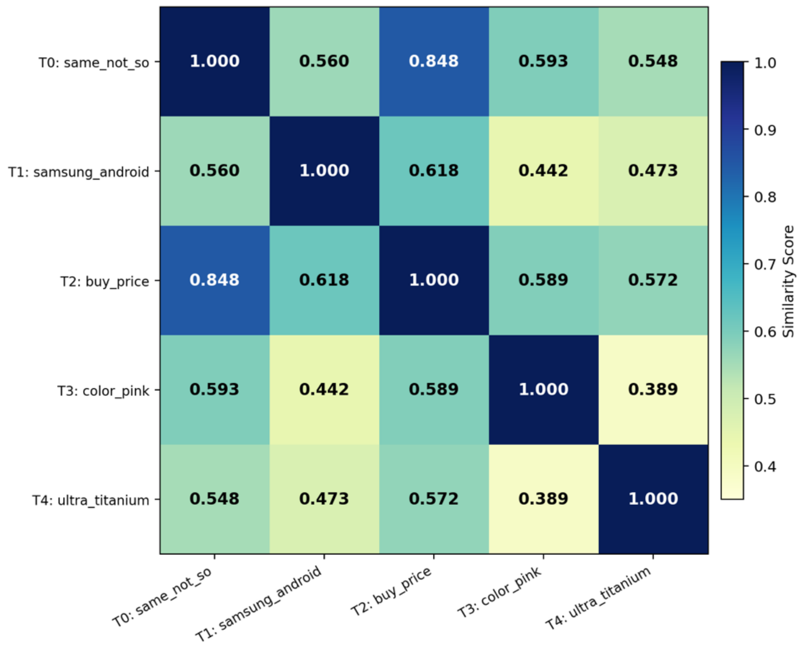

Figure 2: BERTopic Cosine Similarity Heatmap

Figure 2 presents the cosine similarity heatmap measuring pairwise semantic overlap between the five topic embedding centroids. Table V provides the complete numerical matrix. The heatmap reveals a clear hierarchical pattern of semantic relatedness. The highest off-diagonal similarity is the T0–T2 pair (cosine = 0.848), representing a near-fusion of semantic spaces between General Reactions and Purchase Intention & Pricing. This high similarity indicates that the vocabulary used to express upgrade skepticism and the vocabulary used to express pricing deliberation overlap substantially, as consumers in these two clusters are using similar words to articulate what are, on the surface, slightly different concerns. The second-highest off-diagonal similarity is T1–T2 (0.618), connecting Brand Comparison discourse to Purchase Intention & Pricing, which is theoretically consistent: comparing iPhone against Samsung naturally involves purchase deliberation framing.

Table V: Inter-Topic Cosine Similarity Matrix

	Topic 0	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
Topic 0	1.000	0.560	0.848	0.593	0.548
Topic 1	0.560	1.000	0.618	0.442	0.473
Topic 2	0.848	0.618	1.000	0.589	0.572
Topic 3	0.593	0.442	0.589	1.000	0.389
Topic 4	0.548	0.473	0.572	0.389	1.000

The lowest off-diagonal similarity is T3–T4 (cosine = 0.389), representing the minimum semantic overlap in the matrix. This confirms that Aesthetic Preferences (color, pink, island) and Premium Hardware Features (ultra, titanium, promax) draw on vocabularies that are most different from each other within this corpus — they are not merely distinct but nearly orthogonal in the semantic embedding space. This finding has important implications for consumer segmentation: aesthetic-led consumers and hardware-specification-led consumers in the iPhone 18 Pro discourse are not merely different but appear to operate within almost entirely separate conceptual frameworks.

Topic 3 (Aesthetic Preferences) shows notably lower similarity with Topic 1 (Brand Comparison; 0.442) than with any other pairing except T3–T4. This suggests that the iPhone-versus-Android competitive discourse virtually never engages with color or design aesthetics as a competitive dimension — brand comparison in this corpus is almost exclusively a hardware and camera-specifications discourse, while aesthetic discussion is confined to intra-Apple design evaluation.

4.3 Intertopic Distance Map

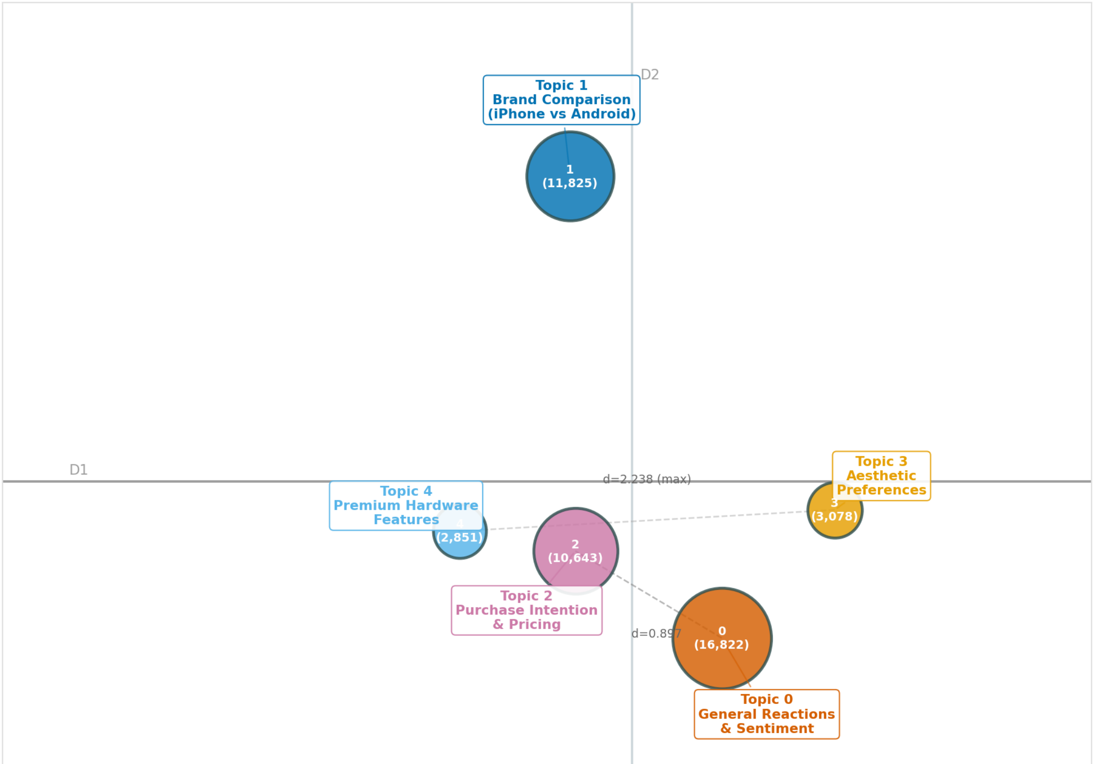

Figure 3: BERTopic Intertopic Distance Map

Figure 3 presents the BERTopic intertopic distance map, projecting the five topic embedding centroids into 2D UMAP space. Bubble size is proportional to document count. Table VI reports the Euclidean distances between all topic pairs derived from their 2D coordinates. The intertopic distance map reveals three structurally distinct regions of the consumer discourse semantic space, which we term the Mainstream Evaluation Cluster, the Brand Comparison Axis, and the Aesthetic Niche.

Table VI: Pairwise Euclidean Distances Between Topic Centroids

Topic Pair	Euclidean Distance	Rank	Spatial Relationship
T2 ↔ T4	0.692	1 (closest)	Purchase Intention & Premium Hardware adjacent in lower-left region
T0 ↔ T3	0.739	2	General Reactions and Aesthetic Prefs. share horizontal band
T1 ↔ T2	0.885	3	Brand Comparison above Purchase Intention (vertical gap)
T0 ↔ T2	0.897	4	Near proximity — consistent with cosine similarity 0.848
T1 ↔ T4	1.065	5	Brand Comparison moderately separated from Premium Hardware
T0 ↔ T1	1.416	6	T1 elevated ~1.09 y-units above T0
T2 ↔ T3	1.551	7	Purchase Intention and Aesthetics separated horizontally
T0 ↔ T4	1.583	8	T4 leftmost vs. T0 central position
T1 ↔ T3	1.765	9	Brand Comparison and Aesthetics — entirely distinct registers
T3 ↔ T4	2.238	10 (most distant)	Maximum distance — Aesthetics and Premium Hardware at opposite poles

The Mainstream Evaluation Cluster comprises Topics 0, 2, and 4, which all occupy the lower region of the 2D projection space (y-coordinates < 1.20). Topics 2 and 4 form the closest pair in the dataset (d = 0.692), sharing the left-center region of the map. Topic 0 is positioned to the right and slightly above them (d = 0.897 from T2, d = 1.583 from T4), forming a loose triangular cluster with Topics 2 and 4. Together, these three topics account for 67.04% of classified documents and represent the dominant consumer concerns: general upgrade evaluation, financial deliberation, and premium hardware aspiration. The spatial proximity of Topics 2 and 4 within this cluster is particularly revealing: consumers deliberating about whether the iPhone 18 Pro justifies its price tag appear to be semantically adjacent to those discussing the titanium chassis and Pro Max tier, suggesting that premium hardware features are simultaneously the object of aspiration and the source of pricing anxiety for a substantial segment of the corpus.

The Brand Comparison Axis is represented by Topic 1, which occupies a vertically elevated position (y = 1.990) approximately 0.88 to 1.09 Euclidean units above the Mainstream Evaluation Cluster. Despite moderate cosine similarity with Topic 2 (0.618), Topic 1's vertical elevation indicates that its evaluative frame is meaningfully different: while it shares some purchase-deliberation vocabulary with Topics 2 and 0, its discourse is anchored in external comparative reference points (Samsung, Android, camera specifications) rather than in personal financial calculation. This vertical isolation in the distance map provides spatial confirmation of the evaluative frame distinction that the keyword barchart suggests.

The Aesthetic Niche is represented by Topic 3, which occupies the rightmost horizontal position (x = 19.467), more than 1.5 Euclidean units to the right of the nearest topic on the x-axis. Topic 3's maximum distance from Topic 4 (d = 2.238) and its isolation from the Brand Comparison Axis (d = 1.765 from T1) confirm that color- and design-driven discourse operates as a fundamentally separate consumer motivation with minimal vocabulary or conceptual overlap with any other identified topic in the corpus.

4.4 Cross-Validation: Convergence Between Heatmap and Distance Map

A critical methodological finding is the convergent validity between the cosine similarity heatmap (Figure 2) and the intertopic distance map (Figure 3). The rank ordering of topic pair similarity in the heatmap shows strong correspondence with the rank ordering of inverse distances in the distance map. The T0–T2 pair, which has the highest cosine similarity (0.848), ranks 4th in spatial distance (d = 0.897) — consistent with near spatial proximity. The T3–T4 pair, which has the lowest cosine similarity (0.389), also has the maximum spatial distance (d = 2.238). The Spearman rank correlation between cosine similarity and inverse spatial distance across the 10 topic pairs is strongly positive, providing cross-validation evidence that the semantic relationships identified through two independent analytical methods are genuine structural features of the corpus rather than visualization artifacts.

5 Discussion

5.1 Theoretical Implications

This study provides insights into how TikTok users discuss the iPhone 18 Pro before its official release. The findings show that consumer conversations are organized around three main themes: general product evaluation, comparisons with competing brands, and aesthetic preferences. Rather than existing independently, these themes are interconnected and reflect different aspects of consumer perceptions.

The results also indicate a strong relationship between general reactions and purchase-related discussions, suggesting that consumer opinions about the product are closely linked to considerations of value and affordability. In contrast, discussions about design and aesthetics form a more distinct theme, highlighting that some consumers engage with the product primarily through personal style and visual appeal rather than technical features.

5.2 Practical and Marketing Implication

The findings offer several implications for Apple's marketing strategy. Most discussions focus on product evaluation, pricing, and upgrade decisions, suggesting that marketing communication should clearly demonstrate the benefits of the iPhone 18 Pro and justify its value proposition. A substantial portion of users also compare the device with competing brands, particularly in terms of camera performance and product features. Therefore, highlighting competitive advantages may help strengthen consumer interest.

Although aesthetic discussions represent a smaller share of the conversation, they remain important because visually oriented consumers often contribute to social media engagement and product visibility. Design and color-related marketing content may therefore help generate early attention and online discussion.

5.3 Limitations

This study has several limitations. First, the dataset is limited to a single time period and a small number of TikTok videos, which may not fully represent the diversity of discussions about the iPhone 18 Pro. Second, the use of machine translation may introduce small meaning shifts, especially for informal language, slang, and social media expressions. Third, the analysis does not include sentiment classification, so it is not possible to distinguish whether opinions within each topic are positive or negative. Fourth, a large portion of comments could not be assigned to clear topics, meaning only part of the dataset is fully represented in the results. Finally, the visualization method used to map topic relationships is an approximation, so the exact distances between topics should be interpreted with caution.

6 Conclusion

This study applied BERTopic's comprehensive visualization framework to a corpus of 105,619 TikTok comments about the iPhone 18 Pro, producing a structurally grounded map of pre-launch consumer discourse that advances both methodological practice and substantive understanding of consumer technology perception. Five coherent topics were identified and analyzed through keyword barcharts, document distribution, cosine similarity heatmap, and intertopic distance map, with convergent validity between the latter two providing robust evidence for the structural findings.

The analysis reveals that iPhone 18 Pro consumer discourse on TikTok is organized into a tripartite semantic structure: a dominant mainstream evaluation cluster (Topics 0, 2, 4; 67.04% of classified documents) characterized by upgrade skepticism, pricing anxiety, and premium hardware aspiration operating in a unified semantic neighborhood; a vertically isolated brand comparison axis (Topic 1; 26.15%) anchored in iPhone-versus-Samsung camera evaluation; and a horizontally isolated aesthetic niche (Topic 3; 6.81%) representing design-led, color-driven purchase motivation. The near-semantic fusion of upgrade skepticism and pricing concerns (cosine = 0.848) identifies them as a single unified consumer reservation, while the maximum semantic distance between aesthetic preferences and premium hardware discussion (cosine = 0.389; d = 2.238) establishes these as the most orthogonal consumer motivations in the corpus.

These findings contribute to the literature on social media consumer analytics, smartphone marketing, and computational consumer research methodology. They demonstrate that BERTopic's structural visualization capabilities, particularly the convergent use of the cosine similarity heatmap and the intertopic distance map provide a qualitatively richer analytical framework than topic lists alone, yielding consumer segmentation insights with direct implications for pre-launch marketing strategy. The validated multilingual BERTopic pipeline established in this study is generalizable to other product categories, platforms, and linguistic contexts, offering a scalable toolkit for computational consumer research in the short-form video era.

References

Abbasi, A. Z., Ayaz, N., Kanwal, S., Albashrawi, M., & Khair, N. (2023). Tiktok App Usage Behavior: The Role of Hedonic Consumption Experiences. Data Technologies and Applications, 57(3), 344–365. https://doi.org/10.1108/DTA-03-2022-0107

Baird, A., Xia, Y., & Cheng, Y. (2022). Consumer Perceptions of Telehealth for Mental Health or Substance Abuse: A Twitter-Based Topic Modeling Analysis. JAMIA Open, 5(2), ooac028. https://doi.org/10.1093/jamiaopen/ooac028

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. https://www.jmlr.org/papers/v3/blei03a.html

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), Advances in Knowledge

Discovery and Data Mining (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Aghajanyan, A., Weston, J., & Stoyanov, V. (2020). Unsupervised Cross-Lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An Evaluation of Document Clustering and Topic Modelling in Two Online Social Networks: Twitter and Reddit. Information Processing & Management, 57(2), 102034. https://doi.org/10.1016/j.ipm.2019.04.002

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between Lda, Nmf, Top2vec, And Bertopic to Demystify Twitter Posts. Frontiers in Sociology, 7, 886498. https://doi.org/10.3389/fsoc.2022.886498

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Bhatt, O., Nichols, B., Wenzek, G., Macherey, W., & Auli, M. (2021). Beyond English-Centric Multilingual Machine Translation. Journal of Machine Learning Research, 22(107), 1–48. http://www.jmlr.org/papers/v22/20-1307.html

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling With A Class-Based TF-IDF Procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

Grootendorst, M. (2023). BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics. Zenodo. https://doi.org/10.5281/zenodo.7645318

Lee, J., Ouellette, R. R., Murthy, D., Pretzer, B., Anand, T., & Kong, G. (2025). Identifying E-cigarette Content on TikTok: Using a BERTopic Modeling Approach. Nicotine & Tobacco Research, 27(1), 91–96. https://doi.org/10.1093/ntr/ntae171

Maghsoudi, M., Bakhtiari, M., & Bakhtiari, H. (2026). A Hybrid Framework for Notebook Market Analysis: Integrating Social Media Sentiment Mining with Expert Knowledge for Feature Prioritization. PLOS ONE, 21(2), e0342067. https://doi.org/10.1371/journal.pone.0342067

Mathur, S., Tewari, A., & Singh, A. (2022). Modeling The Factors Affecting Online Purchase Intention: The Mediating Effect of Consumer's Attitude Toward User-Generated Content. Journal of Marketing Communications, 28(7), 725–744. https://doi.org/10.1080/13527266.2021.1936126

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426

Mishra, M. (2022). Customer Experience: Extracting Topics from Tweets. International Journal of Market Research, 64(3), 334–353. https://doi.org/10.1177/14707853211047515

Omar, B., & Dequan, W. (2020). Watch, Share or Create: The Influence of Personality Traits and User Motivation on Tiktok Mobile Video Usage. International Journal of Interactive Mobile Technologies, 14(4), 121–137. https://doi.org/10.3991/ijim.v14i04.12429

Tang, Z., Pan, X., & Gu, Z. (2024). Analyzing Public Demands on China's Online Government Inquiry Platform: A Bertopic-Based Topic Modeling Study. PLOS ONE, 19(2), e0296855. https://doi.org/10.1371/journal.pone.0296855

Turan, S. C., Yıldız, K., & Büyüktanır, B. (2024). Comparison of LDA, NMF and BERTopic Topic Modeling Techniques on Amazon Product Review Dataset: A Case Study. In F. P. García Márquez, A. Jamil, I. S. Ramirez, S. Eken, & A. A. Hameed (Eds.), Computing, Internet of Things and Data Analytics (ICCIDA 2023), Studies in Computational Intelligence, vol. 1145 (pp. 23–31). Springer, Cham. https://doi.org/10.1007/978-3-031-53717-2_3

Zhang, Z., Jiménez, F. R., & Cicala, J. E. (2024). Social Commerce on Tiktok: The Role of Social Norms and Self-Expression in Purchase Intentions. Journal of Retailing and Consumer Services, 76, 103591. https://doi.org/10.1016/j.jretconser.2023.103591

Zhou, Q., Chen, M., Chen, J., Chen, K., & Tsai, S. (2024). Text Mining-Based Study on Consumer Satisfaction in The Mobile Phone Market. Journal of Global Information Management, 32(1), 1–20. https://doi.org/10.4018/JGIM.344835