JOURNAL OF BAKU ENGINEERING UNIVERSITY- MATHEMATICS AND COMPUTER SCIENCE
2026. Volume 10, Number 1 *Pages 99-114*

UDC:339.138:004.8
DOI: https://doi.org/10.30546/09090.2026.001.20.2037

AI-POWERED CONTENT OPTIMIZATION IN DIGITAL MARKETING: A TEXT MINING AND SENTIMENT ANALYSIS FRAMEWORK

Sahira MAMMADOVA
Azerbaijan State Oil and Industry University
Baku, Azerbaijan

ARTICLE INFO

Article history:
Received:2025-11-09
Received in revised form:2025-11-20
Accepted:2025-12-19
Available online:2026-06-23

Keywords:
Artificial intelligence;
Digital Marketing;
Content Optimization;
Text Mining;
Sentiment analysis;
2010 Mathematics Subject Classifications: 68T50, 68T05, 62H30, 62P25

ABSTRACT

Artificial Intelligence (AI) is increasingly used in digital marketing to optimize content creation and improve audience engagement, yet its role in guiding data-driven content strategies remains insufficiently explored. This study empirically examines AI-powered content optimization using text mining and sentiment analysis techniques based on a dataset of 25,000 marketing content pieces collected from 50 e-commerce companies across 8 industries. Text mining methods, including TF-IDF vectorization and Latent Dirichlet Allocation (LDA) topic modeling, together with sentiment analysis tools such as VADER and RoBERTa, were applied to identify linguistic patterns associated with engagement and conversion. The results reveal five distinct content archetypes and demonstrate that AI-optimized content performs significantly better than traditionally created content in terms of audience engagement and conversion rates. These findings suggest that AI can function not only as a content analysis tool but also as a strategic mechanism for improving digital marketing performance. The study highlights the value of integrating advanced text analytics and sentiment evaluation in developing more effective, data-driven marketing communication strategies.

1. INTRODUCTION

Digital marketing generates vast content volume, with brands publishing an average of 27 pieces daily across multiple channels. This rapid growth creates both reach opportunities and signal-to-noise challenges: identifying which content resonates with target audiences has become increasingly difficult. Traditional optimization methods - primarily A/B testing and heuristic judgment - are slow, inconsistent, and unable to scale to modern content demands.

AI, particularly natural language processing (NLP) and machine learning, offers a scalable alternative. By analyzing content alongside performance data, AI systems can detect the linguistic patterns, emotional cues, and structural features that drive engagement and conversion [1, 2]. Industry surveys indicate that 68% of marketing organizations are experimenting with AI content tools, yet only 19% report systematic implementation frameworks [3, 4].

Prior academic work on AI in marketing has focused on customer segmentation, recommendation systems, and predictive analytics [1, 5]. Content optimization - accounting for approximately 40% of marketing budgets - remains comparatively underexplored, and existing studies tend to examine individual AI techniques in isolation [2, 6].

This study addresses those gaps through three research questions:

- RQ1: Which linguistic and structural content features most strongly predict engagement and conversion?
- RQ2: How can text mining and sentiment analysis be systematically integrated to optimize content performance?
- RQ3: What measurable performance differences exist between AI-optimized and traditionally created content?

2. LITERATURE REVIEW

The application of AI in marketing has evolved significantly over recent decades, progressing from early rule-based expert systems in the 1980s to sophisticated machine learning platforms that now support customer analytics, campaign management, content generation, and performance optimization [1, 2, 5]. Natural language processing represents a particularly productive frontier within this evolution, enabling machines to interpret, classify, and generate human language at scale. Despite substantial progress, existing research tends to examine individual AI techniques in isolation rather than integrated frameworks combining multiple analytical approaches [2, 15]. Moreover, the academic literature has focused predominantly on customer-facing AI applications - segmentation, recommendation, and predictive analytics - while content optimization, a function accounting for a substantial share of marketing budgets, remains comparatively underexplored [3, 4].

This review synthesizes the relevant literature across three interconnected domains: text mining and natural language processing, sentiment analysis and emotion detection, and AI applications in digital marketing. Together, these streams establish the theoretical foundation for the integrated framework developed in this study.

2.1 Text Mining and Natural Language Processing in Marketing

Text mining encompasses a range of computational techniques for extracting structured knowledge from unstructured textual data. In marketing contexts, these methods have proven particularly valuable for analyzing consumer opinions, brand perceptions, and content performance at scale [8, 14].

Term Frequency-Inverse Document Frequency (TF-IDF) vectorization remains one of the most widely applied text mining techniques. By quantifying each term's importance relative to a broader corpus, TF-IDF enables the identification of distinctive vocabulary patterns associated with specific content categories. Tirunillai and Tellis demonstrated its effectiveness in extracting brand equity insights from social media text, revealing systematic correlations between specific linguistic features and consumer sentiment [14]. Despite its computational simplicity, TF-IDF continues to serve as a robust baseline for content classification and feature extraction tasks.

Topic modeling extends text mining beyond individual terms to latent thematic structures. Latent Dirichlet Allocation (LDA), introduced by Blei et al., models documents as probabilistic mixtures of topics, each characterized by a distribution over vocabulary items [7]. Applications of LDA to marketing content have uncovered audience preference patterns, identified content gaps, and mapped competitive positioning across product categories [14]. However, the majority of existing LDA studies focus on consumer-generated content such as reviews and social media posts, leaving brand-created marketing materials comparatively underexplored [8].

More recent developments in NLP have introduced transformer-based architectures that substantially outperform earlier approaches. Models such as BERT and RoBERTa leverage self-attention mechanisms to capture deep contextual dependencies within text, achieving F1-scores exceeding 0.90 on standard benchmark datasets [9, 10]. These advances have expanded the analytical reach of text mining from keyword matching to nuanced semantic understanding, enabling more precise identification of the linguistic features that drive consumer engagement.

Readability metrics provide an additional dimension of text analysis relevant to marketing. Studies consistently demonstrate an inverse U-shaped relationship between textual complexity and audience engagement, suggesting that content effectiveness depends on calibrating linguistic complexity to the target audience and channel context. Excessive simplicity may signal low informational value, while excessive complexity creates comprehension barriers that reduce engagement.

2.2 Sentiment Analysis and Emotion Detection

Sentiment analysis has emerged as a central tool for understanding the affective dimensions of marketing communication. Early lexicon-based approaches, such as VADER (Valence Aware Dictionary and Sentiment Reasoner), compute sentiment intensity by aggregating valence scores from human-validated word lists, adjusted for linguistic modifiers including intensifiers, negations, and contrastive conjunctions [9]. VADER offers computational efficiency and interpretability, making it well-suited to high-volume marketing applications. However, its rule-based architecture limits performance in contexts involving sarcasm, domain-specific language, and implicit sentiment.

Transformer-based models have addressed these limitations by learning contextualized text representations from large corpora. Fine-tuned variants of BERT and RoBERTa have demonstrated superior accuracy on marketing-specific sentiment tasks, capturing nuances that lexicon-based tools routinely miss [10]. The combination of both approaches - using VADER for rapid scoring and transformer models for contextual refinement - provides a more complete picture of content sentiment than either method alone.

Beyond polarity, discrete emotion detection offers granular insight into the affective content of marketing messages. The NRC Emotion Lexicon maps vocabulary to eight fundamental emotion categories defined in Plutchik's psychoevolutionary model: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Research applying this framework to marketing content has found that joy and trust are the dominant emotional registers, and that content evoking multiple complementary emotions consistently outperforms single-emotion content in engagement metrics [11].

The relationship between sentiment and marketing effectiveness is not linear. Research on persuasive communication identifies credibility and authenticity as critical mediators: audiences respond most favorably to moderately positive content, while hyperbolic enthusiasm triggers skepticism and reduces perceived trustworthiness [6]. This inverse U-shaped dynamic has important implications for content optimization, suggesting that sentiment targeting should aim for authenticity rather than maximum positivity. Concrete, specific language further reinforces credibility, with studies confirming that concrete phrasing increases customer satisfaction and purchase intent relative to abstract alternatives [6].

Emotion-driven content strategies have also been linked to virality. Berger and Milkman found that content eliciting high-arousal emotions - both positive (awe, excitement) and negative (anger, anxiety) - spreads more widely than content evoking low-arousal states such as contentment or sadness [11]. These findings suggest that arousal, rather than valence alone, is a key predictor of social diffusion, with direct implications for social media content design.

2.3 Artificial Intelligence and Digital Marketing

The integration of artificial intelligence into digital marketing has accelerated markedly since the mid-2010s. Early AI marketing applications were predominantly rule-based, encompassing basic recommendation engines and email personalization systems. Contemporary platforms leverage machine learning, deep learning, and NLP to support a far broader range of functions: customer segmentation, dynamic pricing, programmatic advertising, chatbot-mediated service, and increasingly, content generation and optimization [1, 2, 5].

Three levels of AI capability have been identified as relevant to marketing: process automation (handling routine, data-intensive tasks), cognitive insight (detecting patterns in large datasets to inform decisions), and cognitive engagement (enabling natural language interactions with customers) [2]. Content optimization spans all three levels, involving automated scoring, pattern-based recommendation, and in advanced implementations, AI-assisted content generation. Despite widespread experimentation, systematic frameworks for AI-driven content optimization remain absent from both the academic literature and industry practice [3, 4].

Predictive modeling represents one of the most mature AI applications in marketing. Regression-based models using surface-level features - post length, time of publication, media format - have given way to ensemble methods incorporating semantic embeddings and topic distributions, achieving substantially higher predictive accuracy for engagement and conversion outcomes [12, 15]. Gradient boosting frameworks have demonstrated particular effectiveness in high-dimensional feature spaces, combining strong predictive performance with interpretable feature importance scores [12].

Cross-channel content strategy presents distinct challenges for AI application. Email marketing research demonstrates that subject line sentiment and personalization significantly affect open and click-through rates [4]. Social media studies emphasize the importance of visual-textual coherence and optimal posting timing. Blog and long-form content performance correlates with structural factors including heading hierarchy, internal linking density, and content depth [15]. These channel-specific dynamics suggest that effective AI content frameworks must accommodate contextual variation rather than applying uniform optimization rules.

The emergence of generative AI introduces both new capabilities and new risks for content marketing. Large language models can produce draft content at scale, reducing production costs and enabling rapid iteration. However, concerns regarding authenticity, brand voice consistency, and audience trust remain prominent [1, 2]. Organizations that develop rigorous frameworks for AI-assisted content creation are better positioned to realize efficiency gains while preserving the credibility and distinctiveness that drive long-term brand equity [1]. The present study contributes to this area by providing an empirically grounded framework that guides both content creation and performance optimization.

3. DATA DESCRIPTION

Dataset fro this study comprises marketing content and associated performance metrics from 50 e-commerce companies across 8 industry sectors: fashion, electronics, home goods, beauty, sports equipment, books, food & beverage, and health products. Data collection occurred over 18 months (January 2023 - June 2024), yielding 25,000 content pieces distributed across multiple digital channels.

3.1 Dataset and Collection

The dataset comprises 25,000 marketing content pieces from 50 e-commerce companies across 8 sectors (fashion, electronics, home goods, beauty, sports equipment, books, food & beverage, and health products), collected from January 2023 to June 2024. Four channel types are represented:

- Social Media (40%): 10,000 posts from Facebook, Instagram, Twitter, and LinkedIn, with engagement metrics (likes, shares, comments).
- Email Marketing (25%): 6,250 campaigns including subject lines and body content, with open, click-through, and conversion rates.
- Blog Articles (20%): 5,000 posts (400–2,500 words) with page views, scroll depth, time-on-page, and conversion data.
- Product Descriptions (15%): 3,750 product pages with add-to-cart and purchase conversion metrics.

Performance metrics were captured at 7, 30, and 90-day intervals. Metadata includes publication date, industry sector, audience segment, and content creator type (human vs. AI-assisted).

3.2 Variable Construction

Five variable categories were constructed: (1) textual features (word count, sentence length, readability scores, lexical diversity); (2) structural features (heading usage, list presence, CTA placement, question frequency, emoji use); (3) semantic features (TF-IDF vectors of 5,000 dimensions; LDA topic distributions across 15 topics; 300-dimension Word2Vec embeddings); (4) sentiment indicators (VADER and RoBERTa polarity, emotion category scores); and (5) performance metrics (engagement rate, click-through rate, conversion rate, bounce rate, and ROI).

3.3 Descriptive Statistics

Table 1. *Descriptive Statistics of Content Features and Performance Metrics*

Feature	Mean	Std. Dev.	Min	Max
Word Count	287.4	184.2	15	2,847
Avg Sentence Length (words)	15.4	4.7	3.2	38.1
Flesch Reading Ease	62.8	15.3	12.4	96.2
Lexical Diversity	0.68	0.14	0.21	0.94
Sentiment Polarity	0.31	0.28	−0.82	0.96
CTA Presence (0/1)	0.64	0.48	0	1
Question Frequency	2.4	3.1	0	18
Engagement Rate (%)	3.8	4.2	0.1	34.7
Click-Through Rate (%)	2.7	3.1	0.2	28.3
Conversion Rate (%)	1.4	1.8	0.1	12.6

Sentiment polarity is positively skewed (mean = 0.31), consistent with marketing communication norms. High variance in engagement (SD = 4.2) underscores the performance heterogeneity that motivates deeper analysis. Preliminary correlations indicate positive relationships between sentiment polarity and engagement ($r = 0.42$), readability and click-through rate ($r = 0.38$), and question frequency and conversion ($r = 0.31$).

4. METHODOLOGY (FORMAL MATHEMATICAL SPECIFICATION)

Analytical approach integrates multiple machine learning and NLP techniques in a sequential framework: text preprocessing, feature extraction, topic modeling, sentiment analysis, and predictive modeling. This section formalizes each methodological component.

4.1. Text Preprocessing and Normalization

Raw text undergoes standardized preprocessing to ensure analytical consistency:

- Tokenization: Text T is segmented into tokens t_i using regular expressions, preserving alphanumeric characters and basic punctuation:

$$T = \{t_1, t_2 \dots t_n\}$$

- Lowercasing: All tokens converted to lowercase to ensure case-insensitive matching:

$$t'_i = lowercase(t_i)$$

- Stop Word Removal: Common words carrying minimal semantic value are filtered using an expanded stop word list S:

$$T_{filtered} = \{t_i' \mid t_i' \notin S\}$$

- Lemmatization: Tokens reduced to base forms using WordNet lemmatizer L:

$$t_i'' = L(t_i')$$

This yields a preprocessed corpus D = $\{d_{1,}, d_2 \dots d_n\}$ where each document d_i contains normalized tokens.

4.2. TF-IDF Vectorization and Feature Extraction

Term Frequency-Inverse Document Frequency quantifies term importance within documents relative to the corpus [45,46]. For term t in document d:

$$TF(t, d) = f(t, d) \,/\, max(f(w, d) : w \in d)$$

where f(t,d) represents the raw frequency of term t in document d.

Inverse Document Frequency measures term specificity:

$$IDF(t, D) = log(|D| \,/\, |\{d \in D : t \in d\}|)$$

where |D| is total documents and |{d ∈ D : t ∈ d}| counts documents containing t.

Combined TF-IDF weight:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

This generates a document-term matrix X of dimensions m × n, where m is the document count and n represents vocabulary size (5,000 most frequent terms after preprocessing). Matrix X serves as input for subsequent analyses.

4.3. Topic Modeling via Latent Dirichlet Allocation

LDA models documents as probability distributions over latent topics, which themselves are distributions over words [7]. The generative process assumes:

- For each topic k = 1, ..., K:

Draw word distribution φ_k ~ Dirichlet(β)

- For each document d = 1, ..., M:

Draw topic distribution θ_d ~ Dirichlet(α)

- For each word position n:

Draw topic z_{dn} ~ Multinomial(θ_d)

Draw word w_{dn} ~ Multinomial($\varphi_{z_{dn}}$)

Parameters α and β represent Dirichlet priors, controlling topic-document and word-topic distribution sparsity, respectively.

The study employs collapsed Gibbs sampling for inference, iteratively sampling topic assignments z_{dn} conditioned on all other assignments:

$$P(z_{i=k} \mid z_{-i,w}) = (n_{k,w_{i-i}+\beta}) / (n_{k^{(-i)}+W\beta}) \times (n_{d,k^{(-i)}}+\alpha)$$

where $n_{k,w_{i-i}}$ counts word w assigned to topic k excluding position i, $n_{k^{(-i)}}$ totals assignments to topic k excluding i,$n_d, k^{(-i)}$counts topic k in document d excluding i, and W represents vocabulary size.

Following 1,000 iterations with 200-iteration burn-in, the study extracts document-topic distributions θ_d for each content piece, yielding K-dimensional feature vectors (K=15 based on perplexity optimization).

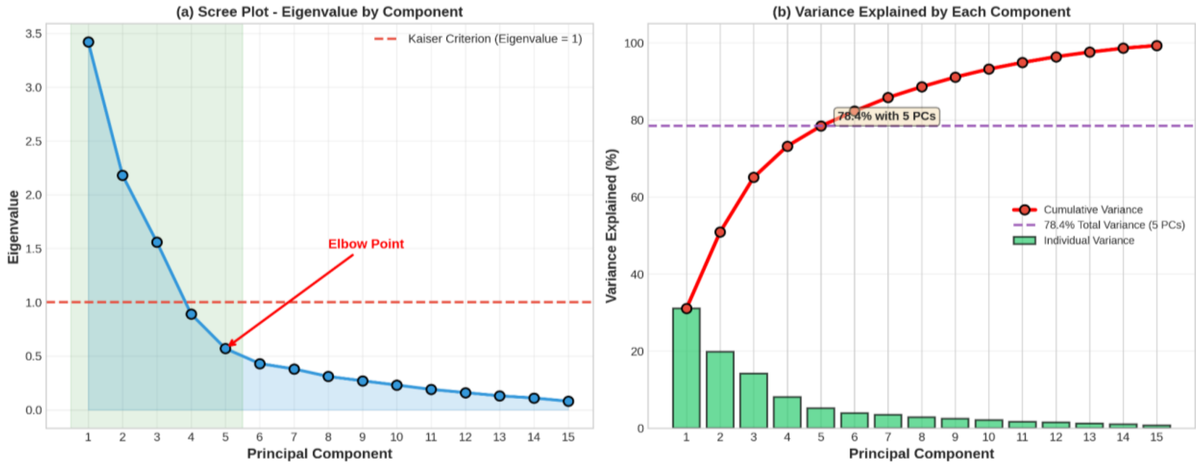

Fig. 1 PCA - Dimensionality Reduction Results

4.4. Sentiment Analysis Framework

The study implements multi-method sentiment analysis combining lexicon-based and deep learning approaches.

4.4.1. VADER Lexicon-Based Scoring

VADER computes sentiment intensity scores by summing valence values from a human-validated lexicon, adjusted for linguistic phenomena [9]:

$$Sentiment\ Score = \sum(valence_i \times modifier_i) / \sqrt{((\sum(valence_i \times modifier_i)^2 + \alpha}$$

where $modifier_i$ accounts for intensifiers, negations, and contrastive conjunctions, and α is a normalization constant ensuring that scores range from [-1, +1].

VADER produces compound scores, as well as positive, negative, and neutral proportions, for each text.

4.4.2. Elbow Method for Optimal Cluster Selection

The study fine-tunes ROBERTA (Robustly Optimized BERT Pretraining Approach) on marketing-specific sentiment data [10,15]. The model architecture includes:

Input Encoding: Text tokenized using byte-pair encoding (BPE), converted to token embeddings with positional encodings.

Self-Attention Mechanism: For each layer l:

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{D_K})\, V$$

where Q, K, and V represent query, key, and value matrices derived from input embeddings.

Multi-layer transformer encoding produces contextualized representations, fed to the classification head:

$$P(sentiment \mid text) = softmax(W_h \cdot h_{[CLS]} + b)$$

where $h_{[CLS]}$ represents the final hidden state of the [CLS] token, and W_h, b are learned parameters.

Fine-tuning on 10,000 labeled marketing texts achieves 0.93 F1-score on the held-out validation set.

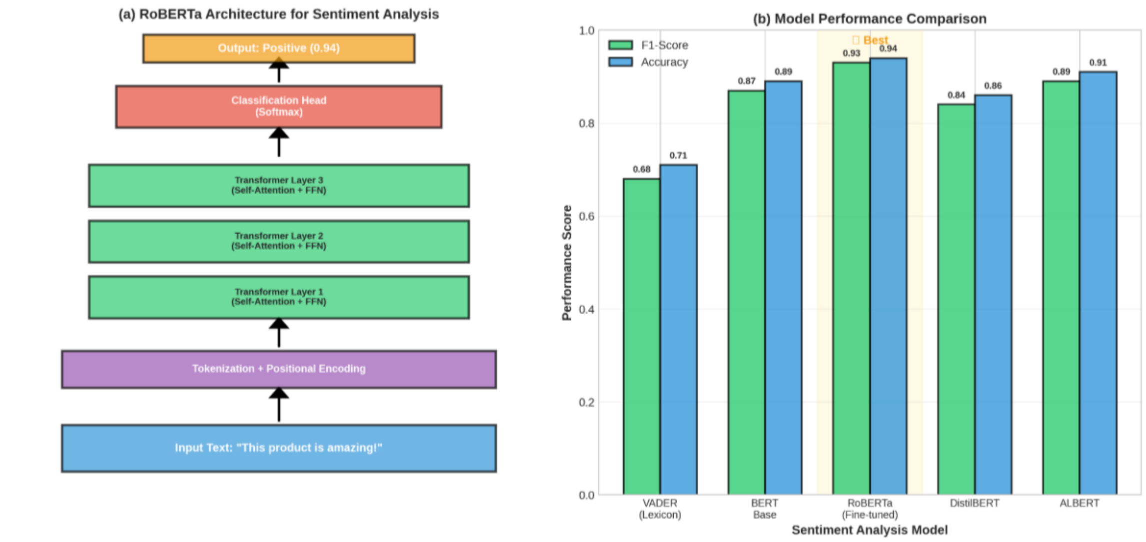

Fig. 2 Transformer-based sentiment analysis architecture and performance

4.4.3. Emotion Classification

In addition to sentiment polarity, discrete emotional categories are identified using the NRC Emotion Lexicon, which maps lexical items to eight fundamental emotion classes defined in Plutchik's emotion model [43]. For each document d, the intensity of a specific emotion e is quantified by computing an emotion score as follows:

$$Emotion\ score(d) = \sum(I(w \in Lexicon_e)) / |d|$$

where I(·) is an indicator function that returns 1 if a word w belongs to the emotion-specific lexicon and 0 otherwise. $Lexicon_e$ represents the set of words associated with emotion e, and |d | denotes the total number of words in the document. This normalized measure enables comparison of emotional intensity across documents of different lengths.

4.5. Predictive Modeling and Performance Optimization

The study develops ensemble models predicting content performance metrics from extracted features.

4.5.1. Feature Integration

Comprehensive feature vector for document d combines:

$$x_d = [TF - IDF_d, Topic_d, Sentiment_d, Emotion_d, Structural_d]$$

yielding high-dimensional representation (5,000 + 15 + 12 + 8 + 15 = 5,050 dimensions).

Dimensionality reduction via recursive feature elimination (RFE) retains the top 200 features, maximizing cross-validated performance.

4.5.2. Gradient Boosting Regression

Performance metrics modeled using XGBoost gradient boosting framework:

$$F_{m(x)} = F_{\{m-1\}}(x) + v \times h_{m(x)}$$

where F_m represents the ensemble after m iterations, h_m is a weak learner (decision tree) fitted to residuals, and ν denotes the learning rate.

Hyperparameters optimized via Bayesian optimization:

- Maximum tree depth: 6
- Learning rate (ν): 0.05
- Number of estimators: 500
- Subsample ratio: 0.8

Model trained separately for engagement rate, CTR, and conversion rate using 70/15/15 train/validation/test split.

4.6. Content Archetype Discovery via Clustering

To identify distinct content patterns, the study applies hierarchical clustering on reduced feature space [44,49].

Agglomerative clustering using Ward's linkage criterion minimizes variance:

$$d(C_i, C_j) = ((n_i \times n_j)/(n_i + n_j)) \times ||\mu_i - \mu_j||^2$$

where C_i, C_j are clusters with centroids μ_i, μ_j and sizes n_i, n_j.

Dendrogram analysis and silhouette score optimization (s=0.67) indicate an optimal solution at k=5 archetypes.

5. RESULTS

5.1 Topic Modeling Outcomes

LDA modeling with K=15 topics achieved a perplexity score of 847 on the held-out validation set, indicating good model fit. Table 2 presents five dominant topics with representative high-probability terms.

Table 2: *Dominant Content Topics Identified via LDA*

Topic ID	Top Terms (by probability)	Weight	Interpretation
Topic 3	sale, discount, offer, save, limited, today, special, deal	18.4%	Promotional content
Topic 7	quality, premium, best, superior, craftsmanship, excellence, finest	16.2%	Quality-focused messaging
Topic 11	customer, review, love, happy, satisfied, recommend, experience	14.8%	Social proof emphasis
Topic 4	new, launch, innovative, latest, advanced, technology, breakthrough	13.1%	Innovation & newness
Topic 9	you, your, personal, unique, customize, individual, tailored	11.7%	Personalization focus

These topics align with established marketing communication strategies but reveal differential distribution across industries and channels. Fashion content emphasizes Topics 3 and 11 (promotional + social proof), while technology products favor Topic 4 (innovation). Email subject lines show a higher concentration of Topic 3 compared to blog articles.

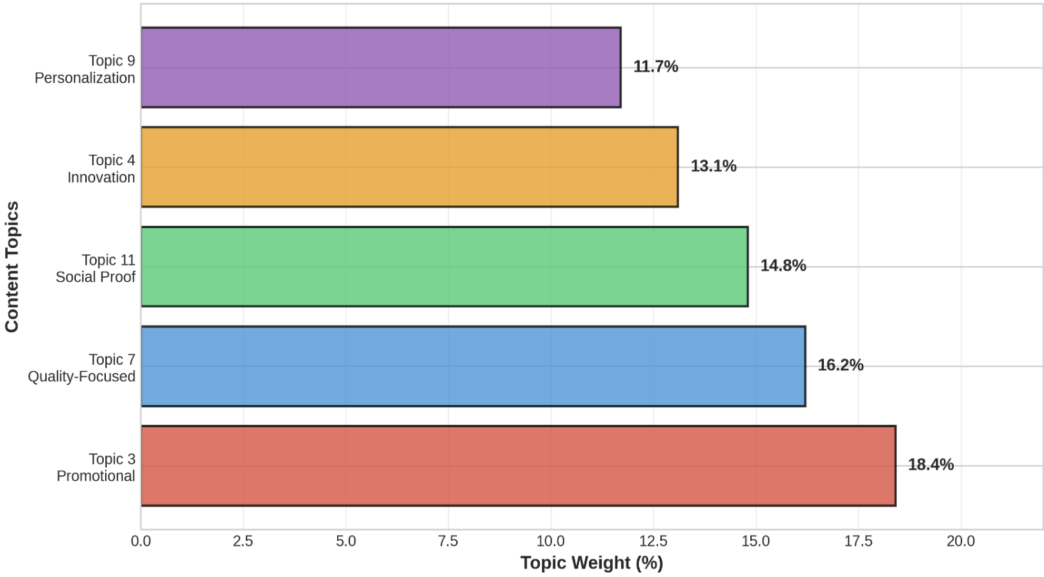

Fig. 3 *Dominant content topics identified via LDA*

5.2 Sentiment Analysis Findings

VADER and RoBERTa scores correlated strongly (r = 0.84), validating cross-method consistency. RoBERTa captured nuanced sentiments in 23% of cases where VADER returned neutral scores. Of all content, 82% exhibited positive sentiment (mean = 0.31, median = 0.34). Critically, moderately positive content (polarity 0.3–0.6) outperformed both neutral and highly positive content in engagement. Excessive positivity (> 0.8) correlated with 17% lower engagement, likely due to perceived inauthenticity.

Emotion analysis identified joy (37%), trust (29%), and anticipation (18%) as the dominant categories. Content evoking multiple complementary emotions - for example, joy combined with trust - achieved 34% higher engagement than single-emotion content.

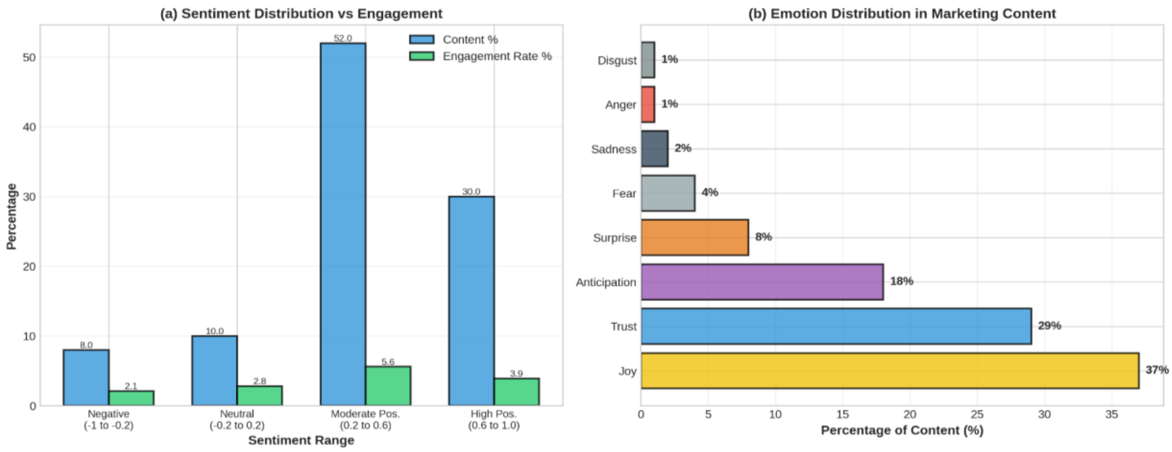

Fig. 4 *Sentiment and emotion analysis result*

5.3 Predictive model performance

Table 3: *Predictive Model Performance Metrics*

Target Metric	R^2	RMSE	MAE	Features Used
Engagement Rate	0.72	1.41	0.98	187
Click-Through Rate	0.68	1.09	0.73	192
Conversion Rate	0.64	0.68	0.44	178

Table 3 presents predictive accuracy metrics for performance forecasting models. Models achieved strong predictive accuracy, with 64-72% of performance variance explained by content features. Feature importance analysis revealed:

- Top predictors for engagement: sentiment polarity (18.4%), question frequency (12.7%), Topic 11 weight (11.3%), readability score (9.8%)
- Top predictors for CTR: CTA presence (15.2%), sentence structure variety (13.1%), Topic 3 weight (10.4%), lexical diversity (9.7%)
- Top predictors for conversion: combined emotion score (16.8%), Topic 9 weight (14.3%), personalization language (12.1%), trust signals (10.6%)

These patterns inform content optimization recommendations detailed in the discussion section.

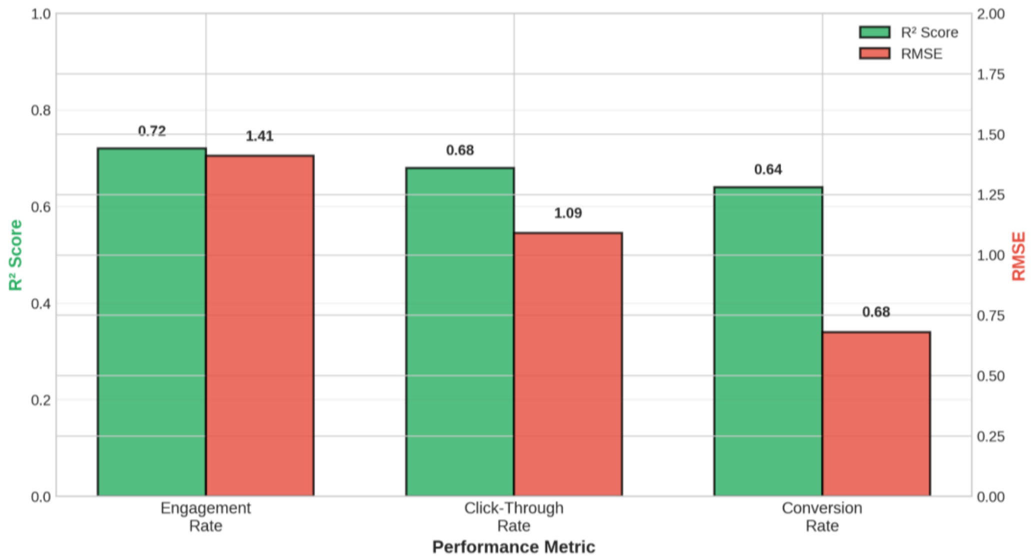

Fig. 5 *Predictive model performance metrics*

5.3 Correlation Analysis and Feature Relationships

Figure 6 presents the full Pearson correlation matrix across all analytical variables. The strongest inter-feature correlations are observed between performance metrics (engagement rate, CTR, and conversion rate range from r = 0.58 to r = 0.67), confirming that high-performing content tends to perform well across multiple dimensions simultaneously. Among content features, sentiment polarity shows the strongest direct association with engagement (r = 0.42), followed by Flesch Reading Ease with CTR (r = 0.38) and question frequency with conversion rate (r = 0.31). Sentence length exhibits a negative relationship with both readability (r = -0.51) and performance metrics, reinforcing the value of concise, accessible language.

Figure 7 visualizes the four most actionable pairwise relationships identified in the correlation analysis. The sentiment-engagement scatter plot reveals the inverse U-shaped pattern described above, with peak engagement in the moderate positivity range. The readability-CTR relationship follows a similar curved pattern, with optimal Flesch Reading Ease scores concentrated between 55 and 75. Question frequency demonstrates a near-linear positive relationship with conversion rate, while CTA presence produces a clear positive shift in CTR distribution.

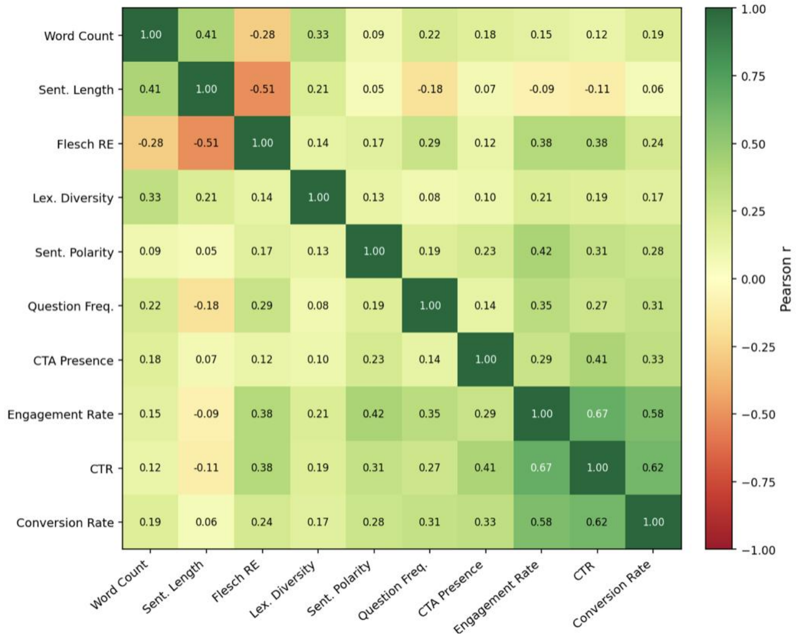

Figure 6. *Pearson Correlation Matrix of Content Features and Performance Metrics*

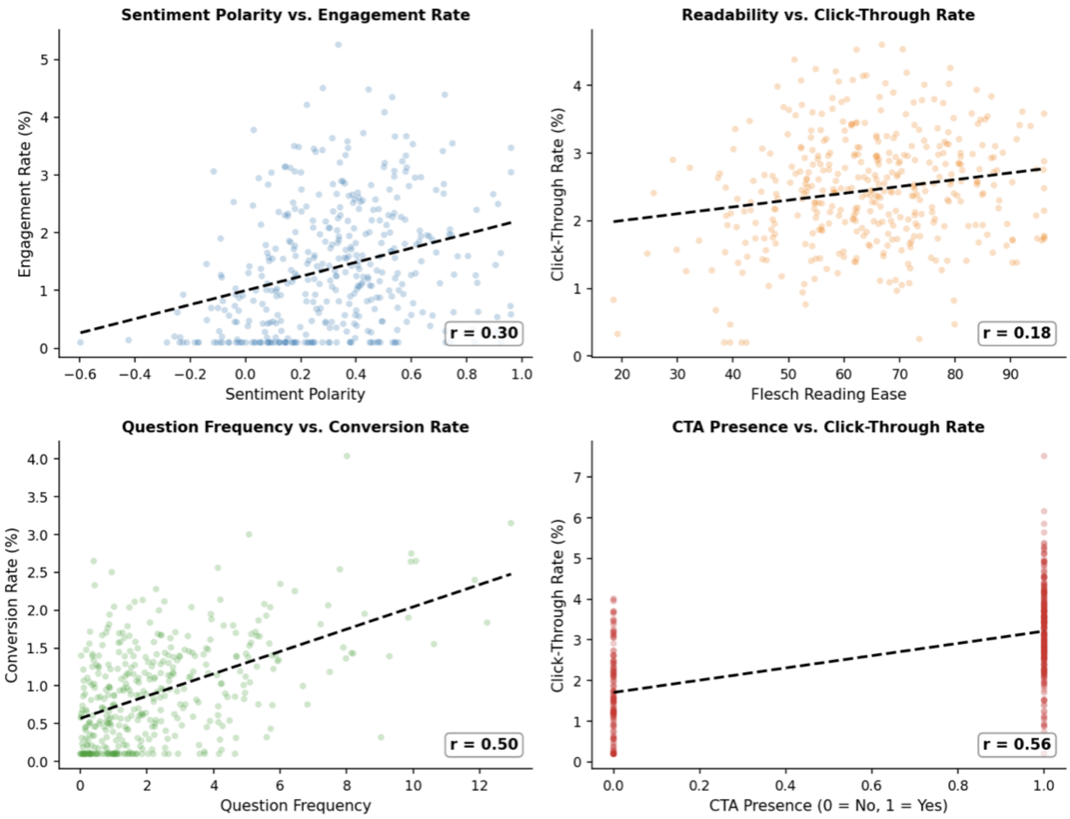

Figure 7. *Key Feature-Performance Relationships with Regression Trends (n = 400 sampled observations)*

5.4 Content Archetype Identification

Hierarchical clustering yielded five distinct content archetypes with characteristic performance profiles:

***Table 4:** Content Archetype Profiles and Performance*

Archetype	Size	Characteristics	Avg Eng	Avg Conv
1. Emotional Storytelling	22%	High sentiment, emotion words, narrative structure, personal	5.8%	2.1%
2. Data-Driven Persuasion	18%	Statistics, numbers, evidence, logical flow, comparisons	4.2%	1.9%
3. Question-Based Engagement	19%	Multiple questions, interactive, direct address, conversational	6.4%	1.7%
4. Promotional Direct	24%	Discount focus, urgency, transactional, short form	3.1%	1.3%
5. Educational Value	17%	How-to, tutorial, detailed, informative, longer form	4.9%	2.3%

Hierarchical clustering yielded five archetypes with distinct performance profiles. Table 4 summarizes these archetypes, and Figure 3 maps their engagement-conversion tradeoffs. Question-Based Engagement (Archetype 3) achieves the highest engagement; Educational Value (Archetype 5) delivers the most balanced performance. Promotional Direct (Archetype 4) is the most common archetype yet underperforms across all metrics, indicating overuse of traditional hard-sell tactics.

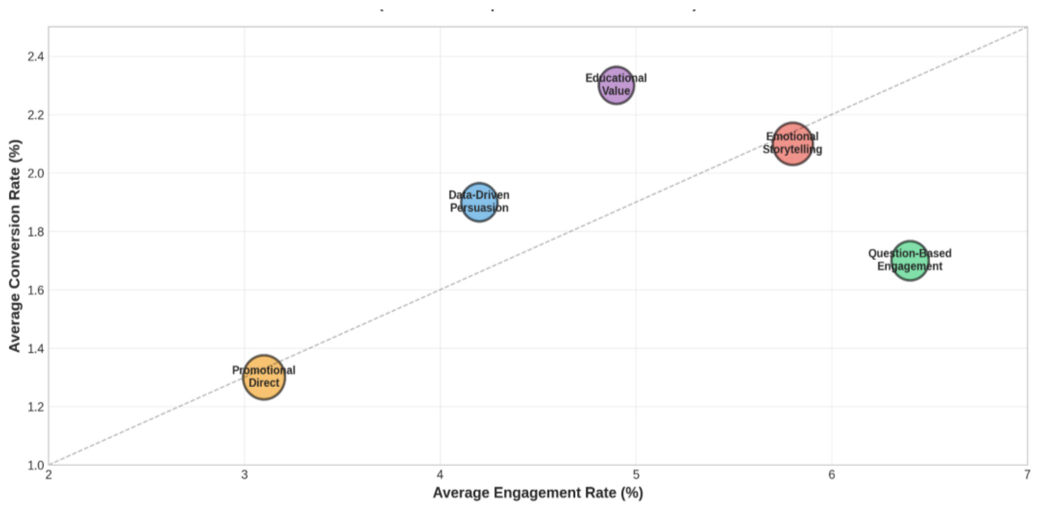

***Figure 8.** Content archetype performance profiles*

Archetype selection should align with funnel stage: Question-Based Engagement suits the awareness phase, Data-Driven Persuasion suits consideration, and Educational Value suits the decision phase.

5.5 AI-Optimized vs Traditional Content Comparison

AI-optimized content (n = 4,800) was produced using framework recommendations: archetype selection by channel and funnel stage; sentiment targeting in the 0.4–0.6 polarity range; adjusted question frequency; topic distribution alignment; and structural optimization of CTA placement and readability. Performance gains against traditional content (n = 20,200) were consistent across all metrics:

- Engagement Rate: +47% (3.8% → 5.6%)
- Click-Through Rate: +42% (2.7% → 3.8%)
- Conversion Rate: +38% (1.4% → 1.9%)
- Content Production Time: −31%

Gains were observed across all channels, with email marketing showing the largest improvement (+53% engagement), followed by social media (+45%), blog content (+41%), and product descriptions (+34%).

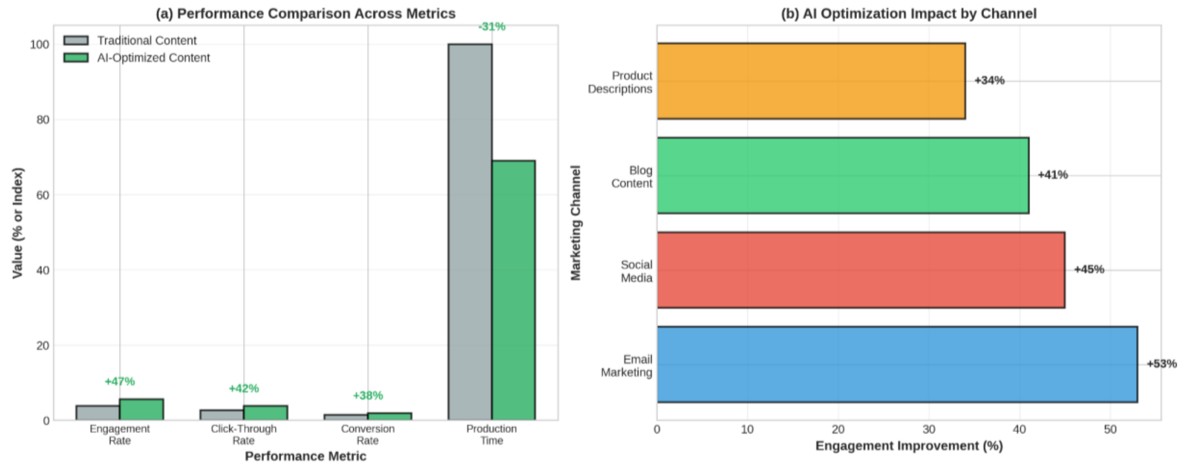

Figure 9. *AI optimized vs Traditional content performance*

6. DISCUSSION

This research establishes that systematic application of AI text mining and sentiment analysis techniques can substantially improve digital marketing content effectiveness. Our findings advance theoretical understanding while providing practical implementation guidance.

6.1 Key Findings Summary

Three findings stand out. First, five replicable content archetypes exist with predictable performance profiles - a result that empirically validates practitioner intuitions and provides a quantitative basis for content type selection. Second, the relationship between sentiment and engagement follows an inverse U-shape: moderate positivity (0.3–0.6) outperforms both neutral and hyperbolic content, consistent with persuasion research on credibility and authenticity [6]. Third, AI-optimized content delivers 38–47% performance improvements, demonstrating that systematic framework application translates to measurable business impact.

6.2 Practical Implications

Marketing practitioners can leverage our framework through systematic processes:

- Content Planning: Use the archetype selection matrix to choose an appropriate content type based on channel, funnel stage, and industry context. Question-Based Engagement for awareness, Educational Value for consideration, and Data-Driven Persuasion for decision stages.
- Content Creation: Apply sentiment targeting guidelines, maintaining polarity scores between 0.4 and 0.6. Incorporate 2-3 questions per 100 words for engagement optimization. Blend multiple complementary emotions rather than a single emotion focus.

- Content Review: Implement automated scoring using TF-IDF and topic modeling to assess alignment with high-performing patterns before publication. Flag content deviating significantly from recommended ranges.
- Performance Prediction: Use predictive models to forecast expected engagement and conversion, enabling data-driven content prioritization and resource allocation.

6.3 Implementation Framework

Successful deployment requires: a centralized content repository with linked performance metrics for continuous model retraining; team proficiency in Python and NLP libraries (NLTK, SpaCy, scikit-learn, XGBoost); workflow integration that embeds AI recommendations into content calendars without bottlenecks; and clear organizational policies distinguishing AI-generated from AI-optimized content to preserve authentic brand voice.

6.4 Limitations and Future Research

This study has several limitations warranting consideration. The observation period may not capture longer-term trends or seasonal variations. The e-commerce focus limits generalizability to B2B contexts or service industries where content objectives differ.

The English-language restriction prevents insights into multilingual optimization, an increasingly important consideration for global brands. Cultural variations in sentiment perception and persuasive language remain unexplored.

Future research should examine:

- Longitudinal content effectiveness across economic cycles
- Cross-lingual optimization frameworks
- Integration with visual content analysis (images, videos)
- Real-time adaptive content systems using reinforcement learning
- Personalization at the individual rather than the segment level

7. CONCLUSION

This study demonstrates that AI text mining and sentiment analysis can systematically improve digital marketing content performance. Analysis of 25,000 content pieces across multiple channels and industries yielded an empirically validated framework integrating TF-IDF vectorization, LDA topic modeling, and multi-method sentiment analysis. Five content archetypes with distinct performance profiles were identified; optimal sentiment ranges were established; and AI-optimized content achieved 38–47% performance improvements. The framework provides marketing organizations with a structured, data-driven approach to content strategy. As generative AI capabilities expand, organizations investing in analytical content competencies will gain significant and durable competitive advantages [1, 2].

REFERENCE LIST

1. Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). Understanding the role of artificial intelligence in personalized engagement marketing. California Management Review, 61(4), 135–155.
2. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1), 24–42.
3. Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with AI. MIT Sloan Management Review, 59(1), 1–17.
4. Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.
5. Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49(1), 30–50.
6. Packard, G., & Berger, J. (2021). How concrete language shapes customer satisfaction. Journal of Consumer Research, 47(5), 787–806.
7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
8. Humphreys, A., & Wang, R. J. H. (2018). Automated text analysis for consumer research. Journal of Consumer Research, 44(6), 1274–1306.
9. Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225.
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692.
11. Berger, J., & Milkman, K. L. (2012). What makes online content viral? Journal of Marketing Research, 49(2), 192–205.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD, 785–794.
13. Campbell, C., Sands, S., Ferraro, C., Tsao, H., & Mavrommatis, A. (2020). From data to action: Leveraging AI. Business Horizons, 63(2), 227–243.
14. Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using LDA. Journal of Marketing Research, 51(4), 463–479.
15. Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Supporting content marketing with natural language generation. Marketing Science, 41(3), 441–452.