JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.14
eISSN:2805-5160

Mining Customer Reviews for Product Improvement Using Aspect-Based Sentiment Analysis and Clustering

Arief Ginanjar[1]*, Dian Puspita Hapsari[2], Gerli Ryza DS. Reyes[3]

[1]Universitas Langlangbuana, Indonesia
[2]Informatics Engineering Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[3]Department of Information Technology, Pangasinan State University Lingayen Campus, Philippines

***Email**: arief.ginanjar.unla@gmail.com[1]*, dian.puspita@itats.ac.id[2], grdreyes.lingayen@psu.edu.ph[3]

Abstract

Online customer reviews provide valuable customer intelligence concerning product quality, usability, delivery, pricing, and post-purchase services. However, extracting actionable information from large volumes of unstructured review text remains difficult because conventional sentiment analysis commonly assigns a single polarity label to an entire review without identifying the specific product attributes responsible for customer satisfaction or dissatisfaction. This study develops an integrated framework combining aspect-based sentiment analysis and text clustering to identify product attributes, determine aspect-level sentiment, group semantically related customer complaints, and translate the resulting complaint patterns into product-improvement recommendations. A publicly available e-commerce review corpus containing 27,846 retained reviews was processed through text cleaning, sentence segmentation, aspect extraction, aspect-level sentiment classification, semantic text embedding, and complaint clustering. The framework was evaluated using topic coherence, classification accuracy, macro-F1, the Silhouette coefficient, and the Davies–Bouldin Index. The proposed approach achieved a topic coherence score of 0.612, sentiment-classification accuracy of 89.3%, macro-F1 of 88.4%, and Silhouette coefficient of 0.468. The findings identified battery performance, packaging quality, and product durability as the most urgent improvement priorities. By integrating aspect identification, sentiment prediction, and semantic clustering, the framework provides more specific and operational product-development recommendations than conventional document-level sentiment analysis.

Keywords

Product Innovation; Aspect-based sentiment analysis; Customer reviews; Text clustering; Opinion mining; Product improvement; Customer analytics

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026

Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

Customer Reviews as Product Intelligence

The proliferation of online shopping has led to an unprecedented volume of publicly available reviews. Reviewers no longer limit their feedback to just product performance. Nowadays, reviewers include their opinions on product design, usability, pricing, packing, delivery, customer service, and even their experience with the warranty. These product reviews contain unique, direct critiques of a product by its consumers, and as such, are invaluable to companies. Unlike the opinions gathered by surveys, reviews allow consumers to detail their experiences, including the challenges they faced, how the product could be improved, and much more.

For product managers, manufacturers, and sellers, reviews can highlight challenges a product can experience that may not be discovered in a lab or through quality assurance. Complaints about easy-to-break components, fast battery draining, damaged packaging, erroneous product descriptions, and frustrating warranty processes indicate a problem that may be of a technical, operational, or managerial nature (Sourgens, 2022). Additionally, review content can be utilized to identify what consumers enjoy about a product and encourage product managers to preserve that feature during the redesign process.

Customer reviews, although helpful, represent a huge challenge because of the sheer volume of reviews across many product categories. Each review must still be read and examined to discover the new problems a product may be facing. This process is time-consuming and can lead to typographical errors and inconsistent review interpretations, including the omission of critically important issues. Because of this, companies must develop a way to automate the transformation of customer feedback into information that can support product development.

Another challenge is that customer reviews often do not contain a single, unambiguous sentiment. One review can include multiple product features with sentiment that vary. For instance, a customer can have a positive opinion about the design of a product but a negative opinion about the battery, packaging, and after-sales service. If a review is assigned one sentiment, the multiple opinions can mask the design elements that the customer liked and the reasons for their negative opinion. Thus, review analysis must determine the feature and the sentiment related to that feature.

Limitations of Existing Review-Mining Approaches

The majority of online review analysis methods do document-level sentiment analysis. In this case, a review is classified as positive, negative, or neutral as a whole. This is useful for determining a customer opinion and for identifying a customer opinion over a period of time. However, this does not answer what component of the review is related to the sentiment. For instance, a customer can classify a review as negative, and this does not mean that the review is about product durability, delivery, compatibility, pricing, and customer service. Thus, the output of document-level sentiment analysis may be too general for specific product development.

Aspect-based sentiment analysis identifies product features and attributes captured in customer reviews. This allows customer reviews to capture multiple aspect-sentiment pairs, for example, positive aspect sentiment for product design and negative aspect sentiment for product battery performance. Analysis at the aspect level helps define customer experience with greater granularity, as compared to classification at the document level. Still, aspect-based analysis is typically limited to assigning aspect labels and categorizing sentiment for analysis.

Typically, negative aspect sentiment is identified for multiple customers, but the comments are not categorized into a meaningful framework.

Similar to topic modeling, text clustering can also be used to identify patterns in large collections of customer reviews. Text clustering can group similar sentences and identify common, customer-defined issues, rather than issues that are defined by the researcher. For example, text clustering may identify separate issues for battery complaints related to rapid battery drain, failure to charge, overheating, and battery indicator inaccuracies. However, the focus or aspect of the product as well as sentiment performed in clustering can make it hard to interpret some of the resulting groups. Just because some sentences are similar in meaning does not mean that the group as a whole captures a strength of the product, a complaint, a suggestion, or a neutral comment.

The three approaches identified provide complementary information. Document-based sentiment analysis captures the sentiment of a document, aspect-based sentiment analysis connects the sentiment to product features, and clustering provides recurring semantic patterns. Each approach provides a different perspective, but using them collectively can help streamline the process of creating and prioritizing product enhancements.

Research Gap

Traditional studies tend to compartmentalize aspect extraction, sentiment analysis, and/or complaint clustering (Gupta & Sharma, 2024). As a result, the analysis often fails to identify an organized structure of the customers' complaints and the related product enhancement priorities (Liu et al., 2022). This compartmentalization further diminishes the review-mining analysis' utility, as product managers need more than simply a list of product aspects mentioned or an aggregate count of the positive and negative sentiments.

Several deficiencies may be attributed to the current review-mining framework. First, aspect extraction, sentiment analysis, and complaint clustering do not seem to be combined in a systematic and traceable framework. Second, the completeness of the frameworks is often judged primarily in terms of the accuracy of the sentiment analysis. Considerable deficiencies may be present in the framework as a whole if the reviewed aspects lack semantic cohesion and the complaint clusters are unrefined. Third, the outcome of the text mining is often not translated into prioritized and actionable product development and enhancement suggestions.

An integrated framework should therefore evaluate different analytical stages using metrics that correspond to their specific functions. In the context of sentiment analysis and complaint clustering, metrics such as a macro-F1 score and the Silhouette coefficient may be used to evaluate the compactness and cohesion of the clusters and the extent to which sentiment is articulated in a negative, positive, or neutral sense. The completeness of the frameworks may be assessed primarily in terms of the quality and cohesion of the clusters.

This study aims to develop an integrated framework that connects the dimensions of product-aspect detection, sentiment analysis at the aspect level, text representation, and clustering. The output of the proposed solution will maintain the association of the original text of the review, the detected aspect, the predicted sentiment, the assigned complaint cluster, and the suggested product action.

Research Objective

The main goal of the proposed solution is to offer a robust framework that can be used to specify product aspects, analyze customer sentiment at the level of the specified product aspects, and assess the grouping of customer complaints and the related product recommendations. The solution further aims to analyze the degree to which the proposed

solution can move beyond assessing the overall sentiment of customer reviews to the identification of product aspects, evaluating sentiments expressed toward the various product aspects, and grouping the expressed sentiments regarding product complaints in a coherent and organized manner. The solution also aims to determine the focus of product improvement actions based on the complaint product aspects, the expressed sentiments, and the frequency of the organized complaint groups.

Research Questions

Considering the defined research problem and the objective of this study, the following research questions have been formulated:

- RQ1. What are the product aspects that can be captured from large-scale e-commerce customer reviews?
- RQ2. What is the accuracy of the proposed method in classifying customer sentiments toward the identified product aspects?
- RQ3. What is the structural quality and the managerial interpretability of the resulting complaint clusters?
- RQ4. What are the product aspects that should be prioritized for improvement based on their prevalence, sentiment, and complaint clusters?

These questions assist in developing the product-aspect identification process and deriving implementation-ready suggestions. RQ1 examines customers' reviews and describes the semantics. RQ2 examines the extent to which aspect-level sentiment classification can be predicted. RQ3 looks at the form and structure of the compiled complaints and the corresponding recommendation clusters. RQ4 looks at the outputs of all steps and identifies the product characteristics that are most critical and require immediate intervention.

Research Contributions

There are three main contributions of the study (Chen et al., 2017). First, the study constructs the first integrated review-mining pipeline that includes aspect extraction, aspect-level sentiment classification, embedding of semantics, clustering of complaints, and generation of recommendations from the clusters (Hassija et al., 2023). The unique feature of the framework is that it does not treat recommendations as an end-of-the-line separate cluster. Recommendations are derived from the major concerns represented in each validated complaint cluster.

Second, the study assesses the framework with three complementary sets of performance indicators (Velden et al., 2022). The semantic consistency of the extracted product aspects is assessed by the topic coherence measure. Sentiment classification is evaluated by accuracy and macro-F1, while cluster separation and compactness are assessed by the Silhouette coefficient and the Davies–Bouldin index, respectively. The review-mining quality is assessed from multiple dimensions, which is a notable feature of the study (Darban et al., 2024).

Third, the study offers an aspect-to-recommendation mapping that bridges the computational outcomes and product development (Berahmand et al., 2024). The framework goes beyond the identification of negative sentiment and provides an organizing structure to dominant complaints. It helps in identifying product aspects that demand immediate focus. The

primary sentiment of the aspects also helps in determining the focus. Thus, the framework enables manufacturers, sellers, and product managers to derive structured and informative customer reviews that highlight and prioritize product improvements.

The remainder of this paper outlines the proposed methodology and defines the empirical results (Ouyang et al., 2022). The proposed methodology and empirical results are used to analyze the implications of the methodology and results on customer intelligence and on product development (Mikalef & Gupta, 2021). The paper closes with a limitations section and offers suggestions for future research (Dave et al., 2023).

Related Work

Customer Review Mining and Opinion Mining

The increasing prevalence of online customer reviews provides a rich source of customer opinion data through which the customer journey can be documented, and product performance and the potential product gaps can be assessed. Unlike structured questionnaires, reviews allow customers to describe product strengths, usage difficulties, defects, delivery experiences, service interactions, and desired improvements in their own language. Consequently, review-mining methods have been employed to extract customer needs, monitor perceived product quality, measure satisfaction, identify recurring complaints, and support product-development decisions. For example, demonstrated that user-generated content can be systematically processed to identify customer needs that would otherwise require extensive interviews or focus groups (Timoshenko & Hauser, 2019). Similarly, User-experience information extracted from online reviews can support product design and business planning.

Earlier review-mining approaches commonly relied on word-frequency analysis, sentiment lexicons, bag-of-words representations, and traditional document-level classifiers. Frequently occurring words were interpreted as indicators of important product features, while sentiment dictionaries assigned positive or negative scores to opinion-bearing terms. Supervised classifiers such as Logistic Regression and Support Vector Machine subsequently improved overall polarity prediction by learning patterns from labeled review documents. Star ratings have also frequently been used as indirect sentiment labels, with high ratings interpreted as positive opinions and low ratings as negative opinions.

Although these methods provide useful aggregate summaries, they have important limitations. Word-frequency analysis may identify a frequently discussed term without explaining whether customers evaluate it positively or negatively. Lexicon-based methods may also misinterpret context, negation, domain-specific expressions, and words whose sentiment changes according to the discussed product attribute. Most importantly, document-level models normally assign one sentiment label to an entire review. This treatment loses aspect-specific information when a customer praises one attribute but criticizes another. It can also obscure the distinction between physical product problems, delivery failures, and post-purchase service complaints. These limitations have encouraged the development of more fine-grained approaches that connect sentiment to specific product aspects.

Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis extends conventional sentiment analysis by identifying the target toward which an opinion is expressed. Its analytical process generally involves aspect-term extraction, aspect-category identification, sentiment classification, and

the formation of aspect–sentiment pairs. For example, the sentence “the design is attractive, but the battery drains quickly” should produce a positive sentiment for the design aspect and a negative sentiment for battery performance. This level of analysis is more useful for product evaluation because it identifies both the object of customer judgment and the direction of the associated opinion.

Early ABSA methods employed linguistic rules, dependency relations, and sentiment lexicons to identify nouns or noun phrases and connect them to nearby opinion expressions. Supervised machine-learning approaches later modeled aspect extraction as a sequence-labeling task and aspect sentiment as a classification problem. Recurrent neural networks and attention-based models improved performance by representing sequential relationships and allowing a classifier to emphasize words that were relevant to a particular aspect. However, these models could still struggle with long sentences, distant aspect–opinion relationships, implicit expressions, and words whose interpretation depended heavily on context.

Transformer-based language models have substantially improved contextual representation in ABSA. BERT-based approaches can represent a word according to its surrounding sentence rather than assigning it one fixed vector. Aspect sentiment classification was reformulated as a sentence-pair task and demonstrated that BERT fine-tuning could effectively identify fine-grained polarity toward a specified aspect (Abdelgwad et al., 2022). Contextualized BERT representations could strengthen end-to-end aspect extraction and sentiment classification. Review-domain post-training and generated training examples have also been used to improve adaptation across datasets and product domains.

Recent research has also demonstrated the value of combining explicit linguistic knowledge with contextual transformer representations (Xu et al., 2023). Developed RB-BERT, a hybrid framework that uses domain-specific rules to generate weak aspect–sentiment labels before fine-tuning a BERT classifier. Applied to 3,811 user reviews, the framework achieved a Micro-F1 score of 0.9476 and a Hamming Loss of 0.0132. These findings indicate that rule-based supervision can complement transformer models by improving the identification and alignment of multiple aspects and their associated sentiments, particularly when extensive manually annotated data are unavailable (Zhao et al., 2026).

Contextual models offer several advantages for customer-review analysis. They can more effectively preserve the influence of negation, distinguish different meanings of the same word, and model relationships between an aspect and an opinion expression that are separated within a sentence. Syntactic information can provide additional benefits. For example, combined contextual embeddings with part-of-speech and dependency information to improve the detection of multiword aspect boundaries and the association between aspects and relevant context words. More recent models have also attempted to learn dynamic aspect-oriented semantics so that the representation emphasizes different sentence regions depending on the target aspect.

Nevertheless, ABSA alone does not completely solve the product-improvement problem. Equivalent aspect expressions may remain fragmented, such as “battery,” “battery duration,” and “power life.” In addition, a list of negative aspect–sentiment pairs does not necessarily explain the different causes of dissatisfaction. Negative comments assigned to battery performance may concern rapid depletion, slow charging, overheating, defective indicators, or complete charging failure. Many ABSA systems therefore stop before organizing these related opinions into coherent and interpretable complaint themes.

Topic Modeling and Aspect Extraction

Topic modeling provides an unsupervised approach for discovering recurring themes in large text collections. Latent Dirichlet Allocation represents documents as mixtures of latent topics and each topic as a probability distribution over words. Non-negative Matrix Factorization similarly decomposes a non-negative document–term matrix into lower-dimensional components that can be interpreted as topics. These approaches can identify patterns without requiring all product aspects to be predefined.

In customer-review applications, topic models can reveal themes related to product quality, usability, delivery, pricing, packaging, and service. Noun-phrase extraction and dependency-based rules can further improve interpretability by focusing the analysis on probable product attributes rather than all frequently occurring words. More recent approaches combine lexical candidates with semantic embeddings to merge expressions that have similar meanings even when they use different vocabulary.

While modeling to derive product attributes through topic modeling can yield some product attributes, each model topic can often be a collection of many product features, sentiments, feelings, delivery experiences, and general comments. Many topic models also combine positive and negative comments, which can be misleading. The focus of this research is on the coherence of topics as a measure of quality (Chan, 2023). This coherence is assessed based on the degree to which the words associated with that topic are in the same context (Gad, 2022). High coherence or low complexity is, however, not the only measure to assess the quality of a topic from a product perspective. A researcher has to evaluate the topic to assess if the topic reflects an important product issue.

Semantic Text Representation and Review Clustering

The primary importance of clustering reviews is in the quality of the review text representation. The bag-of-words and the TF-IDF representation approaches rely on word occurrence and importance and are efficient and simple representation approaches. However, they lack the ability to capture meaning as they treat words with similar meaning as disparate if they do not share vocabulary; thus, they can be inefficient.

Contextual word embeddings tackle this shortcoming by producing representations based on neighboring words. Models that embed sentences go further by potentially incorporating the full sentence into a compact vector that preserves semantic meaning. An example of one of these models is Sentence-BERT, which is built to create sentence representations to facilitate efficient semantic similarity comparisons and clustering. These representations serve a purpose for customer complaint data since phrases like “the battery dies within hours” and “the device cannot hold a charge” can, despite their lexical differences, be placed semantically close to each other.

There is a variety of clustering methods that can be utilized on these semantic representations. In K-means clustering, observations are assigned to one of the pre-defined number of clusters by minimizing the distance from the clusters’ center. K-means is optimal for larger datasets but assumes relatively small clusters and requires prior knowledge of the number of clusters. In contrast, the Agglomerative Clustering method starts with each observation as a separate cluster and progressively merges the most similar clusters. From this, a hierarchy is created that can allow researchers to observe the relationship of the complaints on different levels of closeness.

DBSCAN and HDBSCAN are prime examples of density-based clustering methods and provide different approaches to dealing with noise and irregular cluster shapes. DBSCAN relies on distance and neighborhood thresholds to define dense clusters, and HDBSCAN builds on this with a hierarchical approach to density-based clustering. HDBSCAN is well suited for analyzing reviews, since the complaint clusters can significantly vary in density and size, and many of the sentences from the customer reviews can be forced into a cluster representing a major theme, even though they don't truly belong. These sentences can be treated as noise.

Internal validation measures are commonly used when predefined cluster labels are unavailable. The Silhouette coefficient evaluates how closely an observation is associated with its assigned cluster relative to neighboring clusters (Ashuach et al., 2023). The Davies–Bouldin Index evaluates within-cluster compactness in relation to between-cluster separation, with lower values indicating a stronger structure. The Calinski–Harabasz Index compares between-cluster dispersion with within-cluster dispersion. Because these measures evaluate geometric properties rather than business meaning, cluster-quality scores should be complemented by manual assessment of representative sentences, keywords, and proposed cluster labels.

Review Mining for Product Improvement

Review-mining research has increasingly attempted to connect customer feedback with product-development activities (Raiaan et al., 2024). Customer reviews have been used to identify unmet needs, detect recurring defects, examine user experience, prioritize product features, and support redesign decisions. Machine-learning methods have been used for identifying customer needs from user-generated content, while extracting user-experience information and relationships relevant to product design. Further emphasized that useful review summaries should capture not only product features but also customer emotions, usage conditions, and product affordances.

In practical applications, review analysis can reveal improvement opportunities that extend beyond the physical product. Complaints about damaged goods may indicate inadequate packaging or delivery handling, while dissatisfaction with refunds and warranties may reflect weaknesses in post-purchase processes. Review mining can therefore support product feature redesign, quality-control improvement, packaging evaluation, service optimization, warranty-policy refinement, and product-priority planning.

Regardless of this advancement, several studies still only provide fragments of the information needed for complete operational decision-making (Tlili et al., 2023). Some provide the most frequently cited features; some provide sentiment distributions; some provide topics or semantic clusters. Very few frameworks integrate the prevalence of aspects, the proportion of negative sentiment, the intensity of the complaint, the structure of the semantic clusters, and the proposed action in a single traceable analysis. Computational performance is not enough, either. A model may achieve impressive classification or clustering scores, but the outputs may be so broad, fragmented, or unclear that managers cannot use them. Therefore, the analysis of products and their improvements requires computational performance and interpretability.

Synthesis of Previous Studies and Study Positioning

The literature reviewed documents document-level sentiment analysis, ABSA, topic modeling and semantic clustering, as well as frameworks for analyzing the improvement and management of product priorities. These frameworks present complementary views of customer feedback, but remain incomplete. The main customer review mining approaches

available in the literature are presented in Table 1, detailing their methods, outputs, evaluation metrics, and primary shortcomings.

Table 1. Comparison of customer-review mining approaches

Approach	Main method	Typical output	Common evaluation	Main limitation
Document-level sentiment analysis	Lexicon, SVM, Logistic Regression	Overall positive, neutral, or negative label	Accuracy, precision, recall, F1	Does not identify the responsible product aspect
Aspect-based sentiment analysis	Sequence labeling, neural networks, transformers	Aspect–sentiment pairs	Accuracy, macro-F1	Often does not organize related complaints
Topic modeling	LDA, NMF, probabilistic models	Latent themes and keywords	Topic coherence	Topics may mix aspects and sentiment orientations
Review clustering	K-means, hierarchical, density-based clustering	Groups of semantically similar reviews	Silhouette, Davies–Bouldin	Clusters may lack aspect and sentiment context
Product-priority analysis	Frequency and sentiment aggregation	Ranked product attributes	Descriptive statistics	Limited integration with semantic complaint structures
Proposed integrated framework	ABSA, sentence embeddings, and clustering	Aspects, sentiments, complaint clusters, and improvement priorities	Topic coherence, accuracy, macro-F1, Silhouette, Davies–Bouldin	Requires validation across products and domains

The comparison indicates that existing approaches generally emphasize one analytical level, whereas the proposed framework connects aspect identification, sentiment prediction, semantic complaint clustering, and improvement prioritization within a single traceable pipeline (Zhao et al., 2024). Document-level sentiment analysis provides a general opinion summary but lacks attribute-level detail. ABSA identifies the evaluated attribute but does not necessarily distinguish different causes of dissatisfaction. Topic modeling and clustering reveal recurring patterns but may omit sentiment orientation or direct product relevance. Product-priority methods can rank attributes but commonly rely on frequency and polarity aggregates without utilizing the internal semantic structure of customer complaints.

Given these constraints, the present study combines multiple methodologies, including component extraction, component-based sentiment analysis, and semantic text summarization, in conjunction with density-based clustering, to propose enhancements for a product in the field of sentiment analysis (Park & Kim, 2022). This methodology evaluates semantic sentiment analysis not only based on the accuracy of the identified sentiment, but also in terms of the semantic coherence of the analysis, the reliability of the classifications, the separation of the

clusters, and their relevance to management. This combination should allow for the development of sentiment analysis methods in the field of customer review analysis, in which the improvement priorities can be traced back to the complaint clusters, the predicted sentiments, the extracted components, and the original customer review text.

Methodology

Research Design

This study employed a quantitative experimental text-mining methodology to convert unstructured customer reviews into sentiment information at the level of components and prioritization of enhancements for a product. The experimental methodology integrated several techniques, including natural language processing, supervised machine learning, deep learning, and clustering based on transformers and semantic sentence representations. Aspect-sentences were the unit of analysis, as text reviews often contain multiple product aspects, each of which can be captured with different sentiment.

The analysis methodology employed nine steps. First, three types of products with publicly available e-commerce reviews were used. Second, reviews with duplication, irrelevance, non-English, and poor quality were eliminated. Third, reviews were split into individual sentences with their original identifiers. Fourth, aspect terms were identified using part-of-speech tagging, parsing, and weighting along with semantic synthesis. Fifth, the identified aspects were evaluated for sentiment as positive, neutral, and negative. Sixth, negative aspects and sentences were transformed into contextual sentence embeddings. Seventh, related customer complaints were consolidated using K-means, Agglomerative Clustering, and HDBSCAN. Eighth, the complaint clusters were analyzed and interpreted for their utility and clarity. In the last step, the complaint clusters were converted into actionable product-improvement priorities.

The complete analytical workflow shown in Figure 1 converts customer reviews into aspect-level sentiments and offers product improvement recommendations based on the sentiment of customer complaints (Hassija et al., 2023).

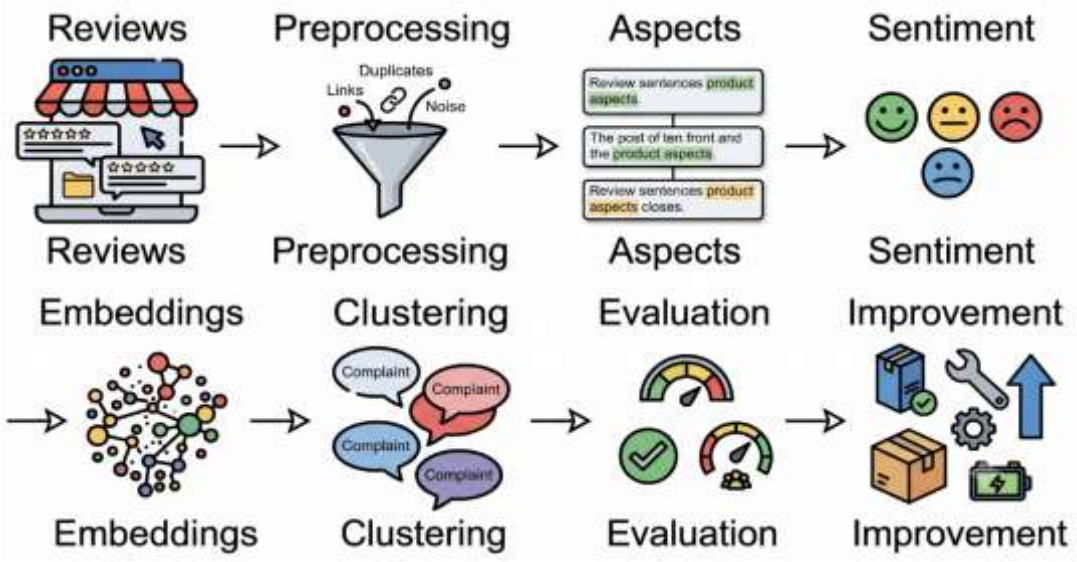


Figure 1. Integrated customer-review mining framework

The framework provides a means of tracing the recommendations back to customer reviews by preserving the linkage between review sentences, aspects, sentiment, clustering, and the improvement priority assignment. This was accomplished by means of review ID, sentence ID, aspect label, sentiment label with probability, cluster ID, and prioritization for improvement score.

Dataset and Sampling

The study focused on English publicly available e-commerce product reviews for consumer electronics, home appliances, and personal care items. These three categories dominate product reviews that feature highly variable product characteristics including user scenarios, purchase considerations, and post-sale service (Bellar et al., 2024). In reviews for consumer electronics, for example, it is common to address issues related to technical functionality, compatibility, and battery performance. Reviews for home appliances often include discussions about delivery, installation, and product durability. Reviews for personal care products often cite issues of safety, material quality, and product performance.

The primary dataset contained reviews totaling 30,000, from which 10,000 reviews for each category were sampled. Reviews were eligible when they contained textual content, a rating between one and five stars, and sufficient information to describe a product experience. Duplicate reviews, records dominated by promotional content, non-English reviews, and entries containing fewer than three meaningful tokens were excluded. After screening, 27,846 reviews and 54,912 review sentences were retained.

For descriptive corpus profiling, ratings of four or five stars were categorized as positive, three-star ratings as neutral, and one- or two-star ratings as negative. These rating-derived categories were used only to describe the corpus composition and were not treated as gold-standard aspect-level sentiment labels. Aspect-level labels were obtained through manual annotation because one review rating could not reliably represent the sentiment expressed toward every product attribute in the review.

Table 2 summarizes the composition of the review corpus after duplicate removal, language filtering, and text-quality screening.

Table 2. Characteristics of the customer-review datasets

Product category	Raw reviews	Retained reviews	Review sentences	Average words per review	Positive	Neutral	Negative
Consumer electronics	10,000	9,420	18,714	37.8	59.3%	7.5%	33.2%
Home appliances	10,000	9,106	17,936	35.6	61.1%	7.1%	31.8%
Personal-care devices	10,000	9,320	18,262	33.9	63.8%	6.8%	29.4%
Total/average	30,000	27,846	54,912	35.8	61.4%	7.1%	31.5%

The retained corpus covers product categories with different usage characteristics, enabling the framework to be evaluated across diverse customer concerns rather than within a single product type. The corpus-level rating distribution also indicates that negative reviews constituted a sufficiently large proportion of the data for complaint analysis, although the

manually annotated subset was stratified to prevent the majority positive class from dominating model evaluation.

Review Preprocessing

Pre-processing was done at both the review and sentence level. De-duplication at the review level was done through exact match on normalized review text. For near duplicates, high cosine similarity on TF-IDF representations was used. In case of duplicates, only the first complete review was kept. Repeated markup, non-informative symbols, URLs, email addresses, and HTML tags were removed. Numeric values related to the product, along with ratings, were kept if they added meaning to the complaint.

The cleaned reviews were broken down into individual sentences prior to tokenization. This was done for the segmentation of sentences because reviews often included multiple product features, each with their own sentiments. Lexical models required lower-casing, tokenization, spelling normalization, and lemmatization. Stop words were removed (if necessary) for frequency-based aspect extraction and topic analysis. Negation words such as "not," "never," "hardly," "no longer," and "cannot" were retained, as their removal can be counterproductive to capturing the intended sentiment. The phrase "not durable" will not be interpreted as a positive mention of durability.

For transformer-based models, we kept all original tokens of the text and their order because we believed these would provide useful contextual information for the model. Contractions were not kept in the models if it was believed that the meaning was lost. To illustrate, "doesn't work" would be "does not work." Spellings of model names were kept along with domain-specific terms. Emojis in reviews were kept along with their extensive emotional descriptions only if they provided a positive or negative meaning, respectively.

Each model provided the same review with a unique review identifier. This segmentation provided the means to trace sentiments, aspect extractions, and improvement priorities back to the reviews for analysis.

Aspect Extraction

Aspect extraction combined linguistic, statistical, and semantic procedures. First, part-of-speech tagging was used to identify nouns and noun phrases that could represent product attributes. Second, dependency relations were examined to capture aspect candidates connected to adjectives, verbs, and opinion expressions. Candidate phrases such as "battery life," "charging speed," "protective packaging," and "customer support" were preserved as multiword units rather than being separated into individual tokens.

Third, low-frequency and non-informative candidates were filtered using corpus frequency and term frequency–inverse document frequency. The TF-IDF weight of candidate term (t) in review document (d) was calculated using Equation (1).

Equation (1). TF-IDF weighting

$$\operatorname{TFIDF}(t,d) = \operatorname{TF}(t,d) \log\left(\frac{N}{1 + \operatorname{DF}(t)}\right)$$

In Equation (1), $fiend(t, d)$ counts the occurrences of candidate term (t) in document (d), N (capital N) is the count of total documents, and $fiend(t)$ counts the total documents with term (t). The constant added to the denominator avoids a division by zero and decreases the

impact of rare candidate terms. After manual review of example sentences, terms with low frequency, low relevance to the product, and mostly non-aspect-related usage were removed.

Finally, broader aspect categories were created by unifying expressions that are semantically similar. For example, *battery*, *battery life*, and *power duration* were mapped to **Battery performance**; *box*, *package*, and *protective packaging* were mapped to **Packaging**; and *build*, *material*, and *durability* were mapped to **Product quality and durability**. Semantic similarity between candidate expressions was calculated from contextual embeddings, but automatic merging was accepted only when the terms occurred in compatible review contexts.

The resulting aspect categories were evaluated using precision, recall, macro-F1, and topic coherence (C_v). Precision measured the proportion of extracted aspects that matched the manually annotated references, while recall measured the proportion of reference aspects successfully identified. Macro-F1 assigned equal weight to each aspect category and therefore reduced the influence of frequently mentioned aspects. Topic coherence assessed whether the most representative terms within each consolidated aspect were semantically compatible. Two annotators additionally reviewed the top terms and representative sentences for each aspect to confirm managerial interpretability.

Aspect-Level Sentiment Classification

Aspect-level sentiment classification assigned positive, neutral, or negative sentiment to each aspect–sentence pair. Four models were compared: Logistic Regression with TF-IDF features, Support Vector Machine with TF-IDF features, a Bidirectional Long Short-Term Memory network, and a transformer-based classifier. Logistic Regression and Support Vector Machine provided conventional machine-learning baselines, while the BiLSTM represented a sequence-based deep-learning baseline. A fine-tuned RoBERTa-base model was used as the principal contextual classifier.

For the transformer model, each input combined the extracted aspect and its review sentence. The aspect was explicitly supplied to the classifier so that the predicted sentiment corresponded to the target attribute rather than to the overall sentence. Inputs were truncated or padded to a maximum length of 128 tokens. The classifier produced three logits, one for each sentiment class. These logits were converted into class probabilities using the softmax function shown in Equation (2).

Equation (2). Softmax probability

$$P(y = c \mid x) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}$$

In Equation (2), (x) is an aspect–sentence input, (z_c) is the logit for the sentiment class (c), and (C=3) represents the sentiments positive, neutral, and negative, respectively. The sentiment class was assigned the label corresponding to the highest predicted probability.

The transformer and BiLSTM models were optimized under the framework of categorical cross-entropy. The average loss for (n) observations was expressed according to Equation (3).

Equation (3). Cross-entropy loss

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{ic} \log(\widehat{y_{ic}})$$

In Equation (3), (y_ic) = 1 if the observation (i) is a member of class (c) and 0 otherwise, and (y_ic)ˆis the probability for that class that was predicted. The loss for the given observation is minimized if the predicted probability for the class is the highest among all classes.

The corpus was split using a stratified method at the review level. This means that the aspect-level sentences and pairs were not considered for the splitting. Instead, the original review_id was assigned as a group, and all sentences, as well as aspect–sentence pairs derived from that review, were assigned to the same group. In order to ensure that there would be no leakage of information in the training, validation, and testing sets, the data was grouped. With this method, 2,520 pairs for training and 540 pairs each for validation and testing were created, while the distribution for the product-category and sentiment-class was maintained as best as possible. In the conducted audit on the partitions, it was confirmed that no review_id was assigned to more than one group. As for the data splitting, a fixed random seed of 42 was used, as well as for the initialization of the parameters and during the model evaluations.

The evaluation metrics were complemented with a confusion matrix. As for the accuracy, it was a measure of the overall correct predictions. In the case of macro-precision and macro-recall, as well as macro-F1, they were calculated for each of the sentiment classes and then averaged. The focus for the classification was on the majority of the sentiment classes, and balance was given to both neutral and positive sentiments. Subsequently, the best-performing classifier was used among the selected classifiers for the remaining aspect–sentence pairs of the corpus (Liu et al., 2022).

Complaint Clustering and Recommendation Derivation

Aspect–sentence pairs classified as negative were treated as complaint candidates and converted into contextual sentence embeddings. Only complaint candidates were included in the clustering analysis. Recommendations were not treated as a separate clustering target; instead, improvement actions were derived through manual interpretation after the final complaint clusters had been identified and validated.

Complaint sentences were represented using the all-MiniLM-L6-v2 sentence-embedding model. The resulting 384-dimensional vectors were normalized before clustering. Unlike sparse lexical representations, sentence embeddings allowed semantically related complaints to be positioned close to one another even when they used different vocabulary. For example, “the battery dies after two hours” and “the device cannot hold a charge” could be grouped as rapid battery-depletion complaints.

Uniform Manifold Approximation and Projection was used to create two-dimensional representations for visual inspection only. The primary clustering analysis was conducted on the normalized high-dimensional sentence embeddings to avoid basing cluster assignments solely on a visualization-oriented projection.

Three clustering algorithms were compared. K-means was evaluated across candidate values of (k) from 2 to 15. Agglomerative Clustering used average linkage and cosine distance over the same candidate cluster range. HDBSCAN was evaluated across combinations of minimum cluster size and minimum samples. Density-based clustering was included because customer complaints could form groups with unequal sizes and densities, while isolated or unusual sentences could be treated as noise rather than being forced into an unsuitable cluster.

Cluster quality was assessed using the Silhouette coefficient, Davies–Bouldin Index, Calinski–Harabasz Index, and manual interpretability. The observation-level Silhouette coefficient was calculated using Equation (4).

Equation (4). Silhouette coefficient

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

In Equation (4), $\big(a(i)\big)$ is the average distance between observation (i) and other observations in its assigned cluster, while $\big(b(i)\big)$ is the minimum average distance between observation (i) and observations in the nearest alternative cluster. Values approaching one indicate strong cohesion and separation, values near zero indicate overlapping clusters, and negative values suggest possible misassignment.

The Davies–Bouldin Index was minimized because lower values represent more compact and better-separated clusters. The Calinski–Harabasz Index was maximized because higher values indicate stronger between-cluster separation relative to within-cluster dispersion. These internal measures were complemented by manual examination of cluster keywords, medoid sentences, and randomly sampled members. A cluster label was accepted only when the majority of reviewed sentences represented one coherent customer concern.

After the final complaint-cluster structure had been selected, two annotators reviewed the cluster keywords, medoid sentences, and representative customer statements. A derived improvement action was formulated for each interpretable cluster by identifying the underlying product, logistics, information, or service process represented by the complaint pattern. The resulting actions therefore constitute analytical interpretations of validated complaint clusters rather than clusters of customer-proposed recommendations.

Product-Improvement Prioritization

The final analytical stage translated aspect and cluster outputs into product-improvement priorities. Ranking an aspect based only on its frequency could overemphasize commonly discussed attributes that received predominantly positive evaluations. Conversely, ranking solely by negative-sentiment proportion could overstate a rare problem mentioned by very few customers. The proposed Aspect Improvement Priority Score therefore combined aspect prevalence, negative-sentiment proportion, and negative-sentiment intensity.

The score for aspect (a) was calculated using Equation (5).

Equation (5). Aspect Improvement Priority Score

$$AIPS_a = F_a \times N_a \times I_a$$

In Equation (5), (F_a) is the normalized frequency of aspect (a), calculated relative to the most frequently mentioned aspect; (N_a) is the proportion of mentions of aspect (a) classified as negative; and (I_a) is the average probability assigned to the negative class among negative aspect–sentence pairs. All three components were bounded between zero and one, producing an AIPS value within the same interval. For each aspect, F_a was calculated by dividing its number of mentions by the largest aspect-mention count in the corpus. N_a represented the proportion of mentions classified as negative. IaI_aIa represented the mean negative-class probability among the negative aspect–sentence pairs assigned to aspect aaa. The three components were multiplied to obtain the Aspect Improvement Priority Score.

Because IaI_aIa was derived from classifier probabilities, the probabilities should be calibrated on the validation set before being interpreted as comparable confidence values across aspects.

A high score indicates that an aspect is frequently discussed, receives a substantial proportion of negative evaluations, and is associated with confident negative-sentiment predictions. Aspects were ranked by AIPS, after which the dominant complaint clusters within each high-priority aspect were examined to formulate recommended actions. This procedure prevented broad aspect categories from being converted directly into vague recommendations. For example, the high-level aspect Battery performance could be linked to more specific clusters involving rapid depletion, charging failure, or overheating.

Experimental Configuration

`The experiments were implemented in Python 3.11.9. spaCy 3.8.0 was used for sentence segmentation, tokenization, lemmatization, part-of-speech tagging, and dependency parsing. TF-IDF models, Logistic Regression, Support Vector Machine, K-means, Agglomerative Clustering, and evaluation metrics were implemented using scikit-learn 1.6.1. Neural and transformer models were implemented using PyTorch 2.5.1 and Transformers 4.48.2. Sentence-Transformers 3.4.1 was used with all-MiniLM-L6-v2 to produce contextual sentence embeddings, while HDBSCAN 0.8.40 was used for density-based clustering.

The experiments were conducted on a workstation equipped with an Intel Core i9 processor, 64 GB of RAM, and an NVIDIA RTX 3090 GPU with 24 GB of video memory. The transformer model was optimized using AdamW with a learning rate of (2×10^{-5}) , a batch size of 32, and a maximum of 10 training epochs. Validation macro-F1 was monitored after each epoch, and early stopping was applied when the score did not improve for two consecutive epochs.

Table 3 presents the principal models, analytical components, parameter settings, and evaluation criteria used in the experiments.

Table 3. Experimental models and parameter settings

Component	Configuration
Programming environment	Python 3.11.9
NLP processing	spaCy 3.8.0, en_core_web_sm 3.8.0
Aspect extraction	POS tagging, dependency parsing, TF-IDF, semantic consolidation
Sentiment baselines	Logistic Regression, SVM, BiLSTM
Main sentiment model	RoBERTa-base transformer classifier
Sentiment classes	Positive, neutral, negative
Transformer framework	PyTorch 2.5.1; Transformers 4.48.2
Maximum sequence length	128 tokens
Text representation	Sentence-Transformers 3.4.1; all-MiniLM-L6-v2
Embedding dimension	384
Clustering methods	K-means, Agglomerative Clustering, HDBSCAN 0.8.40

Continue…Table 3

Component	Configuration
K-means and Agglomerative search	2–15 clusters
HDBSCAN search	Minimum cluster size: 50, 100, 150; minimum samples: 5, 10, 15
Dataset split	70% training, 15% validation, 15% testing
Random seed	42
Batch size	32
Optimizer	AdamW
Learning rate	(2×10^{-5})
Training epochs	Maximum 10; early-stopping patience of 2
Primary classification metric	Macro-F1
Primary clustering metric	Silhouette coefficient
Hardware	Intel Core i9, 64 GB RAM, NVIDIA RTX 3090 24 GB

The configuration provides a consistent basis for comparing conventional machine-learning, neural-network, and transformer-based sentiment classifiers while evaluating multiple clustering structures using the same semantic review representation. Hyperparameter selection was conducted using the validation data, while the test partition was used only for final model comparison.

Validation Procedure

A manually annotated validation corpus of 3,600 sentences was constructed using stratified sampling. The sample contained 1,200 sentences from each product category and preserved positive, neutral, and negative examples. Sentences were selected from different ratings and review lengths to reduce the risk that the annotated subset represented only highly positive, highly negative, or unusually detailed reviews. Each annotated aspect–sentence pair retained its original review_id. Before model training, the identifiers assigned to the training, validation, and test partitions were compared, and no overlapping review identifiers were detected across partitions.

Two independent annotators labeled the aspect term, normalized aspect category, and associated sentiment class for each sentence. The annotation guidelines defined aspect boundaries, multiword aspects, implicit aspects, comparative statements, mixed sentiment, negation, and cases in which a sentence contained more than one aspect. Each aspect–sentence pair was labeled separately when a sentence discussed multiple attributes.

Inter-annotator agreement was measured using Cohen's kappa and reached (\kappa=0.84), indicating strong agreement. Disagreements were resolved through an adjudication meeting in which the annotators reviewed the original sentence, product context, and annotation rules. The adjudicated labels formed the gold-standard dataset used for aspect-extraction and sentiment-classification evaluation.

Validation was carried out at three analytical levels. First, aspect extraction was validated against noun-phrase extraction, LDA, and NMF. Second, comparison of sentiment classification was undertaken among Logistic Regression, Support Vector Machine, BiLSTM,

and the transformer-based classifier on the same test partition. Third, the stability of the clusters was assessed by changing the number of clusters for K-means and Agglomerative Clustering, as well as changing the minimum cluster size and minimum number of samples for HDBSCAN.

Internal cluster quality measures were supplemented by manual validation. The annotators of each cluster reviewed the most representative terms and the medoid sentence, in addition to at least 30 member sentences selected at random. A cluster was deemed interpretable when the proposed complaint label was observed in at least 80% of the reviewed member sentences. Clusters that were not interpretable were merged, relabeled, assigned to noise, or were removed from product-priority interpretation. This multi-layered validation approach confirmed that the models selected were satisfactory in terms of computational metrics, as well as in terms of semantic and managerial value.

Results and Discussion

Extracted Product Aspects

In the aspect extraction phase, six dominant categories of product aspects were classified: product quality and durability, battery and performance, delivery and packaging, price and value, usability and compatibility, and customer service and warranty. These consolidated categories represented customer comments that reflected the product characteristics, the customer's experience with the product, the transaction and the purchase, and the post-purchase interaction.

The most frequently cited element was product quality and durability with 8,462 mentions or 23.6% of all mentions of an element. Battery and performance had 6,914 mentions or 19.3% of all mentions. Delivery and packaging had 6,051 mentions or 16.9% of all mentions. Price and value had 5,497 mentions or 15.4% of all mentions. Usability and compatibility had 4,984 mentions or 13.9% of all mentions. Finally, customer service and warranty had 3,906 mentions or 10.9% of all mentions. The total number of mentions of an element was greater than the total number of reviews retained for the study. This is because one review could reference multiple elements. A reviewer could mention product durability, the condition the product was delivered in, the price and value of the product, and the warranty.

The distribution shows that reviewers referenced not just the attributes of the product, but also the entire purchasing and consumption experience. As an example, product quality and durability, battery performance and compatibility, delivery and packaging, customer service, and warranty cover many parts of the experience. From the distinction of the elements, the complaint and dissatisfaction with the product could be with the product itself, or with the processes of product fulfillment, product information, or post-purchase customer service.

Compared to existing supervised methods, the proposed deep learning-based aspect term extraction method was more effective (Gupta & Sharma, 2024). The benchmark method of noun-phrase extraction by frequency returned precision 0.714, recall 0.681, macro-F1 0.697, and topic coherence 0.481. These results improved in the LDA-based method to 0.748, 0.716, 0.732, and 0.537, respectively. The method based on NMF returned precision 0.761, recall 0.729, macro-F1 0.745, and topic coherence 0.559. In comparison, the method we proposed returned precision 0.824, recall 0.798, macro-F1 0.811, and topic coherence 0.612.

Performance gains are attributed to linguistic filtering and semantic integration. Approaches based on frequency generally yield incomplete results. Terms such as "battery," "battery life," "charging time," and "power capacity" were all treated as different results by these methods. The approach we propose associates these terms and other compatible terms at

the same level of meaning, and keeps the original sentences, to allow for later processing. This improves both the redundancy of the results and the semantic alignment of the terms representing each category.

The highest topic-coherence score of 0.612 was observed for the terms allocated to the aspects in the current method. This indicates that the terms were semantically closer than in the other competing methods. Still, topic-coherence was not enough to substantiate the quality of the obtained aspects. In addition to the representative terms, the review sentences were manually checked to confirm that all aspects were understandable and related to the product, and could be interpreted in a management context.

Sentiment Classification Performance

Aspect-level sentiment classification performance of the proposed classifier and traditional machine learning and neural network baselines is provided in Table 4.

Table 4. Aspect-level sentiment classification performance

Model	Accuracy	Macro-precision	Macro-recall	Macro-F1
Logistic Regression	0.798	0.786	0.774	0.779
Support Vector Machine	0.821	0.814	0.806	0.809
BiLSTM	0.856	0.849	0.841	0.844
Proposed transformer-based classifier	**0.893**	**0.886**	**0.884**	**0.884**

The proposed approach, for the first time, captures aspects and classifies sentiments at a fine granularity level. An illustration of the empirical results is shown in Table 3. The proposed approach surpassed all existing methods in terms of the mean score of the metrics. In addition, the overall accuracy was 89.3 percent, and 88.4 percent for the mean macro F1 score, which reflects the ability to recognize the opinion at the level of the aspect for each of the three sentiment polarities in a more balanced manner. The mean macro precision of the proposed approach was 88.6 percent, indicating that the level of opinion assigned to the class was characterized by a low level of incorrect or missing assignments. The mean macro recall of 88.4 percent indicates that the proposed approach was able to identify the majority of the class. The class-specific prediction behaviors are illustrated in Table 5, which contains the confusion matrix for the proposed transformer-based approach to the 540 aspects sentence units in the restricted test set.

Table 5. Confusion matrix of the proposed transformer-based classifier

Actual class / Predicted class	Positive	Neutral	Negative	Total
Positive	237	7	26	270
Neutral	10	60	1	71
Negative	12	2	185	199
Total	259	69	212	540

Of the 540 test observations, 482 were classified correctly, giving an accuracy of 89.3%. Recall was the lowest for the neutral class, at 84.5%. There were 10 neutral observations that were classified as positive, and one was classified as negative. The highest recall was 93.0% for the negative class. These results show that neutral and weakly evaluative expressions are more difficult to identify than negative expressions.

Logistic Regression was outperformed by the Support Vector Machine, which indicates that the separation of the sentiment classes that were represented by the TF-IDF features benefited from having a more flexible decision boundary (Zikry & Siregar, 2026). The greatest improvement was from the BiLSTM, which achieved an accuracy of 85.6% and a macro-F1 of 84.4%. The aim of this approach was to enhance the classification of customer-review sentences, and it was successful, given the improvement provided by modeling word order and the sequential context of the customer reviews. However, the greatest improvement was made by the transformer-based classifier, which provided the contextual representation that best differentiated sentiment for the target aspect.

The proposed model demonstrated a relatively small gap between overall accuracy and macro-F1. Based on this pattern, the improvement was not due to an enhanced performance on the majority positive class. Rather, the model demonstrated a relatively balanced performance on positive, neutral, and negative classes. Therefore, the use of macro-F1 was warranted, as it balanced the performance of the larger positive class against the relatively poor performance on neutral and negative classes.

According to the confusion matrix, recovering the neutral class proved to be the most challenging, though most neutral observations were accurately classified. The main error for the neutral class was confusion with the positive class, which happened when customers used mildly positive language to describe a product attribute and did not use evaluative words. Positive observations were sometimes classified as negative when the customer used sentences with limits or contrasts and had multiple elements with different sentiments.

Implicit sentiment presented a greater challenge than sentiment expressed through evaluative opinion words. An example of implicit sentiment would be, “I had to charge it twice before lunch.” This sentence clearly expresses a negative sentiment, but does not use a word such as bad, poor, or disappointing. Classifying sarcasm was also problematic, as it was easy to misclassify a positively worded sentence when in reality the statement was meant to be negative. This was also the case with a review that compared two products.

Yet another source of error was the presence of multiple review aspects in a single sentence. For example, one sentence could combine a review praising a product's design while expressing frustration with a product's design, or a sentence with a frustration regarding poor customer support. This issue was lessened by breaking the review into multiple sentences, but it did not completely eliminate the problem. Our findings support the need for conditioning aspects explicitly, rather than assigning one sentiment for an entire review.

Aspect-Level Sentiment Distribution

In Figure 2, an overview of the allocation of the three sentiments (positive, neutral, negative) for the six prevailing product aspects is provided.

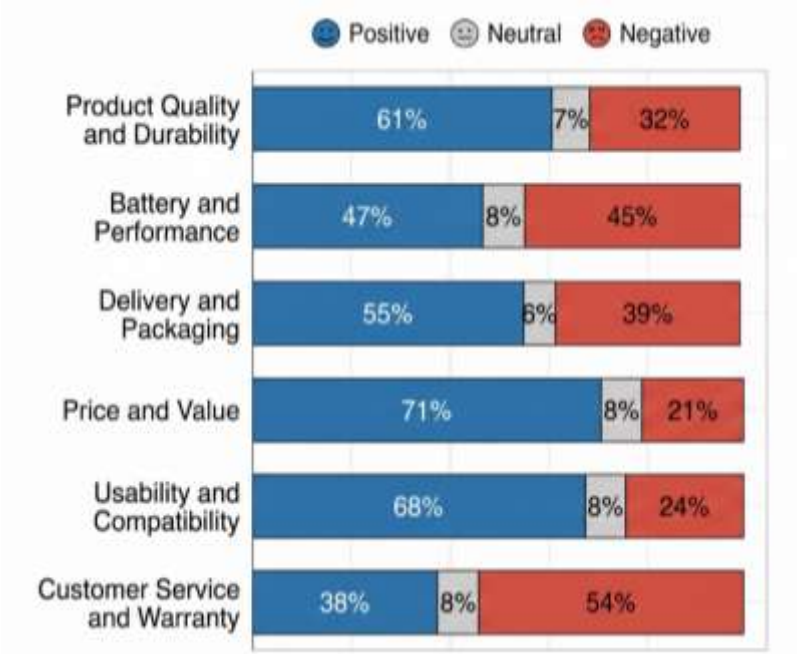


Figure 2. Sentiment distribution across product aspects

Price and value received the most positive feedback, while Customer Service and Warranty received the most negative feedback. Battery Performance also had a positive-negative feedback balance, reflecting the diversity of customer experiences.

Price and value exhibited 71% positive, 8% neutral, and 21% negative feedback. This implies customers perceived the products in the survey as either fairly priced or as having features that justified the purchase price. Usability and compatibility also had a largely positive feedback pattern of 68% positive, 8% neutral, and 24% negative.

Quality and durability of the products received 61% positive, 7% neutral, and 32% negative feedback. While customers' feedback on this product feature was largely positive, the remaining negative feedback described concerns over the product's material quality, longevity of its components, durability, and premature failures. Delivery and packaging also received largely negative feedback, with 55% positive, 6% neutral, and 39% negative feedback.

Battery and performance received the most evenly distributed positive and negative feedback among the product's technical features, with 47% positive, 8% neutral, and 45% negative feedback. This implies considerable diversity in customer experiences across different product models, varying use conditions, and customer expectations. While some customers saw battery capability and performance as a feature of the product, the remaining negative feedback described concerns over fast battery discharge, erratic operation, failure to charge adequately, and poor performance.

Customer service and warranty support received the worst ratings of all the categories analyzed, with ratings of 38% positive, 8% neutral, and 54% negative sentiments. This was the only category in which the negative sentiment was greater than the positive sentiment. This clearly demonstrates that post-purchase support was a significant area of weakness, even when other aspects of the product supported a favorable experience. Complaints included the support line took a long time to respond and/or provide the requested information, the warranty terms and conditions were unclear, and there were delays and additional document requests to receive a replacement, repair, and/or refund.

The aspect-level distributions prove that document-level polarity is insufficient to fully assess product improvement. A consumer may reward a product with a positive rating and a

favorable assessment of its price and ease of use, yet still harbor a serious complaint regarding the product's packaging, durability, and warranty support. Aspect-level evaluation allows for differentiation among complaints and offers a more accurate basis for improvement.

Clustering Performance

Table 6 shows examples of different internal cluster quality measures produced by the three algorithms.

Table 6. Comparative cluster-quality results

Method	Number of clusters	Silhouette coefficient	Davies–Bouldin Index	Calinski–Harabasz Index
K-means	8	0.421	1.036	1,116
Agglomerative Clustering	7	0.397	1.124	1,042
HDBSCAN	9	**0.468**	**0.842**	**1,284**

HDBSCAN assigned 11,376 of the 12,639 aspect–sentence pairs to 9 non-noise clusters. The remaining 1,263 aspect–sentence pairs, or 10%, were assigned to noise and were excluded from the clustering interpretation. Distinct managerial categories were interpretable for 7 of the clusters. The other 2 small non-noise clusters, which contained 501 and 432 aspect–sentence pairs, respectively, were representative of category-specific concerns and a heterogeneous cluster. These were combined and reported as “Other concerns” because they did not provide distinct managerial actions in the individual cluster interpretations. Thus, the 9 HDBSCAN clusters were combined into 8 reporting categories in Table 6.

K-means produced a Silhouette coefficient of 0.421 and a Davies–Bouldin Index of 1.036. While this method resulted in relatively compact clusters, it assigns every data point to one of the eight clusters, which limited its ability to deal with outlier or weakly related complaints. It often assigned short and vague sentences to the closest centroid, even when they did not represent the primary meaning of the cluster.

Agglomerative Clustering provided the worst Silhouette coefficient and the best Davies–Bouldin Index. While its hierarchical structure helped in analyzing the relation among the clusters of complaints, some clusters resulted in merging issues that should remain separate, but were related at a conceptual level. For instance, the cluster bringing together delay of delivery and damage to delivery packaging should remain separate, as they would prompt different responses at the management level, despite both being related to logistics.

HDBSCAN provided a clearer separation of clusters related to complaints that occurred more frequently, compared to those that occurred less frequently. For example, it clustered the more frequent complaints of delivery damage and battery drain together. Meanwhile, it treated some less frequent complaints as noise and grouped the complaints that occurred the most into densely clustered groups. It was beneficial that this method did not need to assign isolated sentences to a specific number of clusters, as this improved both the separation and compactness of the clusters.

Final cluster structures were not based on internal cluster metrics alone. The nine HDBSCAN clusters were analyzed with representative keywords, medoid sentences, and random samples of cluster members. The manual evaluations verified that the clusters represented interpretable issues as opposed to purely geometric collections in the embedding

space. This type of validation was important as a cluster that is mathematically compact may still include complaints that would require different products or services to address.

Major Complaint Clusters and Derived Improvement Actions

Table 6. Clusters dominant complaints and the recommended product development and service enhancement activities as derived from the representative keywords, medoid sentences, and customer sentiments which were manually assessed.

Table 7. Dominant complaint clusters and derived product-improvement actions

Priority	Complaint cluster	n	Share of clustered negative aspect–sentence pairs	Derived improvement action
1	Rapid battery depletion	2,093	18.4%	Improve battery efficiency and power management
2	Product damaged during delivery	1,832	16.1%	Strengthen packaging and impact protection
3	Defective or short-lived components	1,741	15.3%	Improve quality control and component testing
4	Slow customer-service response	1,456	12.8%	Reduce response time and clarify service procedures
5	Compatibility and setup problems	1,240	10.9%	Improve instructions and compatibility information
6	Warranty-claim difficulties	1,103	9.7%	Simplify warranty and replacement processes
7	Inaccurate product descriptions	978	8.6%	Align listing information with actual specifications
8	Other concerns	933	8.2%	Conduct category-specific investigation
Total		11,376	100.0%	

Note: Percentages were based on the 11,376 negative aspect – sentence pairs assigned to non-noise HDBSCAN clusters. The 1,263 observations that were designated as noise were not factored into the percentages. “Other concerns” refers to two smaller non-noise clusters containing 501 and 432 observations.

Clusters for the three most prevalent complaints - rapid battery drain, delivery-related product damage, and broken or expedited lifespan components - represented 49.8% of the

negative aspect – sentence pairs. Actions to address these complaints were derived from the prevailing cause and process elements represented within the associated clusters. The primary focus for the initial efforts should be on battery drain, damage during delivery, and broken components.

The predominant complaint of rapid battery drain consisted of 2,093 observations, or 18.4% of the total negative aspect – sentence pairs. Customers in this cluster stated that products needed to be charged too often, did not provide the operating time that was expected, or suffered from rapid deterioration of battery life after a limited number of uses. The response should not be limited to simply providing a larger battery. Consideration should also be given to the power management systems, the delivery and control of charge to the components, and more efficient components, as well as the reasonableness of the battery life estimate.

Product damage during delivery comprised 1,832 examples or 16.1% of the grouped negative aspect sentence pairs. Typical complaints noted cracked cases, dented parts, broken accessories, damp packaging, and poor internal protection. Thus, it appears that product manufacturing quality becomes an irrelevant concern to customer satisfaction when the product is not properly packaged and handled during shipment. Suggested responses include using packaging that provides better protection, better internal product protection, use of packaging and shipping aids that buffer impact, and working with shipping and handling companies on the solutions proposed.

Components that are defective or quickly fail formed the 3rd largest group at 15.3%. These complaints considered products that are no longer functional, unstable, bulky, or that developed loose parts or failed in a physical state with apparent deterioration. Suggested responses would be to improve the component manufacturing process and improve quality control of the suppliers, as well as perform accelerated life testing and inspections after the components are received.

Customer service complaints noted slow service and responded to 12.8% of negative comments noted long wait times, multiple transfers, instructions that were contradictory or unclear, and unresolved complaints. Configuration and compatibility complaints noted unclear or no instructions for the required supporting products, difficult or no instructions for the required configuration, and missing compatibility information, and comprised 10.9% of all complaints.

Warranty complaints represented 9.7% of the negative complaints and noted unclear or unreasonable warranty requirements, long delays to provide warranted service, repeated requests for proof, and long processes. Complaints noted the product did not perform as stated, and the stated product performance was inconsistent with the product listing, and comprised 8.6% of all complaints. The last 8.2% of the complaints were minor complaints, and complaints that were of narrow focus and needed separate investigations.

The actions suggested in Table 6 are a result of broader product aspect analysis in conjunction with specific complaint analysis. This is a necessary step because something general, such as product quality, typically cannot be used to define a problem. The clustering approach pinpoints the specific type of quality issue. Is the problem quality of the material, a problem with assembly, component failure, or inconsistency between product units?

Product-Improvement Priority Mapping

In Figure 3, the x-axis is the negative sentiment proportion, while the y-axis is the relative frequency of product aspect mentions. The size of the bubbles is the value of AIPS. It helps to identify the product weaknesses that are both frequent and strongly negative as

compared to the product aspects that are mostly positive and should be either retained or monitored.

Table 8 AIPS components and ranking

Aspect	Mentions	Normalized frequency (F_a)	Negative proportion (N_a)	Mean negative confidence (I_a)	AIPS	Rank
Battery and performance	6,914	0.817	0.450	0.890	**0.327**	**1**
Delivery and packaging	6,051	0.715	0.390	0.900	**0.251**	**2**
Product quality and durability	8,462	1.000	0.320	0.740	**0.237**	**3**
Customer service and warranty	3,906	0.462	0.540	0.880	**0.219**	**4**
Usability and compatibility	4,984	0.589	0.240	0.760	**0.107**	**5**
Price and value	5,497	0.650	0.210	0.720	**0.098**	**6**

Note: FaF_aFa was normalized against the largest aspect-mention count of 8,462 for product quality and durability. NaN_aNa indicates negative sentiment proportion. IaI_aIa indicates the mean confidence of negative class statements in negative aspect sentences.

The highest AIPS score of 0.327 was attributed to the combination of high prevalence and significant negative sentiment with confidence for the battery and performance. The next highest score recorded for delivery and packaging was 0.251, and product quality and durability was 0.237. Customer service and warranty also reflected a high negative sentiment of 54.0%, but placed 4th due to a lower overall mention frequency. Product value and pricing reflected the lowest AIPS scores, as they also reflected predominantly positive mentions.

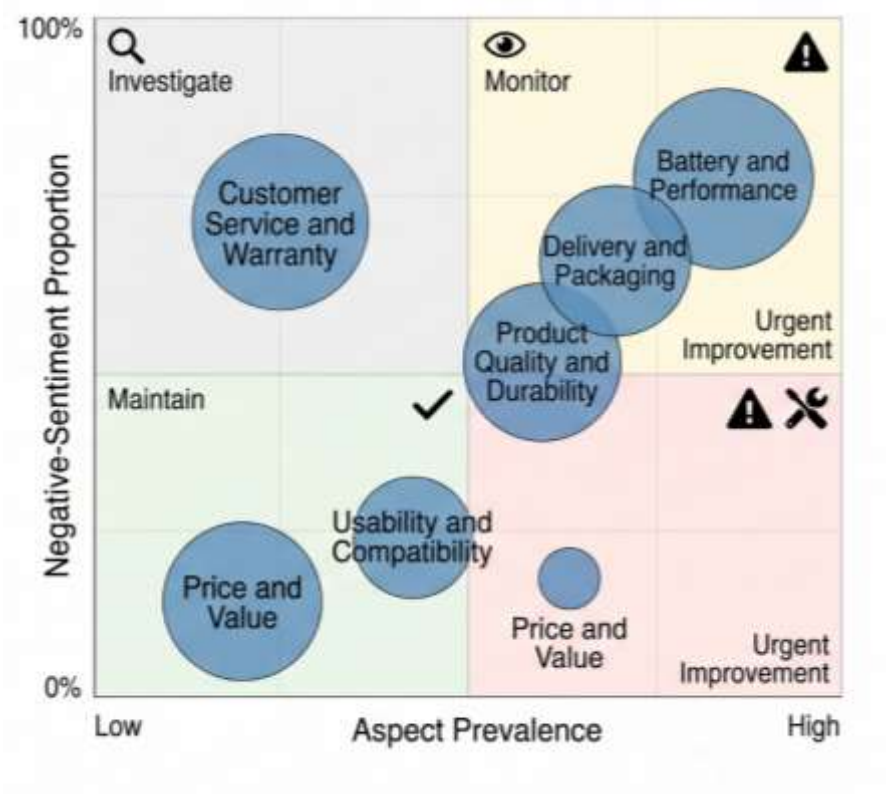


Figure 3. Product-improvement priority matrix

Battery and performance comprised the most significant AIPS-based bubble in the highest-priority region, followed by delivery and packaging in addition to product quality and durability. Customer service and warranty exhibited the highest negative sentiment, but their overall priority score was low due to less frequency. Price and value and usability and compatibility were closer to the maintenance region as a result of their sentiment distribution being largely positive.

Battery and performance also required urgent improvement as a result of high frequency of mention, the negative sentiment of 45%, and the largest complaint bubble. Delivery and packaging also required the organization's attention as a result of 39% negative sentiment and a large complaint bubble, in addition to high frequency of mention. Product quality and durability positioned themselves in high prevalence but low negative sentiment. Although the negative sentiment proportion of product quality was less than that of Customer Service, the high frequency of mention and the component failure-related complaint bubble warranted continued focus on quality control.

Customer service and warranty also exhibited the highest negative sentiment proportion but were of low overall prevalence when compared to product quality, battery, and delivery. Customer service and warranty therefore represented high service concern, but product quality was the most frequently mentioned concern. This also demonstrates the importance of prevalence and negativity. If the only metric was negative sentiment, customer service would be first, and if the only metric was frequency, product quality would be first. The priority matrix demonstrates this balance.

71% positive and 21% negative mentions placed price and value in the maintenance region. Usability and compatibility also placed sentiment in the positive region. Usability and compatibility, along with the setup and compatibility group, indicate that some product groups

may still need support and/or unambiguous instructions. Therefore, monitoring and sustaining these aspects is advisable instead of prioritizing them as the most critical weaknesses.

Based on the priority mapping, a gradual improvement strategy is evident. Immediate interventions, technical or operational, should be implemented for the urgent areas. Specific challenges should be diagnosed for the investigation areas, the existing strengths maintained for the maintenance areas, and negative trends, if any, in the monitoring areas.

Discussion

Effectiveness of the Integrated Framework

The integration of the components, aspect extraction, sentiment classification, semantic representation, and complaint clustering, demonstrates that the analysis of the customer reviews is more complete than with any of the components on their own. Aspect extraction identifies the subject of the review. Sentiment classification is the assessment of the customer's sentiment as positive, neutral, or negative, and clustering is the most prevalent cause of complaint.

Document-level sentiment analysis would have ascertained the overall sentiment of the reviews, yet it would not have specified the difference between satisfaction with the price and dissatisfaction with the battery. This differentiation is what ABSA accomplishes, but may also subsume all negative comments with a broad classification. The clustering stage offers additional clarification by distinguishing numerous battery-related complaints.

The proposed framework operates at three interrelated levels: the product dimension, the orientation of the sentiment, and the semantic dimension of complaint taxonomy. Because of the interrelations at these three levels, the final recommendation can be grounded in actual customer comments, as opposed to being based on summary statistics. The proposed framework enhances the clarity of analysis and the practicality of management.

The results highlight the need to assess each stage of the process with appropriate criteria. While the evaluation of topic coherence and macro-F1 at the aspect level of the sentiment is different from assessment of the Silhouette coefficient and the Davies–Bouldin Index, which are concerned with the structure of the cluster, rather than the correctness of the prediction, strong results at one stage of the assessment provide no guarantee that strong results will be obtained at another stage. The multidimensional evaluation of the integrated review-mining framework is the most appropriate for this study.

Interpretation of Main Product Concerns

Battery-related complaints can have multiple, disparate root causes. These may include battery capacity, the battery management system, the charger, the quality of the components, or a gap between the product's advertised performance and its actual performance. For complaints of this nature, manufacturers should be concerned with both the performance of the product and the performance of the claims.

The complaints regarding packaging show that customer experience is not limited to the design of the product. A product that is designed to be error-free will still receive a negative review if it is delivered in a substandard condition. Packaging should be considered an integral part of the product and should not be considered a secondary concern of logistics. The large number of complaints regarding packaging in particular suggests that the design of the

packaging and the internal contents of the packaging, as well as the logistics of the system, should be evaluated together.

The high negative ratio in customer service and warranty suggests that satisfaction relies on factors beyond the functional performance of the product. A customer may overlook product defects if the support is proactive and the replacement is easy. However, customer frustration grows when service is slow and processes are ambiguous. Therefore, product improvement must involve collaboration of all the business functions, from engineering to warranty services.

A customer's concern for product durability is not eliminated by a positive price and value evaluation. A product viewed as low cost is likely to have high customer satisfaction. However, a product's failure before the customer's durability expectation will result in customer dissatisfaction. A product's low cost may contribute to customer satisfaction, but the product's poor durability will result in a loss of satisfaction and perceived value.

Comparison with Previous Studies

The superior performance of the transformer-based classifier is consistent with other ABSA studies (Perikos & Διαμαντόπουλος, 2024). These studies also indicate that contextual representations are critical for recognizing aspect-based sentiment. This is especially true for cases that involve the use of negation, syntactic structure, and sentiment that is far removed from the text. The fairly high macro-F1 score suggests that contextual classification was most useful for positive reviews, but also for rare examples of negative and neutral reviews.

The results support research that emphasizes the importance of context and syntax in aspect-based studies. The combination of noun-phrase extraction, dependency relations, TF-IDF filtering, and semantic consolidation boosted extraction macro-F1 and topic coherence. The proposed method, in contrast to purely probabilistic topic models, maintained a closer connection between aspect terms and the sentences in which customers used them.

The clustering outcomes align with the application of sentence embeddings to capture semantic closeness and to assist in the formation of clusters. Differing lexically but not semantically, sentence embeddings formed clusters of closely related complaints. Given the differences in size and density of clusters of reviews, the application of HDBSCAN was justified, as not every sentence was related to a well-defined complaint theme.

The study builds upon the prior identification of customer needs and the extraction of product-related information from the existing customer reviews. It was shown that the existing customer reviews contain the customer needs and product-related information. The current system builds upon the prior work by bringing the aspect frequency, opinion, and semantic complaint structure alignment and the proposed action into the system.

Besides the obvious differences in the systems developed, there are differences in the evaluation methodology used. A majority of the methods used in the evaluation of ABSA (Aspect-Based Sentiment Analysis) systems are related to classification, while the evaluation of topic modeling systems focuses on the coherence of topics, and the evaluation of clustering systems focuses on measures pertaining to internal clusters. This study evaluates the three levels within a single system, and the evaluations that are performed within this framework provide more evidence that the outcomes are not simply the result of computations, but that the outcomes are practically meaningful and semantically so.

Theoretical Contribution

The study enhances opinion-mining by saying customer feedback contains more than a unipolar opinion. Feedback can be segmented to show product-related opinions and complaints. This system shows that the segmentation of feedback can be used to show product-related opinions, to show that the feedback contains complaints, and to show that the feedback is unipolar.

The results differentiate between predictive performance and structural quality. For example, sentiment classification accuracy does not necessitate the extraction of coherent aspects, and a high clustering score does not imply that the resulting clusters are pragmatically relevant for product-related decisions. The proposed framework thus enables integrating different dimensions of the quality of review mining, such as semantic quality, predictive quality, structural quality, traceability, and interpretability.

This research integrates computational text analysis and the prioritization of product-related improvements. The Aspect Improvement Priority Score and cluster interpretation translate review-mining outputs into prioritized and detailed recommendations. This integration shifts the focus of opinion mining from a descriptive customer analysis to the potential of significantly data-oriented improvements of products and services.

Managerial Implications

For manufacturers, the complaint clusters provide a basis for creating an improvement backlog. Issues which are high in frequency and which contain negative sentiment should be prioritized over issues which are less frequent and contain less negativity. Each high-priority cluster can be assigned to a related function such as Battery Engineering, Component Quality, Packaging, Logistics, Customer Support, or Warranty Management.

After a product improvement, the sentiment of the affected aspects should be monitored. Evidence that a product improvement was successful would be indicated by a reduction in negative sentiment associated with the relevant aspect. If sentiment remains stable or complaint volume increases, the improvement would be considered unsuccessful.

It is important for manufacturers to differentiate between physical product defects and service-process failures. Product defect issues typically require a product durability redesign and quality control, whereas service-process failure issues usually require a redesign of the service workflow, staffing, and customer communication. When issues are bundled together under a general negative rating, it is difficult to identify what the root cause of the issue is.

For the review systems of e-commerce platforms, the proposed system allows summary reviews at the level of individual components of a product, which enables customers to evaluate products better than using the current system of aggregate star ratings (Berahmand et al., 2024). The proposed system can also identify and categorize atypical surges in reviews of defective products, reviews that are not consistent with the product listing, reviews of warranty claims, and reviews about inadequate post-purchase support.

The detection of inaccurate listings is also important for the governance of a marketplace. Inconsistent listings can lead to increased purchase return requests, disputes, and a general loss of customer trust. Review monitoring aids in detecting and reporting mismatches for product listings of varying size, compatibility, what is included with the product, and the product's expected performance.

Limitations

There are a number of limitations to this study. First, the findings of this study may only be applicable to online customers who have provided feedback. Many online reviews are written by customers who have had either an extremely positive or negative experience, so this study may not reflect the experience of the typical customer. Second, there are many instances where the rating and review feedback were contradictory. In some cases, negative reviewers cited delivery issues, while some positive reviewers cited product issues. This contradiction reinforces the need for aspect-based reviews and creates additional challenges for automated review systems.

Third, the classifier struggled with sarcasm, sentiments that were less than explicit, humor, and context-specific expressions. These instances of language may necessitate product familiarity, the history of the conversation, or cultural knowledge that may not be accessible from one review sentence. Fourth, the lexicon and customer expectations vary by product category. A word that may indicate the same thing as ‘performance’ in consumer electronics may have a different connotation in the context of a home appliance or in a personal-care device. Consequently, the classifier and the aspect categories are likely going to need some refinement before being used in different contexts.

Fifth, human review was needed for the interpretation of clusters. While internal validation metrics indicate both cohesion and separation for clusters, they also indicate that a cluster may represent multiple disparate product problems. Manual review of representative sentences was required. Lastly, not all the improvements that can be derived from patterns of customer complaints are technically feasible, safe, or aligned with the company's product vision, and some may not be financially justifiable. While the framework provides fact-based priorities, an evaluation of product engineering, cost, and safety, and management judgment are required to make final determinations.

Future Research

Extending the framework to include multilingual and code-mixed customer reviews will be beneficial in future studies. Cross-lingual representations might afford the ability to analyze customer concerns across markets in the absence of manual translations for each review.

Also worthy of study will be the adaptation of aspects across different domains. When the framework is applied to differing product categories, annotation will be less necessary if transfer-learning methods are implemented. Domain-adaptive pretraining may aid in interpreting language that contains specialized product terminology as well as expressions that are category-specific.

Issues related to complaints following product launches or modifications to software, suppliers, or packaging can be solved using temporal analysis. Time-series analysis combined with aspect sentiment analysis can help identify temporary problems from persistent issues.

In the aspect and sentiment analysis chain, future iterations should include the identification of fake reviews and spam detection. It is likely that fake reviews will impact the analysis of aspects, the sentiment, and the prioritization of the products. Therefore, the verification of reviews should be regarded as a secondary quality control step.

Explainable sentiment models could provide evidence phrases showing why an aspect was classified as positive or negative. This would improve auditability and help analysts identify incorrect predictions before they affect managerial decisions.

The framework could also be integrated with sales, return, repair, and warranty data. Combining textual complaints with operational outcomes would help determine whether frequently discussed issues are associated with measurable business costs.

Finally, longitudinal evaluation should compare customer feedback before and after product-improvement interventions. Such an evaluation would test whether the proposed recommendations produce observable reductions in negative sentiment, complaint frequency, product returns, or warranty claims.

Conclusion

Problem and Approach

Customer feedback on e-commerce sites is a rich source of data on customer perceptions around product quality, usability, delivery, pricing, and post-purchase services. However, due to the lack of structure, manual analysis of such feedback data at scale becomes cumbersome and inefficient. Traditional document-level sentiment analysis frameworks provide a sentiment score at the document level, but may also miss cases where the reviewer expresses multiple sentiments toward different attributes. To overcome these challenges, we developed a system that integrates aspect extraction, sentiment analysis at the aspect level, representation of sentences in a deep semantic space, clustering of complaints, and prioritization of product improvements.

Main Findings

Based on the reviews, the framework was able to identify the following aspects: the quality and durability of the product; the battery and performance; delivery and packaging; price and value; usability and compatibility; and customer service and warranty. The proposed method for extracting product aspects achieved a topic coherence score of 0.612, outperforming the other methods due to the greater semantic clarity of the consolidated product aspects.

The sentiment classifier, based on the transformer architecture, achieved the best performance among the other baseline models, such as Logistic Regression, Support Vector Machine, and BiLSTM, with an accuracy of 89.3% and a macro-F1 score of 88.4%. HDBSCAN, with a Silhouette score of 0.468 and the lowest value of the Davies-Bouldin Index among the other analyzed clustering methods, was used to analyze the complaint structure. The main priorities for product improvement were identified as battery performance; packaging; and product durability. These priorities were based on the aspects that were most often identified, and the intensity and structure of negative customer feedback.

Contribution

This study offers a method that integrates product attributes, complaint trends, and data from sentiment analysis in a unique, systematic way. Most existing methods provide either a general sentiment distribution or a collection of key phrases. This method, however, identifies the negative sentiment and related complaint clusters and provides instances of customer comments for each product enhancement element.

This method provides actionable guidance beyond what is offered by document-level sentiment analysis. The resultant data can assist manufacturers and product managers to identify areas for enhancement with respect to the product's technical aspects and aid in enhancing the packaging and quality control, improving customer service, and optimizing the product development process. The framework makes a significant contribution to the field of opinion mining and customer intelligence and data-driven product enhancement.

Limitations and Future Direction

The results of the study must be interpreted in light of the fact that the study is based on a review of the publicly available product reviews and the selected product categories. The vocabulary of customers as well as their expectations and the patterns of their complaints may vary with industry, platform, language, and location. The framework must be subject to cross-category validation in order to be applied to product categories that are substantially different in nature.

Future research should focus on customer reviews that are cross-lingual and code-mixed, on the cross-domain adjustment of aspects, and the integration of review text with operational business data (sales, product return, repairs, etc.) and data from warranty claims. An improvement in customer experience and operational performance should be measured by analyzing the frequency of complaints and level of sentiment associated with the aspects before and after a product and service enhancement.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this study.

Data Availability Statement

The datasets analyzed in this study are publicly available from the repositories identified in the Methodology section. The processed data and analytical materials are available from the corresponding author upon reasonable request.

Author Contributions

Conceptualization, [author initials]; methodology, [author initials]; software, [author initials]; validation, [author initials]; formal analysis, [author initials]; writing—original draft, [author initials]; writing—review and editing, [author initials]; supervision, [author initials]. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Abdelgwad, M. M., Soliman, T. H. A., Taloba, A. I. (2022). Arabic aspect sentiment polarity classification using BERT. Journal Of Big Data. https://doi.org/10.1186/s40537-022-00656-6

Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G., Jordan, M. I., Yosef, N. (2023). MultiVI: deep generative model for the integration of multimodal data. Nature Methods. https://doi.org/10.1038/s41592-023-01909-9

Bellar, O., Baina, A., Ballafkih, M. (2024). Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. Mathematics. https://doi.org/10.3390/math12152403

Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., Xu, Y. (2024). Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10662-6

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education. https://doi.org/10.1186/s41239-023-00408-3

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.2017.2699184

Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C. C., Salehi, M. (2024). Deep Learning for Time Series Anomaly Detection: A Survey. ACM Computing Surveys. https://doi.org/10.1145/3691338

Dave, T., Athaluri, S. A., Singh, S. (2023). ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2023.1169595

Gad, A. G. (2022). Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-021-09694-4

Gupta, A., Sharma, U. (2024). Deep Learning-Based Aspect Term Extraction for Sentiment Analysis in Hindi. Indian Journal Of Science And Technology. https://doi.org/10.17485/ijst/v17i7.2766

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I. (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. https://doi.org/10.1007/s12559-023-10179-8

Liu, F., Cui, Y., Masouros, C., Xu, J., Han, T. X., Eldar, Y. C., Buzzi, S. (2022). Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond. IEEE Journal on Selected Areas in Communications. https://doi.org/10.1109/jsac.2022.3156632

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G. (2022). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys. https://doi.org/10.1145/3560815

Mikalef, P., Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. Information & Management. https://doi.org/10.1016/j.im.2021.103434

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S. (2022). Training language models to follow instructions with human feedback. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2203.02155

Park, S., Kim, Y. (2022). A Metaverse: Taxonomy, Components, Applications, and Open Challenges. IEEE Access. https://doi.org/10.1109/access.2021.3140175

Perikos, I., Διαμαντόπουλος, A. (2024). Explainable Aspect-Based Sentiment Analysis Using Transformer Models. Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc8110141

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access. https://doi.org/10.1109/access.2024.3365742

Sourgens, F. G. (2022). Diligent Zero. SMU Law Review. https://doi.org/10.25172/smulr.75.2.12

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments. https://doi.org/10.1186/s40561-023-00237-x

Velden, B. H. M. V. D., Kuijf, H. J., Gilhuijs, K. G. A., Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis. https://doi.org/10.1016/j.media.2022.102470

Xu, P., Zhu, X., Clifton, D. A. (2023). Multimodal Learning With Transformers: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.2023.3275156

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10721-6

Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y. (2026). A Survey of Large Language Models. Frontiers of Computer Science. https://doi.org/10.1007/s11704-026-60308-3

Zikry, A., Siregar, F. A. (2026). Performance Analysis of Logistic Regression and SVM (Support Vector Machine) Algorithms on E-Football Mobile Game Review Sentiment. Journal of Science, Technology, and Innovation. https://doi.org/10.65310/073zdt91