

Zero-Shot Complaint Classification and Style-Controlled Response Generation via Large Language Models for Emotion-Aware E-Commerce Review Management [†]

Yi Lin ¹, Chien-Hung Lai ^{2,*}  and Tzu-Shuang Liu ¹

¹ Department of Business Administration, Takming University of Science and Technology, Taipei 11451, Taiwan; linyi@takming.edu.tw (Y.L.); 0007493@gs.takming.edu.tw (T.-S.L.)

² Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan

* Correspondence: ee12245@ntut.edu.tw; Tel.: +886-2-27712171 (ext. 2200)

[†] Presented at 8th International Conference on Knowledge Innovation and Invention 2025 (ICKII 2025), Fukuoka, Japan, 22–24 August 2025.

Abstract

We developed a large language model-powered system that classifies complaint categories and adapts response styles for e-commerce reviews. By integrating sentiment clustering, zero-shot classification, and style-conditioned prompt engineering, it enables context-aware, emotionally aligned reply generation for enhancing automated customer interaction and reputation management.

Keywords: sentiment clustering; complaint classification; LLM; response style adaptation; prompt engineering

1. Introduction

The explosion of user-generated reviews on e-commerce platforms has brought both opportunities and challenges for businesses. These reviews serve as real-time customer feedback, revealing satisfaction levels, emotional tone, and expectations for products and services. The result of Ref. [1] proposed a sentiment clustering system capable of identifying emotional polarity and grouping reviews by topic using a bidirectional encoder representation from transformers (BERT)-based embedding and a K-means pipeline. While the system effectively quantified emotional feedback and grouped it thematically, it lacked deeper contextual reasoning and stylistic adaptability in responses.

Building upon previous sentiment clustering work, this research adds a semantic classification layer and style-conditioned response generation for more precise and human-aligned review engagement. While detecting sentiment is a fundamental step in understanding customer dissatisfaction, it is not sufficient. Negative reviews are often diverse in content—some stem from sizing issues, others from poor quality, late delivery, or even customer service frustration. Furthermore, human communication is nuanced; not every negative comment demands a serious or apologetic tone—some may benefit from humor, empathy, or directness.

To address this, this research presents a system that classifies emotional polarity, identifies the main reason for dissatisfaction, and determines the most appropriate style for replying, using LLMs to automatically generate personalized brand responses. The system forms the backbone of the second-stage system for intelligent review management.



Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 2 February 2026

Citation: Lin, Y.; Lai, C.-H.; Liu, T.-S. Zero-Shot Complaint Classification and Style-Controlled Response Generation via Large Language Models for Emotion-Aware E-Commerce Review Management. *Eng. Proc.* **2025**, *120*, 29. <https://doi.org/10.3390/engproc2025120029>

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

This research aims to develop an emotional and context-aware response system which (1) leverages zero-shot classification in Ref. [2] via LLMs to determine the core complaint category in a review, (2) infers an appropriate response style based on emotion and topic, (3) automatically generates contextually accurate and stylistically aligned brand replies using LLM prompts, and (4) integrates these results with previously developed sentiment and clustering outputs for comprehensive analysis.

The system developed presents a zero-shot prompt-based complaint classification module requiring no fine-tuning, a schema for style-adaptive response generation aligned with emotional tone and complaint category, and a structured data pipeline combining clustering, sentiment, complaint type, and LLM outputs into a unified dataset, demonstrating how LLMs can be prompted effectively to align content with style and function.

2. Complaint Classification Using LLMs

2.1. Zero-Shot Classification

A comprehensive framework combining prompt-based LLM with multimodal emotion-aware summarization was proposed in Ref. [3]. The use of aspect-aware attention is consistent with the focus of this study, which is to use pitch-guided cues for zero-shot complaint category classification of e-commerce reviews.

Traditional text classifiers require labeled data and fine-tuning, which can be costly and inflexible. This research extends prior sentiment clustering approaches by incorporating a semantic classification layer and style-conditioned response generation framework, enabling more precise and human-aligned review engagement. The proposed method employs zero-shot learning with LLMs to dynamically infer root causes of dissatisfaction directly from unstructured customer review text, eliminating the need for model retraining. This approach is flexible to category changes, adaptable across domains, and easily integrated into a prompt-based response generation pipeline.

2.2. Prompt Structure for Complaint Classification

A central component of this research is the use of a prompt. From a cognitive perspective, Prompt is the conditional input. In most LLM architectures (such as Together artificial intelligence (AI) in Ref. [4]), the core of the model is the conditional probability distribution, as shown in Equation (1).

$$Y = P(X) \quad (1)$$

where X is the input text, which is the prompt, and Y is the output generated by the model, which is the response. The essence of a prompt is to transform the problem description into X , then let the model maximize the conditional probability $P(X)$ based on X .

More specifically, prompt forms a conditional probability space; given a prompt X , the goal of LLM is to predict the probability of each possible next Token y_t , as shown in Equation (2).

$$Y = P(y_1, y_2, \dots, y_{t-1}, X) \quad (2)$$

where $y_1, y_2, \dots, y_{t-1}$ are the texts generated so far, and X is the entire prompt as one of the conditions. The key is that the prompt changes the initial hidden state and attention mechanism, so that subsequent generation is strongly guided by X .

Therefore, the prompt design can be modeled as an optimization problem. Assuming that the model output is expected to meet certain expected attributes (for example, the response should be “professional”), the goal is to find an optimal prompt X^* that can make the generated text Y meet the expected attribute A to the greatest extent, expressed as (3):

$$X^* = \arg E_{Y \sim P(X)} [Quality(Y; A)] \quad (3)$$

where $Quality(Y; A)$ is a scoring function that evaluates whether the output Y meets the attribute A . Therefore, it can also be understood that prompt engineering is essentially a problem of condition shaping.

Based on review, inspection, and business-relevant concerns, we define six primary categories of customer dissatisfaction.

1. Size or fit problem
2. Material or quality issue
3. Price or value dissatisfaction
4. Shipping or delivery issue
5. Customer service problem
6. Other

These categories balance coverage and interpretability for both analysis and response design.

2.3. Prompt Template Design

To enable accurate classification, this research proposed a consistent and instruction-following prompt structure for LLMs, as shown in Figure 1. In addition, this research can also automatically select different tones to generate responses based on the conditions in Table 1, according to the estimated emotional scores of customer reviews, where s_i represents the sentiment score calculated based on the customer review of this record.

You are a review analysis expert.

Below is a customer's product review. Your task is to identify the MAIN complaint category expressed by the reviewer. Choose ONLY ONE from the list of possible categories.

Review:
"{review_text}"

Available Categories:
– Size or Fit Problem
– Material or Quality Issue
– Price or Value Dissatisfaction
– Shipping or Delivery Issue
– Customer Service Problem
– Other

Respond only with one category name from the list.

Figure 1. Prompt proposed in this research.

Table 1. Criteria for selecting corresponding tones.

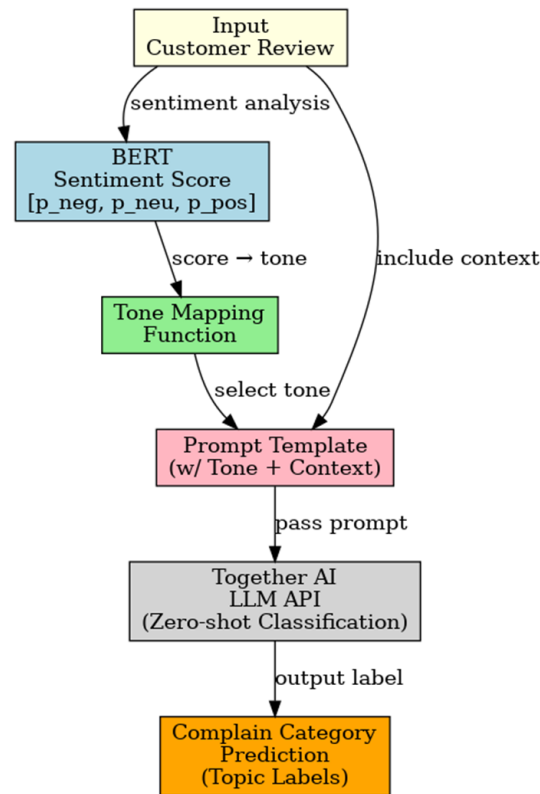
Sentiment Score Range	Corresponding Tone	Label [1]	Stars [1]
$s_i < -0.6$	Apologetic	0	1
$-0.6 \leq s_i < -0.3$	Empathetic	0	2
$-0.3 \leq s_i \leq 0.3$	Professional	1	3
$0.3 < s_i \leq 0.7$	Encouraging	2	4
$s_i > 0.7$	Humorous	2	5

2.4. Inference Process and Data Integration

By applying the above prompt to each review in the dataset using Together AI, each response will be recorded as a new field in the dataset, which is named “complaint_category”. The proposed method enables high-semantic-resolution categorization of previously unlabeled free-text reviews. Representative classification outputs are presented in Table 2, while Figure 2 illustrates the end-to-end workflow of the LLM-based classification module.

Table 2. Categorized output.

Customers' Reviews	"Complaint_Category"
The sweater is cute but way too tight around the arms.	Size or Fit Problem
This dress fell apart after one wash. Poor quality.	Material or Quality Issue
Still waiting for delivery after 2 weeks.	Shipping or Delivery Issue

**Figure 2.** Complete processing flow of the LLM classification module.

3. System Architecture

3.1. Overview

The complete system is composed of the following five modules.

1. Text preprocessing and embedding module
2. Sentiment analysis module
3. Complaint category classification module (LLM-based zero-shot)
4. Response style selector
5. LLM prompt construction and response generator

3.2. Text Preprocessing and Embedding Module

Prepare review data for semantic and sentiment analysis by cleaning and embedding each review into a high-dimensional space. Let x_i be a raw review. The defined preprocessing logical function is expressed in Equation (4).

$$\tilde{x}_i = \text{Clean}(x_i) = \text{Lowercase}(\text{RemovePunctuation}(x_i)) \quad (4)$$

Based on Ref. [1], we applied a pretrained BERT model to extract the [CLS] token embedding as the sentence representation, as shown in (5).

$$v_i = \text{BERT}_{CLS}(\tilde{x}_i) \in R^{768} \quad (5)$$

3.3. Sentiment Analysis Module

Following the fundamental sentiment estimation framework [1], the sentiment probability distribution y_i across classes can be computed as shown in Equation (6). Subsequently, a scalar sentiment score is defined in Equation (7), where p_{pos} and p_{neg} represent the positive and negative class probabilities, respectively. The sentiment labeling scheme is adopted from Ref. [1], as presented in Equation (8), where s_i represents scalar sentiment score in Equation (7).

$$y_i = \text{softmax}(W \cdot v_i + b) \in R^3 \quad (6)$$

$$s_i = p_{pos} - p_{neg} \in [-1, 1] \quad (7)$$

$$\text{Label}(s_i) = \begin{cases} 0, & s_i < -0.3 \\ 1, & -0.3 \leq s_i \leq 0.3 \\ 2, & s_i \geq 0.3 \end{cases} \quad (8)$$

3.4. LLM-Based Complaint Category Classification Module

Given a review x_i , this research defined a prompt template $P(x_i)$ shown in Equation (9), where C is a set of category labels and expressed in Equation (10), then the LLM is instructed to return a categorized answer shown in Equation (11), where $\hat{c}_i$ represents the returned answer, arg is the LLM. This is a zero-shot inference task with constrained output.

$$P(x_i) = \text{Prompt}(x_i, C) \quad (9)$$

$$C = \{c_1, c_2, \dots, c_k\}, k = 6 \quad (10)$$

$$\hat{c}_i = \text{arg}P(x_i) \quad (11)$$

3.5. Response Style Selector

Let S denote the space of available response styles. The appropriate tone for addressing complaints varies depending on their nature. For instance, a humorous tone may effectively mitigate frustration in cases involving minor fit issues, whereas a serious and professional tone is imperative when handling complaints related to product defects or service failures. This module predicts the optimal response style by leveraging both sentiment analysis scores and complaint categorization, as formalized in the mapping defined in (12). The style selection function is defined in Equation (13), where C_i represents the i -th complaint category and s_i represents scalar sentiment score in Equation (7), which implements the rule-based mapping specified in Table 1.

$$S = \{\text{Apologetic}, \text{Empathetic}, \text{Professional}, \text{Encouraging}, \text{Humorous}\} \quad (12)$$

$$\text{style}_i = f(s_i, C_i) \in S \quad (13)$$

3.6. LLM-Based Brand Response Generator

This module generates a personalized and context-aware response to each customer review using LLM, based on the emotion score, complaint category, and selected response style illustrated in Equation (14), where x_i is the review text, the response tone will be selected by (13), and sentiment descriptions corresponding to s_i are illustrated in Equation (15). The template to generate prompts is shown in Figure 3.

$$P_i = \text{Prompt}(x_i, s_i, \text{complaint_category}, \text{style}_i) \quad (14)$$

$$SentimentDescription = \begin{cases} \text{very negative}, s_i < -0.6 \\ \text{somewhat negative}, s_i < -0.3 \\ \text{neutral}, s_i \leq 0.3 \\ \text{positive}, s_i > 0.3 \end{cases} \quad (15)$$

```

You are a brand representative.

A customer left the following review:
\ "{{row['review_text']}}"

Their feedback falls under the category:
{row['complaint_category']}
Their emotional tone is:
{sentiment_description(row['sentiment_score'])}
Your task is to write a reply in a
{row['response_tone']} tone that addresses their
concern, reflects {row['response_tone']}
communication style, and maintains a positive brand
image.

```

Figure 3. Template to generate prompts.

4. Implementation Result

To implement the system of this research, the same dataset, Women's Clothing E-Commerce Reviews from Kaggle in Ref. [5], was chosen. The implementation pipeline is shown in Figure 4.

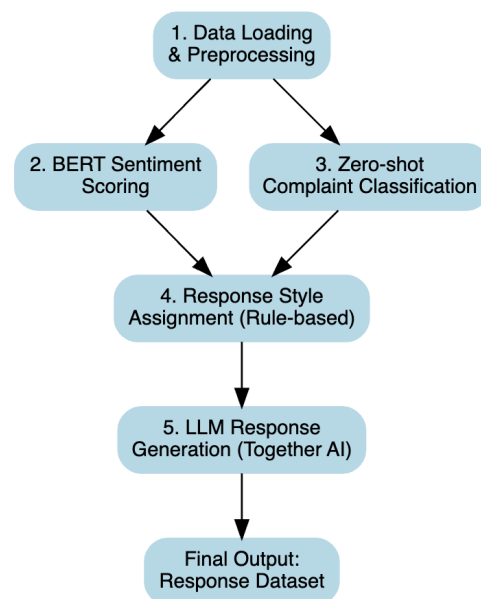


Figure 4. Proposed implementation pipeline.

When analyzing the effectiveness and quality of each implemented module, special attention should be paid to the zero-shot complaint category classification, rule-based response style selection, and the content of responses generated by LLM. The analysis results combine distributional metrics (counts by category/style), qualitative review samples, and human interpretation of the adequacy of responses. The system automatically generates the text content of the response through the LLM, as shown in Figure 5, and evaluates the effect with the Bilingual Evaluation Complementarity (BLEU) score in Ref. [6], as shown in Figure 6. The research results show the expected results after being implemented according to the theoretical inference.

I've passed this dress by a number of times online, but ordered it on a whim on the mothers day promo. based on the reviews, i ordered my regular size, medium. it fit perfectly. i do think the cut is on the trim side, but it isn't tight or fitted. i probably couldn't eat a big meal in this though as it would probably show in this dress. the fabric is buttery soft! the button placket down the front is very clean and lays flat. there are two side pleats that are a different fabric and are ever so	Size or Fit Problem	Empathetic	Dear [Customer Name],\n\nThank you for taking the time to share your experience with our dress. We're sorry to hear that you feel the cut is on the trim side, and we understand your concern about not being able to eat a big meal comfortably in it.\n\nWe appreciate your feedback and will certainly pass it along to our design team. In the meantime, we'd be more than happy to help you find a size that provides a more comfortable fit for your needs. Please feel free to reach out to our customer service team, and we'll do our best to assist you.\n\nOnce again, thank you for your purchase and for sharing your feedback. We value your business and are committed to ensuring your satisfaction.\n\nBest regards,\n[Your Name]\n[Your Position]\n[Brand Name] Customer Service Team
I bought this online in dark green after trying on the burgundy. i usually wear a medium (5'2", 120) but i looked better in the small in this jacket. the combination of very deep green, black and tan is beautiful. the cut is very flattering with the jacket open or snapped closed. it is nice enough to wear with black pants or skirt but looks just as good with jeans. i am in my 60's and this look is chic and fresh without being too "young" for my age. it's also very well made. i love it.	Other	Professional	Dear Customer,\n\nThank you for taking the time to share your positive feedback about our Dark Green Jacket. We're thrilled to hear that you found the perfect fit in the Small size and that you appreciate the flattering cut and versatile style of the jacket. It's always a pleasure to know that our products can help our customers feel confident and chic.\n\nWe're also glad to hear that you find the combination of deep green, black, and tan to be beautiful. It's always our aim to create pieces that are not only stylish but also timeless.\n\nWe're also proud of the quality of our products, and we're glad to hear that our Dark Green Jacket is well-made.\n\nIf you have any other questions or need assistance with anything else, please don't hesitate to reach out. We're here to help!\n\nBest regards,\n[Your Name]\n[Your Position]\n[Your Brand Name] Customer Service Team

Figure 5. Mapping of customer reviews and generated responses.

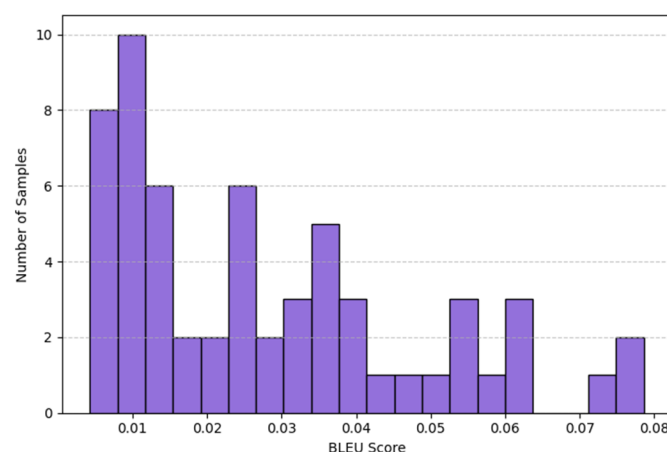


Figure 6. Evaluation results using the BLUE score.

5. Conclusions

This study presents a modular, fully automated framework that integrates sentiment analysis and LLMs to enable real-time, emotionally attuned responses to e-commerce customer reviews. Expanding upon prior sentiment clustering work, two key enhancements are introduced: A: a zero-shot complaint classification module leveraging prompt-based LLM inference to identify dissatisfaction causes without supervised data, and B: a response generation mechanism that adapts tone and content according to emotional polarity and complaint type.

These contributions collectively enable scalable and context-sensitive brand communication. The system not only reduces the manual burden of customer service but also improves consumer trust and brand perception by providing stylistically appropriate, empathy-aware replies. Furthermore, the integration of interpretable complaint categories offers actionable insights into product and service optimization.

As digital commerce continues to expand, the proposed framework offers a viable blueprint for developing emotion-aware AI agents that support automated reputation management, intelligent customer engagement, and data-driven product feedback cycles.

Author Contributions: Conceptualization, C.-H.L. and Y.L.; methodology, C.-H.L.; investigation, T.-S.L.; resources, T.-S.L. and Y.L.; writing—original draft preparation, C.-H.L.; writing—review and editing, C.-H.L. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at https://drive.google.com/drive/folders/10J0PxvD1k-hUHTpmasOGGSKdwchQRK0?usp=drive_link (accessed on 29 January 2026).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lai, C.-H.; Yaonan, H.; Lin, Y.; Liu, T.-S. Research on Designing a Composite Machine Learning System for Real-Time Response to Negative Online Reviews: A Case Study Based on the Negative Reinforcement Model of Digital Marketing. In Proceedings of the 8th International Conference on Knowledge Innovation and Invention, Fukuoka, Japan, 22 August 2025.
2. Zero-Shot Classification of Crisis Tweets Using Instruction-Finetuned Large Language Models. Available online: <https://arxiv.org/abs/2410.00182> (accessed on 20 November 2025).
3. Large Language Models Meet Text-centric Multimodal Sentiment Analysis: A Survey. Available online: <https://arxiv.org/abs/2406.08068> (accessed on 20 November 2025).
4. Together AI. Available online: <https://www.together.ai/> (accessed on 20 November 2025).
5. Women's E-Commerce Clothing Reviews. Available online: <https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews> (accessed on 20 November 2025).
6. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.