FINE-TUNING LARGE LANGUAGE MODELS FOR MULTI-TASK CONSUMER DATA ANALYSIS IN FASHION DESIGN PROCESS

A CASE STUDY OF CHINESE WOMEN'S FASHION MARKET

HAOZE ZHOU

University of Florence, Italy
haozezhou.edu@gmail.com
Orcid 0009-0004-8155-9697

ZHIJIAN ZHANG

Brera Academy of Fine Arts, Italy
zhijianzhang@fadbrera.edu.it
Orcid 0009-0005-1166-0391

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.
Competing Interests: The Author(s) declare(s) no conflict of interest
DOI: 10.36253/fh-3610

Abstract

Artificial intelligence (AI) is rapidly growing within the fashion industry, with current attention primarily focused on image transformation and generation. However, the application of text comprehension AI in design processes, particularly in market research, remains insufficiently explored. This research fine-tunes RoBERTa models to construct an analytical framework including data cleaning, sentiment analysis, and topic classification for Chinese women's fashion analysis.The research analyzed 30,796 user comments from Bilibili. The fine-tuned models achieved strong performance: 95% accuracy for data quality classification, 97.65% for sentiment analysis, and F1-scores ranging from 0.70 to 0.97 across nine topic categories. Analysis of 6,029 high-quality comments revealed that 89.1% of consumers expressed negative or neutral sentiments, with size fit (43.5%) and gender differences (41.3%) being main concerns. The research identified nine systematic industry challenges, including size standards deficiencies, design practices that enforce traditional gender norms at the expense of functionality, and unfair pricing practices.This research shows that fine-tuning Large Language Models works for fashion processes analysis, providing evidence for widespread consumer dissatisfaction. The research fills the gap in applying fine-tuned LLMs to fashion design processes while demonstrating new ways for integrating fashion education with AI, contributing to digital transformation in fashion education and industry development.

Keywords: *Large Language Models (LLMs); Sentiment analysis; Topic classification; Women's fashion; Model fine-tuning*

INTRODUCTION

The fashion industry is closely connected to every individual and represents an essential sector for every nation. According to Statista, global fashion market revenue is projected to reach $920.19 billion in 2025 (Statista, 2025). In recent years, the advancement of AI technologies has transformed numerous industries (Rizzi & Casciani, 2023; Mohammadi et al., 2021). Among these developments, generative AI for image creation has had a major impact on the design industry, shown by its application in textile pattern creation (Jung & Suh, 2023). As industry rules frameworks develop, related legal and ethical issues have gradually emerged (Musmeci, 2024). Compared to the early start of AI research in autonomous driving (Forbes et al., 1995), the development of AI technology in the fashion industry has been relatively slow, initially being primarily utilized for image generation and personalized recommendation systems (Zou et al., 2021).

However, people have started questioning whether the final data is real and whether the creative work is truly original.Given these current conditions and challenges, research on AI technologies specifically designed for the fashion industry domain remains insufficient.

This research will employ Large Language Models (LLMs) to investigate the topic of “Current Development Status of Chinese Women’s Fashion” and attempt to fine-tune AI models within a “fashion industry context” to enhance

the authenticity and reliability of both the analytical processes and outcomes when examining fashion-themed unstructured data(Liu et al., 2024; Fan & Wang, 2024; Liang & Chen, 2024). This methodology can be applied across different stages of design, such as the research phase targeting user sentiment analysis and the analytical phase examining potential topics related to products or brands.

The rest structure of this paper is organized as follows: Section 2 presents a literature review that examines the current development status of AI technologies in the fashion industry and their relationship with users, identifying research gaps in the field. Section 3 introduces the specific experimental methodology, including research design, data collection methods, sample selection, and analytical approaches. Section 4 analyzes the results from both methodological and experimental content perspectives. Section 5 provides a comprehensive discussion of the research findings.

LITERATURE REVIEW

AI AND THE FASHION INDUSTRY

In recent years, AI technology has experienced rapid advancement, with LLMs such as ChatGPT and DeepSeek, based on Transformer architecture, demonstrating significant advantages in "rapid generation," "automated analysis," and "personalized decision-making" (Kalinin et al., 2024; Ranjan et al., 2025). These advantages have different impacts across different stages of product development.

For instance, the design process for apparel products can typically be divided into several stages: the analysis phase (analyzing target groups), where companies like Zara use data mining methods through social media to better understand target consumers' preferences (Cao, 2024; Dang, 2022); the preparation phase (collecting trend information and other related data); the conceptualization phase (where designers organize information and derive inspiration); and the implementation phase (creating fashion illustrations and garment production). For example, Nike enhances consumer engagement and brand loyalty by adding user customization options that allow consumers to participate in the design process (Gan, 2023).

Consumers are both research subjects and end-users of garments (Sarkar, 2011; Schiaroli et al., 2024). Fine-tuned LLMs can extract design preferences from consumer reviews, identify popular design elements, and quantify consumers' perceived value differences across various price points for apparel (Yu et al., 2021).

However, fashion design students currently tend to focus more on the final presentation effects of products and personal artistic expression within the design process, while the user research phase often fails to receive adequate attention (Harvey et al., 2019; Murzyn-Kupisz & Hołuj, 2021). This includes, for example, the current development status and challenges of women's fashion from the consumer perspective.

LLMS AND ROBERTA

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model proposed by Google in 2018 that achieved significant performance improvements across various natural language processing (NLP) tasks (Devlin et al., 2019). RoBERTa (Robustly Optimized BERT Pretraining Approach) includes several key optimizations based on BERT. Current research demonstrates that fine-tuned pre-trained RoBERTa models can achieve superior performance compared to BERT across various downstream tasks, including sentiment analysis, textual entailment, and reading comprehension (Liu et al., 2019).In fashion industry text analysis, RoBERTa demonstrates superior feature extraction capabilities and enhanced performance when processing complex unstructured textual data, due to its stronger language understanding abilities and more stable training processes.

ChatGPT, DeepSeek, and Gemini represent new-generation LLMs that employ generative architectures and possess capabilities in conversational interaction, text generation, and complex thinking. For analytical tasks involving specific niche domains and deep textual understanding, encoder models such as RoBERTa continue to maintain advantages (Nielsen et al., 2024).

Compared to image-based AI models, traditional text understanding models (such as BERT and RoBERTa) are relatively straightforward in terms of data preprocessing and model fine-tuning. For instance, in the apparel development process, different pre-trained models can be selected based on specific objectives and subjected to domain-specific fine-tuning to obtain models that better align with particular domain requirements, allowing more accurate understanding of

fashion industry terms and concepts (Chen et al., 2024). Though rigorous application in fashion market research remains limited.

In summary, to address this research gap, this research will employ a fine-tuned RoBERTa model using user comments about “Current Development Status of Chinese Women’s Fashion” from Bilibili, a major Chinese video platform, as the data source. Through sentiment analysis and topic modeling methodologies, this research examines users' emotional attitudes and focal concerns regarding this topic from a consumer perspective. This research aims to validate the application value of text-based language models in fashion industry market research and explores the following research questions:

- What emotional attitudes do young consumers hold toward the current development status of Chinese women's fashion?
- What are the core issues in Chinese women's fashion development that consumers focus on?
- Based on consumer feedback, what are the primary challenges facing the Chinese women’s fashion industry?

METHODOLOGY

To understand the current status of the Chinese women's fashion market, this research follows a five-step process for data collection and analysis, implemented using Python version 3.9.1. To validate model reliability, we employ Precision, Recall, and F1-score as evaluation metrics. These metrics represent established standards for machine learning model assessment (Goutte & Gaussier, 2005; Sasaki, 2007; Christen et al., 2023).

DATA COLLECTION, SOURCES, PLATFORM SELECTION, AND DATA SCOPE

The sample utilized user comments from Bilibili, a major Chinese video platform, as the data source. Compared to short-video-focused platforms such as TikTok, Bilibili specializes in long-form video content, providing more adequate time and space for in-depth topic discussions (Shang, 2025; Xia, 2025). According to Bilibili’s Q1 2024 investor presentation, the platform's user base primarily consists of individuals born between 1985 and 2009, representing 65% of total users, with female users comprising 48% of the overall user population. Notably, over 80% of students from China's Project 985 and 211 universities—elite higher education institutions designated by the Ministry of Education—are registered Bilibili users (Bilibili Inc., 2024). This demographic profile indicates a highly educated user base, which contributes to more substantive and articulate discussions in comment sections. Furthermore, the substantial female user presence (48%) ensures adequate representation of the target demographic for women's fashion discourse, making Bilibili an appropriate platform for investigating attitudes toward Chinese women's fashion.(Li et al., 2025).

For this research, searches were conducted on Bilibili using the keywords "women's fashion" and "women's fashion status," yielding 1,006 videos containing these terms. From these results, videos were ranked by view count (highest to lowest), and the 13 videos with the highest comment volumes were selected for analysis. The selected videos span a four-year period from May 13, 2021, to April 20, 2025. Comment data collection, including both original comments and replies, was conducted between June 5 and June 7, 2025, resulting in a total of 30,796 initial text data points (Fig. 01)

DATA PREPROCESSING AND CLEANING

Text Preprocessing

The preprocessing phase involved identifying and removing reply formats such as “Reply @username:”; cleaning HTML tags, URLs, email addresses, and phone numbers; normalizing consecutive repeated characters, words, and punctuation marks; and converting emojis to textual descriptions or removing them entirely.

In this process,a selective emoji protocol was applied: emojis conveying emotional states or emphasis were converted to textual equivalents, while decorative emojis inserted for platform reward purposes were excluded.

Intelligent Text Enhancement

For comments lacking subjects, contextual information was intelligently added (e.g., "too small" → “the clothes are too small”). Missing semantic components were added through pattern matching techniques.

Content Filtering

Short comments (fewer than 3 characters) were removed; pure punctuation, pure numbers, and

Selected Bilibili Video Dataset for Chinese Women's Fashion Analysis

No.	Video Title	Views (k)	Comments
1	High Return Rates for E-commerce Women's Fashion: You Really Can't Blame Others	78	1,234
2	Close Down Quickly! What Kind of Distorted Aesthetics Do Current Women's Clothing Stores Have?	5,212	1,768
3	Buy 10 Items, Return 8! Are Clothes Also Becoming Assassins?	523	2,273
4	Men's Cotton Jacket Only 69 Yuan... With Free Shipping... Is Women's Money So Easy to Earn? Reject Pink Tax!	1,639	2,818
5	Don't You Know Why Your Women's Clothing Has High Return Rates??	1,818	5,685
6	Differences Between Women's and Men's Pants	105	445
7	Women's Clothing is Expensive and Slow to Ship, Who Exactly is Discriminating Against Female Consumers?	929	7,721
8	Merchants Should Allow People Who Return Items to Post Buyer Reviews	4,864	1,018
9	Distorted Aesthetics! Current Women's Clothing Size L Can Only Fit Dogs, Isn't Anyone Managing These Merchants?	184	1,209
10	Isn't Anyone Managing the Sizing of Current Women's Clothing Stores??	841	2,543
11	Isn't Anyone Going to Manage Women's Jeans??	106	450
12	Expensive and Poor Quality: Chinese Women's Fashion Trapped in a Dead End	61	124
13	Question Men, Understand Men, Become a More Money-Saving Woman	1,309	2,936

Fig. 01

meaningless phrases were identified and deleted; complete duplicates were removed along with similarity-based deduplication.

Quality Assessment

A quality scoring system (0-6 points) was developed based on: length score (0-2 points), fashion relevance (0-2 points), subject completeness (0-1 point), evaluative content (0-1 point), with reply format penalties (-0.5 points). Using keyword dictionaries, the system decided and found whether comments contained clear subjects or referential objects. A strict filtering strategy was employed, requiring comments to be fashion-related with quality scores above 2 points.

Cleaning Results

Following completion of the above steps, a total of 10,594 processed comment texts were obtained.

ROBERTA-BASED DATA CLEANING

During manual checking of the preprocessed data, several low-quality comments were identified, primarily characterized by: (1) extensive use of emoticons as replacements for textual expression; (2) incomplete sentence structures lacking essential components such as subjects and objects; (3) fragmented expressions with ambiguous semantics or unclear logic. These issues primarily come from the complexity of internet language and the diversity of user expression habits, which traditional rule-based text cleaning methods struggle to effectively identify and address.

To more precisely identify and filter low-quality comments, 569 comments were randomly sampled from the preprocessed data for manual annotation. The annotation criteria were: 0 for meaningless comments (including semantically incomplete, off-topic, or pure emoticon content), and 1 for meaningful comments (complete in meaning and relevant to the research topic).

Based on the manually annotated data, the hfl/chinese-roberta-wwm-ext pre-trained model was employed for fine-tuning. The hfl/chinese-roberta-wwm-ext is a Chinese pre-trained language model based on RoBERTa architecture, developed by the Harbin Institute of Technology-iFLYTEK Joint Laboratory (HFL) (Cui et al., 2021). This model is widely applied in Chinese text classification and sentiment recognition, specifically

optimized for Chinese text processing.

For model training, the data was split into training and validation sets at an 8:2 ratio. This training method was consistently applied to subsequent sentiment analysis and topic classification tasks.

Model Performance Evaluation

For the 'meaningless' class, the model achieved precision of 0.95, recall of 0.91, and F1-score of 0.93 (n=45). For 'meaningful' comments, precision was 0.94, recall was 0.97, and F1-score was 0.96 (n=69). Overall accuracy was 0.95 across 114 test samples (Tab. 01).

Full Dataset Filtering Results

The trained model was applied to all 10,594 preprocessed comments, with a confidence threshold of 0.7 for filtering[1].Ultimately, 6,029 high-quality comments were obtained, resulting in a data retention rate of 56.9% (filtering rate of 43.1%). During the filtering process, the model reached an average prediction confidence of 83.1% across all comments.

1 Confidence represents the model's certainty level when processing comments; higher values indicate greater model confidence in its predictions.

ROBERTA-BASED SENTIMENT ANALYSIS

For the sentiment analysis phase, this research employed the hfl/chinese-roberta-wwm-ext-large pre-trained model for fine-tuning. This model represents an upgraded version of hfl/chinese-roberta-wwm-ext, capable of capturing more complex linguistic patterns and semantic relationships, getting higher accuracy in text understanding and generation tasks. The manually annotated sample for sentiment analysis comprised 1,062 comments, categorized into three sentiment classes: positive, neutral, and negative.

Model Performance Evaluation

All three sentiment classes achieved F1-scores above 0.97, with the model reaching an overall accuracy of 97.65% across 1,062 validation samples (Tab. 02 for per-class details).

Full Dataset Results

The trained model was applied to all 6,029 data points. Among these, positive comments accounted for 10.9% of the total comments (660 comments), negative comments comprised 41.3% (2,487 comments), and neutral comments represented 47.8% (2,882 comments). The model's average confidence in sentiment classification was 90.8%, with high-confidence predictions accounting for 83.6% of the total, while low-confidence

RoBERTa Model Performance for Data Quality Classification

Class	Precision	Recall	F1-score	Support
Meaningless	0.95	0.91	0.93	45
Meaningful	0.94	0.97	0.96	69
Accuracy			**0.95**	**114**
Macro avg	**0.95**	**0.94**	**0.94**	**114**
Weighted avg	**0.95**	**0.95**	**0.95**	**114**

Tab. 01

Class	Precision	Recall	F1-score	Support
Positive	0.9855	1.0000	0.9927	136
Negative	0.9833	0.9652	0.9741	488
Neutral	0.9663	0.9817	0.9740	438
Accuracy			**0.9765**	**1062**

Tab. 02

predictions comprised only 5.2%.

ROBERTA-BASED TOPIC CLASSIFICATION

Due to the increased complexity of multi-label topic classification and building upon the strong performance achieved in sentiment analysis (average accuracy of 97.65% across three sentiment categories), this section continued to employ the hfl/chinese-roberta-wwm-ext-large pre-trained model for topic modeling.

Initially, a subset of comments required selection for manual annotation. For this part, 380 samples were annotated using a mixed sampling strategy. An additional 249 comments were added to ensure training stability and maintain sample diversity and uniqueness. This resulted in a final dataset of 629 manually annotated samples.

The manually annotated samples revealed 9 distinct topic categories: price, quality, style design, size fit, shopping experience, industry environment, gender differences, repurchase behavior, and fabric materials.

Model Performance Evaluation

Optimal Performance by Topic Category (Tab. 03)

Among the nine topics, the model demonstrated the highest precision in identifying fabric-related topics, successfully recognizing all fabric-related comments. However, recognition accuracy for quality and style topics was relatively lower at only 0.75. The price topic exhibited the worst overall performance with an F1-score of only 0.69. Most topics achieved F1-scores ranging between 0.77-0.87, indicating relatively balanced overall performance.

Final Model Performance (Tab. 04)

Full Dataset Results

Total Sample Size: 6,029

The fine-tuned RoBERTa model was applied to 6,029 high-quality comments for topic classification, identifying a total of 8,862 topic labels with an overall coverage of 147.1%. Size Fit appeared 2,620 times (43.5%), Gender Differences appeared 2,492 times (41.3%), Style Design appeared 1,115 times (18.5%), Industry Environment appeared 752 times (12.5%), Shopping Experience appeared 604 times (10.0%), Price appeared 517 times (8.6%), Quality appeared 353 times (5.9%), Repurchase Behavior appeared 271 times (4.5%), and Fabric appeared 138 times (2.3%) (Fig. 02).

Analysis revealed that over half of the comments (56.8%, 3,426 samples) focused only on single topics. An additional 35.5% (2,141 samples) addressed two topics at the same time, while only 6.9% (413 samples) discussed three topics. Overall, 92.3% of consumers focused on one or two specific concerns in their comments (Fig. 03).

RoBERTa Model Performance for Data Quality Classification

Topic Category	Precision	Recall	F1-score
Quality	0.7500	0.7895	0.7692
Price	0.8889	0.5714	0.6957
Industry environment	0.9048	0.7037	0.7917
Gender differences	0.8333	0.9091	0.8696
Style design	0.7500	0.8333	0.7895
Shopping experience	0.9091	0.7143	0.8000
Size fit	0.8667	0.7879	0.8254
Repurchase behavior	0.9091	0.8333	0.8696
Fabric materials	1.0000	0.9333	0.9655

Tab. 03

Overall Performance Metrics for Multi-Label Topic Classification Model

Averaging Method	Accuracy	Precision	Recall	F1-score
Micro Average	0.6293	0.8544	0.7377	0.7918
Macro Average		0.8503	0.7293	0.7801

Tab. 04

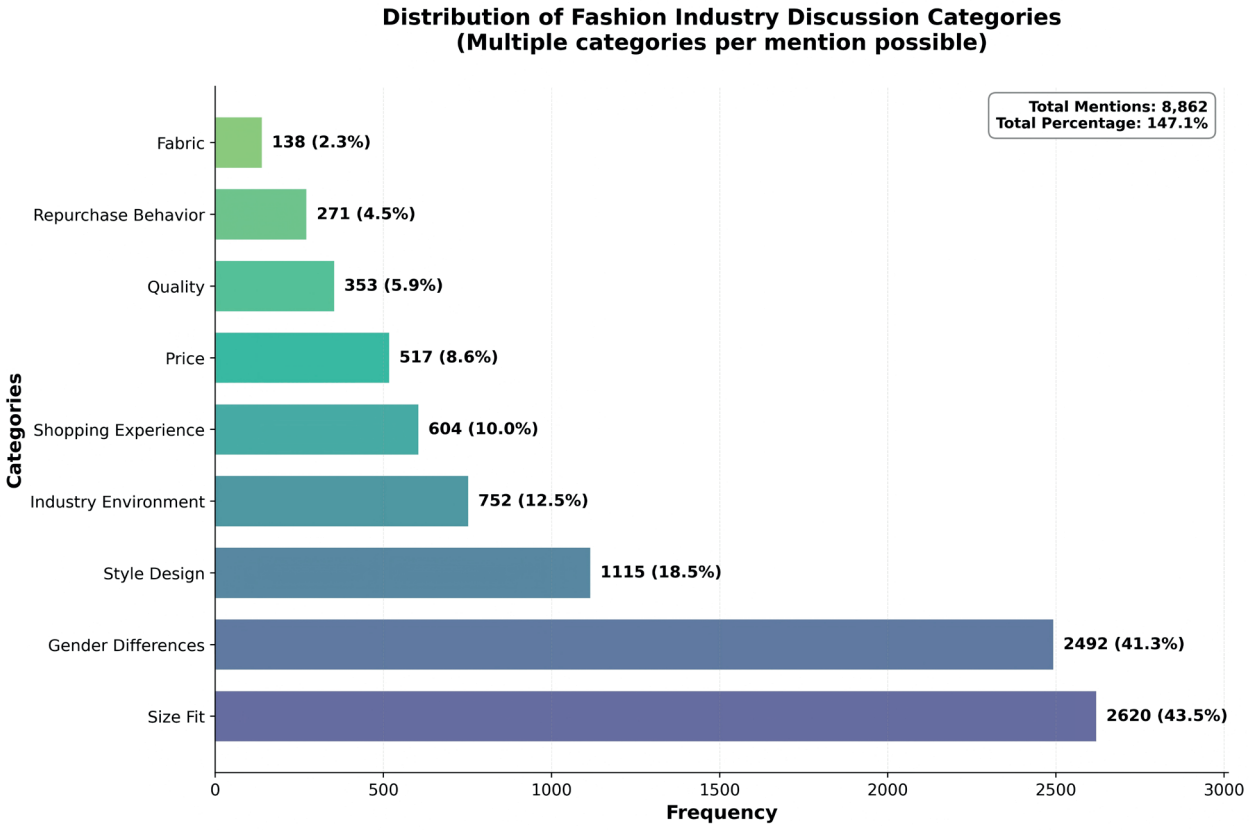
Distribution of Fashion Industry Discussion Categories
(Multiple categories per mention possible)
Total Mentions: 8,862
Total Percentage: 147.1%
Fabric
Repurchase Behavior
Quality
Price
Shopping Experience
Industry Environment
Style Design
Gender Differences
Size Fit
138 (2.3%)
271 (4.5%)
353 (5.9%)
517 (8.6%)
604 (10.0%)
752 (12.5%)
1115 (18.5%)
2492 (41.3%)
2620 (43.5%)
Categories
0
500
1000
1500
2000
2500
3000
Frequency

Fig. 02

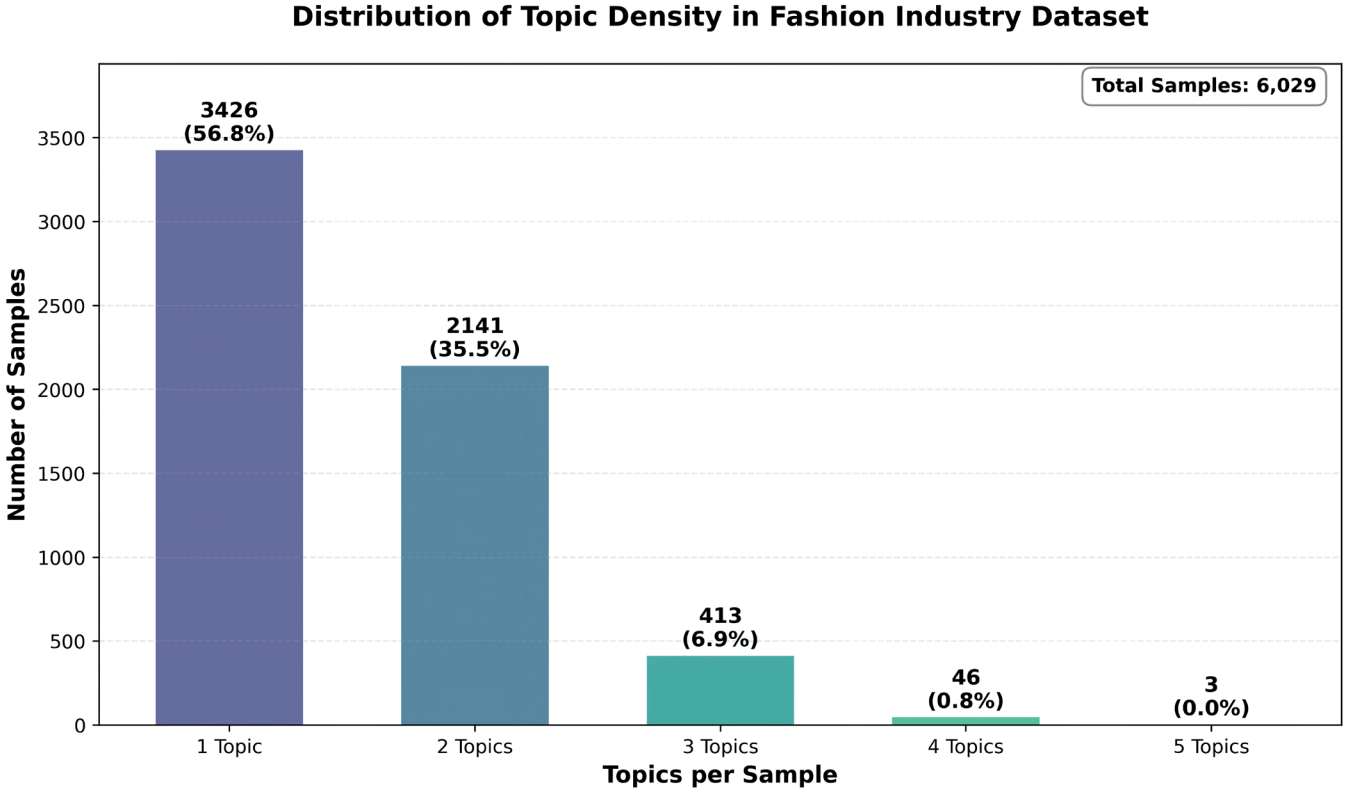
Distribution of Topic Density in Fashion Industry Dataset
Total Samples: 6,029
3426
(56.8%)
2141
(35.5%)
413
(6.9%)
46
(0.8%)
3
(0.0%)
Number of Samples
0
500
1000
1500
2000
2500
3000
3500
1 Topic
2 Topics
3 Topics
4 Topics
5 Topics
Topics per Sample

Fig. 03

RESULTS

ROBERTA-BASED DATA CLEANING

The results demonstrate that initial filtering using self-defined criteria successfully removed 65.60% of low-quality comments. However, following further data cleaning with the RoBERTa model, an additional 43.1% of comments were identified as invalid content and removed. This phenomenon is mainly attributed to the complexity and diversity of Chinese comments on the Bilibili platform(Shang, 2025). As a breeding ground for numerous internet slang expressions, the platform has heavy use of online casual language, shortened expressions, and context-specific popular phrases in user comments. Ultimately, the fine-tuned RoBERTa model achieved 95% classification accuracy on the validation set, demonstrating its reliability and effectiveness in understanding complex online linguistic environments. The model successfully captured meaning features of comment content, accurately distinguishing between meaningful and meaningless comments, thereby establishing a high-quality data foundation for subsequent sentiment analysis and topic modeling tasks.

ROBERTA-BASED SENTIMENT ANALYSIS

Model Performance and Method Validation

The results demonstrate balanced performance across all three sentiment categories, with F1-score variations of only 1.87 percentage points between categories, indicating strong generalization capabilities and class balance..These findings demonstrate that fine-tuned RoBERTa models can accurately classify consumer sentiment (97.65% accuracy) in fashion-specific contexts, with balanced performance across sentiment categories.

Sentiment Distribution Characteristics and Industry Insights

The analysis reveals clear sentiment distribution patterns in discussions regarding the current development status of the Chinese women's fashion market: neutral and negative sentiments dominate, while positive sentiment accounts for only 10.9% of the total. This distribution reflects the predominantly negative attitudes held by young Chinese consumer groups toward the current state of women's fashion industry development. These findings indirectly confirm the existence of systematic industry problems, examined in detail in Section 5.2.

ROBERTA-BASED TOPIC CLASSIFICATION

Due to the significantly higher complexity of the nine-class multi-label task compared to previous binary and three-class classification tasks, individual comments may simultaneously address multiple topics. This complexity creates an inverse relationship between classification difficulty and accuracy, putting higher demands on the model's generalization and learning capabilities.

Performance Analysis

Compared to the previous data cleaning and sentiment classification models, the multi-label nine-class model demonstrated lower performance across precision, recall, and F1-scores, exhibiting relatively careful overall performance. The model displayed typical "high precision, low recall" characteristics, achieving an average precision exceeding 85% while maintaining an average recall of approximately 73%, reflecting a careful predictive tendency of "preferring false negatives over false positives."

Variation Analysis

The model exhibited significant variations in recognition capabilities across different topic categories, primarily attributable to two factors: first, with only 503 training samples split across 9 topic categories, per-category data was severely imbalanced. Second, inconsistent labeling quality caused uneven learning outcomes across topics.

4.3.3 Full Dataset Results

When the fine-tuned model was applied to all 6,029 samples, the results revealed core issues of consumer concern in the Chinese women's fashion market. The data indicate that size fit (43.5%) and gender differences (41.3%) constitute the two most frequently discussed topics, together accounting for 84.8% of comment content. In contrast, fabric (2.3%), repurchase behavior (4.5%), and quality (5.9%) demonstrated relatively low mention frequencies, suggesting limited consumer perception and attention toward these aspects. Analysis revealed that over half of the comments (56.8%) focused exclusively on single topics, indicating that consumers maintain clear and concentrated attention on specific issues, reflecting the high consistency and common nature of major

problems within the industry.

CONCLUSION

METHODOLOGICAL CONTRIBUTIONS

This research applies text mining techniques to analyze unstructured consumer data in the fashion industry. Specifically, it employs a fine-tuned large language model based on the hfl/chinese-roberta-wwm-ext-large pre-trained model for sentiment analysis and multi-label topic classification.The results demonstrated the feasibility and accuracy of LLMs in design decision support, providing practitioners with demand-driven, customized AI experimental solutions (Bertacchini, 2023). This approach not only enhances the data-driven nature and clarity of design decisions but also advances methodological innovation and integration between AI technology and fashion design fields.

FINDINGS

Analysis of 6,029 Chinese social media comments revealed significant structural challenges in the women's fashion industry. Nearly 90% of consumers expressed non-positive sentiments, with sizing issues (43.5%) and gender representation concerns (41.3%) emerging as the most prominent pain points. Notably, while consumers demonstrate clear awareness of these problems through frequent and specific complaints, the issues remain prevalent in current industry practice. This disconnect between consumer recognition and industry status quo suggests systemic challenges that extend beyond surface-level concerns.

Combined with the previous sentiment analysis results, the data reveals several existing problems in the Chinese women's fashion market:

Absence of Size Standardization

Consumers universally face challenges with inaccurate sizing and poor fit. The current market lacks unified sizing standards, with different merchants and shopping platforms operating independently, resulting in significant variations of the same size across different brands (Workman,1991). More critically, influenced by societal aesthetic trends pursuing "pale, thin, and youthful" ideals (Liu & Li, 2024; Guo et al., 2023), women's clothing sizing is designed extremely small, leading to the unfortunate phenomenon of adult women being forced to choose children's clothing. This phenomenon reflects the industry's severe lack of recognition regarding women's body diversity (Hu et al., 2016).

Gender Stereotypes in Fashion Design

Women's fashion design exhibits deeply entrenched gender biases and stereotypes, primarily shown in: design ideas that force gender norms such as "women should wear skirts"(Cai, 2023); functional design deficiencies, including the removal or reduction of pocket designs with the excuse that "women have handbags"(Bolon, 2025); and severe pink tax issues(Brand & Gross, 2020), where female versions of products of the same type and quality are priced significantly higher than male versions. This pricing unfairness lacks reasonable cost reasons.

Forced Design Differentiation

This artificial differentiation has prompted consumer resistance, with increasing numbers of young consumers embracing gender-neutral fashion as a form of social practice (Jiang & Michelsen, 2024). Among these consumers, many women actively choose men's clothing for superior practicality and value. When consumers systematically prefer men's alternatives for everyday needs, this pattern reveals fundamental failures: inferior quality, reduced functionality, and poor value despite higher prices (the "pink tax"). The preference for men's clothing represents not merely individual choice but resistance against gendered design that sacrifices women's actual needs for prescribed femininity.

Vicious Cycle in Industry Ecosystem

The women's fashion industry is trapped in a negative feedback loop: (1) poor product quality, high prices, and inaccurate sizing lead to very high return rates; (2) rising return rates compress merchant profits, forcing further reductions in production costs; (3) platforms implement stricter return and exchange policies to protect consumer rights; (4) merchants transfer costs through extended pre-sale periods and increased selling prices; (5) product quality deteriorates further.

Deteriorating Shopping Experience

Consumers face multiple challenges throughout the entire shopping process: (1) very long pre-sale periods (often exceeding one month) during product selection; (2) receiving obviously second-hand returned merchandise; (3)

price manipulation practices in promotional campaigns (artificially raising prices before applying discounts). These issues seriously damage consumer confidence and patience in women's fashion shopping.

Systemic Price Discrimination

The pink tax phenomenon is especially obvious in the women's fashion industry, where female-related products are regularly priced higher than comparable male products (Fu, 2024). This price differential often lacks reasonable cost reasons and primarily reflects excessive taking advantage of women's purchasing power and the commercialization of gender bias.

Collapse of Quality Assurance Systems

In pursuit of short-term profit maximization, merchants universally compress production costs, resulting in the failure of quality control mechanisms in garment manufacturing processes. This cost compression manifests not only in declining craftsmanship standards but also in the relaxation of product durability and safety standards.

Distorted Market Competition Mechanisms

Quality-oriented merchants face structural disadvantage: fast fashion margins (16%) far exceed those of traditional quality retailers (7%) (Sull & Turconi, 2008), enabling a "bad money drives out good" dynamic (Akerlof, 1970) that further deteriorates the industry ecosystem.

Systemic Decline in Raw Material Quality

Closely linked to quality issues, cost compression directly results in narrowed fabric selection ranges and declining quality. Merchants are forced to choose cheaper and lower-quality raw materials, creating a comprehensive quality decline chain from source to end product.

INDUSTRY SIGNIFICANCE

As one of the world's largest garment production and consumption markets (Circular Fashion: Prospects for China's New Textile Economy), the developmental dynamics and structural issues of China's women's fashion industry hold significant international reference value. The data insights and analytical methods provided by this research not only offer precise guidance for international brands to improve product design but also provide scientific evidence for the global fashion industry to understand the Chinese market and formulate localization strategies.

CAPTIONS

[Fig. 01] Selected Bilibili Video Dataset for Chinese Women's Fashion Analysis.

[Fig. 02] Distribution of Fashion Industry Discussion Categories.

[Fig. 03] Distribution of topic density in Fashion Industry Dataset.

REFERENCES

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Bertacchini, A., & Pantano, P. S. (2023). Synergies in the evolution of artificial intelligence and fashion: A prospective analysis. *Fashion Highlight*, *1*(2), 82–88. https://doi.org/10.36253/fh-2503

Bilibili Inc. (2024). *Q1 2024 Bilibili Inc. investor presentation*. https://ir.bilibili.com/media/0czhzr4m/q1-2024-bilibili-inc-investor-presentation_cn.pdf

Bolon, C. (2025). *The missing void: Lack of pockets in womenswear and its effect on daily life* [Undergraduate honors thesis, University of Northern Colorado]. Scholarship & Creative Works @ Digital UNC. https://digscholarship.unco.edu/honors/125

Brand, A., & Gross, T. (2020). Paying the pink tax on a blue dress: Exploring gender-based price-premiums in fashion recommendations. In R. Bernhaupt, C. Ardito, & S. Sauer (Eds.), *Human-centered software engineering: HCSE 2020*. *Lecture Notes in Computer Science* (*Vol. 12481*, pp. 153–166). Springer. https://doi.org/10.1007/978-3-030-64266-2_12

Cai, X. (2023). Gender stereotypes in female contemporary dress aesthetics. *Lecture Notes in Education Psychology and Public Media*, *5*, 290–295. https://doi.org/10.54254/2753-7048/5/20220537

Cao, J. (2024). Enabling ZARA's operational innovation and value creation with artificial intelligence. *Advances in Economics, Management and Political Sciences*, *86*, 81–87. https://doi.org/10.54254/2754-1169/86/20240948

Chen, K., Luo, Y., Hu, H., & Wang, H. (2024). Intelligent clothing style generation method based on style editing. *Artificial Intelligence and Robotics Research*, *13*(3), 636–647. https://doi.org/10.12677/airr.2024.133065

Christen, P., Hand, D. J., & Kirielle, N. (2023). A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, *56*(3), Article 73. https://doi.org/10.1145/3606367

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Dang, D. (2022). *Artificial intelligence: AI in fashion and beauty e-commerce Zara, Sephora* [Bachelor's thesis, LAB University of Applied Sciences]. Theseus. https://urn.fi/

URN:NBN:fi:amk-2022112524103

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fan, Y., & Wang, Y. (2024). Design and development of large language model applied to fashion analysis. In G. Montagna & C. Carvalho (Eds.), *Human factors for apparel and textile engineering: AHFE 2024 International Conference*. *AHFE Open Access* (*Vol. 134*). AHFE International. https://doi.org/10.54941/ahfe1004917

Forbes, J., Huang, T., Kanazawa, K., & Russell, S. (1995). The BATmobile: Towards a Bayesian automated taxi. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (*Vol. 2*, pp. 1878–1885). Morgan Kaufmann Publishers Inc.

Fu, Y. (2024). Investigating factors influencing acceptance of pink tax on personal care products among Chinese urban women. *Communications in Humanities Research*, *42*, 118–125.

Gan, J. (2023). Analysis of Nike's brand marketing strategy. *Advances in Economics, Management and Political Sciences*, *48*, 1–6. https://doi.org/10.54254/2754-1169/48/20230410

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D. E. Losada & J. M. Fernández-Luna (Eds.), *Advances in information retrieval: ECIR 2005*. *Lecture Notes in Computer Science* (*Vol. 3408*, pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Guo, X., & Duan, Z. (2023). Semantic evolution of internet slang and social mentality representation. *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)*, *6*, 133–145. https://doi.org/10.16088/j.issn.1001-6597.2023.06.012

Harvey, N., Ankiewicz, P., & van As, F. (2019). Fashion design education: Effects of users as design core and inspirational source. In *Proceedings of PATT 37: Developing a Knowledge Economy Through Technology and Engineering Education* (pp. 203–211). University of Malta. https://assets-002.noviams.com/novi-file-uploads/iteea/pubs/PATT37Malta2019Proceedings5-2ed95141.pdf

Hu, X., & Zhou, J. (2016). Study on somatotype characteristics and differences of female youth from Liaoning Province and Guangdong Province in China. In V. Duffy (Ed.), *Digital human modeling: Applications in health, safety, ergonomics and risk management. DHM 2016*. *Lecture Notes in Computer Science* (*Vol. 9745*, pp. 28–37). Springer. https://doi.org/10.1007/978-3-319-40247-5_3

Jiang, X., & Michelsen, M. (2024). Unisex fashion as a social practice: A comparative study between young heterosexual consumers in the United States and China. *Cultural Sociology*, *0*(0). https://doi.org/10.1177/17499755241272866

Jung, D., & Suh, S. (2023). Development of customized textile design using AI technology. *Journal of the Korean Society of Clothing and Textiles*, *47*(6), 1137–1156. https://doi.org/10.5850/JKSCT.2023.47.6.1137

Kalinin, A., Jafari, A., Avots, E., Ozcinar, C., & Anbarjafari, G. (2024). Generative AI-based style recommendation using fashion item detection and classification. *Signal, Image and Video Processing*, *18*(12), 9179–9189. https://doi.org/10.21203/rs.3.rs-4517638/v1

Li, X., Zhu, C., Xie, Y., Wang, L., Luo, X., Wu, S., & Zhang, C. (2025). Quality assessment of Chinese robot-assisted rehabilitation shorts. *Digital Health*, *11*, Article 20552076251391841. https://doi.org/10.1177/20552076251391841

Liang, Z., & Chen, J. (2024). An AI-BERT-Bi-GRU-LDA algorithm for negative sentiment analysis on Bilibili comments. *PeerJ Computer Science*, *10*, Article e2029. https://doi.org/10.7717/peerj-cs.2029

Liu, H., Tang, X., Chen, T., Liu, J., Indu, I., Zou, H. P., Dai, P., Galan, R. F., Porter, M. D., Jia, D., Zhang, N., & Xiong, L. (2024). Sequential LLM framework for fashion recommendation. In F. Dernoncourt, D. Preoțiuc-Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 1276–1285). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-industry.95

Liu, Y., & Li, X. (2024). "Pale, young, and slim" girls on red: A study of young femininities on social media in post-socialist China. *Feminist Media Studies*, *24*(4), 744–759. https://doi.org/10.1080/14680777.2023.2226830

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692

Mohammadi, S. O., & Kalhor, A. (2021). Smart fashion: A review of AI applications in the fashion & apparel industry. *arXiv*. https://arxiv.org/abs/2111.00905

Murzyn-Kupisz, M., & Hołuj, D. (2021). Fashion design education and sustainability: Towards an equilibrium between craftsmanship and artistic and business skills? *Education Sciences*, *11*(9), Article 531. https://doi.org/10.3390/educsci11090531

Musmeci, N., & Pantano, P. S. (2023). Ethical challenges in the evolution of artificial intelligence and fashion: A prospective analysis. *Fashion Highlight*, *1*(2), 90–96. https://doi.org/10.36253/fh-2502

Nielsen, D. S., Enevoldsen, K., & Schneider-Kamp, P. (2025). Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. *arXiv*. https://arxiv.org/abs/2406.13469

Ranjan, A., & Upadhyay, A. K. (2024). Value co-creation by interactive AI in fashion e-commerce. *Cogent Business & Management*, *12*(1), Article 2440127. https://doi.org/10.1080/23311975.2024.2440127

Rizzi, G., & Casciani, D. (2023). A.I. into fashion processes: Laying the groundwork. *Fashion Highlight*, *1*(2), 12–20. https://doi.org/10.36253/fh-2490

Sarkar, S. (2011, July). The design process in fashion product development. *Fibre2Fashion*. https://www.fibre2fashion.com/industry-article/5723/the-design-process-in-fashion-product-development

Sasaki, Y. (2007). *The truth of the F-measure* (Teach Tutor Mater). School of Computer Science, University of Manchester. https://people.cs.pitt.edu/~litman/courses/cs1671s20/F-measure-YS-26Oct07.pdf

Schiaroli, V., Fraccascia, L., & Dangelico, R. M. (2024).

How can consumers behave sustainably in the fashion industry? A systematic literature review of determinants, drivers, and barriers across the consumption phases. *Journal of Cleaner Production*, *483*, Article 144232. https://doi.org/10.1016/j.jclepro.2024.144232

Shang, Z. (2025). Shifting platform governance: Examining participatory content moderation on a Chinese platform Bilibili. *Information, Communication & Society*, 1–21. https://doi.org/10.1080/1369118X.2025.2520004

Statista. (2025, August). Fashion market worldwide - Revenue. *Statista Market Insights*. Retrieved December 26, 2025, from https://www.statista.com/outlook/emo/fashion/worldwide?currency=USD

Sull, D. N., & Turconi, S. (2008). Fast fashion lessons. *Business Strategy Review*, *19*(2), 4–11. https://doi.org/10.1111/j.1467-8616.2008.00527.x

Workman, J. E. (1991). Body measurement specifications for fit models as a factor in clothing size variation. *Clothing and Textiles Research Journal*, *10*(1), 31–36. https://doi.org/10.1177/0887302X9101000105

Xia, L., Bao, J., Yao, K., Zhang, J., & Yu, W. (2025). Evaluation of the quality and reliability of Chinese content about orthognathic surgery on BiliBili and TikTok: A cross-sectional study. *Scientific Reports*, *15*, Article 28967. https://doi.org/10.1038/s41598-025-13941-0

Yu, L., & Bai, X. (2021). Implicit aspect extraction from online clothing reviews with fine-tuning BERT algorithm. *Journal of Physics: Conference Series*, *1995*(1), Article 012040. https://doi.org/10.1088/1742-6596/1995/1/012040

Zou, X., & Wong, W. (2021). Fashion after fashion: A report of AI in fashion. *arXiv*. https://arxiv.org/abs/2105.03050