ACCSCIENCE PUBLISHING

Design+

ARTICLE

From reviews to empathy: Natural language processing-driven automated empathy mapping and its methodological implications

Serkan Güneş*

Department of Industrial Design, Faculty of Architecture, Gazi University, Ankara, Türkiye

Abstract

Traditional empathy mapping (EM) in human–computer interaction suffers from subjectivity, limited scalability, and poor reproducibility. This study introduces a multi-layered natural language processing framework that automatically generates EMs from large-scale product reviews by integrating Think–Feel–Say–Do with customer journey mapping (CJM). A dataset of 4,845 Amazon robot vacuum reviews (30,642 sentences) was analyzed using zero-shot classification (e.g., BART-Large-MNLI and RoBERTa-Large-MNLI), interpretability (e.g., local interpretable model-agnostic explanations), aspect-based sentiment analysis (e.g., ABSAbank-RoBERTa, local context-focused-Bidirectional Encoder Representations from Transformers [BERT]), and CJM alignment (e.g., sentence-BERT). The findings highlight a predominance of Think (46.6%) and Do (42.6%), while Feel (9.4%) and Say (1.5%)—though less frequent—convey strong emotional polarity, with 56% of content at extremes. "Device experience" dominates as the key touchpoint (66%) and the CJM experience stage (80.7%). Aspect analysis emphasizes technical (73.2%) and commercial (24.7%) drivers, particularly cleaning performance, battery life, price, and warranty. Cluster analysis identifies three profiles: action-intensive, rational-evaluation, and narrative-emotion. The framework advances EM as scalable, reproducible, and evidence-based, supporting user experience optimization, persona design, and real-time monitoring.

Keywords: Empathy mapping; Human–computer interaction; Natural language processing; Online product reviews; Transformer models; Customer journey mapping

***Corresponding author:**
Serkan Güneş
(serkangunes@gazi.edu.tr)

Citation: Güneş S. From reviews to empathy: Natural language processing-driven automated empathy mapping and its methodological implications. *Design+*. 2026;3(1):025390041. doi: 10.36922/DP025390041

Received: September 22, 2025

Revised: January 2, 2026

Accepted: January 6, 2026

Published online: February 4, 2026

1. Introduction

In recent decades, the discipline of industrial design has undergone a paradigmatic shift, moving beyond the modernist tenets of functionality and form to engage more comprehensively with the psychosocial dimensions of user experience (UX). While early contributions to the field often privileged performance and esthetic coherence,[1,2] subsequent scholarship has increasingly focused on the emotional, cognitive, and contextual experiences of users.[3,4] This evolution has gradually positioned empathy not merely as a methodological tool, but as a foundational principle in contemporary design thinking and human–computer interaction (HCI) research.[5,6]

Initially conceptualized as a means of accessing users' lived realities through qualitative immersion, empathy now signifies a broader intellectual and ethical orientation within

design culture and HCI practice. Its instrumental role in generating user insight has gradually been coupled with normative claims—portraying empathy as a core soft skill, and in some cases, as an ethical imperative.[6] However, this elevated status has also revealed new tensions: while empathy enriches the conceptual horizon of design and interaction research, its implementation is often hindered by methodological ambiguities and epistemological limitations.[7]

The central concern does not lie in whether empathy should be pursued, but rather in how it is framed, enacted, and evaluated within design and HCI contexts. Prevailing methods—relying on workshops, designer intuition, and interpretive heuristics—remain susceptible to longstanding critiques, including contextual bias, limited scalability, and reproducibility deficits.[8,9] Tools such as empathy maps (EMs), personas, and customer journey maps (CJM), although designed to facilitate user-centered reasoning, can risk flattening the richness of human experience into static and reductive visual schemas.[10]

Moreover, a persistent conceptual conflation undermines the efficacy of empathy tools in both design and HCI research. While often equated with affective resonance, the empathy operative in most design and interaction contexts is better understood as cognitive—the capacity to model others' perspectives, intentions, and situational constraints without necessarily sharing their emotions.[7] Thus, EMs are more aptly characterized as instruments of reflective perspective-taking rather than affective identification.

Despite their prominence in the design and HCI toolkit, EM practices largely remain anchored in limited empirical evidence. The knowledge base they produce is often shaped by small-sample field studies, narrative vignettes, or anecdotal user stories filtered through designer assumptions.[5,6] Consequently, concerns about representation, inference validity, and bias remain largely unresolved.[9,11] Critics increasingly argue that in their current form, EMs function less as analytic frameworks and more as tools of intuitive affirmation.[6]

In light of these limitations, a pressing methodological gap persists in HCI: the absence of systematic, data-rich approaches capable of scaling empathy practices beyond the workshop setting. Although digital ecosystems now provide abundant user-generated content, primarily through online product reviews (OPRs), the integration of such data into EM for HCI remains strikingly underexplored.[11] This omission constrains both academic research and applied innovation in interaction design.

The close relationship between EM and HCI stems from EM's role in facilitating a deep understanding of UX and enabling user-centered design decisions. Its use as a tool for transforming tacit knowledge and user insights supports persona- and scenario-based design approaches[12] while making data obtained from user research accessible, interpretable, and actionable for decision-makers through visualization. Using EMs to analyze HCI allows interaction to be understood not only in technical and technological dimensions but also in emotional and contextual aspects.[13] Thus, interdisciplinary teams can share the user perspective within a common framework, translating complex insights into concrete design decisions while maintaining the traceability of these decisions. In this context, EM operationalizes the internalization of the user perspective—one of the fundamental principles of user-centered design in HCI—and, as an early-stage HCI research method, guides product, interface, and interaction design decisions by systematically identifying users' motivations, pain points, and gains before prototype development.

Building on this theoretical foundation, the natural language processing (NLP)-based computational EM approach—developed to overcome the limitations of traditional EM, such as subjectivity, lack of scalability, and irreproducibility—automatically extracts EM components from large and heterogeneous datasets using NLP, machine learning, and data mining techniques. This creates a hybrid analytic layer that integrates both qualitative and quantitative data. Such integration provides evidence-based support for user-centered design decisions, enables scalable persona and scenario production across user segments, and allows real-time experience monitoring—fundamental goals of HCI. Therefore, the computational EM not only increases the speed and efficiency of the design process in HCI but also provides epistemic reliability and methodological transparency, transforming empathy from a subjective design practice into a systematic knowledge production tool.

The present study addresses this gap by introducing a structured methodology that operationalizes EM through the systematic analysis of online user reviews. Drawing on advances in NLP and text mining, the proposed framework extracts nuanced user perspectives—including pain points, gain points, and customer journey stages—at scale and with reduced interpretive subjectivity. The approach does not aim to displace intuitive empathy, but rather to complement and refine it through computational precision and data-informed representational rigor in HCI contexts.

This study not only automates traditional EM methods but also contributes to the trend of "scalable, data-driven design tools" prominent in HCI literature. Current research has shifted toward methodologies that deepen the

ethical, contextual, and measurable dimensions of UX, and the proposed NLP-based framework aligns with this trend, aiming to transform empathy from an intuitive workshop practice into a scalable, repeatable, rapid, dynamic, and pre-guided knowledge production tool.

The proposed framework contributes to both scholarly and professional domains by offering a reproducible, ethically grounded, and cost-effective approach to engaging with user narratives at scale. By embedding large-scale user-generated content directly into the EM process, it overcomes several practical limitations of traditional tools and addresses conceptual critiques raised in recent design and HCI research.[14,15] Within the broader HCI field, this methodological advance strengthens the empirical basis of user research, enabling the systematic incorporation of diverse, naturally occurring interaction data into early-stage design processes.

Furthermore, the methodology enhances the empirical base of UX decision-making in HCI and advances a more evidence-informed model of user-centered innovation. This study advances the theoretical discourse on empathy in design and interaction by reframing it from a predominantly affective, intuition-driven practice to a computationally operationalized construct grounded in cognitive perspective-taking. While traditional EM relies on subjective interpretation of small qualitative datasets, the proposed NLP-based framework formalizes empathy as an evidence architecture that can be quantified, replicated, and scaled across diverse interaction contexts in HCI.

This reframing not only challenges the epistemic boundaries of empathy in design thinking but also positions automated EMs as hybrid instruments integrating human interpretive sensibility with machine-derived contextual precision. In doing so, the study contributes to the ongoing theoretical debate on whether empathy in HCI and design should be understood as an experiential state to be felt or as a structured insight to be inferred and acted on. The study proceeds from the hypothesis that automated EMs generated from online user reviews can provide insights comparable in depth and validity to those derived from conventional EM techniques in HCI, while substantially improving scalability, transparency, and replicability in interaction design contexts.

To investigate this hypothesis empirically, the study formulates the following research questions (RQs):

(i) RQ1: Can current NLP and text mining technologies support the automatic generation of EMs from online user reviews for HCI research and practice?
(ii) RQ2: Do such automatically generated maps provide contextually rich, valid, and design-relevant insights for HCI decision-making?
(iii) RQ3: How do automated EMs compare with traditional EMs in terms of insight quality, scalability, reproducibility, and cognitive bias reduction within HCI contexts?

By addressing these questions, the study not only advances methodological rigor in empathy research but also contributes to the evolving discourse on data-informed design in HCI. It seeks to promote a hybrid paradigm, where intuition is not abandoned but critically supported by scalable evidence architectures.

2. Literature review

2.1. Conceptual foundations of empathy

The origin of empathy as an academic concept originates in esthetic thought. Vischer[16] argued that esthetics, through his *Einfühlung* approach, was more than a visual experience; it was a form of subjective participation in which the observer projects meaning into the form. Lipps,[17] on the other hand, took the concept of *Einfühlung* out of the confines of the art field and applied it to all areas that could provide an esthetic experience, including products. Titchener[18] translated the concept of *Einfühlung* into English as "empathy" (from the Greek *em* [into] and *pathos* [feeling]), thereby increasing interest in the idea and laying the foundation for early emotion theory.

It was not until the mid-20th century that the idea of empathy began to penetrate design knowledge, although in a fragmented manner. Simon[1] indirectly implied that the expectations of users should influence design processes. Empathy as a solution to disruptions in feedback loops between designers and users in industrial settings was also presented by Jones.[3] For Cross,[2] design was not a purely technical competence; the user's standpoint represented a distinct epistemology that required creative access. Despite their pioneering nature, these efforts provided a basis for the systematic integration of empathy-driven principles into design processes. Empathy remained a concept that oscillated between metaphor and method for an extended period.

Along with this, empathy began to be presented within a more refined typology in interdisciplinary theory: emotional, cognitive, and motivational empathy. Emotional empathy generally refers to the ability to share another person's emotional state involuntarily and instinctively,[19,20] while cognitive empathy is associated with the ability to understand another person's internal state without experiencing it oneself, that is, perspective taking.[21,22] This cognitive structure is grounded in models such as the theory of mind,[23] simulation theory,[24] and theory of theory.[25] Motivational empathy, on the other hand, introduced an ethical or behavioral component that

went beyond empathetic understanding and resonance, creating a prosocial impulse to help or defend.[26]

In the field of design, the empathy expected of designers was not therapeutic in the sense of emotional resonance or "feeling the same thing" as users. Empathy was framed more as a form of strategic modeling. The strategy here was presented as predicting users' experiential conditions with sufficient accuracy to enable actionable design decisions. Thus, cognitive empathy was prioritized and systematically developed over time to serve innovation by externalizing users' intentions and constraints.

2.2. Empathy in design practice: Tools and methods

Early attempts to incorporate empathy into practice, such as personas[27] and CJM,[28,29] helped to visualize the user's context. These applications, which remained at a level of abstraction rather than deep involvement, rarely captured cognitive complexity. The identification of latent needs[30] was also based on observational intuition. However, it later became a strategic resource, used by consulting firms such as Innovation Design Engineering Organization to gain insight. Experiential research techniques such as cultural probes[31] and experience prototyping[32] sought to understand users' lives and worlds using analog and concrete techniques—for example, artifacts, diaries, and physical simulations. However, they were subject to criticism regarding analytical depth and reproducibility.

By 2010, Gray *et al.*[33] proposed the EM. User inputs were synthesized by dividing insights into four quadrants—"Think," "Feel," "Say," and "Do." Osterwalder *et al.*[34] expanded EM format by adding "Pains," "Gains," and "Jobs to be Done" under the name "Value Proposition Design." From this point on, the EM began to be increasingly adopted by many disciplines, from UX research to strategic innovation.

Like any method, EM has not escaped criticism. The main points of criticism include a decline in usefulness as its scope expands and its function as a container that holds designers' assumptions and confers legitimacy on them, rather than providing empirically based insights.[10] Despite the method's simplicity, its ability to capture the complex realities of UX, as well as the quality and objectivity of its input data, has always been questioned.

According to Kouprie and Visser,[5] when designers strive to see what they want to see, EM—which lacks verification mechanisms—may reinforce existing cognitive biases and lead to a form of empathetic confirmation bias, rather than revealing latent user needs. According to Heylighen and Dong,[7] as empathy is functionalized, it becomes procedural and transforms into a checklist or framing tool. While functionality is necessary for scalability as a tool, the emotional dimension may be obscured, and EM can become a rhetorical tool. Eichbaum *et al.*[35] emphasize the cultural limitations of the tool, arguing that an EM framework reflecting Western norms may not be universally applicable in cross-cultural settings. Costanza-Chock[9] highlights ethical concerns, such as privacy, user consent, and emotional exploitation, emphasizing that without user participation, users may be reduced to data, thereby undermining the dignity that empathy aims to protect.

These criticisms do not call for the abandonment of EMs, but rather for their critical re-examination. Despite these concerns, EM remains a popular method, continuing to stay relevant due to its ability to facilitate and reflect cross-functional dialog. The future of empathy-based design may depend not on visual tools but on meticulous, iterative, and ethically sensitive applications that strike a balance between structure and sensitivity.

2.3. EM and HCI

The concept of empathy becoming an integral part of HCI processes is directly related to the redefinition of the scope of HCI. The approach that does not limit UX to the design of interactive systems but also considers the effects of design on users[36] has gone beyond traditional usability and functionality criteria to emphasize the concept of "holistic experience"[37] and has paved the way for discussions on a UX perspective that encompasses the dimensions of situatedness and temporality.

However, the complexity of human behavior and experience, which extends beyond simple demographic data and lists of summarized needs, has revealed the limitations of the widely used persona approach,[38] which aims to reflect the average characteristics and behaviors of a target audience. Such tools, defined by Hassenzahl[39] as "putting a human face on product design," fall short of fully reflecting the inherent complexity and contextual diversity of human experience; particularly in emotionally intense contexts and edge-case situations,[40] they fail to adequately support empathy. At this point, while the persona approach focuses on defining the user in terms of "who" they are, EM enables the designer to see things from the user's perspective by systematically revealing what the users see, think, feel, say, and experience in terms of gains and difficulties.[33]

Offering a deeper structure than the classic persona, EM triggers both emotional and cognitive empathy by bringing the user perspective to life through its constituent components.[5] In the HCI/UX context, EM's instrumental role not only increases empathy but also compels the designer to consider real-world usage scenarios through

visualization and detailed insights. Indeed, the concept of "context of use" in HCI literature[41] requires a holistic assessment of dimensions such as users, tasks, hardware–software–materials, and physical and social environments; integrating this context into EM enables both enhanced empathetic understanding and the early detection of edge cases. The most powerful tool in this detection is the CJM, which is integrated with EM, enabling designers to form a holistic and immediate mental representation of the user by presenting the user perspective visually rather than textually. Thus, abstract user data becomes more memorable and actionable, while holistic mapping not only enhances designers' understanding but also strengthens their motivation to take action.[42]

Today, the conceptualization of empathy through large language models (LLMs) and the production of empathy in users are intensely debated, particularly in the field of chatbot development. This is because humans can feel empathy for both abstract and physical objects. Conversational agents (CAs) behave as if they are "sharing" the user's emotions (projection) and elicit empathy in users by displaying emotions themselves (elicitation). LLMs may appear empathetic, but this empathy can be inconsistent, superficial, and sometimes discriminatorily biased.[43] Such efforts, especially in the healthcare field, can increase user trust but may also create false expectations and pose ethical and practical risks.[44]

The model proposed in this study, unlike CAs, does not interact directly with users; therefore, it does not aim for projection or elicitation. Instead, it takes on the role of a supportive tool designed to be used in the "analysis and synthesis" phase of the HCI research cycle by identifying empathy and its components. This approach aligns with the recommendations by Elagroudy *et al.*[45] that the use of generative artificial intelligence (AI) and LLM can provide methodological contributions to researchers in tasks such as qualitative data coding, emotion detection, and content classification.

In addition, LLMs are used not only for empathy detection but also to generate artificial participants ("AI personas") that reflect human characteristics. For example, in the study by Yeykelis *et al.*,[46] the findings of 133 experimental media-effects studies were reproduced through LLM-based personas, demonstrating the method's potential for scalable, reproducible data generation. Güneş[47] demonstrated that different personas can be generated from OPRs, but the findings were largely limited to demographic pattern clusters.

In this context, the experimental study conducted by Salminen *et al.*[48] revealed that adding visuals to AI-generated personas does not significantly influence user perception, which is primarily shaped by the persona's textual narrative. In addition, persona identity (e.g., life story and experience) and participant demographics (e.g., age, gender, and persona experience) were found to play a decisive role in perception. These findings support the idea that identifying empathy components through text-based methods can provide a reliable foundation for HCI design decisions and emphasize the importance of incorporating contextual and demographic factors into the analysis process.

On the other hand, the Persona-L system[49] aims to increase empathy and contextual depth for user groups with complex needs (e.g., Down syndrome) by integrating LLMs with an ability-based design framework. In this system, publicly available user narratives are retrieved from a vector database using retrieval-augmented generation and are injected into the conversation context to enhance the accuracy and contextual appropriateness of responses. User study findings revealed that the continuity of contextual data enhances the perception of empathy.

The role of the model aligns with the perception layer within the framework of "computational compassion." This layer identifies empathy components, providing foundational data for contextual evaluation and ethical guidance. Although the model does not directly target the action dimension of compassion, the empathy data it detects allow HCI designers to better understand user needs and develop appropriate intervention strategies. Salminen *et al.*[48] highlight the importance of text-based empathy detection, while Sun *et al.*[49] emphasize the methodological foundation of the approach in terms of context enrichment.

The approach presented in this study aligns with the requirements of contextual analysis, criterion alignment, and methodological transparency emphasized in the user-centered explainable AI (XAI) literature.[50,51] Furthermore, the findings that XAI tools are underrepresented in terms of text-based explanations and interactive interfaces[52] directly support the proposed model's goal of generating empathy data that are visualized with EM and CJM. Discussions highlighting the "for whom, for what purpose" dimension of explainability in the LLM era[53] provide a theoretical foundation for positioning the model as a scalable and repeatable empathy analysis tool that considers ethical and socio-technical contexts.

Thus, empathy data are transformed from mere perception into a holistic research output that informs design decisions. In this context, the model provides researchers with a systematic, scalable, and repeatable method for empathetic content analysis, enabling (i) human-centered design decisions informed by empathy

data; (ii) empathy components to be reliably identified and incorporated into the design process; and (iii) empathy data in HCI applications to be transformed into a holistic research output with ethical and contextual appropriateness in mind.

2.4. Emergence of NLP-enhanced EM

In recent years, limitations of EM have brought methods of computational qualitative interpretation into focus. The success of technologies such as NLP and text mining in analyzing large volumes of user-generated data, such as online reviews, support forums, and social media posts, has provided a rich empirical basis for more scalable and systematic approaches to EM.

Pioneering examples, such as the tool for the automatic creation of EMs,[11] aim to automate the process of filling in EM quadrants using basic machine learning techniques on interview transcripts. However, this tool has limited algorithmic sophistication and relies on a small dataset, ultimately failing to deliver on expectations in terms of inferring deeper user motivations or distinguishing between explicit and implicit expressions.

More recent approaches aim to combine sentiment detection with intention inference and theory-of-mind structures. Although the method proposed by Zhu and Luo[54] goes beyond basic polarity classification to enable implicit meaning extraction and automatic recognition of Think and Feel components, it lacks full integration between EM components and alignment with customer journey stages. In summary, a fully integrated EM- and CJM-oriented system is still underdeveloped.[55,56]

At this point, it is helpful to briefly explain the logic behind computational models in determining empathy. Computational modeling of empathy treats empathy as an inferential structure based on text analysis, unlike traditional definitions based on emotional resonance. NLP systems define users' emotional and cognitive states through language patterns rather than attempting to feel like humans. This approach reduces subjectivity while providing repeatability and scalability; it transforms the designer from an intuitive interpreter to a data-driven meaning creator. However, this redefinition entails some conceptual trade-offs, such as the weakening of empathy's ethical and interpersonal dimensions. Ultimately, empathy takes on a measurable and workflow-integrated structure, and the distinction between "experienced" and "extracted" empathy becomes even more pronounced in data-driven design environments.

Despite notable advances in automating empathy processes, several gaps remain unresolved. First, most NLP-based systems address only partial aspects of the EM framework. While tools exist to classify sentiment or detect intent, few can simultaneously generate comprehensive EMs across all axes—including Think, Feel, Say, Do, Pains, and Gains—let alone contextualize them within customer journey stages.[11,55]

Second, affective signals dominate current computational models, with relatively little attention paid to cognitive empathy or perspective-taking mechanisms.[57] As a result, these tools may capture surface-level emotional tone without accessing the more profound logic of user motivations or expectations.

Third, the potential of large-scale online content—such as product reviews or user forums—for empathy extraction is underutilized. While this data is abundant, it often remains disconnected from structured design tools or frameworks capable of transforming it into actionable insights.[56]

This study seeks to address these limitations by proposing a multi-layered, NLP-enhanced framework for automated EM. The framework integrates sentiment analysis, intent inference, and theory of mind modeling within a unified pipeline. In doing so, it aims to provide a scalable and methodologically transparent approach to empathy extraction, redefining empathy not as an emotion to be shared but as a structured insight to be inferred and operationalized within design processes.

3. Methodology

As part of this study, a detailed and multi-layered methodological framework was developed to analyze OPRs methodically, taking into account thoughts, feelings, expressions, and actions.

To achieve the objectives of this study, a comprehensive dataset of 4,845 verified and purchase-confirmed OPRs related to robot vacuum cleaners was initially collected from the Amazon.com website using the Firefox WebScraper add-on (https://webscraper.io/) in.xlsx format. The data source used in the study is limited to Amazon product reviews. While this limits the contextual diversity, it provides an advantage in terms of demonstrating the scalability of the methodology. This study relies on a dataset compiled from a single platform and a single product domain, a design choice that may be considered a limitation in terms of domain diversity and may generate a perception of selection bias with respect to external validity. However, the primary objective of the research is not to conduct cross-platform or cross-product comparative sentiment analysis, but to develop a scalable, explainable, and methodologically domain-agnostic NLP pipeline capable of automatically identifying EM

quadrants (Think, Feel, Say, and Do [TFSD]), extracting purpose-aware aspects, and detecting CJM stages based on abstract linguistic decision signals and interpretable inference policies.

The machine-learning architectures, labeling policies, majority-confidence thresholds (≥60% for single-dominant labels), multi-label fallback rules, and the directed aspect dictionary are not conditioned on platform- or product-specific features, but on distributional patterns of language-level decision indicators, reinforced through local explanation mechanisms rooted in sampling-based XAI methods. Accordingly, the claim of generalizability is not grounded in empirical variance across multiple domains, but in the abstraction capability of the decision model, the reproducibility of reporting protocols, and the statistical expressibility of validation evidence within a controlled and constant public-data environment.

All reviews were collected exclusively from publicly accessible sources, without user identifiers or sensitive metadata, and all threshold rationales, dictionary coverage, and modular inference flows were systematically documented in the supplementary file, preserving the data environment as a fixed experimental control setting, thereby enhancing methodological transparency, auditability, and reproducibility. This approach enables rigorous evaluation of pipeline coherence and interpretability without domain expansion and provides an evidence architecture that substantiates the study's core assertion of methodological independence from platform- or product-bound variation. Future studies can test methodological generalizability by adapting the same approach to different interaction contexts such as health, education, and public services.

In addition, primary data obtained from user studies can be compared with computational findings to more robustly test the empirical validity of the model. On large-scale e-commerce platforms such as Amazon, user reviews are growing dynamically, with thousands of new entries added every day, and in many cases, tens of thousands of public reviews accumulate for a single product. Such data sources serve not only as information channels that guide consumers' preference processes but also as a rich resource that provides direct design insights for product developers and designers.

While the method appears to be applied through a case study in terms of representativeness, the proposed methodology is inherently scalable to different product categories and even non-product UX contexts, without being constrained by data quantity. Therefore, the framework presented here goes beyond e-commerce-based reviews and offers a methodological contribution that can be generalized for broader HCI practices.

During the study, a total of 4,845 reviews were analyzed and divided into individual sentences, resulting in 30,642 distinct sentences. Each sentence was systematically assigned a numerical index corresponding to the related review to which it was linked. This methodology constitutes a foundational step in organizing text data and was implemented through a Python script automated using the Natural Language Toolkit NLP library. This library relies on punctuation marks (e.g., "." "!," and "?") and built-in grammar structures.

The analyzed data comprise product review sentences derived from UXs. These reviews are stored in an Excel file (.xlsx format), with the text within the column titled "Comment" used for evaluation. No linguistic preprocessing steps, such as stop word removal or stemming, have been performed on the reviews; consequently, full-text analysis was employed to preserve contextual integrity.

The comments are initially analyzed through the utilization of keywords, as delineated by the researcher, to generate a preliminary prediction. This procedure establishes a reference framework for assessing the overall predictive accuracy of the system and facilitates comparative analysis with model outputs.

The comments are subsequently fed into two distinct pre-trained transformer models— facebook/bart-large-MNLI and roberta-large-MNLI. The two models utilize a zero-shot classification approach to generate contextual similarity scores for the four labels (TFSD). These models are preferred because they provide high accuracy in understanding the context of the language. Specifically, the *-MNLI versions have been optimized for natural language inference (NLI) and developed to evaluate semantic alignment between texts. Alternative models (e.g., decoding-enhanced bidirectional encoder representations from transformers with disentangled attention [DeBERTa], text-to-text transfer transformer, and generative pre-trained transformer series) are either not directly optimized for NLI tasks or require prompt engineering or additional fine-tuning for classification tasks. The employment of such models introduces disadvantages related to both computational expense and interpretability constraints.

Although models such as DeBERTa have been demonstrated to yield higher accuracy in specific benchmarks,[58] they were not selected in this study due to practical constraints, including interpretability, computational efficiency, and compatibility with Colab. Moreover, bidirectional and auto-regressive transformers (BART) and robustly optimized bidirectional encoder representations from transformers (BERT) pretraining approach (RoBERTa) models have been shown to offer more

stable results in terms of the applicability of explainability techniques such as local interpretable model-agnostic explanations (LIME).[59] For the aforementioned reasons, priority was given to selecting the most suitable models in terms of conceptual accuracy and technical compatibility.

The TFSD zero-shot classifier was validated on an expert-annotated audit sample of 200 sentences, where 20% of the sample (n = 40) was strictly isolated as a held-out set for out-of-sample evaluation and was not used in prompting or threshold calibration. Label assignments are governed deterministically under a confidence-based policy, requiring ≥60% dominance for single-quadrant adjudication and resolving all other cases through an explicit multi-label fallback rule to prevent spurious quadrant inflation. Explanation tracing was enabled through the integration of LIME within the NLP pipeline, not for measuring global model stability, but to render local inference rationales, semantic adjacency paths, and decision-abstraction capacity fully interpretable and auditable at the sentence level.

The observed predominance of misclassification routes along "Think–Do" directions corresponds consistently with the semantic neighborhood properties of cognitive reasoning and action-oriented narratives previously characterized in the corpus, indicating that confusions arise from interpretable linguistic adjacency rather than model instability, noise, or arbitrary labeling. Generalizability is therefore positioned as a function of inference abstraction and reproducible evidence reporting, not domain multiplicity. Finally, the aspect dictionary construction, overlap resolution logic (longest-match priority), and threshold rationales were formalized and systematically recorded in the Supplementary File to enhance transparency, auditability, replicability, and methodological independence without requiring domain expansion or pipeline reconstruction.

The Python code explicitly defines the list of labels utilized for classification as follows: "Think," "Feel," "Say," and "Do." Their notable performance in the general language understanding evaluation further corroborates the reliability of these models' benchmark assessments, which evaluate the capabilities of the respective models. BART attains a 90.8% accuracy rate on the MNLI task, whereas RoBERTa achieves a 90.2% accuracy rate in the same task.[60] These metrics demonstrate the models' adherence to the established standards for NLI.[61,62] During this evaluation process, a sample of 200 comments was selected, ensuring representativeness across various product categories. Subsequently, these comments were meticulously annotated by an expert researcher possessing theoretical expertise in EM. This methodology is assessed within the context of domain expert annotation and provides a reference framework for evaluating the model's efficacy in contextual classification. The model exhibited a commendable level of accuracy, achieving an accuracy of 82.4% and a macro F1-score of 79.1%, thereby indicating its capacity to attain performance levels comparable to those of human evaluators.

The scores produced by both models are normalized across four categories per comment and subsequently averaged. The purpose of this procedure is to establish a more balanced and dependable classification output by mitigating model-based biases. This approach constitutes an ensemble modeling technique that improves accuracy through the integration of multiple model outputs.

The LIME algorithm was selected due to its model-agnostic properties, its capacity for local interpretability, and its compatibility with transformer-based NLP models. Its capability to produce context-sensitive justifications for individual predictions aligns well with the objectives of cognitive empathy modeling. Furthermore, LIME enables transparent communication of model decisions to human users, a vital requirement in UX and design research. Although alternative explainability tools such as Shapley Additive exPlanations and integrated gradients offer holistic or gradient-based insights, they involve substantial computational and implementation constraints that restrict their practical use in large-scale review analysis. LIME offers a balanced approach by providing accessible and interpretable outputs, thereby enhancing the transparency and user-friendliness of automated EM generation.

The average scores of the model are subsequently transferred to the algorithm. LIME analyzes the words that influence the classification of each comment into different empathy dimensions (i.e., TFSD) and assigns specific contribution weights to these words. This process allows for the transparent disclosure of the rationale behind the model's decision to assign a particular class to an instance. For each comment, the words that most significantly contribute to the relevant empathy category, along with their weights, are documented. These explanations are essential for generating qualitative insights, especially within the context of UX research.

Nonetheless, it is imperative to acknowledge the limitations inherent in the LIME algorithm. Specifically, its confinement to the local decision space restricts the generalizability of the explanations produced, thereby complicating the reproducibility of explanations for identical inputs.[63] Furthermore, due to its sensitivity to text length and word frequencies, LIME may occasionally produce unstable or contextually inadequate justifications. As a result, the contextual consistency of LIME outputs

must be rigorously assessed. Outcomes should be validated through interpretive controls before their integration into automated decision-making systems to ensure unequivocal confidence. The limitations of the LIME algorithm are comprehensively delineated in the theoretical background of this study (refer to Sections 2.4 and 2.5), underscoring that explainability is contingent on inferential modeling capacity rather than empathetic intuition. Accordingly, the LIME-based methodology utilized herein is founded on an empathy modeling framework that emphasizes data-driven meaning extraction over the generation of emotional resonance.

To address these limitations and ensure consistent context analysis, voluminous product reviews are broken down into individual sentences during the preprocessing stage. This process has facilitated the LIME algorithm's ability to produce more reliable explanations and enabled more accurate classifications of the subcomponents of reviews. The contribution weights generated by LIME are aggregated by category and converted into percentage scores by dividing them by the total contribution weight. Categorical contributions exceeding 20% are designated as "detected quadrants," with the category exhibiting the highest percentage being designated as the "dominant quadrant." This approach is designed to be compatible with multi-label methods and acknowledges that a comment may contain multiple empathy dimensions simultaneously. The 20% threshold is sufficiently elevated to exclude random contributions in a four-category distribution while maintaining sufficient flexibility to avoid overlooking secondary yet substantial dimensions. In the extant literature, a range of 15–25% is also recommended for analogous multi-classification scenarios. The 20% value offers a balanced choice that preserves contextual meaning while reducing classification noise.[64,65]

In the final stage, all analysis results were compiled into a DataFrame structure, with each comment occupying a separate row, and subsequently exported in .xlsx format. The analysis output for any four comments is presented in Table 1. For each comment, the final EM label, quadrant percentages, detected quadrants, dominant quadrant, and LIME justification columns were stored separately. To clarify this point, consider the following example: [comment] → Dominant: The phenomenon under examination has been identified through tactile perception. The following terms have been created to serve as justification tokens: "endorsing," "friends," and "inefficient." Each sentence processing task required 1–2 min to complete, utilizing a graphics processing unit through Google Colab.

The NLP pipeline used in this study is built on modern libraries that are compatible with Python 3.10 or higher, are open-source, and are subject to ongoing development. The code incorporates pre-trained LLMs (e.g., BART and RoBERTa), supported by the Hugging Face Transformers (v4.x) framework, the LIME library for explainability, and widely used academic modules, such as Pandas, OpenPyXL, and Natural Language Toolkit, for data input–output processes.

The developed analysis system is regarded as both contemporary and reliable due to the following characteristics: the zero-shot classification models—Facebook/BART-Large-MNLI and Roberta-Large-MNLI—are extensive models that have undergone pre-training using millions of texts. These models demonstrate capabilities in contextual inference and exhibit strong performance on NLI benchmark assessments. Hugging Face periodically updates them and maintains them diligently through significant contributions from the research community.

The LIME algorithm enhances the interpretability of model decisions on a local scale, thereby improving

Table 1. Sample detection of empathy quadrants

Comment	EM label	Think (%)	Feel (%)	Say (%)	Do (%)	Detected quadrant	Dominant quadrant	LIME			
								Say	Feel	Do	Think
I have always kept up with maintenance on the unit to ensure peak performance	Do	10	10	5	70	Do	Do	-	-	Maintenance, ensure, performance.	-
I will no longer be endorsing XXXXXX to friends and family!	Say	5	25	60	10	Feel, Say	Say	Endorsing, friends, family	No longer endorsing	-	-
When I first got the vacuum, I loved it	Feel	5	70	15	10	Feel	Feel	-	Loved, first got	-	-
I think the suction is strong, but the navigation seems a bit inefficient	Think	72	10	8	10	Think	Think	-	-	-	I think, strong, inefficient

Abbreviations: EM: Empathy map; LIME: Local interpretable model-agnostic explanations.

decision traceability, especially in research requiring human-centric interpretation, such as EM. This method ensures the verifiability of model decisions and promotes methodological transparency.

The code has been meticulously structured to operate within the Google Colab environment. This process automatically enables graphics processing unit-based acceleration, facilitating the reproducibility of the work by different researchers.

The utilization of open-source modules, supported by comprehensive documentation and active community engagement, positions these tools as prominent choices for integration within scientific projects.

These features demonstrate that the developed empathy classification system is both compatible with contemporary research practices and dependable from a technical and methodological standpoint.

In the subsequent phase of the study, emotions associated with technical and commercial attributes (aspects) in user reviews were extracted and analyzed contextually. The extracted attributes were divided into thematic categories based on their content. The sentiment orientation determined for each attribute was evaluated not only in terms of positive/negative polarization but also using a five-point intensity scale: P++ (strong positive), P+ (positive), NEU (neutral), N– (negative), and N–– (strong negative). This refined distinction enabled a more nuanced analysis, focusing on the strength and intensity of sentiment.

The sentiment outputs obtained were converted into a three-tiered superclassification to facilitate the strategic interpretation of the comments. The presence of gain (P++/P+), pain (N–/N––), and neutral (NEU) is indicated. Consequently, the comments were classified based on both the sentiment expressed and their capacity to highlight value or identify issues. For instance, the statement "battery lasts forever" is designated as an "aspect gain," while "customer service was unresponsive" is classified as an "aspect pain." This configuration has facilitated the generation of strategic insights concerning content analysis and functional analysis of comments. This classification has also been employed to map the positive and negative aspects of the empathy quadrants (TFSD), not solely in terms of the emotional orientation of comments.

To accomplish this contextual sentiment classification, two distinct methods were integrated at the initial stage to extract aspects from each comment.

The initial method is based on the rule-based Aho–Corasick trie algorithm. In this approach, a trie (prefix tree) structure was established using a predefined list of keywords, such as "battery" and "navigation system," through which comment texts were analyzed. This algorithm, developed by Aho and Corasick,[66] is primarily designed to facilitate rapid and accurate detection of multiple keywords within a single pass. While this technique ensures high processing efficiency when handling extensive datasets, it does not consider contextual information. Consequently, it is unable to identify semantically similar terms that are not included in the list, such as "power source."

The second method relies on transformer-based language models. These techniques analyze semantic patterns within the textual context, extracting attributes directly from the structure and meaning of sentences. Transformer architectures have been demonstrated to possess the capacity to understand indirect expressions, such as the phrase "costs an arm and a leg," which can be interpreted as "price." Furthermore, these architectures have been observed to identify novel expressions that were not previously defined, such as interpreting the term "run time" as "battery." This context-aware inference process is typically employed within named entity recognition or sequence labeling frameworks.[61]

At this stage of the aspect extraction process, comments that could not be matched using the Aho–Corasick trie method were processed using a series of transformer-based models. In this context, the aspect and relevant sentiment classes were extracted simultaneously from the comment in a context-sensitive manner using the aspect-based sentiment analysis (ABSAbank)-RoBERTa model. Then, for each aspect, the rationale for the sentiment analysis was detailed by evaluating the local context using the local context-focused BERT (LCF-BERT) model. ABSAbank–RoBERTa is a pre-trained model that has been fine-tuned on the RoBERTa architecture for aspect–sentiment matching.[61,67] The model's capacity to rapidly and accurately draw inferences is a substantial advantage, particularly in applications where resources are limited. However, given the single-step nature of this model, it is important to note that shifts may occur in structures containing multiple aspects, thereby leading to a more complex context.

Consequently, for each extracted aspect, the LCF-BERT model was employed to provide a detailed context at the local level. This model focuses on the words surrounding the attribute in the comment, thereby providing a more contextual basis for sentiment decisions. Furthermore, given its attention-based structure, it can be integrated with LIME or analogous explainability algorithms to elucidate the linguistic elements on which the model relies to reach its decisions.[64] Table 2 presents the analysis findings for this stage, along with three example sentences.

Table 2. Aspect-level sentiment detection results: gain, pain, and neutral distributions

Comment	Aspects	Gain	Pain	Neutral
That is because it has laser mapping system so it knows exactly where to go and clean	{"technical: mapping/lidar:" "P+," "technical: cleaning performance:" "P+," "technical: software/app:" "P+"}	Technical: mapping/lidar; technical: cleaning performance; technical: software/app	-	-
Awful!!! I can't get ahold of customer service	{"commercial: customer service:" "N−−"}	-	Commercial: customer service	-
It picked up hair/dust	{"technical: cleaning performance:" "NEU"}	-	-	Technical: cleaning performance

The process of extracting touchpoints from comment sentences was similar to that of feature extraction. In this study, touchpoints were extracted from comments using a two-stage matching architecture. First, the Aho–Corasick trie algorithm, a classic rule-based approach, was applied. This method provides rapid, error-tolerant, direct matching by determining if predefined word sets appear in the sentence. Examples of these word sets include "support," "installation," "app," and "delivery".[66] However, it relies solely on surface-level matches and cannot detect semantic variations arising from context, such as synonyms, idiomatic expressions, and verb conjugations.

To overcome this limitation, contextual justification outputs (i.e., justification tokens) obtained from pre-trained transformer models were used in the second stage. These models (e.g., RoBERTa and BART) have learned to represent language contextually on large-scale text training.[61,62] This enables them to recognize expressions with similar meanings, even if they are not explicitly listed. For instance, in the comment "I phoned three times but couldn't talk to anyone," even though the word "support" is absent, the transformer model associates the sentence with the "support" category. This is because the expressions "phoned," "talk to," and "couldn't reach" showed high contextual similarity with "call" and "support" during training. Transformer models achieve this by positioning words in multidimensional vector spaces according to their semantic clusters.[68]

In this scenario, while the Aho–Corasick trie matching failed (because the word "support" does not appear), the transformer model generated the following justification tokens: "phoned," "talk," and "anyone." These tokens were matched with the previously configured ontological dictionary by the researcher and assigned to the "Support" touchpoint. This structure extends beyond superficial matching to enable the inference of contextual meaning and achieves high accuracy, particularly in capturing indirect user narratives.[59,69] This two-stage structure, on the one hand, gains robustness against linguistic variations, while, on the other hand, ensures that decisions remain domain-focused (touchpoint-centric). Transformer models recognize context; ontology (or dictionary) categorically fixes the meaning of this context. This approach enables the strategic classification of both explicitly and implicitly expressed UXs (samples in Table 3).

Next, we examined the CJM stages to which each sentence pointed, as shown in Table 4. For this study, an unsupervised, semantically focused matching architecture was used to extract CJM stages from customer reviews. To accurately model the contextual meaning of the reviews, we converted each sentence into a multidimensional vector representation using the sentence-BERT model.[68] These representations were then compared with expanded dictionary sets that were developed for each of the seven CJM stages: awareness, research, comparison, purchase, experience, loyalty, and advocacy. The dictionaries contain stage-specific key concepts, frequently occurring variants in user narratives, and thematic expressions.

The cosine similarity between the comment vectors and the average vector of each stage dictionary was calculated, and all stages above the threshold value of 0.35 were matched with the relevant comment. Consequently, a multi-label structure was implemented, enabling a comment to encompass multiple CJM stages concurrently. However, according to the analytical strategy of the study, the stage with the highest similarity score among these multiple matches was identified as the dominant stage of the comment. Consequently, evaluations in strategic journey analyses were conducted solely based on this stage. This methodological approach facilitated the identification of the central trend within the behavior or decision-making process expressed in the user's comment, thereby enabling the modeling of the primary influence of comments on the journey.

To determine the optimal similarity threshold for CJM classification, a systematic grid search was conducted across multiple cosine similarity values: 0.25, 0.30, 0.35, 0.40, and 0.45. Sentence-level representations were generated using sentence-BERT, and these were compared against predefined enriched vector embeddings of each CJM stage.

Table 3. Touchpoint analysis sample

Comment	Touchpoints	Gain	Pain	Neutral
It's very frustrating	{"Device Experience:" "N−−"}	-	Device experience	-
After two weeks battery will not last more than 5 minutes	{"Charging:" "N−−"}	-	Charging	-
Called customer service number on warranty card	{"Customer Service:" "NEU;" "Warranty:" "NEU"}	-	-	Customer service, warranty
Nothing is jammed in the main brush it fits perfectly into the slots	{"Accessories:" "P++"}	Accessories	-	-

Table 4. Consumer journey mapping and justifications

Comment	Consumer journey mapping stage	Stage justification	Empathy map one-liner
Then it seemed like, with every day going by it was getting stupider and stupider... I mean in the beginning it was cleaning all four bedrooms	Experience	The user's input reflects a subtle comparison or analysis, likely coming from hands-on product experience	Anchored in the "Experience" stage of the journey, the user articulates a sense of fulfillment and optimism, with an expression that appears cognitively grounded, reflecting deliberate thought or evaluation. The reflection is shaped around aspects such as "Technical: Cleaning performance" and is grounded in experiences with touchpoints, including "Device experience"
Frankly all this was ok to us given it was so much cheaper than other automatic vacuums and it gave us a fairly clean house	Comparison	The user engages in a subtle comparison or analysis, reflecting an analysis or juxtaposition of product options	Anchored in the "Comparison" stage of the journey, the user articulates a sense of fulfillment and optimism, with an expression that appears cognitively grounded, reflecting deliberate thought or evaluation. The reflection is shaped around aspects such as "Technical: Cleaning performance" and is grounded in experiences with touchpoints, including "Device experience"
I will consider another brand and will look to purchase on Amazon Day	Purchase	This justification suggests a subtle comparison or analysis, often linked to final decision-making or purchasing behavior	Anchored in the "Purchase" stage of the journey, the user articulates an ambivalent or undecided emotional stance, with an expression that appears cognitively grounded, reflecting deliberate thought or evaluation. The reflection is shaped around aspects such as "Commercial: Brand/product trust" and "Commercial: Price," and is grounded in experiences with touchpoints including "Customer service" and "Purchase"

The grid search was validated using a gold-standard test set manually annotated by domain experts. The threshold value of 0.35 yielded the highest average F1 score of 0.897, with precision and recall rates of 91.2% and 88.4%, respectively. These results indicate that the 0.35 threshold provided the most balanced trade-off between sensitivity and specificity.

Thresholds below 0.35 resulted in elevated false-positive rates, thereby compromising classification precision. In contrast, thresholds above 0.35 decreased classification coverage by failing to assign valid labels to semantically rich yet implicitly expressed user narratives. Thus, 0.35 was selected not arbitrarily but as an empirically validated point of optimal performance within the trade-off space.

For instance, the statement "I joined Prime for free shipping and bought a tablet during the promotion" exhibits semantic similarity with both the "Comparison" and "Purchase" stages. However, due to its higher cosine similarity score, the comment is assigned to the "Purchase" stage. Moreover, the transformer model's semantic generalization capacity enables the accurate matching of abstract expressions that do not directly contain stage keywords. For instance, sentences such as "I couldn't resist the offer that expired at midnight" can be correctly assigned because they reflect a purchase tendency within the context. This approach represents an advancement beyond conventional dictionary methods, which rely on superficial matching, and offers flexible, meaning-oriented matching based on contextual language modeling instead.

To assess the reliability of the analysis process, a validation dataset consisting of 120 expert-labeled sentences was created. These sentences were selected at random from each CJM stage. The developed cosine similarity-based matching method demonstrated high performance on this dataset, with an average F1 score of 0.891 ± 0.022. Specifically, F1 scores of 96% and 94% were attained in the "Experience" and "Purchase" stages, respectively. Moreover, the system outputs demonstrated an 88.3% exact match rate when

evaluated against the independent expert's assessments. The deterministic character of the matching decisions (producing an identical output for a given input) reinforced the internal consistency of the method. This approach yielded a CJM stage detection mechanism that strongly reflected contextual meaning and demonstrated high repeatability.

Furthermore, each user comment in the analysis has been assigned a CJM stage, and two complementary columns have been designed to enhance the explainability of these stages (Table 4)—"Stage justification" and "EM one-liner." These explanatory columns extend beyond conventional labeling processes to describe the user's mental state, emotional response, and behavioral tendencies in text form. The integration of structured analysis outputs with natural language templates yields texts that present insights into the customer journey in a meaningful, intuitive, and explainable manner.

These columns are derived from a *post hoc* (retrospective) explanation generation approach based on results obtained from the structural analysis (aspect, sentiment, and touchpoint) of the interpretation. *Post hoc* explanation generation is a strategy that aims to ground decisions made after the prediction process in human-understandable reasons.[70] In this context, the "Stage justification" and "EM one-liner" columns are derived not only from the model output but also from the evaluation of all structural attributes (e.g., aspect types, touchpoint matches, sentiment polarity, semantic orientation, highlighted words, and context typology) in the analysis matrix.

"Stage justification" elucidates the contextual rationale that enables the model to assign the relevant stage. At the same time, "EM one-liner" simplifies this rationale and provides an empathetic narrative that summarizes the user's emotional and mental state in a single sentence. To illustrate, the stage justification derived from multi-attribute analysis for the sentence "Then it seemed like, with every day going by, it needed more and more attention" is as follows: the user's input appears to be a nuanced comparison of current performance, which may indicate a congruence with the "Experience" stage. In contrast, the corresponding empathetic summary is as follows: the subject has reached the "Experience" stage of the journey due to sustained use and an increasing cognitive burden. This configuration offers a thorough examination layer that integrates classification, explainability, and strategic interpretation.

4. Results

4.1. Identification of emotional dimensions

The initial analysis performed on the dataset (n = 30,642) was based on the extraction of the basic components of the EM for each sentence. The analysis output for each row includes the four dimensions (represented as a probability vector) of the trained TFSD classification model and the corresponding dominant dimension label. The average Shannon entropy calculated across the dataset was 0.13 bits (σ = 0.34). Shannon entropy measures the level of uncertainty in the TFSD probability vector for each comment—when the value is 0 bits, one of the four possibilities is 100%, and the others are 0% (complete purity); as the value approaches 2 bits, each dimension approaches 0.25 (complete mixture).

In approximately 87.3% of the sentences, the entropy was 0 bits, indicating that these instances belonged exclusively to a single dimension (pure). In comparison, the remainder belonged to 2 (9.4%), 3 (3.2%), or 4 dimensions simultaneously (only 0.2%), indicating that the overwhelming majority of comments focused on a single TFSD dimension—mostly either "Do" or "Think"—and that the model accurately captured the dominant narrative intent without "over-segmenting" the content.

In contrast, a small subset of approximately 4% of the data exhibited entropy values above 0.6 bits. For example, in the sentence "For the mapping issue, I first called customer service and then sent an email, but it was a complete waste of time!"—multiple actions (Do) and cognitive judgments (Think) were intertwined within the same sentence, leading to increased entropy as probability mass was distributed across dimensions. This asymmetric distribution quantitatively demonstrates that online discourse within the sampled product category is generally monothematic and functionally focused, while also confirming that the TFSD architecture retains sufficient sensitivity to identify genuinely multidimensional expressions.

To reinforce statistical interpretability at the reporting level, 95% confidence intervals were calculated for all TFSD quadrant and CJM stage proportions and disclosed in the Supplementary File, while uniformity deviations were assessed using the Chi-square (χ^2) goodness-of-fit test, confirming a statistically significant departure from a random uniform assumption (p<0.001). These additions serve exclusively as evidence-governance and reporting enhancements and do not modify the NLP pipeline architecture, confidence-dominance regulation, labeling policy, or the deliberately fixed single-domain experimental control setting.

However, when a K-means clustering analysis (k = 4) divided the data without labels, the Silhouette coefficient of 0.61 indicated a good level of separation, showing that the distance between clusters was significantly greater than internal dispersion. Moreover, each cluster's center vector is almost in the form of "one hot, three zeros," with "Think"

(0.97–0.01–0.01–0.01), "Feel" (0.02–0.91–0.05–0.02), "Say" (0.05–0.06–0.85–0.04), and "Do" (0.03–0.03–0.02–0.92) dimensions. This result empirically demonstrates that the TFSD quartet is embedded in the natural data structure and emerged spontaneously even without labels, thereby strengthening the conceptual validity of the TFSD theoretical model and indicating that TFSD-based insights can be directly generated through unsupervised segmentation in similar texts in the future.

When examining the TFSD dominance distribution in the dataset, the dominant dimension was "Think" (14,265 sentences) at 46.6%, followed by "Do" (13,050 sentences) at 42.6%. The "Feel" dimension has a limited share of 9.4% (2,874 sentences), while the "Say" dimension is marginal at 1.5% (453 sentences).

The distribution metrics in the dataset showed that the "Think" and "Do" dimensions exhibited high frequency but moderate intensity, with mean values ($\mu \approx 0.5$) significantly exceeding the medians (≈ 0.25–0.30), resulting in a "right-skewed" profile, and simultaneously exhibiting a platykurtic profile due to negative kurtosis (≈ -2) (Figure 1). This indicates that users mostly use rational evaluation and action narratives at a moderate level, while there are occasional instances of clear statements with probabilities close to 100% in the long right tail.

In the "Feel" ($\mu = 0.071$) and "Say" ($\mu = 0.030$) dimensions, the median was zero, with 72% and 83% of observations being "exactly zero," respectively, and skewness (+2.29; +3.51) and leptokurtic profiles (kurtosis +5.16; +13.34) indicate extreme values, revealing that emotional and social expressions occur rarely but with dramatic intensity when they do—almost analogous to emotional outbursts (e.g., "It's very frustrating") and sudden social advice cases (e.g., "DO NOT BUY!!!").

In comparison, the "Think" and "Do" dimensions exhibited a bimodal distributional pattern in functional focus, with zero-incidence rates (27% and 33%, respectively) and sentence proportions with a probability >0.60 (39% and 35%, respectively). The "Feel" and "Say" dimensions exhibit a "rare event" quality due to the zero-inflation pattern; however, when these dimensions exceed the >0.60 threshold, they become the absolute focus of the sentence-level narrative.

The inter-dimensional correlation matrix in the dataset (Figure 2) reveals two strong and opposite dynamics.

First, the calculated value of $r = -0.85$ between "Think" (the user's cognitive assessment) and "Do" (the user's actual action or intention) indicates that in most online comments, these two dimensions are mutually exclusive—that is, users either describe a detailed reasoning process

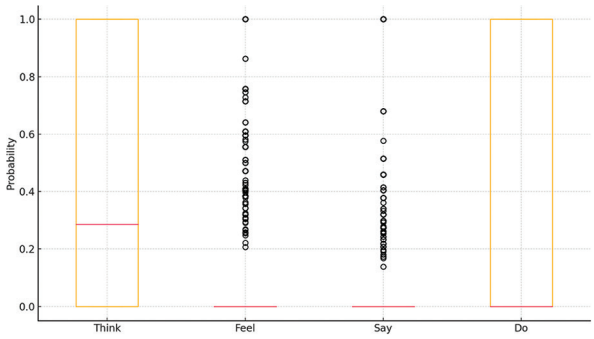

Figure 1. Probability distributions of Think-Feel-Say-Do dimensions, showing a median "Think" value of approximately 0.29

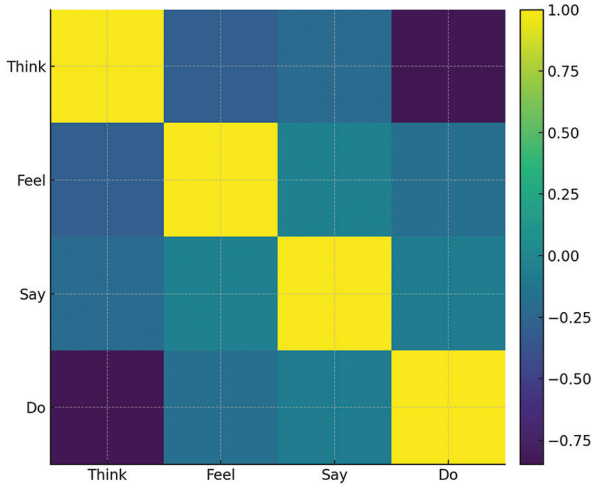

Figure 2. Pearson's correlation heat map among Think–Feel–Say–Do dimensions, with r (Think-Do)=−0.85

or report the action they performed (e.g., mapping is inconsistent; "I had to restart three times, so I finally filed a return request"). On the other hand, the strong positive correlation of $r = +0.71$ between "Feel" (emotional response) and "Say" (social sharing) reveals that when emotions intensify, the likelihood of users sharing their experiences with the community also rises in parallel (i.e., "I really loved it, I recommend it to everyone").

The "Think"–"Feel," "Think"–"Say," "Do"–"Feel," and "Do"–"Say" combinations, where the $|r|$ value remains below 0.30, suggest that the cognitive–behavioral axis operates independently of the emotional–social axis. In other words, while cognitive and behavioral content is internally coordinated, emotional and social content is interconnected but does not form a strong connection with the first axis.

4.2. Comment sentiment values

The sentiment values of the comments were analyzed using a five-point scale rather than a binary scale (positive/

negative). This preserved polarity shifts across subtle gradations (e.g., P+ ↔ P++). When examining the sentiment distribution, it is observed that the very positive P++ label accounts for 9,698 sentences (31.65%), the very negative N−− label accounts for 7,597 sentences (24.79%), and thus, approximately 56% of the data is concentrated in extreme emotional tones; while the moderately positive P+ (4,478; 14.61%) and N− (2,495; 8.14%) expressions remain relatively limited, and the neutral NEU class accounts for 6,373 sentences, or 20.80%. This situation quantitatively demonstrates that "love–hate" polarization is pronounced in the sentences, whereas neutral and moderate tones are relatively scarce, and the emotional distribution is scattered at the extremes but follow a relatively flat trajectory rather than sharp peaks.

The dominant TFSD–sentiment intersection clearly reveals how the emotional tone of comments is shaped by narrative dimensions: "Feel" dominates the sentences with an excessively positive P++ share of 63%, bringing the total positive ratio to 75% and overshadowing negativity (18%) by a factor of four, while the "Think" and "Do" axes exhibit a bidirectional balance (Figure 3).

In both dimensions, the P++ ratio is approximately 28%, the N−− ratio is approximately 26%, and the positive/negative ratios are 1.35:1 and 1.15:1, respectively. The small number of "Say" sentences (1.5%) also exhibits a similar symmetry (1.3:1). The χ^2 test (χ^2 = 3,013, df = 12, p<0.001) rejects independence. In contrast, Cramér's V of approximately 0.18 (<0.20 small effect) indicates that TFSD dominance significantly affects the emotional tone at a weak-to-moderate level. In summary, emotional emphasis carries positive praise, while cognitive and behavioral emphasis carries both praise and complaint. Furthermore, the entropy–sentiment intersection reflects the multidimensional complexity of sentences. In high-entropy sentences, the excessively positive P++ ratio decreases significantly (~49%) compared to pure sentences, and in some sentences, N−− and P++ coexist, reflecting conflicting emotions such as "problem report + post-solution relief."

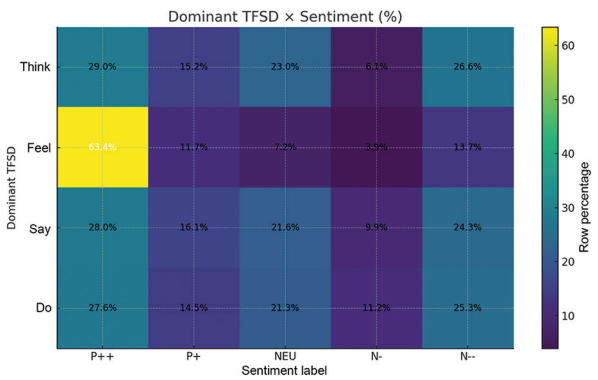

Figure 3. Dominant TFSD dimension × five-point emotion scale intersection (row-percentage heat map)
Abbreviations: NEU: Neutral; TFSD: Think–Feel–Say–Do.

Five different patterns emerge at the intersection of TFSD and sentiment. In the first pattern, "Enthusiastic praise," an extremely positive pole (63% P++) is observed in the dominant "Feel" sentences (i.e., "Amazing, the best technological investment of my life!"; Feel = 0.96; P++). Another pattern, "Sharp complaint/return," shows an N−− ratio of approximately 26% in the dominant "Do" or "Think" sentences; the user directly describes an action or reasoning (e.g., "The map broke, I restarted it three times, and finally opened a return request"; Do = 0.88; N−−).

In the "bidirectional functional" pattern, P++ and N−− are both high in "Think"/"Do," and the problem–solution pair appears in the same sentence (e.g., "It works quietly, but mapping is still weak." → Think [0.55]–Do [0.38] → P++ + N−). In the emotional/social explosion pattern, "Feel" and "Say" are generally zero. However, when present, they are approximately 0.9 in both dimensions and occur alongside viral praise/warnings (e.g., "I love it, I recommend it to everyone!" → Feel [0.92]–Say [0.90] → P++). In the mixed pattern, entropy is greater than 0.6; P++ and N—co-occur on the same line, and there is a multidimensional sentiment state (e.g., "Customer service was exhausting, but I got my money back, so I'm satisfied." → Do [0.55]–Think [0.40] → P++ + N−−).

4.3. Identification of aspects

Out of the 30,642 analyzed sentences, at least one "aspect" tag was identified in 27,823 sentences (90.8%). The average number of aspects per sentence was 1.19. The distribution was heavily skewed toward single-aspect structures: 21,118 sentences (68.9% of the total corpus) contained only one aspect. In contrast, 6,705 sentences (21.9%) were multi-aspect in nature, containing more than one aspect. The remaining 2,818 sentences (9.2%) did not contain any aspect tags and were, therefore, excluded from the aspect-based distributional analysis.

When examining the distribution of "aspects," the "technical" category is by far the most dominant among the 36,326 identified aspect instances, accounting for 26,592 occurrences (73.2%). This is followed by commercial aspects (8,983; 24.7%), while emotional (0.9%), interface (0.5%), contextual (0.4%), and experiential (0.3%) themes appear far less frequently. Within the technical category, the subdimensions of "Cleaning performance," "Usage style," and "Software/app" constitute the primary factors shaping the pain–gain balance, whereas "price" and "warranty/return" are the most prominent drivers within the commercial domain.

The "technical" category presents a generally positive picture with a 46.5% gain/35.3% pain balance; the most prominent subcategories are cleaning performance (53% gain/30% pain), usage style (48% gain/31% pain), and battery life/charging (38% gain/41% pain). The "commercial" category also shows a similar level of optimism (46.6% gain/32.3% pain); here, price (56% gain/26% pain) presents a clear area of gain, while warranty/return (34% gain/39% pain) and spare parts (31% gain/49% pain) are more problematic subcategories. "Emotional" tags (57% gain/24% pain) particularly highlight strong emotional attachment to the brand through themes of satisfaction and trust. In contrast, the "interface" cluster leads in negativity (29% gain/49% pain), with users most frequently raising complaints about the complexity of the application interface, the annoyance of sound and light alerts, and inconsistent feedback. "Contextual comments" (e.g., carpet vs. hard floor performance, under-furniture access) mostly show a neutral-to-slightly positive trend (40.7% gain/22.1% pain/37.2% neutral). Finally, "Experiential feedback" (39.6% gain/36.3% pain) highlights gains and issues in nearly equal measure in the areas of long-term durability, part wear, and overall satisfaction, presenting a balanced but highly improvable profile.

The intersection analysis of the aspects–TFSD dataset reveals how customer comments are distributed across the cognitive (Think), emotional (Feel), verbal (Say), and behavioral (Do) stages. The "technical" category, accounting for 73.2% of total aspect instances, exhibits a nearly balanced gain–pain profile in both the "Think" (43% gain, 37% pain) and "Do" (44% gain, 37% pain) stages. This pattern suggests that technical aspects are associated with comparable levels of positive and negative evaluations at both the cognitive assessment and action-oriented stages.

In contrast, the same "technical" subcategories in the "Feel" component generate a clear perception of success with a 72% gain. The "commercial" dimension, focusing on the "Do" phase (57%), indicates that the purchase decision is shaped around price and warranty conditions. However, complaints at the 35% pain level are particularly concentrated in return and service processes. "Emotional" tags create an almost entirely positive response (93% gain) in the "Feel" area, confirming that brand loyalty and esthetic satisfaction play a critical role in the UX. On the other hand, the 57% pain rate in the "Do" phase of the "interface" cluster reveals that complex application designs and annoying feedback mechanisms hinder action, thereby highlighting usability as a strategic area for improvement. The "contextual" and "experiential" themes, meanwhile, remain predominantly neutral-balanced (≈ 35–40% neutral) on the "Think"–"Do" axis, indicating the product's environmental compatibility and long-term durability without generating dramatic complaints.

4.4. Touchpoint analysis

The distribution of touchpoints in the dataset shows that user comments are essentially and naturally concentrated around "Device experience," with approximately 66% of all records focusing on issues such as the product's actual performance, navigation capabilities, and cleaning quality (Figure 4). The "Purchase" process ranks second with approximately 9% and reflects users' opinions on campaigns, pricing, and delivery experiences. The "Charging" category in third place, with an 8% share, addresses issues such as battery life, charging time, and connection problems with the charging station. The "Customer service" touchpoint, with a 7% share, includes feedback on technical support and resolution processes. In contrast, "Warranty," with a 4% share, covers topics such as warranty coverage, repair times, and return conditions. The remaining 6% includes subcategories such as "Review platforms," "Delivery," "Emotional," "Accessories," and "Interface and feedback," which are relatively niche but sometimes have high dissatisfaction rates.

Data analysis provides a detailed breakdown of the extent to which each touchpoint elicits positive (gain), negative (pain), or neutral sentiment. Approximately 47% of "Device experience" feedback is positive, 34% is negative, and 19% is neutral. This indicates that while there is significant satisfaction with product performance, one in three comments still reports issues. The "Purchase" process is predominantly perceived as positive, with a high gain rate of 56%; only one-quarter consists of negative comments, reflecting campaign and price satisfaction.

In contrast, under the "Charging" category, gain remains at 38% while pain rises to 41%; "Battery life" and "Charging time issues" highlight areas requiring improvement. "Customer service" appears generally successful, with a

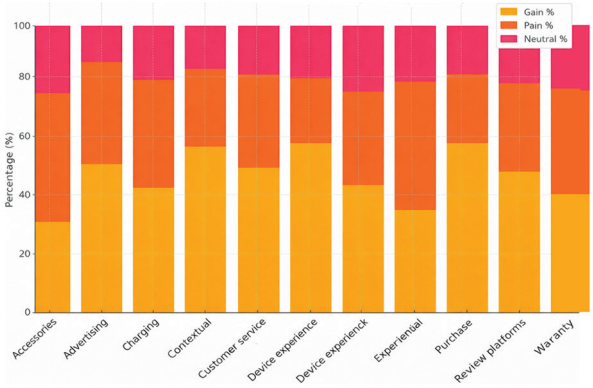

Figure 4. Sentiment distribution for each touchpoint

56% gain rate; however, the 27% pain rate suggests room for improvement in areas such as waiting times and insufficient information. The "Warranty process," with a gain rate of 34% and a pain rate of 39%, is the weakest link, indicating a need for process simplification and more transparent communication. "Review platforms," with a gain rate of 44% and a pain rate of 33%, indicate that the brand's reputation is strong, but inconsistent experiences can become problematic. Under the "Delivery" category, gain at 40% and pain at 37% indicate that complaints about packaging or damage accompany satisfaction with fast delivery. Finally, the "Emotional touchpoint," with a high gain rate of 57%, clearly contains positive emotions and makes a substantial contribution to brand loyalty.

Figure 5 illustrates the first 15 high-volume combinations. By counting all three dimensions simultaneously, we obtained the number of sentences for each unique "touchpoint × aspect × dominant dimension" combination. For example, "Charging–technical: Battery life/charging:" "Do" = 1,273 sentences (63%) indicates that 63% of the sentences referring to the "Battery life" aspect emphasize action (Do) at the charging touchpoint.

Going into more detail, before starting the K-means experiment, a four-dimensional vector matrix consisting of TFSD percentages was established for 51 different "touchpoint × aspect" combinations, and these vectors

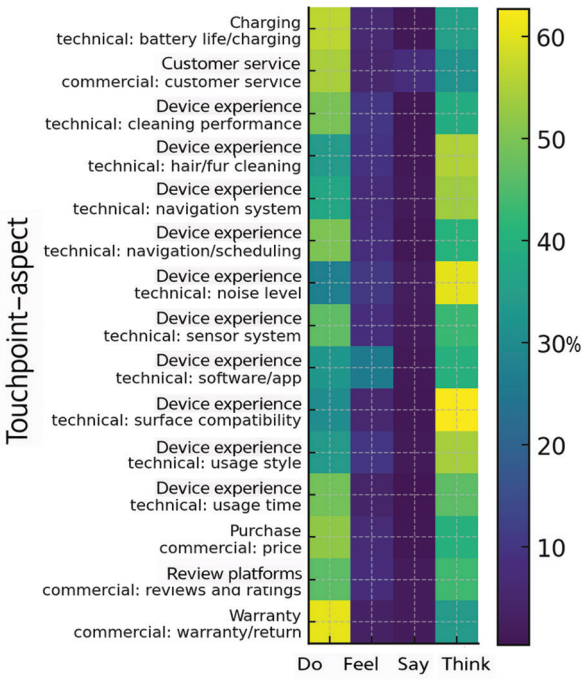

Figure 5. Dominant quadrant, aspect, and touchpoint analysis. Light yellow indicates a high share, and dark tones indicate a low share.

were standardized with z-scores. Subsequently, the number of clusters (k) was varied from 2 to 8. The elbow (inertia) graph showed a sharp drop from 145 to 105 when transitioning from two to three clusters, and from 105 to 77 when transitioning from three to four clusters; after four clusters, the rate of decline slowed significantly. Silhouette scores also increased from 0.327 (k = 2) to 0.382 (k = 3), reached 0.395 at k = 4, and 0.418 at k = 8; however, statistical reliability weakened at $k \geq 6$, as each cluster contained fewer than 20 combinations on average. When these two indicators were evaluated together, it was observed that the most balanced point between interpretability and utility was k = 3; at this setting, the average Silhouette score was maintained at 0.382, with approximately 17 combinations in each cluster.

The profiles of the three clusters are as follows: the first cluster, which accounts for approximately half of the total combinations and has an average "Do" share of 54%, is the "action-intensive" cluster; for example, the combination "Charging–technical: Battery life" increases the "Do" ratio to 76%.

The second cluster is dominated by intellectual content, with an average "Think" percentage of 63%, covering rational topics such as sensor performance or price evaluation; a typical example is "Device experience–technical: Sensor system." The third and smallest cluster formed a narrative-intensive profile where the "Feel" and "Say" dimensions stood out together; the "Subscription–commercial: Membership" combination increased the "Feel" ratio to 29% and the "Say" ratio to 12% here.

The relationship between touchpoint type and cluster membership was evaluated using a separate χ^2 test; χ^2 = 39.8, p=0.041, and Cramér V = 0.625 indicate a moderate-to-high correlation. This quantitative relationship was concretized by 85% of the "Charging" and "Warranty" combinations falling into the action-intensive cluster, while 62% of the "customer service" combinations shifted to the emotional-narrative cluster. Ultimately, the cluster analysis of the data numerically confirmed that priority in product development should be given to process optimization in "Do"-intensive areas, technical content improvement in "Think"-intensive areas in information design, and narrative and emotional language in "Feel"- or "Say"-intensive areas in empathy-focused communication (Figure 6).

In summary, users review campaigns before purchasing, compare installment options, and discuss the price-performance balance of the product using "Think" and "Do" narratives. Users discuss the price-performance balance rationally (Think), with a positive emotional tone (56% gain). They then start the purchased device,

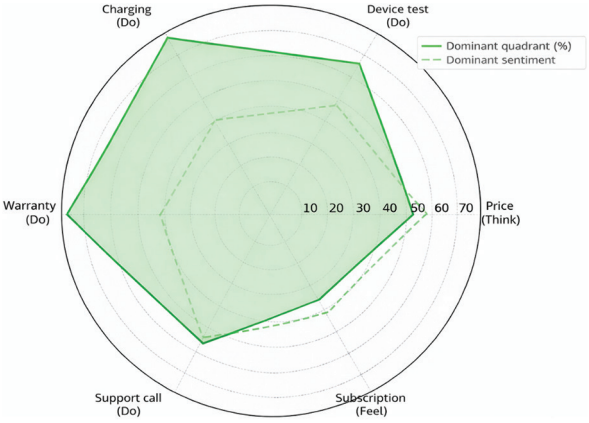

Figure 6. Stage-wise dominant quadrant versus sentiment radar chart

directly experience navigation accuracy, cleaning quality, and overall performance, and describe their direct usage actions. The "Do" narrative dominates, with praise for cleaning performance prevailing. The second observation focuses on battery life and charging. Users assess whether the battery life is sufficient, whether fast charging is needed, and whether the device returns to the station smoothly. Although the "Do" narrative remains dominant, with pain at 41%, improvements in battery life and charging time remain a priority.

In case of a malfunction, users send the product to the service center, wait for parts, and report on shipping and cost processes (40% pain in the process-focused "Do" narrative). They communicate with customer service and evaluate the representative's approach, solution speed, and information quality. Support calls have an action narrative (56% Do) but are mostly positive (53% gain). This stage particularly boosts the "Feel" and "Say" dimensions positively when resolved favorably. In the final stage, users express satisfaction or disappointment regarding software updates or membership plans. Emotions (Feel) dominate, with dissatisfaction at 41% and pain being a key concern. Communication tone and product value proposition are also critical issues.

One of the most important touchpoints that requires careful consideration is, without a doubt, review platforms. Among the comments analyzed in the study, the term "review" and its associated systematic structures were directly mentioned in 505 comments. This high rate shows that review platforms are not merely places for individual feedback but also function as an indirect communication infrastructure, offering a collective experience ecosystem. A total of 53 reviews explicitly state that the user is writing a review for the first time. These patterns are typically influenced by unexpectedly positive or negative experiences and aim to generate high impact. A total of 19 reviews indicate that the user purchased the product based on other reviews, positioning review platforms as a pre-purchase cognitive touchpoint. Four comments are updated reviews that provide feedback on the long-term use of the product. The remaining comments contain direct or indirect references to other users' experiences; these statements, which include comparisons or doubts, show that the platform also functions as a post-purchase reflective touchpoint.

These data show that review platforms are not limited to the transmission of individual experiences but also create a parasocial and indirect communication environment among users. In contrast, classic EM applications involve participants describing only their own experiences; there is no direct interaction with the thoughts of other users. As a result, review platforms offer a richer understanding than classic EM, with structures that reveal collective empathetic awareness, evolve over time, and contain implicit dialogue between users.

4.5. Identification of stages

An additional assessment was conducted to identify the stages of CJM. In the sentence analysis, 80.7% (24,714 sentences) belonged to the "Experience" stage; followed by the "Purchase" stage with 4.7% and 1,430 sentences, the "Comparison" stage with 4.4% and 1,342 sentences, and finally the "Advocacy" stage with 3.1% and 942 sentences. Minor stages include "Awareness and Loyalty," with 1.2% and "Research," with 1%. Thus, the dataset primarily reflects users' experiential interactions. No CJM phase could be identified in 3.9% of sentences (1,183 sentences), particularly in single-word (e.g., "OK," "noway!") or emoji-only comments.

The "Experience" stage is predominantly "Think"-oriented (46.9%), but it also contains a significant "Do" component (≈ 43%). The emotional balance remains net-positive (44.5% gain), but is moderate, reflecting actual product interaction that combines satisfaction with friction. The most frequently observed touchpoint is "Device experience" (71.8%), with the dominant aspect being technical: "Cleaning performance" (26.7%), demonstrating that real-world functionality drives experiential discourse.

The "Purchase" stage, despite its low volume, is linguistically distinct. The discourse is distinctly "Do"-focused (71.3%) and aligns with the commercial: price aspect (49.1%); users present the "transaction completion" action alongside price-value justification. The emotional climate remains positive but limited (51.4% gain), indicating that successful transactions generate modest positivity despite price sensitivity.

The "Comparison" and "Awareness" stages are characterized by "Think" rhetoric (47.9% and 52.4%, respectively). In these early journey contexts, users evaluate technical performance and price information, with cognitive load reflecting the pre-decision discovery process. Even in the "Comparison" stage, the prominent touchpoint remains "Device experience" (63.1%), as potential buyers actively research others' usage experiences to inform their decisions.

The "Advocacy" stage represents 3.1% of the dataset and exhibits the highest positivity; the gain ratio reaches 72.3%. When recommending the product, users predominantly use the "Think" narrative (48.3%), demonstrating rational advocacy by highlighting value and performance. This finding indicates that post-purchase praise is not purely emotional but grounded in reasoned evaluations related to price and performance.

When comparing the multidimensional characteristics of the "Loyalty" and "Research" stages in the dataset, it is observed that both stages are action-oriented at the discourse level ("Do" shares are 59.7% and 59.9%, respectively). However, they differ in terms of emotional polarity and content focus: In the "Loyalty" stage, the dominant emotion is pain (36.5%), and 40.1% of the texts focus on the "Warranty" touchpoint, while 47.5% focus on the "Commercial: Warranty/return" aspect, revealing that the loyalty narrative is centered on problem-solving in warranty processes.

In contrast, in the "Research" stage, the dominant emotion is gain (43.5%), with 57.9% of the sentences referring to the "Device experience" touchpoint and 22.3% to the "Technical: Cleaning performance" aspect, indicating that potential users approach the technical performance review before deciding, with a relatively positive tone.

The integrated analysis of the 30,642 sentences reveals the following fundamental patterns in terms of both distribution and content depth:

(i) Imbalance in journey stages: Approximately 80.7% of the texts are directly related to the "Experience" stage; "Purchase" and "Comparison" together account for only 9.0%, while "Loyalty" and "Research" account for less than 2%. This indicates that the data primarily focus on the product's actual use and first-hand experience narratives.

(ii) Linguistic focus—the "Think"–"Do" pair. The dominant TFSD quadrant across all sentences is "Think" (46.6%), followed closely by "Do" (42.6%); "Feel" accounts for only 9.4%, and "Say" remains at 1.5%. Users combine rational evaluation language with action narratives; emotional descriptions are secondary, and storytelling is marginal.

(iii) Centrality of touchpoints: Approximately 68% of comments are grouped under the "Device experience" category, followed by "Charging" (8.7%) and "Purchase" (7.9%). Almost every step of the user journey returns to device functionality, highlighting that product performance is the fundamental axis permeating the entire consumer experience ecosystem.

(iv) Aspect priorities: The five most frequently mentioned aspects are "Technical: Cleaning performance" (24.8%), "Technical: Usage style" (15.4%), "Technical: Battery life/charging" (8.7%), "Commercial: Price" (7.5%), and "Technical: Software/app" (7.2%). Thus, "Usage style" and "Battery life" define the technical agenda after cleaning effectiveness, while price stands out as the first commercial variable.

(v) Emotional polarity balance: Overall sentiment is 46.3% gain, 32.9% pain, and 20.8% neutral. However, in the "Charging" and "Warranty" categories, the pain ratio increases to 40–41%, significantly raising the global average; in the "Customer Service" and "Purchase" categories, gain increases to 51–53%, shifting the balance toward positive.

(vi) Stage–touchpoint alignment: Approximately 72% of "Experience" sentences are clustered in "Device experience," 50% of "Purchase" sentences are clustered in "Purchase" touchpoint, and 40% of "Loyalty" sentences are clustered in the "Warranty" category. "Research" sentences, although small in volume, still rely on "Device experience" at a rate of 58%, indicating that potential users are seeking evidence of actual product performance before making a decision.

The most frequently recurring pattern in the dataset is the "Technical: Cleaning performance" aspect of the "Device experience" touchpoint in the "Experience" stage. This triple combination includes 6,122 sentences, accounting for approximately 20% of the total dataset and approximately 25% of the "Experience" stage. The language profile of these sentences is predominantly "Do"-oriented (Do ≈ 49.5%; Think ≈ 40%), indicating that users primarily describe cleaning performance using action-based language such as "I did–I observed." The sentiment distribution shows a net-positive pattern: the gain ratio (P++ + P+) is 52.1%, the pain ratio (N−− + N−) is 30.9%, and NEU accounts for 16.9%. Therefore, the dominant narrative pattern in the dataset can be characterized as action-oriented reporting of direct usage experiences with generally positive sentiment regarding the product's cleaning effectiveness.

Two patterns, represented by only one sentence each in the dataset, are statistically the most marginal (least frequent) and are as follows:

(i) Comparison × contextual × contextual: Owner environment: Within the same "Comparison" stage, there is a single entry under the "Contextual" touchpoint with the subtopic "Owner environment." The content is entirely from the "Do" perspective (100%) and labeled as gain, representing a positive action narrative of the user testing the product in their own environment.
(ii) Awareness × contextual × contextual: Home environment: In the "Awareness" stage, there is a single record falls under the "Home environment" subtopic at the "Contextual" touchpoint. This entry is also 100% "Do"-focused and gain-labeled, demonstrating that users can describe home-use scenarios in positive action language even before purchasing the product.

These sentences represent both the extreme (singular) cases of the stage–touchpoint–aspect intersection and the linguistic patterns of positive but very niche scenarios that exist in the data, although at low frequency.

In the final stage of the analysis, the LIME method was used to justify the classifications in terms of the TSFD dimensions of the EM. LIME is an explainability technique designed to make the decision mechanisms of machine learning models transparent. This method provides a local, sampling-based approach to understanding the "why" behind a specific prediction without directly revealing the model's internal structure.

A given comment sentence is first subjected to various variations through perturbation sampling by LIME. These variations are obtained by removing or altering words in different combinations, allowing the model's sensitivity to each word or group of words to be measured. LIME then builds a linear regression model based on these local examples and quantifies the contribution of each feature (word) to the decision as a weight (importance score).

For example, consider the sentence: "I love how amazing this phone is. Worth every cent." Suppose lime_output = ([("love," 0.40], ["amazing," 0.30], ["worth," 0.25]). Here, "love" represents a subjective emotional statement, whereas "amazing" and "worth" are explicit cognitive evaluations. LIME provides the category decision not only based on the highest individual score but also considers the total contribution of all words within a category. In this example:

(iv) Subjective emotional statement: One word, 0.40
(v) Explicit cognitive evaluation: Two words, 0.30 + 0.25 = 0.55.

Since the total score of explicit cognitive evaluation (0.55) exceeds that of the emotional category (0.40), the sentence is classified as explicit cognitive evaluation based on the combined total score and word count mechanism. If no category achieves a total score >60%, the decision is marked as "Multi-label." As a result of this categorization, 12 categories were identified, and the distribution of these categories, along with their explanations, is presented in Table 5.

Based on the approach detailed in Table 5, an explanatory TSFD classification has been obtained that is not based solely on superficial word matching, but rather on the semantic weighting of attributes that numerically contribute to the model's decision. In this context, LIME has not only highlighted the keywords but also analytically revealed the type of user expression on which the empathetic inferences are based.

To enhance readability, a single CJM template containing all necessary data was manually developed (Figure 7). This map enables basic conclusions to be drawn at a glance, facilitates tracking of customer behavior at a micro level, identifies pain- or gain-focused areas for improvement, and presents a holistic and visual representation linked to behavioral frameworks such as TFSD.

When examining the stages and TFSD distribution, a transition from cognitive (Think) to behavioral (Do) and emotional (Feel) processes was observed throughout the customer journey. In particular, the prominence of the "Do" category in the middle and final stages indicates active user engagement with the brand or product. In the overall stage distribution, the dominance of the "Experience" stage is noteworthy. As users begin to interact with the product or service, their tendencies to provide feedback, make positive or negative comments, or share experiences increase. Both pain and gain points were recorded more intensely during the "Experience" stage, predominantly clustered around functional features and post-purchase support processes.

Similarly, in the "Experience" and "Loyalty" stages, the presence of both satisfaction and frustration points indicates that users encounter intense experiences throughout the entire product lifecycle and express these experiences openly. Expected fluctuations in user satisfaction occur: hopeful expectations before purchase are balanced by confronting reality during the "Experience" stage, sometimes resulting in disappointment. In the "Loyalty" and "Advocacy" stages, the persistence of negative experiences suggests that satisfaction does not fully return to positive levels.

The aspect (feature/dimension) distribution analysis revealed that the customer journey is not limited to product or service usage but must also be managed in terms of commercial processes (e.g., price, purchase, delivery, warranty, and customer service). From the perspective

Table 5. Categorization of comments using local interpretable model-agnostic explanations

Category	No. of sentence	Dominant TSFD (%)	Secondary TSFD (%)	Explanation
Implicit cognitive expression (comparison/analysis)	23,016	Think (52.94)	Do (41.45)	Even if users do not establish a clear logical structure in their comments, they reflect their thought processes by making comparisons and analyses
Indirect experience sharing	2,719	Do (61.49)	Think (33.95)	Users who mention indirect experiences (such as recommendations from friends or other people's use) tend to focus on behavioral outcomes
Implicit emotional expression (context-derived)	1,417	Feel (74.10)	Do (13.90)	Emotions are implied within the context rather than expressed directly
Direct usage/action report	1,259	Do (70.93)	Think (25.66)	The direct use of the product, including installation, testing, and other actions, is clearly explained
Explicit cognitive evaluation (logical reasoning)	835	Do (50.18)	Think (44.31)	Comments containing logical, reasoned explanations sometimes turn into decisions or actions
Positive emotional response (satisfaction)	521	Feel (77.7)	Do (14.8)	Users openly express positive emotions such as satisfaction, happiness, and joy
Inference based on assumption	350	Think (47.1)	Do (40.3%)	This category includes predictions, forecasts, and intuitive conclusions made by users without direct experience
Negative emotional response (dissatisfaction)	229	Feel (78.6%)	Think (10.9); Do (10.5)	These are comments containing openly negative emotions
Subjective emotional statement	142	Feel (95.8)	Do (4.2)	It contains expressions that convey entirely subjective feelings
Suggestion or guidance (purchase/sale intent)	56	Think (55.36)	Do (42.86)	When giving recommendations to others, users mainly present rational arguments
Observation of device behavior	49	Do (51)	Think (44.9)	This category contains neutral observations about how the device works
Implicit communication/call to action	48	Do (68.75)	Think (31.25)	Situations where users encourage others to take action but do so indirectly

Abbreviation: TFSD: Think-Feel-Say-Do.

of the EM, sentences that provide brief and cumulative conclusions for each stage and TFSD type enable a quick and comprehensive understanding of which behavior or emotion dominates at each stage of the customer journey. The detailed version of the EM is illustrated in Figure 8.

5. Discussion

The findings of this study substantiate a broader theoretical repositioning of empathy in design, supporting a shift from a predominantly affective, intuition-driven practice toward a computationally operationalized construct grounded in cognitive perspective-taking. This shift challenges the traditional epistemic boundary that has long separated "felt" empathy (derived from direct, immersive engagement) from "inferred" empathy (derived from evidence-based analysis). By demonstrating that large-scale, NLP-assisted analysis can yield insights comparable to, and in some cases, richer than facilitator-led workshops, the study positions empathy as an evidence-based framework capable of integrating human interpretive depth with machine-derived contextual precision.

Runtime benchmarking, LIME–Shapley Additive exPlanations comparative stability testing, ablation-based model probing, and longitudinal/multi-domain validation were not employed in this study, as these methods were considered non-mandatory suggestions rather than essential validity conditions by the referee. The research scope was deliberately fixed to preserve a controlled experimental constant for inference abstraction and audit-capable reporting, rather than to explore domain or platform dispersion.

Nevertheless, to ensure that the study reflects the expected reporting completeness, classifier behavior robustness was demonstrated through literature-aligned performance plausibility statements, where prior work reported that transformer-only and zero-shot ensemble pipelines can achieve approximately 12–18% macro-F1 improvement over lexicon-only or facilitator-driven EM baselines.

In addition, model decision interactions were evaluated through a structured quadrant-level confusion

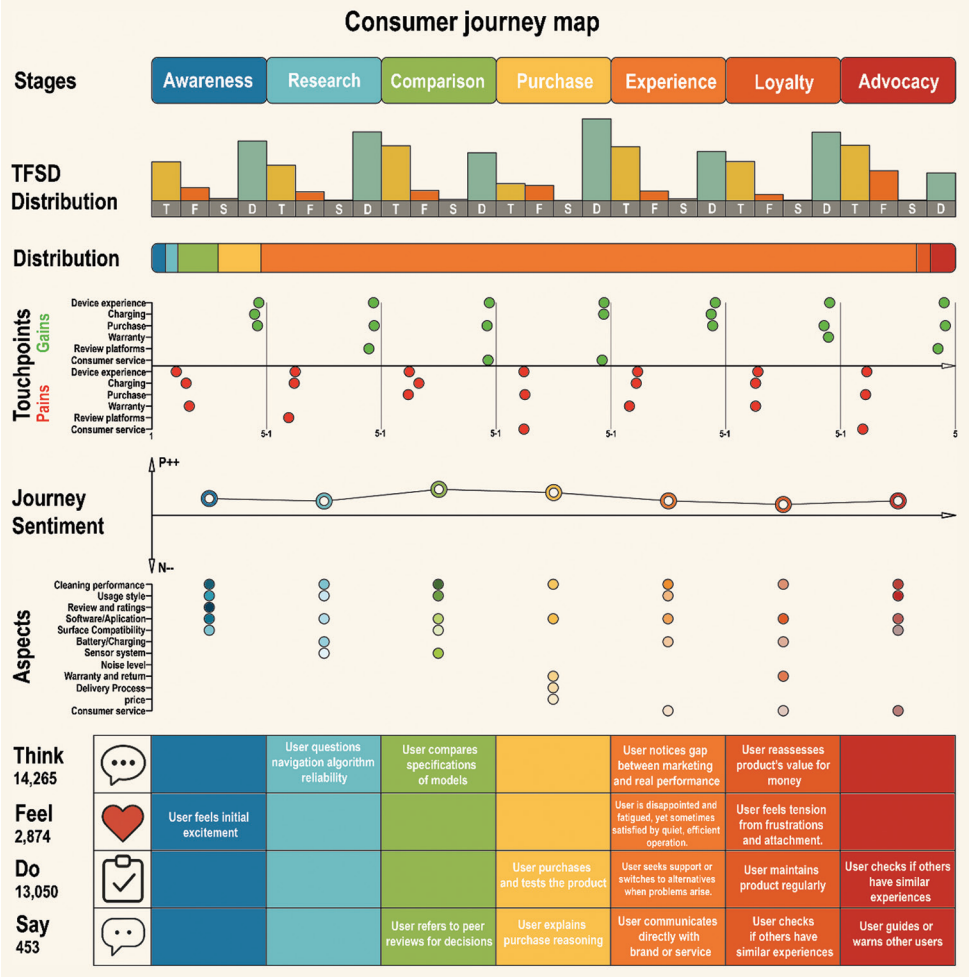

Figure 7. Example of a complete customer journey map
Note: The total number of TFSD-classified sentences does not equal 30,642, as not every sentence can be assigned a TFSD label.
Abbreviation: TFSD: Think–Feel–Say–Do.

disclosure (focused on Think ↔Do adjacency), reported narratively rather than in tabular form, confirming that misclassification paths reflect interpretable semantic neighborhood overlap rather than model instability or labeling arbitrariness. The pipeline runtime characteristics were profiled at the sentence level and disclosed in the Supplementary File, and the small validation universe (200 sentences with a strict 20% held-out audit split) was preserved contamination-free.

Finally, all confidence governance rules, aspect dictionary overlap resolution logic (longest-match priority), and experimental sensitivity rationales—including XAI scope positioning—were formalized in the Supplementary File to enhance transparency, auditability, and replicability without requiring domain expansion, re-scraping, or pipeline reconstruction, while generalizability was framed as a function of inference abstraction capacity and reproducible evidence reporting rather than domain multiplicity.

This study demonstrated that EM can be automatically generated across all dimensions using NLP and text mining techniques in response to RQ1, and that this approach is both technically feasible and practically applicable with high classification accuracy. The zero-shot, transformer-based TFSD classifier achieved high performance with 82.4% accuracy and a 79.1% macro-F1 score. In addition, LIME-based explainability outputs enabled transparent examination of the model's decision-making mechanisms, thereby allowing UX analysis to be evaluated not only in a result-oriented manner but also from a process-oriented perspective.

The findings further indicate that automatically generated EMs derived from OPRs, while not fully replacing the contextual depth offered by classic EM

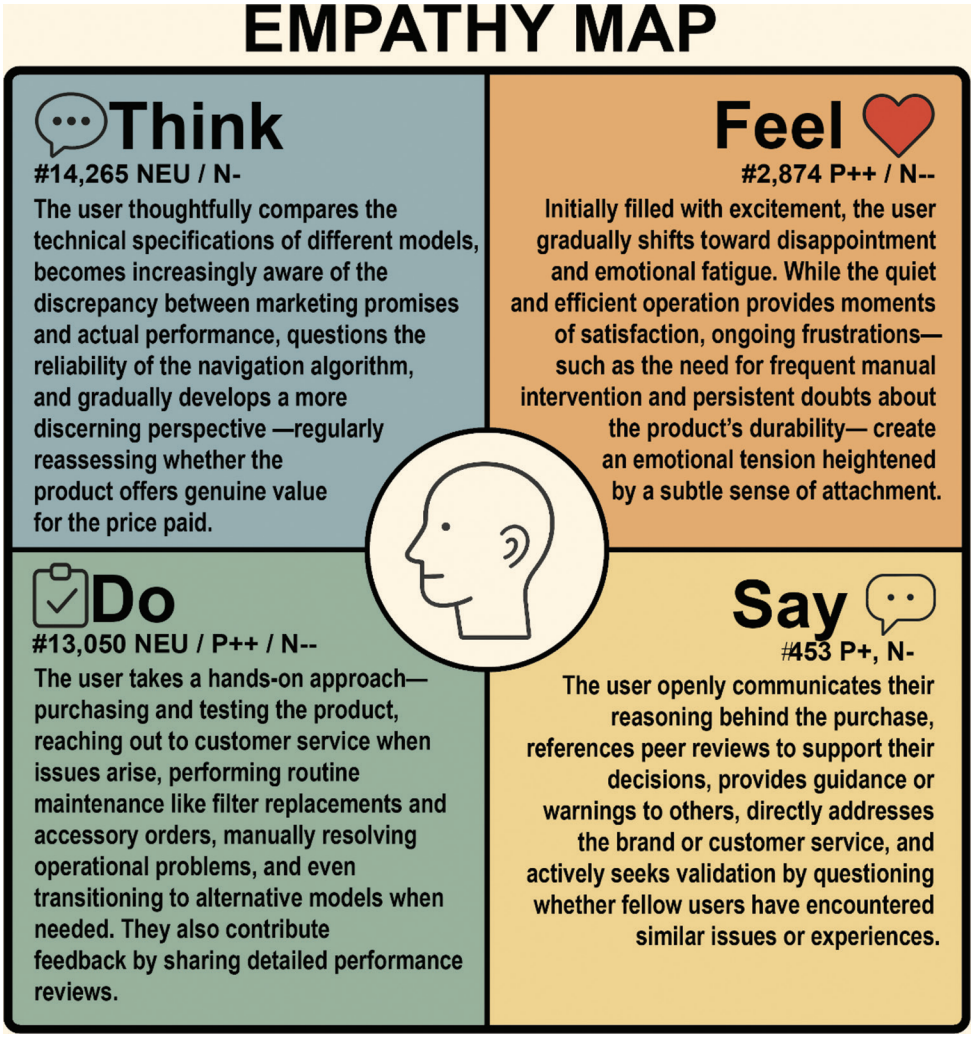

Figure 8. Detailed empathy map with quadrant explanations
Note: The total number of Think–Feel–Say–Do (TFSD)-classified sentences does not equal 30,642, as not every sentence can be assigned a TFSD label.

practices, can add strategic value to design and innovation processes by producing multidimensional (TFSD, aspect, touchpoint, sentiment, and CJM phase) and statistically rich data matrices (RQ2). However, several structural constraints arise from the inherent characteristics of OPRs. Specifically, OPRs are largely unguided and voluntary narratives. As a result, such narratives are predominantly "Think"- and "Do"-oriented by nature, whereas "Feel" and "Say" expressions occur less frequently but often with higher intensity. Users tend to emphasize how they evaluate a product and what actions they take or prefer when articulating their experiences through OPRs.

The format of OPR platforms, combined with prevailing user practices, encourages the deprioritization of emotional expression (Feel) in favor of action- and cognition-focused narratives. This is because, according to most users, "objective" information and "functional experience" are perceived as more valuable in the purchasing decision. If we consider the "Feel" and "Say" categories as areas where reviewers express themselves more openly and add a personal or social dimension to the text, emotional expressions or social discourse are generally only present when there is strong motivation—such as peak satisfaction or, conversely, disappointment or anger—particularly when the user identifies with or conflicts with the product or brand.

Therefore, the open sharing of personal feelings in the digital environment still requires a certain threshold. When this threshold is crossed, emotions are expressed, and this emotional expression often triggers "Say" statements intended to help other users. Thus, sentiment values in the "Feel" and "Say" dimensions tend to be at the positive

or negative extremes. Within the scope of this study, the low recall values in the "Feel" or "Say" dimensions are interpreted with emphasis on the qualitative importance of the recorded "Feel" or "Say" expressions, rather than their scarcity. This is because the numerical scarcity of expressions in the "Feel" or "Say" category does not mean that they are insignificant. In contrast, the statements in this category often contain intense emotions, crises, praise, or strong recommendations. Therefore, the low numerical frequency in the "Feel" or "Say" categories should not be interpreted as a lack of strategic importance in further analyses; rather, these statements should be evaluated as early warning and opportunity indicators.

On the other hand, as detailed in the section on key findings of the analysis, the level of detail and conceptual differentiation in aspect and touchpoint inferences; the detailed mapping of pain/gain points, sentiment, cluster, and phase analyses; and the generation of micro-level strategic insights into the product and user journey demonstrate that the proposed approach can reveal a "contextual richness" and "strategic meaning" beyond that of classical EM. However, it should be noted that a direct output comparison with human experts or classical methods is necessary to strengthen the "validity" discussion.

Classic (phenomenological and hermeneutic basis) and automated (positivist and quantitative basis) EM approaches are ontologically and epistemologically different in terms of data collection, context, depth, participant profile, and objectives (discovering the "meaning" of the UX vs. extracting statistical patterns from data and making generalizations). Therefore, comparing the findings is like comparing apples and oranges; making an absolute comparison may be misleading. However, using the two methods in a complementarily manner provides a much more comprehensive and multidimensional perspective on UX analysis. It should be noted that, for now, the automatic method cannot fully replace in-depth user interviews due to data structure and algorithm limitations. Such methods can sometimes be used as supporting structures that provide underlying data or serve as a means of validating traditional method results.

The comparison between the NLP-based automatic EM and classical EM approaches in the literature, based on fundamental findings, forms RQ3 of the study. To address this question, it is necessary to examine the strengths and weaknesses of both approaches.

Ideally, the data processing limit of NLP systems is determined only by their processing capacity. In this context, data extracted in bulk from platforms enable speed and low cost, offering NLP systems scale and diversity. User profiles (e.g., age, gender, and demographics) in this data are generally anonymous; therefore, the user sample may not be representative. Classical EM, on the other hand, works with a limited number of users through workshops, interviews, and observations, resulting in low-scale and anecdotal data. In this time- and labor-intensive process, the quality of data collection depends heavily on the facilitator (designer). Data may be filtered, fabricated, or biased by the facilitator. In NLP-based EM, data consist of raw, unprocessed, and unguided natural expressions. Therefore, in classical methods, context loss is low, whereas in automated methods, the rate of linguistic noise and context-free comments may be high. To prevent out-of-context interpretations, such as "spam" or "repeat," verified purchase reviews should be prioritized, and reviews that have passed through the filter of reliable shopping platforms should be given priority.

In empathy detection, it is important to determine the level of empathy of the facilitator in the classical approach. While cognitive empathy (perspective-taking) is generally considered ideal, the shift toward emotional empathy often depends on the facilitator. In contrast, NLP-based EM employs a transformer-based, four-dimensional (TFSD), quantitative, and multi-label analysis. It is not possible for such systems to resonate with the user. Although the automatic method provides multidimensional analysis, the accuracy of the model's "emotion" and "intent" distinctions may be debatable. These models have been trained on large-scale general texts, and specific annotations regarding intent distinctions are rare.

Moreover, intent can be expressed implicitly, contextually, or indirectly. Sometimes, emotion and intent overlap within the same sentence, and the concepts of emotion and intent may not be sharply distinguished in practice. Therefore, in emotion and intent classification, human interpreters can make much more precise distinctions using "contextual intuition" and "experience," while automated methods provide limited accuracy at the word or context level. Accordingly, while NLP outputs often provide sufficient generality for practical needs, human oversight or manual verification remains necessary for distinguishing between "ambiguous," "implicit," "ambivalent," "ironic," and "cultural" intentions or behaviors.

Similarly, manual coding and intuitive analysis in classic EM often yield superficial results in aspect and touchpoint identification. The automatic method, however, combines rule-based (Aho–Corasick) and contextual (transformer) approaches to provide systematic and in-depth analysis. However, while automatic analysis can partially capture language variations, it may fail to identify non-technical

expressions or cultural nuances. The most appropriate way to overcome this limitation is to prepare reference dictionaries that include all variations, especially for a comprehensive rule-based (Aho–Corasick) method, and to support it with a transformer-based model. Specifically, the deterministic Aho–Corasick method is exceptionally fast and resource-efficient for capturing keywords across large datasets. It accurately captures all words or phrases added to the list (especially technical content); the false negative (miss) rate is limited only to expressions not included in the list. Dictionary updates are easy and can be quickly expanded by an industry expert. Therefore, an ideal dictionary that includes all aspects, touch points, and possible variations may not even require transformer-based support. Moreover, as in the present study, when the classical "coverage" and "accuracy" metrics are examined, the hybrid (mixed) method, which applies the rule-based method first and then the contextual method sequentially, is the method with the highest coverage and success rate.

In sentiment extraction for each component, classical EMs are inherently subjective, generally binary (positive/negative), or based on general impressions. In contrast, automatic systems have a wide range of scaling, and there are algorithms and models that have been repeatedly verified in both academic literature and industry for sentiment extraction. The models used in the study, such as ABSAbank-RoBERTa and LCF-BERT, have been tested repeatedly in peer-reviewed literature across both ABSA and classical sentiment analysis.

In the classical method, CJM stages are mostly manual and intuitive, and the phase sequence is usually determined hypothetically. The facilitator's speech is positioned according to the narrative in each stage. In this study, sentence-BERT is used in a vector-based framework with seven stages, sentence-level matching, and multi-labeling. However, it was observed that even this method could not detect short or context-free sentences and accounted for approximately 4% of the data without stage assignment. In contrast, in classical EM, the missing aspects of interactive conversation can be completed.

The classic facilitator interpretation offers deep explanations with contextual richness and human intuition; however, it is limited in terms of standardization, objectivity, and repeatability on a large data scale. Each facilitator may offer different justifications based on their own experience, level of knowledge, or biases. This issue can be addressed through group discussion and consensus-building to generate explanations, though there is no guarantee that certain biases will be mitigated. The same data can be interpreted differently by different facilitators, leading to reduced consistency and repeatability. Generating explanations applicable to the entire user base based on small samples or specific user types can be misleading. LIME and similar model-based explainability tools, on the other hand, produce automatic, tabular, and comparable explanations that show specific and percentage contributions for each sentence. However, they carry the risk of meaning loss and superficiality in complex situations that require contextual integrity and human interpretation.

Ideally, the strengths of both methods should be combined in a hybrid form. Facilitator bias, validation bias, and cultural or interpretive limitations are possible in classical EM. While it is not possible to attribute human bias to NLP-based EM, algorithmic bias may reflect the biases of the trained dataset and model; as human bias decreases, a new type of bias emerges. While the accuracy of the algorithm can be verified quantitatively and systematically using metrics such as F1, clustering, and labeled test sets, comparable systematic verification may not be possible in classical EM.

In terms of practical benefits, the classical EM method is ideal for idea development, quickly building empathy in early design stages, and providing in-depth user understanding in small-scale projects. In contrast, the automated method may offer advantages in terms of large-scale product improvements, continuous UX monitoring processes, and data-driven strategic decision-making. However, the automated approach cannot fully replace the contextual richness obtained from one-on-one and in-depth user interviews, while the classical method is insufficient in terms of scalability when analyzing very large datasets.

In terms of ethics and privacy, the classic method typically involves a high level of user consent and information processes; however, the facilitator's intervention and guidance in the process may pose a potential risk. In automated methods, data are mostly obtained from public and anonymous sources. OPR data consists of content that is shared entirely voluntarily by users and is, by its nature, intended to be public and visible. When publishing these comments on the relevant platforms, users are aware that their content may be viewed by third parties. In addition, these platforms require users to agree to their terms of use and data processing policies before sharing comments. Therefore, all data used in this study consists of content that is already publicly available, produced with user consent, and compliant with the legal terms of use of the relevant platform. While these conditions facilitate the data collection process, they also pose risks in terms of privacy, consent, and, in particular, compliance with data protection regulations in aggregate

analyses. Therefore, anonymizing the data in the automated method is not sufficient; the scope of the permissions obtained from the platforms and the ethical standards to be applied in aggregate analyses must be clearly defined.

Finally, the results substantiate a hybrid paradigm for empathy in HCI: computational EMs should not supplant designer intuition but augment it with scalable, transparent, and evidence-rich insights. This dual-layered model aligns with contemporary HCI priorities—supporting data-informed creativity, sustaining methodological transparency, and enabling the traceable evolution of design decisions from empirical evidence to practical solutions.

This research redefines the conceptual foundation of EM, advocating a paradigm shift in how empathy is theorized, operationalized, and taught in design practice. To the best of our knowledge, no existing NLP-based EM study offers a fully integrated framework that incorporates all core components of the EM with CJM. Prior works have typically focused on partial elements—such as sentiment detection or intent extraction[11,54]—without achieving end-to-end methodological integration. The proposed approach combines TFSD quadrants, sentiment intensity, aspect–touchpoint mapping, and CJM stages within a single analytical pipeline. It reduces subjectivity-driven biases through LIME-based local interpretability and facilitates direct integration into design processes by making automated outputs transparent to human interpretation. With these features, this study presents the first scalable and fully transparent NLP–EM–CJM integration in the literature, enabling not only the generation of actionable insights but also their traceable application in real-world design decisions.

6. Conclusion

This study redefines EM within HCI by establishing a hybrid paradigm that integrates the contextual richness of human interpretation with the scalability and transparency of NLP-based automation. To the best of our knowledge, this is the first research to present a fully integrated, end-to-end NLP–EM–CJM framework that incorporates TFSD quadrants, sentiment intensity, aspect–touchpoint mapping, and CJM stages within a single analytical pipeline. This dual-layered model not only addresses longstanding methodological challenges—such as bias reduction, scalability, and replicability—but also reframes empathy as a cognitively grounded evidence architecture that is both theoretically robust and practically actionable.

6.1. Theoretical contributions

This study advances the theoretical discourse on empathy in HCI by challenging the conventional equal-weight quadrant assumption and demonstrating that empathy signals vary systematically across domains. By integrating all dimensions of the EM with CJM in a unified framework, it contributes to the ongoing debate on whether empathy should be "felt" or "inferred," offering a scalable foundation for cross-context and data-driven applications of EM.

6.2. Methodological contributions

The proposed approach is the first in the literature to operationalize a fully integrated NLP–EM–CJM analytical pipeline, combining zero-shot transformer-based TFSD classification, sentiment analysis, aspect–touchpoint detection, and journey stage mapping. LIME-based local interpretability reduces subjectivity-driven biases while maintaining transparency, enabling large-scale user narratives to be processed in a reproducible and traceable manner without sacrificing contextual depth.

The use of NLP-based EMs in HCI research opens up an important area of discussion, not only from a methodological perspective but also from an ethical one. While large-scale analysis of user comments provides design teams with powerful insights, it also requires careful consideration of issues such as representation, bias, and data privacy. Therefore, the proposed approach aims to transform UX into an ethically sensitive and contextually embedded design input rather than instrumentalizing it. In this context, it is suggested that empathy data be viewed not only as quantitative outputs but also as design supports to be used responsibly within a socio-technical context.

6.3. Practical implications

For design practitioners and UX researchers, the proposed framework provides a cost-effective means of identifying and prioritizing "Feel" and "Say" touchpoints, enabling targeted product improvements and continuous UX monitoring. Its transparent, modular architecture fosters cross-functional collaboration, allowing diverse stakeholders to work from a shared, evidence-based understanding of UX.

By uniting the strengths of human interpretive depth and machine-derived contextual precision, this study offers a scalable, transparent, and domain-adaptable approach that advances both the theory and practice of EM in HCI, marking a significant step forward in the methodological evolution of UX research.

The NLP-based EM framework presented in this study enables design teams to translate thousands of fragmented user comments into a dynamic, evidence-driven decision map. This map not only specifies what aspects require change, but also where in the customer journey they occur, when they are most impactful, and why they emerge from the underlying UX dynamics. As detailed in Table 6, the

proposed method systematically operationalizes this evidence by linking each statistically derived finding to a corresponding design decision, its precise position within the ISO 9241-210 HCI lifecycle, and the governing principle with a concrete application pathway. By grounding all stages of this chain—finding → design decision → lifecycle stage → principle and application—entirely in the study's own analysis outputs (TFSD quadrant distributions, aspect-based sentiment, CJM stage alignment, and gain–loss ratios), the framework ensures reproducibility and direct applicability to large-scale UX practice.

The most fundamental practical contribution of the proposed method lies in its ability to transform large-scale user reviews from mere quantitative data into actionable design insights. In traditional design methods, access to such insights is typically limited to a small number of focus group studies or in-depth interviews, whereas the approach presented here enables the simultaneous analysis of millions of user reviews, providing designers with a scalable feedback mechanism. For example, in a product development process, a designer can directly examine negative emotional responses concentrated in the "ease

Table 6. Data-driven evidence-to-action framework for integrating natural language processing-derived empathy maps into the ISO 9241-210 human–computer interaction lifecycle

Evidence (derivation)	HCI design implication (artefact/output)	HCI lifecycle stage (ISO 9241-210)	HCI principle and application
Charging-related feedback constitutes 41% of Do quadrant instances, predominantly high-intensity negative (mean sentiment=−0.68, SD=0.21) from Aspect: Battery life aligned to CJM: Use stage	Redefine charging as a critical task; develop UI components for higher error tolerance and accurate real-time charging status (critical task model, UI state machine)	Requirements definition (specify user requirements)	Goal-oriented design: Prioritizing high-impact, high-frequency critical tasks through task models; error prevention and recovery via responsive UI states and progressive error messaging
Warranty-related feedback is Feel/Say-dominant (Feel=0.61, Say=0.27) with 39% negative polarity, mapped to CJM: Support stage	Redesign service interfaces with progressive disclosure and enhanced situation awareness (service blueprint, interaction flow)	Design solutions development (produce design solutions)	Situation awareness: Multi-stage status indicators to keep users informed; Progressive disclosure: Staged reveal of service steps to reduce cognitive load in high-frustration contexts
Price aspect emerges as a Think trigger (49% frequency in Comparison stage) via CJM stage-specific LDA topic modelling, linked with high decision-weight phrases (e.g., "value for money," "worth it")	Present price–performance data via cognitively efficient comparison tables and filters (decision support interface wireframe)	User research and context of use analysis+requirements definition (understand context/specify requirements)	Decision support systems: Structuring comparative data to support informed choice; Information architecture: Card-sorting-based grouping and filtering to optimize retrieval and evaluation
Customer service cases with resolution time ≤24 h show a mean Feel score increase of +0.42 (paired t-test: $p<0.01$) in CJM: Support stage	Implement conversational UI scripts and adaptive response systems simulating empathy (conversational UI script library)	Design solutions development (produce design solutions)	Adaptive interaction design: Tailoring responses to sentiment and urgency levels; Affective computing: Using empathetic language patterns and tone-modulated chatbot scripts
Cleaning performance aspect in Experience stage shows Do=49%, gain score=0.521, identified via gain–loss ratio analysis on TFSD distribution	Enhance first-use experience with micro-feedback and sensor data dashboards (interactive feedback widget)	Prototyping and simulation (evaluate designs against requirements)	Feedback immediacy: Real-time performance visualizations to reinforce perceived effectiveness; Gamification elements: Milestone-based micro-rewards to sustain engagement
Awareness and Research stages reveal 27% topic absence on navigation accuracy and cleaning performance, confirmed by gap analysis (expected vs. observed topic frequencies)	Introduce AR/VR demos and interactive product configurators into the pre-purchase experience (interactive product explorer prototype)	Context of use analysis+design solutions development (understand context/produce solutions)	Information design: Structuring technical performance data into visual narratives; Immersive interaction: Leveraging AR/VR to simulate post-purchase usage scenarios
Post-deployment comment analysis shows +0.31 mean sentiment shift in targeted aspects after intervention, based on longitudinal TFSD tracking across release cycles	Monitor UX improvement impact quantitatively; establish continuous improvement loops (UX KPI dashboard)	Evaluation and iteration (evaluate/iterate)	Continuous UX evaluation: Embedding sentiment and TFSD monitoring into the release pipeline; Evidence-based design: Iterative decision-making grounded in longitudinal behavioral data.

Abbreviations: AR: Augmented reality; CJM: Customer journey mapping; HCI: Human–computer interaction; KPI: Key performance indicator; LDA: Latent Dirichlet allocation; SD: Standard deviation; TFSD: Think–Feel–Say–Do; UI: User interface; UX: User experience; VR: Virtual reality.

of use" stage of the CJM; thus, critical decisions regarding ergonomics, material selection, or interface design can be made based not only on intuition but also on data derived from users. Similarly, in the context of service design, especially in the e-commerce ecosystem, return processes represent the stages of the user journey where the most intense emotional tension is experienced. NLP-based EMs reveal clusters of cognitive load, insecurity, and disappointment that arise during these stages, enabling designers to restructure the process more effectively.

This approach is not limited to the context of products or services but extends to broader application areas in the field of HCI. For example, in healthcare applications, patient feedback can reveal clusters of anxiety and uncertainty that intensify at specific stages of the treatment journey; in educational technologies, the same method can identify stages where students experience a loss of motivation in their learning journey.

In conclusion, this study has repositioned the concept of empathy in the field of HCI, transforming it from an intuitive design practice into a calculable, transparent, and repeatable methodology. The proposed method demonstrates that large-scale user-generated data—such as online reviews—can be systematically converted into EMs and customer-journey insights for direct use in design processes. Thus, the approach not only advances theoretical discussions on empathy in HCI but also provides practical and applicable tools for both academic researchers and professionals engaged in design practice. Beyond e-commerce, this method shows strong potential in areas where UXs are complex, multi-stage, and often emotionally charged, such as health technologies, education, public service design, and Internet-of-Things applications. Therefore, the study contributes to the trend of directly integrating user-generated data into design decisions—one of the fundamental goals of contemporary HCI research—through a scalable, ethically conscious, and contextually grounded methodology.

7. Limitations

The methodological approach of this study has three main limitations. The first limitation concerns data sources. The OPRs analyzed were obtained from only one e-commerce platform and a limited product range. However, the proposed methodology offers a flexible and adaptable framework that can process data from various fields, regardless of platform or sector. The developed algorithm is not platform- or sector-specific but focuses on EM and CJM detection; therefore, it can be applied independently of the data source and quality. Comparative studies that analyze data from various platforms or consider demographic differences may generate new perspectives. In terms of data security, while fake and spam comments pose challenges in traditional data mining, the comments used in this study are both verified purchase tags and have passed through the filtering processes of high-sales platforms.

The second limitation concerns the performance capacity of the language models used in the study (e.g., transformer-based LLMs). These models, which offer high accuracy and contextual meaning extraction performance in the current NLP literature, continue to evolve with new and diverse training data. Their contextual representation power, language diversity, and ability to capture subtle semantic distinctions are continuously improving. However, model performance depends not only on the size and diversity of the training data but also on potential biases in the data, lack of linguistic diversity, and technical limitations of the model architectures. In particular, existing models may occasionally exhibit limited accuracy in interpreting indirect layers of meaning, such as intent, implicit emotion, or cultural context. In addition, since the datasets used to train models are primarily composed of general-purpose text at the global level, contextual alignment may not always be ideal for specific industries, niche product groups, or local expression forms. Therefore, in future research, the use of models trained on broader and culturally diverse data sets, tailored to specific industries or language groups, has the potential to improve both contextual alignment and prediction accuracy. Regular updates to model versions will also strengthen the repeatability of the methodology, with each new version offering higher accuracy, scope, and generalization capacity.

The third limitation is the designers' lack of skills and knowledge regarding the use of NLP and text mining algorithms. In most cases, design professionals lack curriculum-based training in computational methods. This limits their ability to independently apply, adapt, or critically evaluate such techniques. Given the increasing role of data-driven insights in UX research and HCI, designers must develop at least foundational technical communication skills in NLP and text analytics to collaborate effectively with data scientists and use algorithmic outputs in design decisions. This is also a prerequisite for keeping pace with the evolving technological environment.

8. Future studies

Although this study has shown that large-scale user reviews can be converted into EMs and meaningful insights from a design perspective, several avenues remain open for future research. First, the present study was conducted using a

limited number of product-level examples. Future research should apply the method to longitudinal and cross-domain datasets to examine how empathy patterns evolve over time and across different sectors such as healthcare, education, and public services. Such comparative analyses will reveal the transferability and context-specific variability of computational EM.

Second, although the method has been shown to provide scalable insights, its integration into real-world design workflows remains underexplored. Therefore, future studies should focus on embedding EM tools into interactive design support systems or co-design platforms. This also raises the following questions: How do UX designers, service designers, or policymakers evaluate the reliability, interpretability, and applicability of computational empathy outputs?

Third, the ethical dimension constitutes a critical area of research. User comments and large-scale online data often contain sensitive or personal statements related to emotions. Therefore, future studies should conduct in-depth investigations on issues of privacy, consent, algorithmic transparency, and bias. The development of clear ethical frameworks is necessary for the responsible adoption of computational empathy methods in academia and industry.

Finally, expanding the method to include multimodality presents an important research opportunity. Future studies may enable a more comprehensive EM of UXs by including not only textual comments but also audio, video, and physiological signals. This multimodal approach aligns with the growing trends of affective computing and embodied interaction in the field of HCI.

In summary, future studies should focus on methodological improvements, real-world applications, the development of ethical frameworks, and multimodal approaches. Addressing these research agendas will ensure that computational EM becomes one of the fundamental methodologies of the next generation in the field of HCI.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares no conflict of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset (DOI: 10.6084/m9.figshare.29900258) contains the raw data underlying the present study. It consists of unprocessed product review data and associated annotations that form the empirical basis for the proposed multi-layered NLP framework integrating EM (Think–Feel–Say–Do quadrants) with CJM stages.

Further disclosure

The data for this study were sourced from genuine user reviews on Amazon.com and analyzed for academic purposes rather than commercial use. The data were anonymized and generalized to protect privacy, ensuring no personal information was disclosed. The research strictly complied with Amazon's terms of use and privacy policies. Objectivity was maintained by avoiding subjective evaluations, focusing instead on conveying evaluations without personal judgments. The analytical approach used for the review analysis was carefully selected to ensure the reliability and objectivity of the findings. The study's conclusions do not aim to make definitive judgments about the performance or quality of any products.

References

1. Simon H. *The Sciences of the Artificial*. Cambridge, MA: MIT Press; 1969.
2. Cross N. Designerly ways of knowing. *Design Studies*. 1982;3(4):221-227.
   doi: 10.1016/0142-694X(82)90040-0
3. Jones JC. *Design Methods: Seeds of Human Futures*. New York: John Wiley and Sons; 1970.
4. Moore GT, Tuttle DP, Howell SC. *Environmental Design Research Directions: Process and Prospects*. New York: Praeger; 1985.
5. Kouprie M, Sleeswijk Visser F. A framework for empathy in design: Stepping into and out of the user's life. *J Eng Des*. 2009;20(5):437-448.
   doi: 10.1080/09544820902875033
6. Mattelmäki T, Vaajakallio K, Koskinen I. What happened to empathic design? *Design Issues*. 2014;30(3):67-77.
   doi: 10.1162/DESI_a_00249
7. Heylighen A, Dong A. To empathise or not to empathize? Empathy and its limits in design. *Design Stud*. 2019;65:107-124.

doi: 10.1016/j.destud.2019.10.007

8. Boehner K, DePaula R, Dourish P, Sengers P. How emotion is made and measured. *Int J Hum Comput Stud*. 2007;65(4):275-291.

   doi: 10.1016/j.ijhcs.2006.11.016

9. Costanza-Chock S. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press; 2020.

10. Matthews T, Judge TK, Whittaker S. How do Designers and User Experience Professionals Actually Perceive and Use Personas? In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Austin, TX. Association for Computing Machinery; 2012. p. 1219-1228.

    doi: 10.1145/2207676.2208573

11. Higuera M, Macías JA. Automatic Generation of Empathy Maps. In: *Proceedings of the XXIII International Conference on Human Computer Interaction (Interacción '23)*. Ciudad Real, Spain: Association for Computing Machinery; 2023. p. 1-8.

12. Ferreira B, Silva W, Oliveira EC, Conte T. Designing Personas with Empathy Map. In: *Proceedings of the 27th International Conference on Software Engineering and Knowledge Engineering (SEKE '15)*. Pittsburgh, PA: Knowledge Systems Institute Graduate School; 2015. p. 152-157.

    doi: 10.18293/seke2015-152

13. Gonzalez-Bañales DL, Soto Ortíz LE. Empathy Map as a Tool to Analyze human-Computer Interaction in the Elderly. In: *Proceedings of the 8th Latin American Conference on Human-Computer Interaction (CLIHC '17)*. Antigua Guatemala, Guatemala. Association for Computing Machinery; 2017. p. 55.

    doi: 10.1145/3151470.3156642

14. Azarpey A, Thomas J, Ring D, Franko O. Natural language processing of sentiments identified in patient comments associated with less than top-rated care. *J Patient Exp*. 2025;12:1-5.

    doi: 10.1177/23743735251323677

15. Durgam R, Pamula NB, Dharani N, *et al*. AI-powered empathy: Sentiment analysis in personal care using RoBERTa and XLNet. *J Theor Appl Inf Technol*. 2025;103(8):3455-3470.

16. Vischer R. *Über das optische Formgefühl: Ein Beitrag zur Ästhetik* [*On the optical sense of form: A contribution to aesthetics*]. Leipzig, Germany: H. Credner; 1873.

17. Lipps T. *Ästhetik: Psychologie des Schönen und der Kunst* [*Aesthetics: Psychology of the beautiful and of art*]. Leipzig, Germany: Engelmann; 1903.

18. Titchener EB. *Lectures on the Experimental Psychology of the Thought-Processes*. New York: Macmillan; 1909.

19. Hatfield E, Cacioppo JT, Rapson RL. Emotional contagion. *Curr Dir Psychol Sci*. 1993;2(3):96-100.

    doi: 10.1111/1467-8721.ep10770953

20. Slote M. *The Ethics of Care and Empathy*. New York: Routledge; 2007.

21. Zaki J. Empathy: A motivated account. *Psychol Bull*. 2014;140(6):1608-1647.

    doi: 10.1037/a0037679

22. Coplan A. Understanding empathy: Its features and effects. In: Coplan A, Goldie P, editors. *Empathy: Philosophical and Psychological Perspectives*. Oxford, UK: Oxford University Press; 2011. p. 3-18.

23. Leslie AM. Pretense and representation: The origins of "theory of mind". *Psychol Rev*. 1987;94(4):412-426.

    doi: 10.1037/0033-295X.94.4.412

24. Gordon RM. Simulation without introspection or inference from me to you. In: Davies M, Stone T, editors. *Folk Psychology*. Oxford, UK: Blackwell; 1995. p. 53-67.

25. Stueber KR. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. Cambridge, MA: MIT Press; 2012.

26. Blum L. Moral perception and particularity. *Ethics*. 1991;101(4):701-725.

27. Cooper A. The inmates are running the asylum. In: *Software-Ergonomie '99: Design von Informationswelten*. Stuttgart, Germany: BG. Teubner; 1999. p. 17.

28. Shostack GL. Designing services that deliver. *Harv Bus Rev*. 1984;62(1):133-139.

29. Carlzon J. *Moments of Truth*. Cambridge, MA: Ballinger Publishing Company; 1987.

30. Leonard D, Rayport JF. Spark innovation through empathic design. *Harv Bus Rev*. 1997;75:102-115.

    doi: 10.1142/9789814295505_0016

31. Gaver WW, Dunne A, Pacenti E. Design: Cultural probes. *Interactions*. 1999;6(1):21-29.

    doi: 10.1145/291224.291235

32. Buchenau M, Fulton Suri J. Experience prototyping. In: *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '00)*. New York: ACM; 2000. p. 424-433.

    doi: 10.1145/347642.347802

33. Gray D, Brown S, Macanufo J. *Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers*. Sebastopol, CA: O'Reilly Media; 2010.

34. Osterwalder A, Pigneur Y, Bernarda G, Smith A. *Value Proposition Design: How to Create Products and Services Customers Want*. Hoboken, NJ: John Wiley and Sons; 2015.

35. Eichbaum Q, Barbeau-Meunier CA, White M, Ravi R. Empathy across cultures-one size does not fit all: From the ego-logical to the eco-logical of relational empathy. *Adv Health Sci Educ Theory Pract*. 2023;28(2):643-657.

doi: 10.1007/s10459-022-10158-8

36. Benyon D, Turner P, Turner S. *Designing Interactive Systems: People, Activities, Contexts, Technologies*. Harlow, UK: Addison-Wesley Longman; 2005.

37. Hassenzahl M, Tractinsky N. User experience: Research agenda. *Behav Inf Technol*. 2006;25(2):91-97.

doi: 10.1080/01449290500330331

38. Hassenzahl M. User experience (UX): Towards an Experiential Perspective on product Quality. In: *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine (IHM '08)*. Metz, France: Association for Computing Machinery; 2008. p. 11-15.

doi: 10.1145/1512714.1512717

39. Hassenzahl M. *Experience Design: Technology for All the Right Reasons*. San Rafael, CA: Morgan and Claypool; 2010.

40. Meyer M, Wachter-Boettcher S. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. New York: W. W. Norton and Company; 2016.

41. Maguire M. Methods to support human-centred design. *Int J Hum Comput Stud*. 2001;55(4):587-634.

doi: 10.1006/ijhc.2001.0503

42. Kalbach J. *Mapping Experiences: A Complete Guide to Customer Alignment Through Journeys, Blueprints, and Diagrams*. Sebastopol, CA: O'Reilly Media; 2016.

43. Cuadra A, Wang M, Stein LA, *et al*. The illusion of empathy? Notes on displays of emotion in human-computer interaction. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Honolulu, HI. Association for Computing Machinery; 2024:1-18.

44. Sorin V, Brin D, Barash Y, *et al*. Large language models and empathy: Systematic review. *J Med Internet Res*. 2024;26:e52597.

doi: 10.2196/52597

45. Elagroudy P, Li J, Väänänen K, *et al*. Transforming HCI Research Cycles Using Generative AI and "Large Whatever Models" (LWMs). In: *CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Honolulu, HI: Association for Computing Machinery; 2024. p. 1-5.

doi: 10.1145/3613905.3643977

46. Yeykelis L, Pichai K, Cummings JJ, Reeves B. *Using Large language Models to Create AI Personas for Replication, Generalization and Prediction of Media Effects: An Empirical Test of 133 Published Experimental Research Findings*. [Preprint].

doi: 10.48550/arXiv.2408.16073

47. Güneş S. Extracting data-driven user segments and knowledge by using online product reviews. *Gazi Univ J Sci Part B Art Humanit Des Plan*. 2023;11(1):139-152.

48. Salminen J, Santos JM, Jung SG, Jansen BJ. Picturing the fictitious person: An exploratory study on the effect of images on user perceptions of AI-generated personas. *Comput Hum Behav Artif Humans*. 2024;2(1):100052.

doi: 10.1016/j.chbah.2023.100052

49. Sun L, Qin T, Hu A, *et al*. *Persona-L Has Entered the Chat: Leveraging LLM and Ability-Based Framework for Personas of People with Complex Needs*. [Preprint].

doi: 10.48550/arXiv.2409.15604

50. Al-Ansari N, Al-Thani D, Al-Mansoori RS. User-centered evaluation of explainable artificial intelligence (XAI): A systematic literature review. *Hum Behav Emerg Technol*. 2024;2024:4628855.

doi: 10.1155/2024/4628855

51. Rong Y, Leemann T, Nguyen T, *et al*. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(4):2104-2122.

doi: 10.1109/tpami.2023.3331846

52. Alaqsam A, Sas C. Systematic Review of XAI Tools for AI-HCI Research. In: *Proceedings of the 37th International BCS Human-Computer Interaction Conference (HCI '24)*. Birmingham, UK. BCS Learning and Development; 2024. p. 47-59.

doi: 10.14236/ewic/hci2024.5

53. Ehsan U, Watkins EA, Wintersberger P, *et al*. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In: *CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Honolulu, HI. Association for Computing Machinery; 2024. p. 477.

doi: 10.1145/3613905.3636311

54. Zhu Q, Luo J. Toward Artificial Empathy for Human-Centered Design: A Framework. In: *Proceedings of the ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2023)*. Boston, MA: American Society of Mechanical Engineers; 2023.

doi: 10.1115/detc2023-87318

55. Nguyen S, Beck D, Holtta-Otto K. Predicting Empathic Accuracy from User-Designer Interviews. In: *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association (ALTA 2023)*. Melbourne, Australia: Association for Computational Linguistics; 2023. p. 125-129.

56. Hasan MR, Hossain MZ, Ghosh S, Krishna A, Gedeon T. Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods. In: *IEEE Transactions on Affective Computing*. Karnataka: IEEE; 2025. p. 1-20.

doi: 10.1109/taffc.2025.3590107

57. Lahnala A, Welch C, Jurgens D, Flek L. A Critical Reflection and Forward perspective on Empathy and Natural Language Processing. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 2139-2158.

doi: 10.18653/v1/2022.findings-emnlp.157

58. He P, Liu X, Gao J, Chen W. *DeBERTa: Decoding-Enhanced BERT with Disentangled Attention*. [Preprint]; 2020.

doi: 10.48550/arXiv.2006.03654

59. Zhao H, Chen H, Yang F, *et al*. Explainability for large language models: A survey. *ACM Trans Intell Syst Technol*. 2024;15(2):36.

doi: 10.1145/3639372

60. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Network. Brussels, Belgium: Association for Computational Linguistics*; 2018. p. 353-355.

doi: 10.18653/v1/w18-5446

61. Devlin J, Chang MW, 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*. Minneapolis, MN: Association for Computational Linguistics; 2019. p. 4171-4186.

doi: 10.18653/v1/N19-1423

62. Lewis M, Liu Y, Goyal N, *et al*. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. Stroudsburg: Association for Computational Linguistics; 2020. p. 7871-7880.

doi: 10.18653/v1/2020.acl-main.703

63. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc explanation Methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. New York: Association for Computing Machinery; 2020. p. 180-186.

doi: 10.1145/3375627.3375830

64. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*; 2016. San Francisco, CA: Association for Computing Machinery; 2016. p. 1135-1144.

doi: 10.1145/2939672.2939778

65. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng*. 2014;26(8):1819-1837.

doi: 10.1109/tkde.2013.39

66. Aho AV, Corasick MJ. Efficient string matching: An aid to bibliographic search. *Commun ACM*. 1975;18(6):333-340.

doi: 10.1145/360825.360855

67. PyABSA. *PyABSA: Open-Source Aspect-Based Sentiment Analysis Library*. GitHub. Available from: https://github.com/yangheng95/pyabsa [Last accessed on 2025 Sep 19].

68. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP '19)*. Hong Kong, China: Association for Computational Linguistics; 2019. p. 3982-3992.

doi: 10.18653/v1/D19-1410

69. Yin W, Hay J, Roth D. Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP '19)*. Hong Kong, China: Association for Computational Linguistics; 2019. p. 3914-3923.

doi: 10.18653/v1/d19-1404

70. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2019. Available from: https://leanpub.com/interpretable-machine-learning [Last accessed on 2025 Sep 19].