systems

MDPI

Article

A Staged Framework for Mining User Requirements from Vehicle Online Reviews Based on Large Language Models and Sentiment Weighting

Zuo You [1], Shenglan Peng [1,*], Wei Zhang [1] and Hao Tan [1,2,3,*]

[1] School of Design, Hunan University, Changsha 410082, China; youzuo@hnu.edu.cn (Z.Y.); wwwwww-z@hnu.edu.cn (W.Z.)
[2] State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha 410082, China
[3] Institute of Culture and Media Computing, Hunan University, Changsha 410082, China
* Correspondence: shenglanp@hnu.edu.cn (S.P.); htan@hnu.edu.cn (H.T.)

Abstract

Large-scale online reviews provide a valuable source of user feedback, yet existing methods still offer limited support for transforming massive review corpora into structured and decision-relevant requirement knowledge. Rule-based and manually intensive approaches are difficult to scale, while direct end-to-end use of large language models often faces challenges in maintaining stable requirement structures and practical large-scale deployment. To address these limitations, this study proposes a staged framework for mining user requirements from vehicle online reviews, with a particular focus on supporting early-stage requirement engineering. The framework uses large language models for requirement taxonomy construction and automatic annotation and transfers large-scale requirement categorization to a BERT classifier to balance semantic capability with deployment efficiency. A mini-batch iterative strategy is further introduced to progressively induce requirement categories from review data rather than fully predefining them in advance. In addition, sentiment weighting is incorporated to prioritize requirement categories and better reflect user pain points in subsequent analysis and decision support. Experiments on 467,962 review texts show that the proposed method outperforms several conventional machine learning and deep learning baselines on the studied dataset. Beyond quantitative evaluation, the study also examines the structural characteristics of online review-based requirement identification and explores the applicability of the framework in other review scenarios. Overall, the proposed framework provides a practical systems-oriented workflow for large-scale requirement analysis and contributes a review-based approach to early-stage requirement engineering, including requirement identification, categorization, prioritization, and interpretation.

Keywords: requirement mining; requirement engineering; large language models; vehicle online reviews; sentiment weighting; requirement prioritization

Academic Editor: Vladimír Bureš

Received: 15 April 2026
Revised: 31 May 2026
Accepted: 2 June 2026
Published: 4 June 2026

1. Introduction

Large-scale online reviews have become an increasingly important source of user-generated feedback in the new energy vehicle market [1]. Compared with traditional surveys or interviews, such reviews capture user experience in a more immediate, diverse, and continuously updated manner [2]. They reflect what users notice in actual use, what

they complain about after repeated interaction, and what they expect when comparing products in a rapidly evolving market [3]. In this sense, online reviews provide more than fragmented expressions of satisfaction or dissatisfaction; they also contain rich signals of user requirements that are potentially valuable for early-stage requirement analysis and decision making. However, such signals are not presented in a stable, explicit, or well-structured form [4,5]. Instead, they are often embedded in colloquial expressions, fragmented experiences, mixed sentiments, and uneven descriptions of usage situations [6]. Consequently, despite the rapid growth of review data, there remains a methodological challenge in organizing these weakly structured signals into coherent and scalable requirement knowledge for subsequent analysis and prioritization.

What online reviews provide is not a ready-made set of requirements, but a large number of weakly structured requirement signals embedded in everyday language [7]. In new energy vehicle-related online reviews, these signals are often scattered across discussions of driving performance, intelligent interaction, cabin comfort, charging experience, and routine use [8]. Some are direct complaints about concrete functions, while others are indirect evaluations of inconvenience, mismatch, or unmet expectations [9]. Turning such signals into usable requirement knowledge therefore requires more than extracting keywords or assigning texts to predefined labels. It requires the formation of requirement structures that can organize dispersed user expressions into interpretable analytical units and relate them to subsequent analysis and decision support [10]. Recent research has increasingly recognized the value of online review data in supporting product analysis, including efforts to quantify user perception and model its evolution to inform design decisions [11]. This issue is closely connected to the concerns of requirement engineering [12], where the central task is not only to gather stakeholder inputs, but also to structure them in a way that supports later interpretation, prioritization, and decision making [13]. Under this condition, review-based requirement mining should not be understood as a narrow text processing task alone. It is also a process of requirement elicitation and structuring under conditions of large scale, semantic variation, and high noise.

A review rarely functions as a neutral and self-contained statement of need. It may be written by a driver, yet refer to the experience of passengers. It may describe an issue in the voice of an individual user, while implicitly expressing family-level concerns about safety, comfort, or space. It may mention a product attribute in purely textual terms, while its meaning depends heavily on a concrete context of use such as commuting, shared travel, thermal discomfort, parking, or charging inconvenience [14,15]. Even when requirement categories can be identified from such reviews, their practical relevance is still shaped by the position they occupy in a broader chain that extends from user expression to product judgment and, eventually, design response [16]. This means that online review-based requirement analysis does not concern only what users say. It also concerns whose concerns are embedded in the expression, under what conditions those concerns emerge, and how far identified requirement signals can travel toward downstream decision making [17]. Seen in this way, the problem is not reducible to standalone text classification. It involves the interpretation of requirement structures, the role of actors and usage conditions in shaping those structures, and the boundary between requirement identification and later innovation translation within a broader socio-technical context [18]. This is precisely where the broader analytical value of the problem lies. The task is not simply to read text more accurately, but to organize weakly structured feedback into a form that remains meaningful within a larger system of product use, evaluation, and improvement.

Existing studies have approached requirement mining from online reviews through rule-based methods, traditional machine learning techniques, and, more recently, large language model-based approaches [19]. Rule-based and manually intensive methods

can be effective in narrowly defined tasks, but they are difficult to maintain and extend when review corpora become large and heterogeneous [20]. Traditional machine learning approaches improve processing efficiency, yet they often depend on predefined categories, feature engineering, or substantial annotation effort [21]. Large language models (LLMs) provide stronger semantic capability, but their direct end-to-end use on full review corpora makes it difficult to maintain stable requirement structures and practical deployment at scale [22]. In other words, existing methods tend to address only part of the problem. Some improve interpretability within constrained settings [23]. Some improve processing efficiency [24]. Some improve semantic understanding [25]. What remains insufficiently addressed is how these capabilities can be organized into a coherent analytical framework that is able to transform massive online review data into structured and decision-relevant requirement knowledge across stages of analysis. This gap is particularly important in new energy vehicle-related review scenarios, where user expressions are not only large in volume, but also diverse in topic, usage context, and practical implication. In addition, prior research has given comparatively limited attention to how identified requirements can be further organized and interpreted in ways that support more decision-relevant analysis [26].

To address this challenge, this study proposes a staged framework for mining user requirements from large-scale new energy vehicle-related online reviews. The framework employs LLMs in requirement taxonomy construction and automatic annotation and transfers large-scale requirement categorization to a BERT classifier, thereby providing a practical way to connect semantic capability with scalable deployment. It further introduces a mini-batch iterative strategy that allows requirement categories to be progressively induced from review data rather than being fully specified in advance by experts. On this basis, requirement categories are prioritized through sentiment weighting so that the mining results can better support product improvement and decision making. In addition, the study further examines structural characteristics and application boundaries of online review-based requirement analysis through qualitative discussion and exploratory examination in household appliance review scenarios. The main contributions of this study are as follows:

1. This study proposes a staged framework for mining user requirements from large-scale new energy vehicle-related online reviews. The framework employs LLMs in requirement taxonomy construction and automatic annotation and transfers large-scale requirement categorization to a BERT classifier, thereby providing a practical way to balance semantic capability with scalable deployment.
2. This study introduces a mini-batch iterative strategy for requirement taxonomy construction, which allows requirement categories to be progressively induced from review data rather than being fully specified in advance by experts.
3. This study extends review-based requirement identification to requirement prioritization through sentiment weighting and further complements the framework with qualitative analysis of structural characteristics and an exploratory examination in household appliance review scenarios.

2. Related Work

Research on mining user requirements from online reviews has developed along several technical directions, including rule-based approaches, traditional machine learning approaches, and, more recently, deep learning and large language model-based approaches. These studies have improved the identification of requirement-related information from review texts, but they also reveal continuing limitations in scalability, requirement representation, and the organization of review-based analysis into a coherent framework.

2.1. Rule-Based Approaches

Early studies on review-based requirement identification often relied on manually designed rules and domain knowledge. These approaches typically combine techniques such as part-of-speech tagging, dependency parsing, and domain-specific lexicons to detect requirement-related expressions from user feedback. Groen et al. [27] proposed a requirement engineering framework based on crowdsourced reviews and identified functional and non-functional requirements through manually defined rules. Maalej and Nabil [28] similarly used keyword matching and syntactic patterns to classify app reviews into categories such as bug reports, feature requests, and user evaluations. The main strength of rule-based approaches lies in their interpretability [29]. Because the extraction logic is explicitly specified, the results are easier to trace and explain within limited and well-defined domains. However, the usefulness of rule-based approaches is closely tied to the stability of linguistic patterns and domain assumptions. When review corpora become large, open, and linguistically diverse, manually constructed rules are difficult to maintain and difficult to generalize. This limitation is particularly evident in vehicle-related online reviews, where requirement expressions are often colloquial, indirect, and context-dependent. Under such conditions, fixed rule libraries struggle to cover the range of possible expressions and require continuous adjustment as products, user concerns, and discourse patterns evolve. In this sense, rule-based studies demonstrated the feasibility of extracting requirement-related signals from reviews, but they provided only limited support for large-scale requirement analysis in open review environments.

2.2. Traditional Machine Learning Approaches

Traditional machine learning approaches shifted review-based requirement analysis from manually specified extraction rules toward statistical learning from textual features. This transition made it possible to process larger review corpora more efficiently and to identify requirement-related patterns without relying entirely on explicit rule construction. In this line of work, review texts are typically represented through techniques such as TF-IDF, n-grams, or topic models, and are then classified with algorithms such as Naive Bayes, Support Vector Machines, or tree-based ensemble methods. Subedi et al. [30] provided an early basis for review text classification using conventional classifiers. In requirement mining tasks, LDA has been widely used to discover latent topics from review corpora [31], while SVM and related methods have been applied to assign reviews to predefined requirement categories [32]. More recent ensemble methods have further improved classification efficiency in some settings [33]. Traditional machine learning methods improve scalability and reduce the need for manually specified extraction rules. They are therefore more suitable for processing larger corpora. Nevertheless, their ability to support requirement analysis remains constrained when requirement information is represented. Topic models can identify clusters of co-occurring terms, but the resulting topics often require substantial human interpretation before they can be treated as requirement structures. Supervised classifiers, in turn, depend on predefined labels and manually designed features, which limits their ability to capture more complex semantics in colloquial review texts [34]. As a result, traditional machine learning moved review analysis toward larger-scale processing, but it did not fully resolve how requirement structures themselves could be formed from large volumes of weakly structured user expression. This limitation is especially important in online review settings where requirement-related signals are dispersed across informal language, mixed concerns, and varying situations of use.

2.3. Deep Learning and LLMs

Deep learning approaches further advanced review analysis by learning semantic representations directly from text. Models such as TextCNN improved sentence classification performance by capturing local textual patterns through neural architectures [35], while pre-trained language models such as BERT introduced stronger contextual understanding and substantially improved short-text classification [36]. In parallel, recent studies have integrated topic-related text analysis, sentiment analysis, and time-series trend modeling to quantify multidimensional user perception and support value-oriented design [11]. Such approaches contribute valuable perception-level measurement instruments, including composite indicators that capture how user evaluations evolve over time. More recently, LLMs have attracted growing attention because of their strong semantic abstraction, generative ability, and effectiveness in zero-shot or few-shot settings. Existing studies have shown that these models can support tasks such as information extraction, annotation assistance, and requirement-related text analysis with much stronger language understanding than earlier approaches [37,38]. Even so, stronger semantic capability does not by itself solve the larger analytical problem posed by massive review corpora. Direct end-to-end use of LLMs on full review datasets remains challenging because of inference cost, context window limitations, and output variability across prompts and input batches [39].

Even so, stronger semantic capability does not by itself provide a coherent analytical workflow for large-scale requirement mining. Direct end-to-end use of LLMs on full review corpora remains difficult to organize in a stable, scalable, and practically deployable manners [40]. This difficulty is associated not only with inference cost and context limitations, but also with the challenge of maintaining stable requirement structures, generating consistent annotations, and supporting downstream analysis across large and heterogeneous datasets [41]. Existing work has therefore improved the semantic recognition of review texts, but it has not fully resolved how requirement taxonomy formation, annotation, large-scale categorization, and subsequent interpretation should be organized as a connected analytical process.

Overall, existing studies have substantially advanced the identification of requirement-related information from online reviews. Yet they still provide limited support for organizing large-scale review data into a coherent and scalable requirement analysis framework. This limitation becomes especially important when mined requirements are expected to support not only initial recognition, but also subsequent structuring, prioritization, and interpretation. The present study addresses this gap through a staged framework for review-based requirement mining and further reflects on the structural characteristics and application boundaries of mined requirements.

3. Methodology

This study proposes a staged framework for mining user requirements from large-scale new energy vehicle-related online reviews. The framework is designed to address the methodological challenge identified in the previous sections, namely, how to transform massive and weakly structured review corpora into structured and decision-relevant requirement knowledge in a coherent and scalable manner. To this end, the overall method is organized into four linked stages. The first stage focuses on data collection and preprocessing to construct a reliable review corpus for subsequent analysis. The second stage constructs a requirement taxonomy through a mini-batch iterative strategy supported by LLMs, allowing requirement categories to be progressively induced from review data. The third stage combines LLM-assisted annotation with BERT classification so that semantic capability can be retained while large-scale categorization remains practically deployable. The fourth stage extends requirement identification to sentiment-weighted prioritization,

thereby organizing identified requirement categories into a more decision-relevant form. Figure 1 presents the overall structure of the proposed framework.

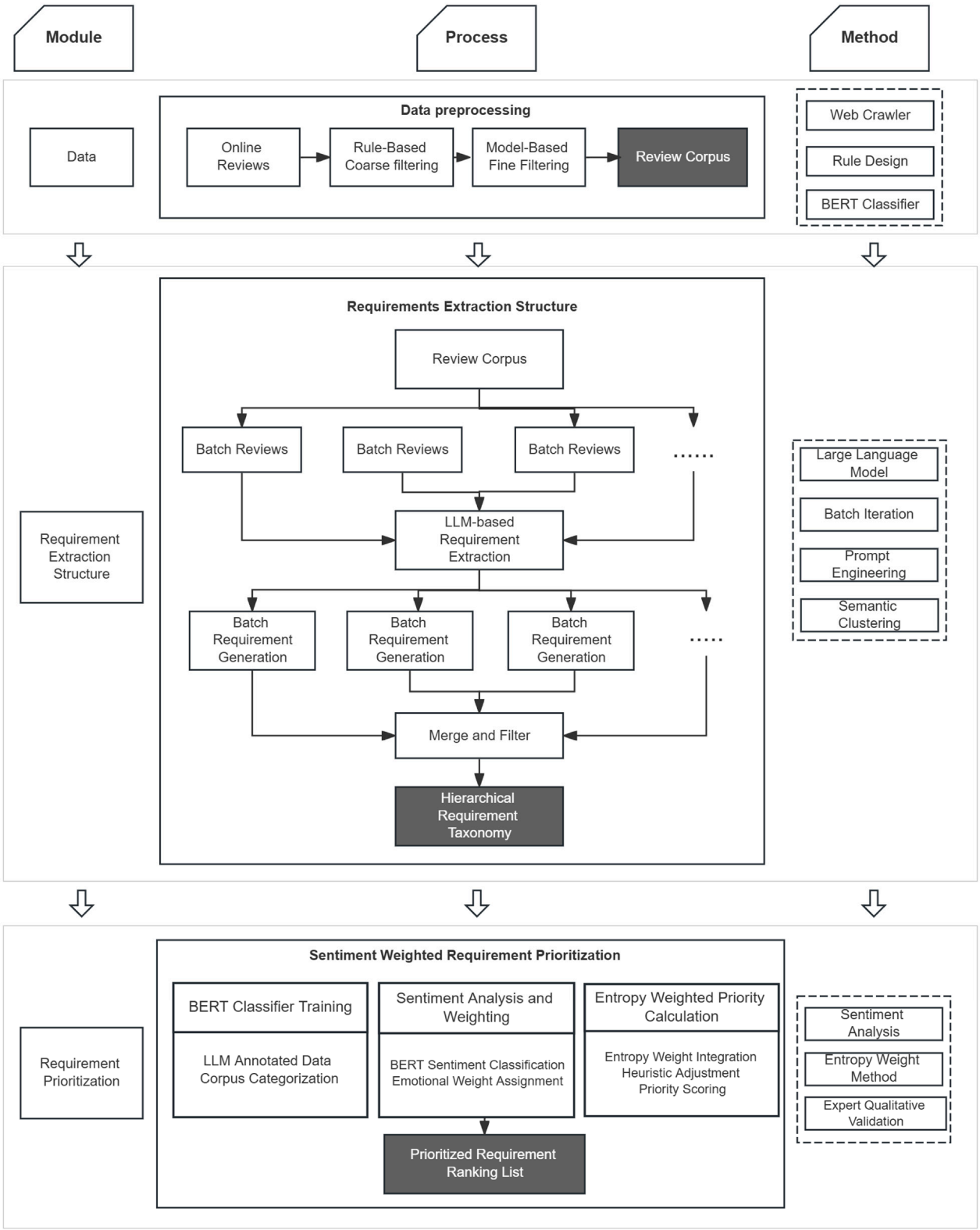

Figure 1. Overall framework of requirement mining method based on LLMs and sentiment weighting.

3.1. Dataset and Preprocessing

The dataset used in this study consists of large-scale new energy vehicle-related online reviews collected from two major automobile discussion platforms in mainland China,

Autohome and Dongchedi. These reviews provide a rich source of user-generated feedback on vehicle performance, intelligent interaction, cabin comfort, charging experience, and everyday use. At the same time, like many naturally occurring online corpora, the raw data are characterized by high volume, heterogeneous expression, and substantial noise. In this study, preprocessing is organized as a two-stage procedure that combines rule-based coarse filtering with model-based fine filtering, with the aim of improving corpus quality while retaining requirement-relevant user expression for later analysis. A total of 467,962 raw review entries were collected from the two platforms. These reviews cover mainstream in-market new energy vehicle models and span the period from June 2024 to June 2025.

In the first stage, rule-based coarse filtering was applied to eliminate invalid texts on the basis of surface-level characteristics. Reviews with highly similar content posted by the same user within a short time interval of less than 24 h were treated as duplicates, and only the earliest entry was retained. Text similarity was measured using an edit distance-based method, with a similarity score above 0.9 treated as duplicate. In addition, three rule-based filters were introduced. First, reviews with empty content or fewer than three valid characters were removed as null or extremely short texts. Second, reviews containing promotional expressions such as advertisement, rebate, fake order, or external contact information were removed through a predefined dictionary of 127 advertising-related keywords. Third, reviews in which garbled characters or special symbols accounted for more than 50 percent of the content were treated as invalid and removed.

Rule-based filtering remains insufficient for identifying texts that are structurally valid but semantically uninformative for requirement analysis. To address this issue, the second stage introduced model-based fine filtering at the semantic level. A manually annotated dataset of 2593 review texts was constructed for this purpose, including 1520 valid entries and 1073 invalid entries. These annotated data were used to train a binary classifier for distinguishing valid requirement-related reviews from invalid or semantically uninformative texts. The trained classifier was then applied to the remaining corpus for fine filtering. After the preprocessing procedure, approximately 35,745 valid reviews were retained for subsequent requirement mining. The binary classifier used here is built upon the same pre-trained BERT backbone as the requirement classifier and is fine-tuned under the same configuration; implementation details are reported in Section 3.3.

3.2. Requirement Label System Construction Based on Mini-Batch Iteration and LLMs

The construction of a requirement label system is a foundational step in requirement mining, and its quality directly affects the accuracy of subsequent classification and priority ranking. Traditional methods rely on domain experts to manually summarize requirement categories, which is costly and difficult to scale. This study proposes an automated method for constructing a requirement label system based on LLMs. By combining a mini-batch iterative strategy with prompt engineering, this approach enables automatic generation and dynamic updating of requirement labels.

3.2.1. Heuristic Mini-Batch Iterative Strategy

Directly applying LLMs to the entire review corpus for requirement taxonomy construction is difficult for two reasons. First, full corpus input is constrained by context length and inference cost, which makes it impractical to process hundreds of thousands of review texts in a single analytical step. Second, requirement taxonomy construction does not depend only on local sentence-level understanding. It also requires the progressive aggregation of dispersed requirement signals into a sufficiently stable and interpretable category structure. Processing reviews one by one is therefore inefficient, while processing the entire corpus at once is difficult to organize and difficult to control.

To address these challenges, this study draws on the mini-batch optimization idea from stochastic gradient descent (SGD) in machine learning and proposes a mini-batch iterative requirement mining method [42]. The core idea is to randomly sample a subset of cleaned reviews, divide it evenly into multiple batches, extract requirements from each batch using the semantic understanding capabilities of LLMs, and merge them into a global requirement set. This process is repeated across multiple iterations so that requirement categories can be progressively induced from review data rather than being fully specified in advance. In this way, taxonomy construction is treated as an incremental analytical process rather than a one-step generation task. The computation process is as follows:

Sampling of the original dataset: calculate the number of batches per iteration and obtain the sampled subset.

Let the full cleaned dataset be $D = \{c_1, c_2, \ldots c_N\}$ where N is the total number of cleaned reviews. Randomly sample a subset $D_s \subset D$, and let each iteration use a sample size of B. The number of batches per iteration is then:

$$K = \frac{|D_s|}{B} \tag{1}$$

Requirement extraction for a single batch: input the k-th batch of reviews into the LLM to generate a local requirement set:

$$r_k = f(B_k) \tag{2}$$

Global merging and deduplication: merge the local requirements with the temporary global requirement set from the current iteration, removing redundant entries:

$$R_t^k = Unique(R_t^{k-1} \cup r_k) \tag{3}$$

where $R_t^0 = R_{t-1}$, the requirement set from the previous iteration.

Multi-iteration bias mitigation: to reduce bias caused by random batch division, the above single-iteration process is repeated for T rounds. Before each iteration, the current global requirement set and batch data are shuffled, and steps 1 and 2 are repeated to obtain the final requirement set R. In this process, Unique is the requirement merging strategy that consolidates semantically identical labels. Some labels may differ substantially in wording but express similar underlying requirements (e.g., "hard seats" and "poor seating comfort"), which traditional keyword-based or text similarity approaches struggle to handle. This study uses prompt design with LLM capabilities to perform effective deduplication.

This study adopts Qwen3-max as the LLM; the batch size is 25. The prompt for the model is as follows:

"You are a professional automotive user research expert, skilled in mining users' core demands from car review abstracts and generating structured demand tags. Please output the results strictly in JSON format without any additional explanatory text.

Task: extract users' core demands from {product_name} user reviews and generate 20–25 demand tags. Each tag must include [Tag Name + Tag Description (explaining user demands and applicable scenarios corresponding to the requirement) + 1 typical review example].

Scenario Instructions: independently identify and extract key dimensions from reviews (e.g., quality, performance, service, price, appearance, functions, etc.), with no restrictions on fixed dimensions. Distinguish between "explicit complaints" (clearly expressed user dissatisfaction) and "implicit expectations" (unspoken but implied user needs). Ensure tags are concise and interpretable and avoid ambiguous expressions.

Format Requirement: each tag must contain [Name + Description + Typical Example].

Please output in the following JSON format: {"Product": "{product_name}", "Total Tags": 20, "Tag List": [{"Tag Name": "Tag 1", "Tag Description": "Describe the user demands and applicable scenarios corresponding to this requirement", "Typical Example": "Real citation extracted from reviews"}]}.

Review Samples: {reviews_text}".

This strategy has three analytical advantages. First, it reduces the practical burden of direct full corpus inference and makes taxonomy construction feasible under large-scale review conditions. Second, it allows requirement structures to emerge progressively from multiple localized observations rather than forcing the corpus into a fixed category system at the outset. Third, the iterative merging process helps mitigate the influence of isolated batches and supports the gradual stabilization of the requirement taxonomy. The mini-batch iterative strategy serves as a mechanism for linking semantic capability to requirement structure formation in large-scale review analysis.

3.2.2. Assessment of Requirement Taxonomy Quality

To improve the scientific rigor, robustness, and interpretability of the requirement taxonomy constructed through LLMs, this study evaluates the resulting category structure from three perspectives, namely, evolutionary convergence, semantic alignment, and expert review [43]. The purpose of this assessment is to examine whether the induced taxonomy is sufficiently stable and interpretable for subsequent classification and prioritization, rather than to treat it as a complete or exhaustive representation of all possible requirements.

To reduce stochastic interference during mini-batch iterations and examine whether the taxonomy gradually approaches a more stable structure, this study monitored the total number of newly added categories, merged categories, and retained requirement categories across successive iterations. The observed results show that the growth of requirement categories is more pronounced in the early iterations, while the marginal increase slows in later rounds and the overall scale gradually stabilizes. This pattern suggests that the mini-batch iterative strategy can progressively approximate a relatively stable requirement structure while reducing the influence of batch-specific variation.

Addressing the common phenomenon of heterogeneous expressions with homogeneous intent in online reviews, such as "stiff seats" and "poor comfort", this study uses the LLM to assist semantic consolidation during taxonomy construction. Three criteria, namely semantic equivalence, granularity consistency, and contextual understanding, are embedded into the prompt design so that requirement labels can be merged at a relatively consistent level of abstraction. Through this semantic alignment process, the resulting labeling system covers core dimensions such as dynamic performance, driving comfort, spatial layout, intelligent interaction, and energy consumption. This helps improve the logical transparency and interpretability of the labeling system for later classification and analysis.

To systematically eliminate potential self-validation bias inherent in automated processes, an expert review mechanism was introduced to form a quality closed loop. A random sample of 300 original reviews was selected for validation. Three researchers with product-engineering backgrounds independently assessed the labeling system. Each reviewer worked independently and judged for every review whether the assigned label was appropriate. The resulting agreement was substantial (Fleiss' $\kappa = 0.71$). The results validated the logical robustness of the framework. Minor semantic overlaps occurred at the boundaries between "comfort" and "functional configuration".

Taken together, these three forms of assessment indicate that the induced requirement taxonomy is sufficiently stable, semantically coherent, and interpretable to support the subsequent stages of annotation, categorization, and prioritization.

3.3. LLM-Assisted Annotation and Classification

The requirement taxonomy constructed in Section 3.2 provides the basis for subsequent large-scale requirement categorization, including dimensions such as dynamic performance, comfort configuration, intelligent interaction, energy consumption, and spatial layout. However, directly applying LLMs to label the entire review corpus would introduce substantial inference overheads, making such a strategy difficult to sustain at scale. For this reason, this study adopts a staged strategy that combines LLM-assisted annotation with BERT classification. The purpose is to retain semantic capability in the annotation stage while enabling more efficient corpus-level categorization in the subsequent stage.

This study uses LLMs to annotate a subset of review data and generate training labels for supervised classification. The annotated subset is then used to fine-tune a BERT classifier for large-scale requirement categorization. We fine-tune a BERT model [44] as the requirement classifier. The same pre-trained BERT backbone is also used for the binary validity classifier in Section 3.1; the two tasks differ only in their training data and label sets, while sharing an identical model architecture and training configuration. To ensure the reproducibility of our experimental results, we fixed the random seed to 42 for all stochastic processes, including data splitting, weight initialization, and model training. The BERT-based models were fine-tuned with a batch size of 16 and a learning rate of 2×10^{-5} and trained for 5 epochs to ensure full convergence. To improve the reliability of this stage, a portion of the training data was manually annotated and compared with the LLM-generated labels as a supplementary check on annotation quality. The comparison suggests that the LLM-generated annotations provide a workable basis for downstream classifier training. This strategy reduces dependence on direct full corpus LLM inference while preserving the semantic advantages of LLM-assisted labeling. Through this stage, the semantically structured taxonomy is transformed into a scalable classification model, making it possible to categorize the full review corpus in the following prioritization stage.

3.4. Sentiment-Weighted Requirement Prioritization

It is necessary to identify how these categories may be organized in terms of relative salience for subsequent analysis and decision support. Sentiment information in user reviews provides an additional dimension for this purpose. Negative sentiment often reflects dissatisfaction, unmet expectations, or user pain points, while positive sentiment reflects more satisfactory aspects of product experience. For this reason, this study introduces a sentiment-weighted requirement prioritization method that combines requirement labeling with sentiment analysis in order to organize identified requirement categories into a more decision-relevant order.

The BERT model is first used to estimate the sentiment polarity of each review. For sentiment polarity estimation, we adopt Erlangshen-Roberta-110M-Sentiment, a RoBERTa-based model from the BERT family released by IDEA CCNL via Hugging Face. This model is used off the shelf without any task-specific fine-tuning on our data [45]. To derive a more systematic weighting scheme and reduce purely ad hoc assignment, an entropy-based weighting method is then combined with business heuristic adjustment to calculate the relative weights of negative, positive, and neutral sentiments. The basic idea is that different sentiment types may exhibit different distributions across requirement categories. A sentiment type with a more uneven distribution across categories may contribute more strongly to differentiating requirement salience and thus may be assigned a higher weight in the prioritization stage.

Assume there are C categories in total. For each review, the sentiment polarity is determined using the BERT model. The counts $N_{neg}(C)$, $\mathrm{N_{neu}(C)}$, $N_{pos}(C)$ are then statisti-

cally obtained, representing the number of negative, neutral, and positive reviews under requirement category c, respectively.

Step 1: Calculate the proportion of sentiments within each requirement category. For requirement category c, the proportion of the s-th type of sentiment $s \in \{neg, neu, pos\}$ is calculated as follows:

$$p_s(c) = \frac{N_s(c)}{\sum_{s \in (neg,neu,pos)} N_s(c)} \tag{4}$$

Step 2: Calculate the normalized weight of sentiments across all requirements.

For a specific sentiment type s, first calculate its sum across all requirement categories:

$$T_s = \sum_{c=1}^{c} N_s(c) \tag{5}$$

Then, calculate the normalized weight of this sentiment for category c:

$$q_s(c) = \frac{N_s(c)}{T_s} \tag{6}$$

Using the normalized weights, the information entropy for sentiment s is calculated as follows:

$$E_s = -\sum_{c=1}^{C} q_s(c) \times ln(q_s(c)) \tag{7}$$

A smaller entropy value indicates a greater distributional variance of that sentiment across different requirements, signifying a stronger capability to differentiate priorities. The formula for the initial weight is derived as follows:

$$w_s^0 = 1 - E_s \tag{8}$$

Since the entropy weight method reflects the distributional variance of different sentiment inclinations across requirements but lacks domain-specific judgment, business heuristic weights are incorporated into the calculation. The final weight coefficient is determined as follows:

$$w_s = p \times w_s^0 + (1 - p) \times w_s^1 \tag{9}$$

where p represents the proportion of the entropy weight, and w_s^1 is the business weight coefficient for the current sentiment. The composite weight w_s obtained from Equation (9) reflects the combined effect of distributional informativeness and business relevance prior for each sentiment type. However, this composite weight is expressed on an absolute scale that is not directly comparable across sentiment types and may, under certain distributions, take negative values. To convert these absolute weights into a unified, dimensionally consistent multiplier that can be used in the linear aggregation of category-level review counts, a relative normalization is introduced. We adopt the neutral sentiment as the baseline, since neutral mentions correspond to instances in which a requirement is discussed but no clear evaluative orientation is expressed, and therefore provide the most natural zero-reference point against which negative and positive sentiments can be measured. The absolute value operator is applied to ensure that the resulting normalized weights remain positive in all cases, which is required for their use as multiplicative coefficients on review counts. The normalized sentiment weights are then defined as follows:

$$\tilde{w}_{neg} = \frac{|w_{neg}|}{|w_{neu}|} \tag{10}$$

$$\tilde{w}_{neu} = 1.0 \tag{11}$$

$$\tilde{w}_{pos} = \frac{|w_{pos}|}{|w_{neu}|} \tag{12}$$

Requirement Prioritization and Ranking: the classifier trained in Section 3.3 is applied to the entire automobile review dataset for requirement category labeling, while the sentiment analysis model identifies sentiment inclinations. For each requirement category c, the priority score $P(c)$ is calculated as follows:

$$P(c) = \tilde{w}\_neg \times N\_neg(c) + \tilde{w}\_neu \times N\_neu(c) + \tilde{w}\_pos \times N\_pos(c) \tag{13}$$

In Equation (13), the coefficients $\tilde{w}\_neg$, $\tilde{w}\_neu$, and $\tilde{w}\_pos$ are exactly the normalized sentiment weights defined in Equations (10)–(12), and $N\_neg(c)$, $N\_neu(c)$, and $N\_pos(c)$ are the per-category counts described above. The priority score $P(c)$ thus reflects two factors simultaneously: the overall volume of user discussion in category c, and the sentiment composition of that discussion as weighted by the cross-category informativeness of each sentiment type. Because the entropy-based weights make sentiment types with more discriminative distributions across categories carry larger weights, and because business heuristics further adjust these weights to reflect the differential importance of negative, neutral, and positive feedback for product improvement, $P(c)$ tends to be higher for categories that combine high discussion volume with sentiment compositions that are informative for requirement-related decision making. Ranking all requirement categories in descending order of $P(c)$ yields the final requirement priority sequence.

By ranking all requirement categories according to their priority scores, the framework produces a more interpretable ordering of requirement salience. This procedure takes into account both the frequency of requirement-related discussion and the distribution of user sentiment across categories. It offers an analytical basis for requirement interpretation and product-related decision support. This prioritization stage converts categorized requirement signals into an ordered decision support output, which is evaluated and interpreted in the Section 4.

4. Results

4.1. Performance Comparison of Review Validity Classification

To evaluate the performance of different models on the binary classification task of review validity, this study used the manually annotated dataset of 2593 automobile reviews described in Section 3.1, comprising 1520 valid entries and 1073 invalid entries. The dataset was randomly split the dataset into training, validation, and test sets at a ratio of 7:2:1. Subsequently, the performance of several distinct models was compared. The experimental results are presented in Table 1, and the schematic diagram of model performance comparison is shown in Figure 2.

Table 1. Performance comparison of different models in review validity classification.

Model	Accuracy (%)
Qwen3-embedding + SVM	83.0
Qwen3-embedding + LightGBM	81.5
Qwen3-embedding + XGBOOST	82.5
TextCNN	79.8
BERT (Proposed)	83.5

Comparative experimental results indicate that the BERT-based approach significantly outperforms other classification methods that integrate word embedding features with traditional machine learning or Convolutional Neural Networks (CNNs). This outcome

validates the inherent advantages of the Transformer architecture in capturing contextual semantic associations within review texts. For the binary classification task of review validity, it is essential for the model to effectively represent inter-textual semantic dependencies—a requirement that BERT fulfills more proficiently, thereby exhibiting superior classification performance in this task.

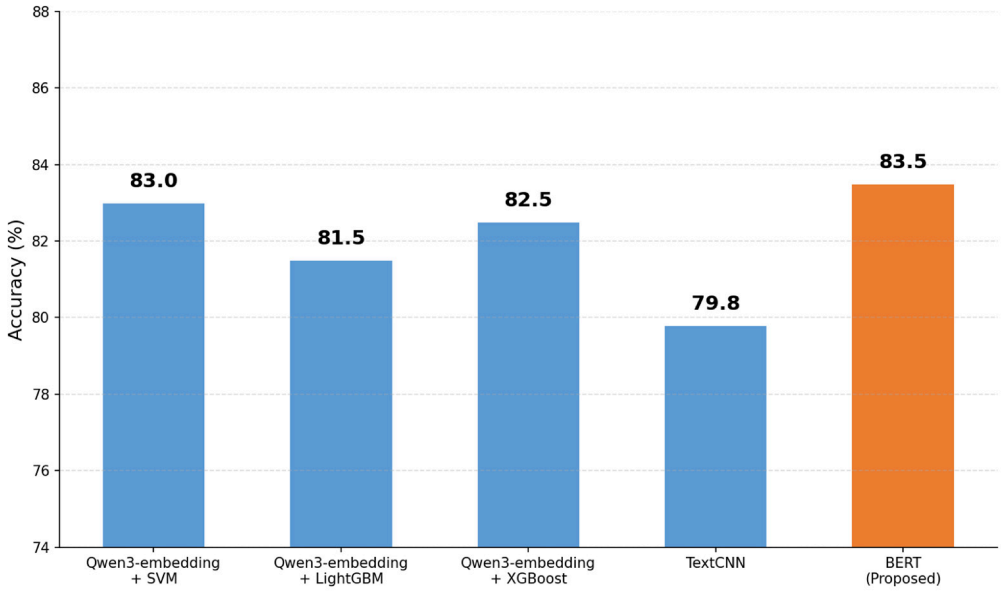

Figure 2. Schematic diagram of model performance comparison.

To further verify the efficacy of incorporating a text classification model for fine-grained screening, this study compares the performance of a rule-only filtering approach against the proposed hybrid method, which combines rule-based coarse screening with model-based fine screening. The experimental results are presented in Table 2.

Table 2. Comparison of data filtering efficiency between different methods.

Filtering Method	Number of Filtered Records
Rule-only Filtering (Baseline)	544
Proposed Hybrid Method (Rule-based Coarse + Model-based Fine Screening)	2862

The experimental results clearly demonstrate that the model-based fine screening layer identifies a significantly larger volume of invalid data than the rule-only approach. A qualitative analysis of the raw crawled comments reveals the prevalence of entries such as, "Picked up my new series on the first day of 2025!" While such comments are structurally valid, they are semantically vacuous for the purpose of requirement mining as they lack substantive feedback on product attributes or user pain points.

Because these types of entries often contain domain-specific keywords and follow standard linguistic patterns, they cannot be effectively filtered through simple rule-based methods or keyword matching. This finding underscores the indispensability of employing a deep learning-based text classification model for fine screening. By capturing high-level semantic context rather than just surface-level features, the model ensures the extraction of a high-quality, task-relevant dataset, which is a prerequisite for accurate downstream requirement analysis.

4.2. Analysis of Requirement Extraction Results

By employing the proposed method based on mini-batch iterations and LLMs to process the automobile review data, a comprehensive requirement taxonomy was identified. These requirements are categorized into primary and secondary dimensions and ranked according to the volume of corresponding reviews, as illustrated in Table 3.

Table 3. Hierarchical taxonomy of user requirements and distribution.

Primary Category	Sub-Dimension	Requirement Description	Typical Comments
Appearance and Interior	Exterior Design	The user has high expectations for the appearance and hopes that the vehicle's design is both fashionable and distinctive, in line with their personal esthetic preferences.	It features a bold, imposing front fascia, paired with a sleek, rounded body and a full, robust rear end. The overall design is seamless and natural, exuding both luxury and sportiness.
	Quality of Interior Materials	Interior material quality: users expect to use high-end and high-quality materials, including the audio system, multiple driving options and high-tech instrument panels, which are both esthetically pleasing and practical. It is suitable for consumers who pursue high-quality interiors.	In pursuit of quality and comfort? Here is your ideal choice.
Dynamic Performance	Handling Performance	Users pursue strong acceleration and excellent handling. Ideal for those who love driving pleasure.	For young people, if a car wants to win them over, it's got to have sporty tuning and a great acceleration experience.
	Power Output	Users have high demands for vehicle acceleration and driving experience, especially on urban expressways and highways.	The roar of the four-cylinder engine resonates deeply with the soul. Follow your inner voice, start the engine, and set off now!
	Engine Performance	User requirements for engine cold-start performance and noise levels.	The new Series makes a single clunk from the chassis when started cold.
Intelligence and Technology	Smart Configurations	Users show great interest in the infotainment system and driver-assistance technologies. They expect a more convenient and safe driving experience.	Equipped with the iDrive intelligent system. Three-zone air conditioner control can be activated remotely via the mobile app.
	Driver-Assistance Systems	Users expect advanced intelligent connectivity, such as driving assistance systems, which improves driving safety and convenience	Which is more practical: professional driver-assistance systems or HUD?
Economics	Price Rationality	Users hope to buy their ideal car at a better price. which is suitable for budget-conscious consumers who desire luxury brands.	If you are interested in the new series, now is a great time to buy, with big discounts. Affordable price and rich equipment.
	Maintenance Costs	Lower costs for routine maintenance (e.g., brake pad replacement) while ensuring service quality.	Guys, who drives over 20,000 km a year and gets car maintenance twice? The cost is really hard to bear.

Table 3. *Cont.*

Primary Category	Sub-Dimension	Requirement Description	Typical Comments
Comfort	NVH Control	Users want low noise during startup and driving, which improves driving comfort.	Guys, does your car make a loud noise when started in the morning?
	Seat Comfort	Users want seat ventilation. It provides better comfort in hot weather.	Seat ventilation is desired.
	Air Conditioning System	Users want the air conditioning to adjust cabin temperature rapidly.	100 Tips: how to maintain a comfortable cabin temperature!! Set your preferred temperature and air direction.
Spatial Performance	Interior Cabin Space	Users need plenty of interior space, especially for family trips and long drives. They want enough legroom and headroom.	The Xiaomi SU7 is built on Xiaomi's bespoke electric platform with a long wheelbase design. Its roomy interior allows you to stretch out comfortably on every journey!
	Space Utilization Efficiency	Users want a flexible, practical interior for daily use, which is useful for families who need plenty of storage space.	They've updated the door handles, extended the center console, and fitted a longer screen in the 3 Series. But I think the rear space feels smaller.
	Storage Convenience	For daily commuting and long-distance travel, users want more small storage compartments for personal items.	Is being sporty just an excuse for having almost no storage? There are barely any cubbies. I can't even fit my glasses or documents up front.

To validate the accuracy of the mined requirements, this study simultaneously employed the Latent Dirichlet Allocation (LDA) topic modeling method to analyze the same dataset of automobile reviews. The topics and their associated high-frequency keywords identified by the LDA model are presented in Table 4.

Table 4. Topic extraction results based on LDA method.

Subject Number	Keywords
Topic 1	Interior, Handling
Topic 2	Functions, Buttons, System, Telematics/Infotainment
Topic 3	Chassis, Shock Absorption/Suspension
Topic 4	Luxury, Headlights, Exterior, Power, Sportiness, Interior
Topic 5	Rims/Wheels, Tires, Modification
Topic 6	Steering Wheel, Braking
Topic 7	Abnormal Noise, Braking
Topic 8	Maintenance, Engine, Engine Oil, Air Conditioning, Inspection
Topic 9	Price, Car Purchase, Final/On-the-road Price
Topic 10	Esthetics, Color, Exterior

A comparative analysis between the LDA results and the proposed LLM-based method reveals an inherent alignment in their underlying requirement logic. For instance, the keywords "Interior" and "Handling" in Topic 1 of the LDA model correspond directly to the "Interior Material Quality" (under Appearance and Interior) and "Handling Performance" (under Dynamic Performance) categories in Table 3. Similarly, Topic 8 primarily reflects maintenance-related demands, aligning with the "Maintenance Costs" dimension within the "Economics" category.

These overlaps demonstrate that, while both methods capture core user concerns, the proposed method exhibits significant superiority: it autonomously synthesizes and categorizes requirements into a structured, hierarchical taxonomy without the necessity of post hoc expert interpretation. While LDA remains at the level of statistical keyword co-occurrence, requiring manual synthesis to define actionable boundaries, our method achieves end-to-end automated requirement induction, significantly enhancing the efficiency and objectivity of the mining process.

To verify the transferability and robustness of the proposed framework, this study extended the analysis to review datasets from the washing machine and air conditioner domains. The extracted requirement taxonomies for washing machines are presented in Table 5 as an example.

Table 5. Hierarchical taxonomy of user requirements for washing machines.

Primary Category	Sub-Dimension	Requirement Description	Comments Example
After-sales Service	After-sales Support	Users value high-quality and efficient after-sales service, including fast response and professional problem-solving. This applies to all customers in need of after-sales support.	Haier is a trusted big brand—their after-sales service is just top-notch.
	Customer Service	Besides product quality, customers also care a lot about the service attitude throughout the shopping experience.	What is this? I demand an explanation from JD. The drainpipe also seems to be defective. Please have after-sales contact me immediately. The phone line is always busy.
Logistics and Delivery	Quality of Logistics Service	Customers want to receive their orders quickly after placing them, which shows they care deeply about logistics efficiency.	The delivery staff is friendly and punctual.
	Delivery Speed	Users expect sellers to process and ship orders promptly to reduce waiting time. This caters to consumers who need products urgently or are sensitive to delivery speed.	JD Logistics is as fast as ever, and the delivery person was very helpful. 5-star review! I haven't installed it yet; I'll leave an additional review after installation and use.
	Packaging Protection	Customers care about the safety of products during delivery and want to avoid damage caused by improper packaging.	Packaging is intact, and great delivery service.
Noise Performance	Quiet Operation/Silence Effect	Users want the washing machine to run with low noise and not disrupt daily life. This caters to users in quiet living environments or those who are sensitive to noise.	The washing machine is great. The noise is acceptable. Really good!
Ease of Use	Operational Convenience	Easy operation is crucial to improving user experience, including but not limited to an intuitive interface, simple steps, and ease of use.	It's not too big, easy to operate, and the noise level is acceptable. The water inlet seal is excellent with no leaks at all, and the price is reasonable.

Table 5. *Cont.*

Primary Category	Sub-Dimension	Requirement Description	Comments Example
Installation Service	Installation Convenience	Users hope to receive prompt and professional on-site installation support after purchase, which is ideal for busy consumers or those without installation expertise.	Received the washing machine, delivery was fast. JD's service is excellent, after-sales support is great, and installation was prompt.
	Professional Installation Service	Users express satisfaction or dissatisfaction with the quality of on-site installation services, including the attitude and professionalism of the installers.	I am very satisfied with the installation service provided by the after-sales team.
	Ease of Handling/Moving	Given the constraints of high residential floors or other special conditions, users prefer models that are lightweight and easy to move and install.	It's so heavy I can hardly shift it even on the balcony, never mind carrying it all the way up the stairs alone.

4.3. Performance Comparison of Requirement Classification

To evaluate the efficacy of various models in the task of requirement categorization, this study compared the performance of traditional machine learning models, deep learning models, and pre-trained language models. We employed Qwen3-max to annotate 2466 pieces of review data and randomly split the dataset into training, validation, and test sets at a ratio of 7:2:1. The experimental results are summarized in Table 6, and the schematic diagram of model performance comparison is shown in Figure 3.

Table 6. Performance comparison of different models in requirement classification.

Model	Accuracy (%)
Qwen3-embedding + SVM	73.8
Qwen3-embedding + LightGBM	68.6
Qwen3-embedding + XGBOOST	69.4
TextCNN	71.0
BERT (Proposed Method)	78.3

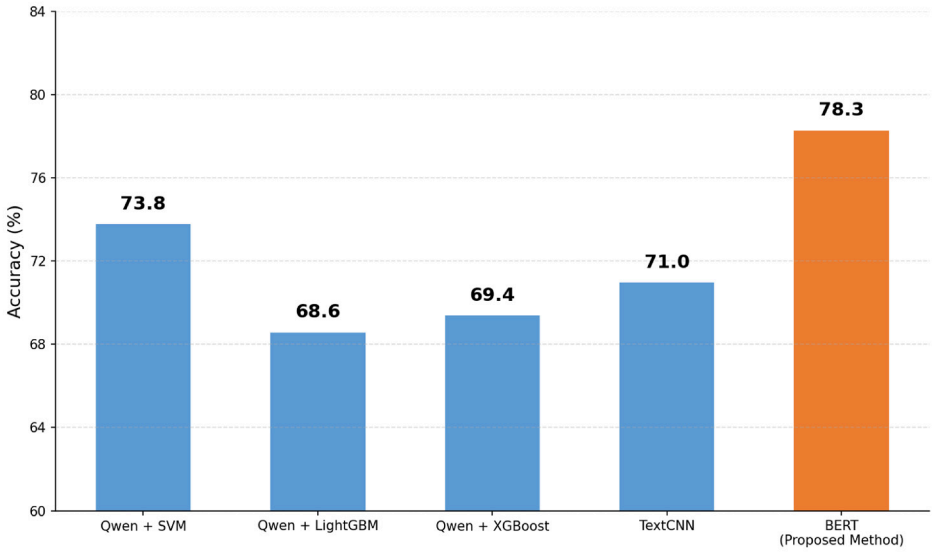

Figure 3. Schematic diagram of requirement classification model performance comparison.

The experimental results indicate that the BERT-based model significantly outperforms all other baselines, achieving a peak accuracy of 78.3%. Compared to traditional machine learning approaches like LightGBM and SVM, BERT demonstrated a substantial performance gain, with accuracy improvements of 9.7% and 4.5%, respectively. Even when compared to advanced deep learning architectures like TextCNN (71.0%), the BERT model maintained a clear advantage. This superiority can be attributed to BERT's ability to model deep bidirectional context, which is essential for disambiguating complex user expressions where the same keyword might imply different requirements depending on the linguistic context. These results confirm that the proposed fine-tuned BERT model provides a highly reliable tool for large-scale automated requirement extraction.

Despite the promising classification performance, roughly one-fifth of review samples still suffer from misclassification errors. This is a meaningful consideration for the practical deployment of the system, and it offers concrete guidance for further optimization. Error case analysis reveals that ambiguous semantic boundaries between similar categories are the major cause of misjudgment. For instance, the comment "The vehicle damage insurance is calculated based on over 290,000 yuan, yet the current new car bare price hardly reaches 270,000 yuan" was mistakenly classified into Economics–Maintenance Costs, while its actual label is Economics–Price Rationality. The sentence simultaneously involves insurance valuation and vehicle pricing information. The model fails to distinguish core semantic focus from incidental descriptive content and confuses associated cost factors with the essential price evaluation intention, thus leading to incorrect category assignment. Such confusion between closely correlated subcategories remains a key obstacle to improving classification accuracy. In practical deployment, this suggests that the framework is best used as a decision support tool, with human verification applied to semantically ambiguous cases.

4.4. Requirement Prioritization and Ranking

In this phase, the requirement categories identified in Section 4.2 were further prioritized by incorporating the sentiment weighting strategy described in Section 3.4. Building upon these procedures, user requirements were re-ranked to reflect both the frequency of review mentions and the affective intensity embedded in textual feedback. As shown in Table 7, the prioritization process integrates requirement extraction, sentiment classification, weight assignment, and weighted aggregation to generate the final requirement ranking. This process-oriented design enables the prioritization results to move beyond frequency-based ordering and to better capture potential user pain points.

To operationalize sentiment weighting, sentiment scores were divided into three categories according to polarity intensity: negative, neutral, and positive. Since the purpose of this study was to identify and prioritize user pain points, negative sentiment was assigned greater analytical importance than neutral and positive sentiment. In addition, an entropy-based weighting approach was employed to incorporate the relative information contribution of each sentiment category into the final prioritization process. Through this combined weighting strategy, negatively oriented user feedback exerted a stronger influence on the final ranking results.

Compared with the quantity-based ranking reported in Section 4.2, the weighted results reveal noticeable shifts in the relative priority of several requirement categories, especially between Intelligence and Technology and Economics. Although the two categories had similar overall review volumes, Intelligence and Technology showed a higher proportion of negative sentiment. Consequently, Intelligence and Technology ranked higher after sentiment weighting was introduced. This finding suggests that the proposed weight-

ing method is effective in identifying requirement categories that are not only frequently mentioned but also more strongly associated with dissatisfaction.

Table 7. Hierarchical taxonomy of automotive user requirements.

Requirement Category	Sub-Requirement Dimension
Appearance and Interior	Exterior Design Quality of Interior Materials
Dynamic Performance	Handling Performance Power Output Engine Performance
Economics	Price Rationality Maintenance Costs Resale Value (Depreciation Rate)
Intelligence and Tech	Smart Configurations Advanced Driver-Assistance Systems (ADASs)
Comfort	NVH (Noise, Vibration, Harshness) Control Seat Comfort Air Conditioning System
Spatial Performance	Interior Cabin Space Space Utilization Efficiency Storage Convenience

5. Discussion

5.1. Implications of the Proposed Framework

The proposed framework contributes a coherent workflow that links requirement taxonomy construction, automatic annotation, large-scale categorization, and prioritization within a connected analytical process. This interpretation is aligned with the view in requirement engineering that requirement analysis involves identification, structuring, and prioritization as connected activities within a larger analytical process [46]. The empirical results provide initial evidence that this organization is workable in practice. The BERT classifier achieved an accuracy of 78.3% on the review dataset, outperforming the four baseline models (SVM: 73.8%, LightGBM: 68.6%, XGBOOST: 69.4%, and TextCNN: 71.0%). Furthermore, the framework retained comparable behavior when applied to washing machine and air conditioner reviews. These results support the central claim that review-based requirement mining can be organized as a connected analytical sequence and that requirement structuring deserves an explicit place in this sequence alongside text classification.

A first implication concerns the functional role of LLMs in large-scale requirement analysis. Their value in the proposed framework lies in upstream tasks that require semantic abstraction and structural formation, particularly requirement taxonomy construction and automatic annotation, which determine the analytical basis for subsequent categorization. The empirical results indicate that the upstream taxonomy was stable enough to support large-scale supervised classification by a lighter model. This suggests that the division of labor between LLMs and BERT is not only a design choice but is also supported by the downstream performance. This observation is consistent with the broader view that large-scale requirement analysis depends on a workflow that can accommodate corpus size, linguistic variation, and downstream analytical demands [47]. The framework therefore clarifies how LLMs can be used more effectively in requirement analysis, with semantic structuring and label formation in the front-end and large-scale execution supported by a lighter model.

A further implication lies in treating requirement taxonomy construction as a central stage of the mining process. In previous studies, requirement categories often serve as predefined analytical containers, and the main task is to assign texts to them [48]. The proposed framework instead allows requirement categories to be formed progressively from the data, before large-scale categorization is carried out. The cross-domain comparison suggests that this data-driven formation step can be adapted to other product domains without requiring a new taxonomy to be hand-designed in advance. This shift gives requirement representation a more central position in the analytical process [49]. Its value lies in linking classification to the formation of a usable and sufficiently stable representation of user requirements derived from colloquial, fragmented, and unevenly distributed user expression.

The prioritization stage further extends the analytical value of the framework. Requirement identification alone does not indicate which issues deserve earlier attention in product improvement [50]. By incorporating sentiment-weighted prioritization, the framework gives identified requirement categories a more decision-relevant form and moves the mining results closer to product judgment and resource allocation. This extension is consistent with the view that prioritization provides an important bridge between requirement identification and subsequent decision making [51]. At the same time, categorization and ranking do not fully capture the complexity of online review-based requirement analysis. The results also point to several structural characteristics and application boundaries that require further discussion to clarify what the framework can reveal and where its explanatory limits remain.

The prioritization results in Section 4 illustrate this decision-relevant form in a concrete way. Intelligence and Technology and Economics had comparable review volumes, yet sentiment weighting moved Intelligence and Technology above Economics because of its higher concentration of negative sentiment. The implication is that the two categories sit at different stages of the relationship between user expectation and product experience. Economics is a relatively stable concern where user expectations and product reality have reached a working balance. Intelligence and Technology is a fast-moving area where new features keep raising expectations faster than experience can mature, and dissatisfaction accumulates around that gap. A volume-based ranking treats the two as equally salient. The sentiment-weighted ranking points to the second one as the place where product iteration and after-sales attention can change the user experience more visibly. The wider value of this stage lies in surfacing categories where dissatisfaction is concentrating relative to how much users talk about them, which is the signal that supports earlier and more targeted action in product improvement.

5.2. Structural Characteristics and Boundary Conditions

The quantitative results establish the feasibility of the proposed framework for large-scale requirement categorization and prioritization. This section discusses how the identified requirement signals should be interpreted under different relational and contextual conditions. The discussion draws on the quantitative findings, supplementary qualitative coding, and exploratory cross-domain comparison, and focuses on three issues, namely, multi-actor requirement expression, scenario dependency, and the boundary.

5.2.1. Multi-Actor Structure of Requirement Expression

The first issue concerns the subject structure underlying requirement expression. The quantitative framework identifies requirement categories from review texts, yet the meaning of these categories depends in part on whose interests and experiences are implicitly carried in the expression of demand. The qualitative coding results suggest that online

reviews in the vehicle context are not restricted to a single driver perspective. A considerable proportion of requirement expressions are associated with front seat passengers, rear seat passengers, children, older family members, or the family as a collective unit of use and decision making. Requirement expression in this setting is therefore socially distributed. Even when a review is written in an individual voice, the concern it conveys may still be shaped by multiple actors within the usage system [52]. Categories such as comfort configuration or spatial layout accordingly carry meanings that extend beyond a single user perspective.

This point adds an important layer to the interpretation of mined requirements. When review-based requirement mining is treated primarily as a text classification task, requirement signals are easily read as if they corresponded to a single and homogeneous user need. A review may record personal experience, while also expressing indirect concerns related to passengers, caregiving needs, or household decision making. Online reviews encode stakeholder relations within a socio-technical setting, and requirement categories retain traces of these relations even after they have been analytically stabilized [53,54]. This observation carries two implications. First, the output of automated requirement identification should be treated as an entry point into a requirement structure that may involve multiple actors and layered interests. Second, the analytical usefulness of the framework depends on whether these categories are interpreted with sufficient attention to the relational conditions from which they arise. The framework is effective for surfacing requirement signals at scale, and its outputs still require interpretation in relation to the multi-actor structure of use.

5.2.2. Scenario Dependency and Context-Sensitive Applicability

A second issue concerns the extent to which requirement interpretation depends on product context. The comparative analysis across washing machines suggests that requirement expressions are not organized in a uniform semantic space. Their meaning is strongly shaped by the physical conditions under which products are used. In the automotive domain, requirement expressions are closely tied to mobile commuting, dynamic interaction, and multi-occupant travel. In household appliance contexts, by contrast, requirement expressions are anchored in relatively stable domestic scenarios such as storage and garment care. This indicates that the same analytical category cannot be assumed to carry the same practical implications across product domains, even when the surface semantics appear comparable. This observation helps clarify the boundary of review-based requirement mining as an analytical framework. The challenge in cross-domain requirement analysis does not arise only from vocabulary variation. More fundamentally, user requirements are embedded in concrete patterns of use. Requirement expressions are therefore shaped not only by what users say, but also by where, how, and under what conditions products are experienced. Requirement meanings in review data are therefore inseparable from contexts of use [55]. A category that appears stable at the textual level may refer to substantially different forms of use, expectation, and failure once it is placed in a different physical setting.

The exploratory comparison with household appliance reviews suggests that the framework retains value beyond the focal new energy vehicle context, but its applicability should be understood in a context-sensitive manner. The framework is able to support requirement identification across domains at the level of analytical procedure, yet the interpretation of resulting categories remains dependent on domain-specific usage scenarios. Transferability is better understood as the capacity of the framework to organize requirement analysis under different contexts, rather than as the invariance of requirement meanings across domains. This distinction is important for avoiding an overly simplified

view of cross-domain extension. It also reinforces a systems-oriented understanding of requirement mining, in which requirement signals gain meaning through their relation to the broader usage environment rather than through textual content alone.

5.2.3. Boundary Between Requirement Identification and Innovation Translation

The proposed framework is effective in identifying requirement categories from large-scale review data, yet a substantial proportion of review expressions remain at the level of symptom description or experience-based complaint, even when they can be reliably mapped to requirement categories. Expressions such as laggy interaction, uncomfortable airflow, or poor preservation performance are informative as indicators of dissatisfaction, but they do not in themselves specify the design variables or engineering parameters required for product intervention. The remaining gap lies in a shift in abstraction. Review texts register situated discomforts and perceived failures, whereas innovation-oriented action requires variables, constraints, and intervention logic. Requirement mining stabilizes the former into analytical categories, while the latter still depends on domain reasoning and interpretive judgment [56].

This boundary can be illustrated more concretely through representative examples that compare raw user expressions, identified requirement categories, and the corresponding design opportunities that still require further interpretation. Table 8 draws on examples from the coded corpus and exploratory cross-domain comparison and shows where semantic translation remains necessary between requirement identification and innovation-oriented action.

Table 8. Examples of the gap between identified requirements and design opportunities.

Domain	Raw Voice of Customer (Raw Voice)	Identified Label (Identified Tag)	Design Opportunity Analysis (Design Opportunity)	Type of Translation Gap
Automotive	"When overtaking on the highway, the acceleration is sluggish; it feels like the power is being suppressed."	Dynamic Performance	Optimize turbocharger engagement timing or recalibrate transmission downshift logic.	Conversion constraints from symptom-level descriptions to engineering parameters.
Automotive	"After installing a child safety seat in the back, the legroom is so cramped that the child kicks print all over the front seat back."	Spatial Layout	Develop kick-resistant backplate materials or increase the travel range of second-row seat rails.	Design deficiencies in multi-agent interaction scenarios and social–spatial complexity.
Refrigerator	"Leafy vegetables yellow and wither only two days after purchase; the preservation effect is poor."	Preservation Performance	Research and develop graded humidity control technology or add physiological lighting with specific wavelengths.	Logical fault line between sensory complaints and technical implementation paths.
Air Conditioner	"Blowing air directly on me at night causes headaches, but turning it off makes it too hot to sleep."	Comfort/Airflow Control	Upgrade "windless" gentle airflow technology or implement infrared thermal imaging-based "avoid-user" functions.	Evolutionary barriers from experiential conflicts to functional innovation synthesis.

As shown in Table 8, user expressions, requirement categories, and design opportunities operate at different levels of abstraction. Review texts often describe symptoms, discomforts, or situated failures. Requirement mining reorganizes these expressions into more stable analytical categories. The further translation into actionable intervention still

depends on engineering parameters, design variables, and domain-specific reasoning. The present results suggest that automated requirement mining is most effective as a front-end layer for requirement discovery, structuring, and preliminary prioritization. Its strength lies in surfacing recurring concerns, organizing dispersed user expression into interpretable categories, and making large-scale requirement signals visible in a form that can support further analysis. Its outputs still require translation before they can function as direct inputs to design reasoning or innovation decision making. The framework therefore occupies a distinct position within a broader requirement and innovation chain. It strengthens early-stage perception and problem framing, while later translation into executable design action continues to depend on human judgment, interdisciplinary knowledge, and domain-specific reasoning.

5.3. Limitation and Future Work

Building on the discussion in Section 5.2, the present framework also has clear limitations that should be acknowledged. The framework has been validated primarily in the context of new energy vehicle reviews, and the examination in household appliance scenarios remains exploratory. Its analytical value therefore lies in showing how requirement signals can be organized across large review corpora, rather than in claiming a domain-independent solution. More importantly, the framework makes requirement structures visible and interpretable, but it does not by itself complete the transition from identified concerns to executable design action. The distance between requirement identification and innovation translation remains a meaningful boundary of the present study. In the same way, sentiment-weighted prioritization should be understood as a way of organizing requirement signals for further judgment, rather than as a final measure of requirement importance.

These boundaries also indicate where the framework can be developed further. One important direction is to examine how the framework performs under a wider range of product contexts, especially where usage scenarios, actor structures, and requirement semantics differ more substantially from those in new energy vehicle reviews. Another direction is to incorporate richer forms of contextual evidence, including multimodal feedback, temporal variation, and more explicit descriptions of use situations, so that requirement interpretation is less dependent on text alone. A further step concerns the connection between requirement mining and design reasoning. Future work may strengthen this connection through human–AI collaboration, expert knowledge integration, and more explicit support for translating requirement categories into design variables and innovation opportunities.

6. Conclusions

This study addresses the challenge of mining user requirements from large-scale online reviews in a way that connects semantic understanding with scalable analysis. Focusing on new energy vehicle review data, it proposes a staged framework that integrates requirement taxonomy construction, automatic annotation, large-scale requirement categorization, and sentiment-weighted prioritization into a coherent analytical process. Within this framework, LLMs are used in upstream tasks that require semantic abstraction, while large-scale categorization is transferred to a BERT classifier. The results suggest that the contribution of the framework lies not only in its quantitative performance, but also in its ability to organize review-based requirement mining as a connected requirement analysis workflow rather than a standalone text classification task.

The study further indicates that the value of the framework should be understood together with its structural characteristics and application boundaries. Online review-based requirement analysis involves multi-actor usage relations, context-dependent requirement

expression, and a remaining gap between identified requirement categories and later innovation translation. These findings suggest that the proposed framework is most effective as a front-end layer for requirement discovery, structuring, and preliminary prioritization. Beyond its theoretical and methodological contributions, the proposed framework offers concrete value for industrial deployment. In the new energy vehicle industry, it can support continuous monitoring and dynamic prioritization of user feedback across dimensions such as economic concerns, dynamic performance, and intelligent technology, helping product teams identify categories in which dissatisfaction is accumulating faster than discussion volume alone would suggest. It can also assist after-sales teams in organizing user complaints by category and severity. The exploratory examination in household appliance reviews further indicates that the framework can be adapted to other consumer product domains in which review-based feedback is abundant.

Overall, this study provides a practical framework for large-scale online review-based requirement analysis and offers a foundation for future work on broader applicability, richer contextual evidence, and closer integration with downstream design reasoning.

Author Contributions: Author Contributions: Conceptualization, S.P. and H.T.; methodology, Z.Y. and S.P.; validation, Z.Y., W.Z. and S.P.; formal analysis, Z.Y. and S.P.; investigation, Z.Y. and W.Z.; data curation, Z.Y. and W.Z.; writing—original draft preparation, Z.Y.; writing—review and editing, S.P. and H.T.; visualization, Z.Y. and W.Z.; supervision, S.P. and H.T.; project administration, H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Program of Hunan Province (Grant No. 2025JJ60818), the Science and Technology Innovation Program of Hunan Province (Grant No. 2025RC1029), and the National Natural Science Foundation of China (Grant No. T2192933).

Data Availability Statement: Data will be made available on request.

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

References

1. Xu, Z.; Wu, Y.; Tang, L.; Gui, S. From user-generated content to quality improvement: A multi-granularity analysis of customer satisfaction and attention in new energy vehicles using deep learning. *Comput. Ind.* **2025**, *173*, 104380. [CrossRef]
2. Zhang, Y.; Guo, W.; Chang, Z.; Ma, J.; Fu, Z.; Wang, L.; Shao, H. User requirement modeling and evolutionary analysis based on review data: Supporting the design upgrade of product attributes. *Adv. Eng. Inform.* **2024**, *62*, 102861. [CrossRef]
3. Sun, H.; Guo, W.; Shao, H.; Rong, B. Dynamical mining of ever-changing user requirements: A product design and improvement perspective. *Adv. Eng. Inform.* **2020**, *46*, 101174. [CrossRef]
4. Porayska-Pomsta, K.; Mellish, C. Modelling human tutors' feedback to inform natural language interfaces for learning. *Int. J. Hum.-Comput. Stud.* **2013**, *71*, 703–724. [CrossRef]
5. Rapp, A.; Curti, L.; Boldi, A. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum.-Comput. Stud.* **2021**, *151*, 102630. [CrossRef]
6. Xiang, H.; Li, W.; Hong, Y.; Li, C. A novel requirement elicitation and evaluation framework for product-service systems based on contextual matching and hybrid decision-making. *Comput. Ind. Eng.* **2024**, *194*, 110391. [CrossRef]
7. Bakar, N.H.; Kasirun, Z.M.; Salleh, N.; Jalab, H.A. Extracting features from online software reviews to aid requirements reuse. *Appl. Soft Comput.* **2016**, *49*, 1297–1315. [CrossRef]
8. Jiang, Z.; Liu, N.; Zhang, X.; Chu, Y.; Zhang, L. Proactive social learning and green product consumption: Evidence from new energy vehicle sales in China. *Technol. Forecast. Soc. Change* **2025**, *219*, 124232. [CrossRef]
9. Yang, Y.; Li, Q.; Li, C.; Qin, Q. User requirements analysis of new energy vehicles based on improved Kano model. *Energy* **2024**, *309*, 133134. [CrossRef]
10. Cong, Y.; Yu, S.; Chu, J.; Su, Z.; Huang, Y.; Li, F. A small sample data-driven method: User needs elicitation from online reviews in new product iteration. *Adv. Eng. Inform.* **2023**, *56*, 101953. [CrossRef]
11. Peng, S.; Ye, D.; Tan, H. Modeling User Requirement for Value-Oriented Design: A Multi-Dimensional Perception Evidence from the Automobile Market. *Systems* **2026**, *14*, 251. [CrossRef]
12. Zhao, Z.; Zhou, Z. A systematic literature review on model-based requirements engineering. *J. Syst. Softw.* **2026**, *237*, 112836. [CrossRef]

13. Wang, H.; Xin, Y.-J.; Deveci, M.; Pedrycz, W.; Wang, Z.; Chen, Z.-S. Leveraging online reviews and expert opinions for electric vehicle type prioritization. *Comput. Ind. Eng.* **2024**, *197*, 110579. [CrossRef]
14. Peng, C.; Carlowitz, S.; Madigan, R.; Marberger, C.; Lee, J.D.; Krems, J.; Beggiato, M.; Romano, R.; Wei, C.; Wooldridge, E.; et al. Conceptualising user comfort in automated driving: Findings from an expert group workshop. *Transp. Res. Interdiscip. Perspect.* **2024**, *24*, 101070. [CrossRef]
15. Lee, S.C.; Nadri, C.; Sanghavi, H.; Jeon, M. Eliciting User Needs and Design Requirements for User Experience in Fully Automated Vehicles. *Int. J. Hum.-Comput. Interact.* **2022**, *38*, 227–239. [CrossRef]
16. Zhuang, W.; Wang, J.; Wu, X.; Chen, C.; Xiao, W.; Jiang, C.; Yao, J.; Bao, J.; Li, X. Parsing and prioritizing users' evolving requirements with large language models: A case study in personalized rehabilitation assistive devices design. *Comput. Ind. Eng.* **2026**, *214*, 111878. [CrossRef]
17. Huang, S.; Desmet, P.M.A.; Mugge, R. Introducing the Fundamental User Needs (FUN) Scales: Assessing Need Satisfaction and Frustration in Design-Mediated Interactions. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 11996–12013. [CrossRef]
18. Von Hippel, E. Lead users: A source of novel product concepts. *Manag. Sci.* **1986**, *32*, 791–805. [CrossRef]
19. Rostam, Z.R.; Kertész, G. Advances in Pre-trained Language Models for Domain-Specific Text Classification: A Systematic Review. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 124. [CrossRef]
20. Hill, C.; Irshaidat, F.; Johnson, M.; Fresneda, J. An analytical assessment of sentiment analysis trends and methods through systematic review and topic modeling. *Decis. Anal. J.* **2025**, *17*, 100644. [CrossRef]
21. Taha, K.; Yoo, P.D.; Yeun, C.; Homouz, D.; Taha, A. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Comput. Sci. Rev.* **2024**, *54*, 100664. [CrossRef]
22. Jiang, G.; Yang, S.; Wang, Y.; Hui, P. When trust collides: Exploring human-LLM cooperation intention through the prisoner's dilemma. *Int. J. Hum.-Comput. Stud.* **2026**, *209*, 103740. [CrossRef]
23. Xu, T.; Huo, Y. Interpretable fake review detection based on multimodal information and human–computer collaboration. *Appl. Soft Comput.* **2026**, *191*, 114722. [CrossRef]
24. Bai, S.; Shi, S.; Han, C.; Yang, M.; Gupta, B.B.; Arya, V. Prioritizing user requirements for digital products using explainable artificial intelligence: A data-driven analysis on video conferencing apps. *Future Gener. Comput. Syst.* **2024**, *158*, 167–182.
25. Chandran, N.V.; Anoop, V.; Asharaf, S. Semantic Text Kernels: A hybrid interpretable framework for deep semantic analysis in textual data. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113568. [CrossRef]
26. Zhang, M.; Sun, L.; Wang, G.A.; Li, Y.; He, S. Using neutral sentiment reviews to improve customer requirement identification and product design strategies. *Int. J. Prod. Econ.* **2022**, *254*, 108641. [CrossRef]
27. Groen, E.C.; Kopczyńska, S.; Hauer, M.P.; Krafft, T.D.; Doerr, J. Users—The hidden software product quality experts?: A study on how app users report quality aspects in online reviews. In *Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference (RE)*; IEEE: New York, NY, USA, 2017; pp. 80–89.
28. Maalej, W.; Kurtanović, Z.; Nabil, H.; Stanik, C. On the automatic classification of app reviews. *Requir. Eng.* **2016**, *21*, 311–331. [CrossRef]
29. El Mrabti, S.; Jaouad, E.-M.; Hachmoud, A.; Lazaar, M. An explainable machine learning model for sentiment analysis of online reviews. *Knowl.-Based Syst.* **2024**, *302*, 112348. [CrossRef]
30. Subedi, I.M.; Singh, M.; Ramasamy, V.; Walia, G.S. Classification of Testable and Valuable User Stories by using Supervised Machine Learning Classifiers. In *Proceedings of the 2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Wuhan, China, 25–28 October 2021*; IEEE: New York, NY, USA, 2021; pp. 409–414.
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Mujahid, M.; Kina, E.; Rustam, F.; Villar, M.G.; Alvarado, E.S.; De La Torre Díez, I.; Ashraf, I. Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *J. Big Data* **2024**, *11*, 87. [CrossRef]
33. Liang, D.; Yi, B. Two-stage three-way enhanced technique for ensemble learning in inclusive policy text classification. *Inf. Sci.* **2021**, *547*, 271–288. [CrossRef]
34. Zhang, W.; Zhou, Y.; Liu, S.; Zhang, Y.; Shang, X. Weakly supervised text classification framework for noisy-labeled imbalanced samples. *Neurocomputing* **2024**, *610*, 128617. [CrossRef]
35. Tahery, S.; Aftabi, S.Z.; Farzi, S. A GA-based algorithm meets the fair ranking problem. *Inf. Process. Manag.* **2021**, *58*, 102711. [CrossRef]
36. Min, B.; Ross, H.; Sulem, E.; Veyseh, A.P.B.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.* **2023**, *56*, 1–40. [CrossRef]
37. Raza, M.; Jahangir, Z.; Riaz, M.B.; Saeed, M.J.; Sattar, M.A. Industrial applications of large language models. *Sci. Rep.* **2025**, *15*, 13755. [CrossRef] [PubMed]
38. Kojima, T.; Gu, S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv* **2022**, arXiv:2205.11916.

39. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 106. [CrossRef]
40. Li, H.; Cao, P.; Wang, X.; Li, Y.; Yi, B.; Huang, M. Pre-training enhanced unsupervised contrastive domain adaptation for industrial equipment remaining useful life prediction. *Adv. Eng. Inform.* **2024**, *60*, 102517. [CrossRef]
41. Tan, J.; Wang, D.; Sun, J.; Liu, Z.; Li, X.; Feng, Y. Towards assessing the quality of knowledge graphs via differential testing. *Inf. Softw. Technol.* **2024**, *174*, 107521. [CrossRef]
42. Lyu, J.; Xie, J.; Yuan, S.; Zhou, Y. A minibatch stochastic gradient descent-based learning metapolicy for inventory systems with myopic optimal policy. *Manag. Sci.* **2025**, *71*, 5572–5588. [CrossRef]
43. Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A.S.; Ceder, G.; Persson, K.A.; Jain, A. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418. [CrossRef]
44. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
45. Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv* **2022**, arXiv:2209.02970.
46. Pasch, S.; Ha, S.-Y. Human–AI Interaction and User Satisfaction: Empirical Evidence from Online Reviews of AI Products. *Int. J. Hum.-Comput. Interact.* **2026**, *42*, 5570–5591. [CrossRef]
47. Hou, X.; Zhao, Y.; Liu, Y.; Yang, Z.; Wang, K.; Li, L.; Luo, X.; Lo, D.; Grundy, J.; Wang, H. Large Language Models for Software Engineering: A Systematic Literature Review. *ACM Trans. Softw. Eng. Methodol.* **2024**, *33*, 220. [CrossRef]
48. Alhoshan, W.; Ferrari, A.; Zhao, L. Zero-shot learning for requirements classification: An exploratory study. *Inf. Softw. Technol.* **2023**, *159*, 107202. [CrossRef]
49. Cai, M.; Yang, W.; Du, Y.; Tan, Y.; Lu, X. Automatic requirements elicitation from user-generated content: A review of data, methods, and representations. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111110. [CrossRef]
50. Sandiwarno, S.; Niu, Z.; Nyamawe, A.S. SES-Net: A Novel Multi-Task Deep Neural Network Model for Analyzing E-learning Users' Satisfaction via Sentiment, Emotion, and Semantic. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 4910–4933. [CrossRef]
51. Suffian, M.; Kuhl, U.; Bogliolo, A.; Alonso-Moral, J.M. The role of user feedback in enhancing understanding and trust in counterfactual explanations for explainable AI. *Int. J. Hum.-Comput. Stud.* **2025**, *199*, 103484. [CrossRef]
52. Tejo, L.C.; Silva, P.A. Engaging with ethics in the HCI design process: A study of ethical design practice in different work contexts. *Int. J. Hum.-Comput. Stud.* **2025**, *205*, 103634. [CrossRef]
53. Kapoor, K.; Ziaee Bigdeli, A.; Dwivedi, Y.K.; Schroeder, A.; Beltagui, A.; Baines, T. A socio-technical view of platform ecosystems: Systematic review and research agenda. *J. Bus. Res.* **2021**, *128*, 94–108. [CrossRef]
54. Millerand, F.; Baker, K.S. Who are the users? Who are the developers? Webs of users and developers in the development process of a technical standard. *Inf. Syst. J.* **2010**, *20*, 137–161. [CrossRef]
55. Yang, B.; Liu, Y.; Liang, Y.; Tang, M. Exploiting user experience from online customer reviews for product design. *Int. J. Inf. Manag.* **2019**, *46*, 173–186. [CrossRef]
56. Jin, J.; Jia, D.; Chen, K. Mining online reviews with a Kansei-integrated Kano model for innovative product design. *Int. J. Prod. Res.* **2022**, *60*, 6708–6727. [CrossRef]