Expert Systems With Applications 306 (2026) 130865

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automating customer feedback analysis in E-commerce: A multi-Model approach

Laleh Davoodi [a,b], József Mezei [a,*], Shahrokh Nikou [c], Leonardo Espinosa-Leal [d,e]

[a] Faculty of Human and Social Sciences, Åbo Akademi University, Turku, Finland
[b] Faculty of Technology, University of Turku, Turku, Finland
[c] Faculty of Industrial Design Engineering, Delft University of Technology, Delft, The Netherlands
[d] Graduate School and Research, Arcada University of Applied Sciences, Helsinki, Finland
[e] VTT Technical Research Centre of Finland Ltd, Espoo, Finland

ARTICLE INFO

Keywords:
Machine learning
Aspect-Based sentiment analysis
Customer feedback analysis
Large language models
E-Commerce

ABSTRACT

Understanding customer satisfaction in e-commerce is crucial for businesses to remain competitive. While traditional feedback analysis methods are labour-intensive and subjective, machine learning advances have enabled more efficient and scalable sentiment analysis. However, existing models struggle with aspect-based sentiment analysis (ABSA), particularly in detecting implicit aspects and handling mixed sentiments. This paper presents a multi-model machine learning pipeline designed to enhance ABSA by integrating fine-tuned Large Language Models (LLMs) with BERT and RoBERTa-based models. The pipeline consists of an LLM-generated synthesized annotated feedback model, a BERT-based aspect detection model, a RoBERTa-based ABSA model, and an LLM-based ABSA model for handling implicit aspects and mixed sentiments. Additionally, a RoBERTa-based model is employed for overall sentiment detection. By leveraging both manually annotated and synthetic data, the pipeline improves sentiment classification accuracy and aspect coverage, even in data-scarce environments. The results demonstrate that combining multiple models enhances detection accuracy compared to single-model approaches. This study provides a scalable and effective solution for e-commerce feedback analysis, offering businesses valuable insights for improving customer experience and decision-making.

1. Introduction

In the ever-evolving world of e-commerce, understanding the key factors influencing customer satisfaction and the online shopping experience is crucial for businesses to operate and stay competitive in the market (Rose et al., 2011). Convenience, reliability, transparency, and a seamless experience are important factors influencing customer satisfaction, from browsing products on e-commerce platforms to the (post-) purchase stage (Kumar & Anjaly, 2017). Therefore, it is necessary to have a holistic approach to understanding what drives satisfaction and which pain points (aspects) can turn a positive shopping experience into a negative one (McColl-Kennedy et al., 2019). The identification of these influencing aspects not only provides insight into customer behaviour and preferences but also offers businesses the knowledge needed for optimising their services, improving user experience, and ultimately gaining a competitive presence in the market (Grewal et al., 2009).

The rise of machine learning and data analytics has provided businesses with powerful tools to better understand customer behaviour and preferences through feedback analysis (Liu et al., 2020; Rane et al., 2024). Every review, rating, and piece of customer feedback is a rich source of information that, when analysed effectively, can reveal critical insights into what customers value most and where they encounter difficulties (Mudambi & Schuff, 2010). Analysing customers' feedback allows businesses to go beyond sales numbers and better connect with the customer's emotional experience, uncovering hidden issues that may not be easily known to business owners (Meyer & Schwager, 2007).

Traditional manual feedback analysis is time-consuming, labour-intensive, and prone to subjective biases. Analysts may interpret customer sentiment differently, leading to inconsistent results. Furthermore, as businesses scale, the amount of feedback becomes overwhelming, making it difficult to analyse all reviews effectively. While there are a number of models available for assessing overall customer sentiment of a customer review (Fang & Zhan, 2015), most of them often struggle with aspect-based sentiment analysis (ABSA) (Nazir et al., 2020), particularly when customer reviews contain mixed sentiments (e.g., positive feedback about shipping but negative feedback about product quality) or implicit aspects (where a sentiment is implied but not explicitly stated). Addressing these challenges requires advanced machine

* Corresponding author.
E-mail address: jozsef.mezei@abo.fi (J. Mezei).

https://doi.org/10.1016/j.eswa.2025.130865
Received 8 February 2025; Received in revised form 10 December 2025; Accepted 14 December 2025
Available online 16 December 2025

learning models that can accurately detect aspects, classify sentiment, and provide structured insights.

Although several machine learning approaches have been developed for sentiment analysis (Do et al., 2019), many existing methods suffer from limited generalizability, low accuracy in aspect detection, and an inability to handle complex sentiment structures. Most feedback analysis solutions rely on a single model, which may not capture all aspects of user sentiment effectively. This highlights the need for a comprehensive, multi-model approach that leverages the strengths of various machine learning models to enhance accuracy, scalability, and reliability in aspect detection and sentiment classification.

Despite the rapid adoption of machine learning in customer feedback analytics, current e-commerce sentiment analysis methods remain fragmented. Most studies address isolated subtasks such as sentiment polarity or explicit aspect extraction. This fragmentation leads to limited scalability, an inability to detect implicit opinions, and a lack of interpretability across models. Hence, there is a pressing need for an integrated, explainable pipeline that combines the complementary strengths of multiple models to extract both explicit and implicit customer sentiments in a unified framework. Based on these challenges, this study addresses the following research questions:

- **RQ1.** How can transformer-based and large language models be integrated into a unified, end-to-end pipeline for automated aspect-based sentiment analysis in e-commerce?
- **RQ2.** To what extent does a multi-model architecture improve the accuracy, interpretability, and scalability of sentiment and aspect detection compared with single-model approaches?

To address these questions, we developed a comprehensive pipeline consisting of multiple machine learning models. This pipeline has the potential to empower businesses to gain detailed insights into customer opinions at both granular and macro levels, enabling targeted improvements and strategic decision-making. The models employed in the pipeline are as follows (Psorakis et al., 2011):

- Large Language Model (LLM) for Generating Annotated Feedback: A fine-tuned model generates annotated customer feedback text, including detected aspects, facilitating the creation of synthetic datasets for training purposes.
- BERT-Based Aspect Detection Model: Utilizing the BERT_Review model to enhance aspect detection accuracy through deep contextual understanding and identifying aspects within customer feedback.
- RoBERTa-Based ABSA Model: A specialized model for identifying sentiments of different aspects within customer feedback.
- LLM-based ABSA Model: This fine-tuned LLM model processes raw customer text to output detected aspects, associated sentiments, and overall sentiment, combining aspect detection and sentiment analysis into a single step. We used this model to identify mixed sentiment in ABSA as it performs better than the base model. Moreover, the BERT_Review model identifies only explicit aspects in the text. Therefore, we used the fine-tuned ABSA LLM-based model to detect implicit aspects that BERT_Review might miss. Merging the outputs from both models results in a comprehensive set of aspects, enhancing overall coverage and accuracy in feedback analysis.
- RoBERTa-Based Model for Overall Sentiment Detection: A model designed to determine the overall sentiment of customer feedback.

Through these steps, this pipeline provides a more effective approach to analysing e-commerce feedback compared to existing models, offering businesses a data-driven solution for improving customer experience. To evaluate this pipeline, customer reviews were collected from the Trustpilot platform. Using this data and the steps described above, the effectiveness of this approach is evaluated through extensive experiments comparing model performances.

The rest of the paper is structured as follows: Section 2 presents a literature review and the necessary background. Section 3 outlines the methodology and dataset. Sections 4, 5, and 6 describe the three stages of the proposed pipeline. The final integrated pipeline is presented in Section 7. Finally, Section 8 discusses the findings, academic and practical contributions, and limitations.

2. Literature review

This section provides a review of existing literature on customer feedback analysis, sentiment classification, and aspect detection, highlighting the limitations of current approaches and the need for a multi-model machine learning pipeline for improved feedback analysis in e-commerce.

2.1. Understanding the role of customer feedback in E-commerce

Customer feedback analysis is an integral aspect of modern e-commerce, serving as a vital tool for understanding consumer preferences, enhancing user experiences, and driving strategic business decisions. The rapid growth of online retail has contributed to increasing the volume of customer feedback, making customer feedback increasingly valuable information for businesses to efficiently analyse this data to make strategic decisions and remain competitive in the market. Moreover, customer feedback is invaluable to e-commerce businesses for several reasons.

Understanding customer opinions can help businesses tailor their services to meet customer expectations better (Gajewska et al., 2020). Consequently, this can lead to providing improved user interfaces, offering personalised shopping experiences, and enhancing customer service and loyalty (Ameen et al., 2021). Moreover, customer feedback towards product experience can inspire new product features or entirely new products, keeping businesses at the forefront of market trends and consumer demands. Furthermore, the insights from customer feedback analysis can serve as a critical component of data-driven strategies, and businesses can make informed decisions regarding product development, marketing strategies, and overall business direction.

While the benefits of customer feedback are clear, the process of manual feedback analysis presents several significant challenges, such as inaccuracy, subjective opinion, and time required (Hemmatian & Sohrabi, 2019). As e-commerce platforms generate vast amounts of feedback daily, manual data exploration and actionable insight extraction are time-consuming and labour-intensive (Tsytsarau & Palpanas, 2012). Moreover, manual analysis of feedback with human involvement is inherently subjective, with the potential for bias. Most crucially, identifying specific aspects in feedback (e.g., product features, delivery experience) and determining the sentiment associated with each aspect are complex tasks. Manual analysis might miss nuanced sentiments, such as mixed feelings within a single piece of feedback. In addition, the scalability of data introduces critical issues. As businesses grow, the volume of feedback increases, making the manual analysis slow and impractical.

To address these challenges, businesses are increasingly turning to automated feedback analysis solutions (Lee & Bradlow, 2011; Yom-Tov et al., 2018). The adoption of automated feedback analysis technologies enables businesses to leverage the full potential of customer feedback for continuous improvement and strategic decision-making (Akter & Wamba, 2016; Kyaw et al., 2023). Solutions, such as advanced machine learning models, enable the processing and analysis of large volumes of feedback efficiently and accurately, and businesses can gain real-time insights, reduce the risk of bias, and handle the complexity and scale of modern e-commerce feedback (Chen et al., 2023; Newman, 2013).

2.2. How aspects influence the customer journey and experience

The analysis of customer feedback is not just about understanding what is working well but also about identifying pain points and gaps in the customer experience (McColl-Kennedy et al., 2019). A data-driven

approach helps businesses to determine specific issues, and detailed feedback analysis enables businesses to identify which aspects of the e-commerce journey influence customers' shopping experience. As such, this understanding helps companies to prioritise improvements that will have the most significant impact on customer satisfaction (Zineb et al., 2021).

In an online shopping context, customers rely on different features and aspects when making their purchase and post-purchase decisions (Han & Ryu, 2012). Unlike traditional brick-and-mortar stores where shoppers can physically see, touch, and try products, online shopping is largely based on trust, convenience, and the information presented on the platform (Huyghe et al., 2017). Moreover, different stages are associated with online shopping and their impact on customer satisfaction. For instance, how easily a customer can find what they need, the transparency of product information, the quality of customer service, the reliability of shipping, and the ease of returns all significantly influence their perception of the online shopping experience (Otim & Grover, 2006; Pham & Ahammad, 2017).

By understanding how each aspect affects the overall customer experience, businesses can make informed decisions on where to invest resources. For example, if feedback suggests that a poor return process is a major source of dissatisfaction, efforts can be made to improve return policies and processes (Meyer & Schwager, 2007). All this information can be used to provide, e.g., personalise recommendations, improve communication strategies, and tailor offerings to meet the needs of various customer groups better, thereby enhancing satisfaction and loyalty (Rane et al., 2023). Furthermore, these aspects of the e-commerce process are interdependent, playing a role in forming customer perception and determining satisfaction level (Wang, 2008).

In other words, understanding the aspects that influence the online customer journey helps companies to identify and address potential issues that may affect customer satisfaction (Lemon & Verhoef, 2016). It also provides a competitive advantage by understanding their customers' expectations, enabling them to differentiate their offerings through superior customer service, more transparent pricing, faster shipping, or a better return process (Lindecrantz et al., 2020; Rane et al., 2023).

In summary, the customer journey in e-commerce is multifaceted and influenced by various aspects, all of which play a crucial role in shaping customer experience and satisfaction (Wagner et al., 2020). Understanding these aspects and analysing customer feedback provides a strategic advantage for businesses. Each identified aspect plays a crucial role at different stages of the e-commerce journey, encompassing both tangible and intangible elements that collectively impact customer satisfaction. Addressing these aspects in a customer-centric way can significantly enhance the overall online shopping experience (Sheth et al., 2023).

Beyond these aspects, human factors also add another layer of influence, shaping how customers experience, interpret, and respond to different elements of the online shopping journey. Human factors broadly refer to the study and application of knowledge about human behaviour, abilities, and limitations to improve the design of systems, tools, and environments for effective and safe human use (Strawderman & Koubek, 2008). In service and logistics contexts, considering human factors is essential because both customers and employees are integral parts of the system; their cognitive, emotional, and behavioral responses directly affect performance, satisfaction, and operational outcomes (Sanders & McCormick, 1998). Moreover, beyond the technical and functional attributes of e-commerce systems, the role of human factors, such as cognitive effort, emotional response, and user interaction, also profoundly shapes the online shopping experience.

In service and logistics research, human factors are defined as the psychological, cognitive, and emotional characteristics that affect how individuals interact with technological systems, services, and organizations (Ali et al., 2025; Ogbeyemi et al., 2024). For instance, Ogbeyemi et al. (2024) highlight the importance of incorporating human-related variables such as worker fatigue and task engagement into distribution and service models. Meanwhile, Ali et al. (2025) emphasize emotional intelligence and interpersonal awareness as key drivers of supply chain collaboration and visibility. Together, these perspectives suggest that human cognitive and emotional processes deeply influence user experience in digital commerce. In this study, we focus on how these human factors are manifested through real customer feedback, examining how specific aspects, such as trust, customer service, and product quality, reflect customers' authentic emotional and cognitive evaluations during the online shopping journey.

2.3. Machine learning approaches in feedback analysis

Machine learning has revolutionised feedback analysis in e-commerce by providing powerful tools for processing and interpreting online customer data, identifying patterns, and creating a personalised experience (Policarpo et al., 2021). In e-commerce and customer feedback analysis, machine learning tools are used, and algorithms are employed to automate the analysis of customer feedback, significantly enhancing efficiency and accuracy (Koufaris et al., 2001). By leveraging techniques such as natural language processing (Paik et al., 2001), sentiment analysis (Yi & Liu, 2020), and aspect-based sentiment analysis (Wankhade et al., 2022), these models can automatically identify key aspects of feedback, determine associated sentiments, and generate comprehensive insights. For instance, Zhao et al. (2018) proposed a knowledge-discovery process to mine affective words describing customers' emotional responses to apparel products. This early line of research illustrates how sentiment analysis was traditionally approached through handcrafted lexicons and expert evaluation rather than large-scale automated modelling.

Automation of customer feedback analysis provides several advantages, such as reducing the time and resources needed for manual processing (Niranjanamurthy et al., 2013), mitigation of human bias, and effortlessly scaling to handle increasing data volumes (Singh et al., 2022). This automation not only provides real-time insights into customer opinions but also enables businesses to quickly identify and respond to emerging trends and issues, thereby improving customer satisfaction and fostering a more responsive and adaptive business environment (Yom-Tov et al., 2018).

Recent studies have also demonstrated the potential of large language models in service-oriented systems beyond e-commerce. For instance, LLMs have been employed to model and improve human satisfaction and emotional understanding in healthcare services (Rosario et al., 2024; Xia et al., 2024). These studies illustrate how LLMs can interpret emotional tone, generate empathetic responses, and support personalized service delivery, thereby transforming user interaction and satisfaction analysis in complex service environments. In line with these developments, the present study extends the application of LLMs to the e-commerce domain, where similar human-centred goals, understanding user sentiment, satisfaction, and trust, are pursued through automated feedback analysis. This parallel underscores the broader relevance of LLM-based approaches in extracting affective and contextual insights across different service sectors.

The development of machine learning models for feedback analysis in e-commerce leverages a variety of techniques and models, such as Support Vector Machines (Yu et al., 2011), Naïve Bayes classifiers (Bayhaqy et al., 2018), and more sophisticated deep learning models like BERT and RoBERTa networks (Zalutska et al., 2023), to optimise accuracy and efficiency. These models are trained on large datasets to identify patterns and make predictions about the content of feedback. Moreover, deep learning models, especially transformer-based architectures such as BERT and its variants, have substantially advanced aspect-based sentiment analysis by capturing contextual and semantic relationships more effectively than traditional machine learning approaches. However, these improvements come with significant costs, as such models typically require large annotated datasets and considerable computational resources (Wankhade et al., 2024).

However, while deep learning models such as neural networks achieve strong predictive performance, they often lack transparency and interpretability. As noted by Wu et al. (2016), artificial neural networks are typically regarded as black-box models since their internal representations are learned from input-output data rather than explicit rules. This lack of explainability motivates our modular design, which retains interpretability through separated model interfaces and transparent sentiment mapping.

Implementing a comprehensive machine learning pipeline for feedback analysis involves designing an end-to-end system that integrates multiple machine learning models to process and analyse customer feedback efficiently (Rachman et al., 2021; Van De Schoot et al., 2021). Prior studies have highlighted the importance of a well-structured pipeline that includes stages for data collection, pre-processing, model training, and evaluation (Biswas et al., 2022). This approach requires the integration of various models for different tasks: aspect detection and (aspect-based and overall) sentiment classification. Workflow and data processing steps are carefully orchestrated to ensure seamless transitions between stages, maximising the pipeline's overall efficiency and accuracy. Implementing a machine learning pipeline enables businesses to harness the full potential of automated feedback analysis, providing real-time insights and facilitating continuous improvement in customer satisfaction and operational effectiveness. In the following, we discuss the details of two main components of our pipeline: aspect detection and sentiment classification.

2.3.1. Aspect detection in customer feedback

Aspect Term Extraction (ATE) in customer feedback is crucial for understanding detailed customer opinions. ATE is the process of identifying specific components or features mentioned in reviews relevant to a domain (Kumar et al., 2022). The detection of aspects in customer feedback and the analysis of data enable businesses to identify strengths and weaknesses regarding various aspects of their product/service offerings (Bagheri et al., 2013). For instance, in the sentence "awful communication and slow dispatch", the terms "communication" and "dispatch" are aspect terms. ATE generally consists of extracting all aspect-related terms (e.g., "dispatch") from the text and grouping aspect terms with similar meanings into broader categories(Davoodi, 2023). Aspect terms are classified into two categories: explicit and implicit. Explicit aspects are directly mentioned in the text, while implicit aspects are inferred without direct mention (Alqaryouti et al., 2020).

Rana and Cheah (2016) classify ATE techniques into supervised (Shu et al., 2017), semi-supervised (Ansari et al., 2020), and unsupervised approaches (Tulkens & Van Cranenburgh, 2020). Existing research often relies on manual or semi-automatic methods for aspect annotation and aspect extraction, which can introduce subjectivity and inconsistency. Therefore, further exploration is needed to enhance automated aspect extraction algorithms, aiming to improve the precision and recall of aspect identification in e-commerce reviews (Dogra et al., 2024).

Unsupervised aspect extraction techniques utilize unannotated data to identify implicit and explicit aspects in reviews without the need for training. These methods, such as statistical analysis, topic modelling, and dependency parsing, are widely applied across different domains and languages. Semi-supervised techniques, on the other hand, leverage both labelled and unlabelled data to extract aspects, using methods like Recurrent Neural Networks (RNN) (Aydin & Güngör, 2020) and lexicon-based approaches (Obaidat et al., 2015). Lastly, supervised techniques rely solely on labelled data for training, employing models like Conditional Random Fields (CRF) (Rubtsova & Koshelnikov, 2015) and Long Short-Term Memory (LSTM) (Giannakopoulos et al., 2017) to extract both implicit and explicit aspects.

Recent advancements in aspect extraction have seen a shift towards transformer based models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and LLAMA (Touvron et al., 2023), which outperform traditional models due to their ability to capture contextual relationships more effectively. Research indicates that BERT, with its bidirectional encoding, captures both forward and backward contexts, making it particularly effective for nuanced aspect extraction tasks (Devlin et al., 2018). Pre-trained LLMs, such as LLAMA, designed with fewer resources but high efficiency, further demonstrate that transformer- based models provide superior performance, especially in low-resource settings, by leveraging pre-trained knowledge and fine-tuning for domain-specific tasks (Touvron et al., 2023). These models not only improve accuracy but also require less domain-specific training data, making them more flexible and scalable across different domains.

In previous studies, various techniques have been used with LLMs, including prompt engineering (Brinkmann et al., 2023; Ma et al., 2023), and fine-tuning (Biana et al., 2024). Prompt engineering, such as one-shot prompting, focuses on carefully crafting input prompts to guide the model in generating more accurate or relevant outputs without altering the model itself. In contrast, fine-tuning involves retraining the model on specific datasets to adjust its internal parameters, making it more specialized for specific tasks or domains. While prompt engineering is faster and more flexible, fine-tuning provides deeper customization and improved task-specific performance. Techniques like Low-Rank Adaptation (LoRA) are used for parameter-efficient fine-tuning, reducing computational costs while preserving model performance (Hu et al., 2021; Zhang et al., 2024c).

However, there has been limited published research specifically examining fine-tuning in the context of the e-commerce domain (Palen-Michel et al., 2024). Zhou et al. (2023) used structured prompts to guide LLMs in identifying specific product attributes from customer queries. Each prompt consists of three components: assumption, example, and question. The LLM is prompted to assume the role of an e-commerce expert tasked with identifying attributes such as "Brand," "Color,' "Size," and "Product Type" from a given query, with examples of correctly extracted attributes provided. Zhang et al. (2024a) developed a multi-task e-commerce shopping assistant which leverages the power of large language models to address various online shopping tasks. These tasks include concept understanding, knowledge reasoning, user behaviour alignment, and multilingual abilities. The LLMs were enhanced using a specialized instruction dataset, EshopInstruct, containing 65,000 samples that reflect e-commerce scenarios. This dataset helped fine-tune the LLMs to handle specific domain knowledge, such as extracting attributes, processing product-related queries, and making product recommendations.

2.3.2. Sentiment analysis and aspect-based sentiment analysis in E-commerce

Sentiment analysis in e-commerce involves evaluating customer feedback to determine the overall sentiment, positive, negative, or neutral, expressed in reviews (Zheng et al., 2014). The sentiment analysis technique provides businesses with a high-level understanding of customer satisfaction, needs, demands, and areas needing improvement. Generally, sentiment analysis is categorized into three levels: document-level, sentence-level, and entity/aspect-level analysis (Hu & Liu, 2004). Document-level sentiment analysis assesses the overall sentiment of an entire document, focusing on the sentiment expressed in the text as a whole rather than on individual sentences. The sentence-level analysis predicts the sentiment of each sentence independently, treating each sentence as a distinct entity. ABSA goes a step further by identifying sentiments related to specific aspects of a product or service, offering a more nuanced view of customer opinions (Vanaja & Belwal, 2018). ABSA includes two subtasks as follows: aspect term extraction (ATE) and aspect sentiment classification (ASC). Combining aspect detection with sentiment analysis allows businesses to gain more profound knowledge on which features of a product/service are more appreciated or criticized, enabling targeted enhancements (Wankhade et al., 2022).

Several techniques have been proposed in the literature to perform ASC. These methods vary from traditional lexicon-based approaches to modern deep learning models, with each offering different levels of

accuracy and efficiency depending on the context and data used. Lexicon-Based Approaches calculate sentiment by aggregating word polarities using sentiment dictionaries. Words are assigned a sentiment score, and the overall polarity for an aspect is derived based on these scores (Mowlaei et al., 2020; Siddiqua et al., 2024). However, these methods struggle with context and negation handling. To avoid this limitation, machine learning models were introduced for ASC tasks. These models, including SVMs, random forests, and logistic regression, are widely applied to classify sentiments. In these methods, labelled aspect-sentiment pairs are used to train classifiers that predict sentiment polarity (Arthamevia et al., 2021; Villaneau et al., 2018). However, traditional machine learning methods struggle to effectively capture the complex non-linear relationships between features and sentiment polarity.

Recent work leverages deep neural networks, including LSTM, RNNs, and transformers, like BERT, and LLMs, which are fine-tuned to classify sentiment based on context (Hammi et al., 2023; Sirisha & Bolem, 2022; Šmíd et al., 2024; Sun et al., 2019). However, the LSTM and CNN models had two key limitations. Firstly, models trained on a specific dataset did not generalize well to other datasets or domains. Secondly, these models struggled to achieve high accuracy. As a result, transformer-based models have gained prominence in recent years as more effective solutions for ABSA tasks(Mughal et al., 2024). With the advent of transfer learning and attention-based mechanisms, networks pre-trained on large datasets have demonstrated remarkable performance breakthroughs (Ruder et al., 2019; Troya et al., 2021; Vaswani et al., 2017). These models have significantly advanced the state of the art in ABSA by capturing deeper semantic and syntactic relationships. However, non-fine-tuned LLMs perform adequately but lag behind specialized models in complex tasks such as ABSA (Simmering & Huoviala, 2023; Zhang et al., 2024b).

There are some proposals in the literature that use LLMs for ABSA tasks in customer reviews, but they are not explicitly focused on the e-commerce domain. For instance, Qian et al. (2024) used 8405 Tripadvisor reviews of college football stadiums, spanning 2011 to 2023, annotated with various customer experience constructs. The dataset was further enriched with synthetic reviews generated by ChatGPT, enhancing the model's training and ensuring accurate outputs. Two GPT-3.5 models were fine-tuned for AE to identify customer experience aspects, and another for Customer Experience Classification (CC) to categorize these aspects based on emotional and social responses. A fine-tuned RoBERTa model handled sentiment scoring, classifying experiences as positive, neutral, or negative (Qian et al., 2024).

2.4. Data augmentation

Data Augmentation involves generating synthetic data from an existing dataset by introducing small variations or combining distant examples, which helps models remain invariant to certain changes. It plays a key role in the training of deep neural networks, as its transformations are interpretable and offer insights into model weaknesses, making it a valuable tool for improving model performance (Shorten et al., 2021). In the literature, various methods are employed to augment data and generate synthesized training data using machine learning. These techniques are crucial in domains where labelled data are scarce, expensive, or time-consuming to collect. Training data is essential for the quality of supervised machine learning, where the addition of more data significantly improves the performance (Bayer et al., 2022).

Data augmentation methods widely used in computer vision (CV), such as flipping, cropping, or tilting, do not apply to NLP tasks. The main reason is that in textual data, the exact order of characters carries critical syntactic and semantic meaning (Zhang et al., 2015). Therefore, in terms of text augmentation, several practices are commonly used in the literature to enhance NLP models, especially when dealing with limited data. Wei and Zou (2019) introduced the EDA (Easy Data Augmentation) technique, designed to improve performance in text classification tasks. EDA is especially beneficial for smaller datasets, where training with only 50 % of the available data using EDA achieves the same accuracy as training with the full dataset. Back-translation (Sennrich et al., 2016) is another technique used in the literature where a sentence is translated into another language and then translated back into the original language. Contextual word embedding augmentation was introduced by (Kobayashi, 2018) to augment text data. This technique replaces words with contextually similar words using word embeddings. Pre-trained language models like BERT or GPT can be used to predict appropriate replacements for words in the sentence (Kobayashi, 2018). Another commonly used techniques include paraphrasing (Prakash et al., 2016), noise injection (Coulombe, 2018) or Adversarial Data Augmentation (Gupta, 2019).

Furthermore, LLMs, such as GPT, can be used to generate new text samples based on existing examples due to their ability to generate high-quality synthetic data that closely resembles human-generated text. Various techniques are used in the literature that employ LLMs to generate synthesized data. Prompt engineering is one of the standard techniques to instruct the LLM to generate synthetic data. This can involve providing a few examples (few-shot learning) or using specific instructions to guide the output (Maheshwari et al., 2024; Patil & Gudivada, 2024). Moreover, LLMs are used in combination with smaller, task-specific models. For instance, after generating synthetic reviews with an LLM, a secondary model might filter or refine the output to ensure it adheres to business-specific constraints or quality standards. These techniques are widely used to improve NLP models by diversifying the training data, making models more robust to variations in text input. Each method can be applied depending on the task, such as classification, machine translation, or question answering.

3. Methodology and data

As shown in Fig. 1, this research is structured into several phases, combining various machine learning models and utilizing synthesized data to enhance the ABSA task. The first phase involves scraping customer feedback data from the Trustpilot platform, followed by data processing to clean the data for model training. Afterwards, feature extraction and aspect selection are conducted. The selected aspects are manually annotated, resulting in a dataset labelled for aspect terms and their associated sentiments. To enhance the dataset, synthetic data generation is employed using LLMs to produce additional annotated data, focusing on under-represented or infrequent aspects. The fine-tuning process is leveraged to generate synthesized customer feedback that contains predefined entities relevant to the feedback.

Firstly, as shown in Fig. 2, a fine-tuned LLM generates synthetic feedback that mirrors real customer feedback. These generated data points are used to supplement the manually annotated dataset, expanding the overall dataset size and addressing the data scarcity problem. The model is fine-tuned to detect multiple aspects, including implicit ones, and the synthesized data serves as training material for downstream models.

Secondly, as presented in Figs. 3 and 7, the final integrated pipeline is built to automate the entire process, from scraping customer feedback to generating detailed aspect-based sentiment insights. Aspects such as Trust, Shipping, Product Availability, Pricing, and Overall Sentiment are identified, and the feedback is presented in a clear, actionable format for business decision-making. The final pipeline utilizes two models for aspect detection: the BERT_Review model detects explicit aspects within the text, while the Fine-Tuned ABSA LLM model focuses on identifying implicit aspects that the BERT model might miss. The union of the detected aspects from both models is reported as the final set of detected aspects, improving coverage and accuracy in feedback analysis.

For sentiment detection, the pipeline employs two different models: a RoBERTa-based ABSA Model, efficient at detecting the sentiments associated with specific aspects, and a fine-tuned LLM ABSA model for handling more complex feedback that contains both positive and negative sentiments in the same entry. Finally, a separate RoBERTa-based

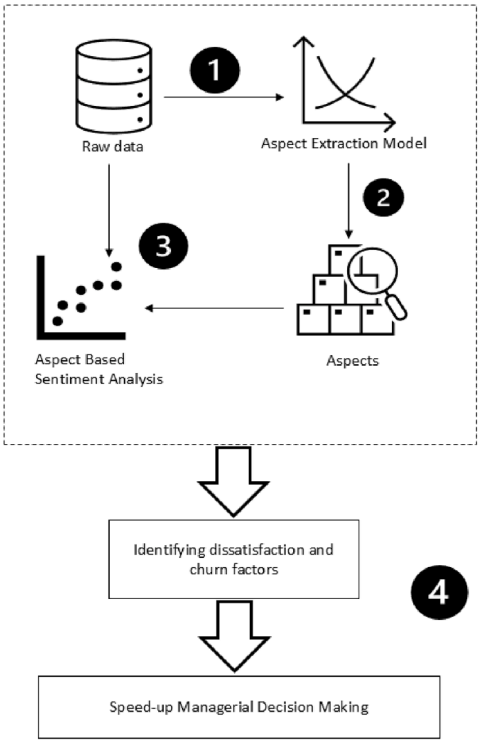

Fig. 1. Final pipeline.

model is used to detect the overall sentiment of the feedback. This model aggregates sentiment data from individual aspects to provide a holistic sentiment label, helping businesses quickly gauge overall customer satisfaction. In the following section, we describe in detail the collected data and the models used in the pipeline.

3.1. Data gathering and annotation

We used two different datasets for this research. The first dataset (Davoodi, 2023) is a manually annotated dataset designed for the aspect extraction task that uses IOB (Inside, Outside, Beginning) tagging [1]. The second manually annotated dataset (Davoodi et al., 2025) is designed for the ABSA task and consists of the sentiments associated with each aspect and the overall sentiment of the feedback [2]. The data used in the study was gathered from the leading online review platform, Trustpilot.

3.1.1. The final set of aspects

In our research, based on the state-of-the-art literature on e-commerce customer experience and feedback analysis, we have identified 14 key aspects that are deemed to shape the e-commerce customer experience and satisfaction collectively. The following aspects provide a comprehensive framework for businesses to better understand their customers.

1. **Shipping:** the speed, cost, and reliability of delivery (Huang et al., 2019).
2. **Trust:** the perceived credibility and reliability of the store (Rita et al., 2019).
3. **Item Quality:** the inherent quality of products offered (Rita et al., 2019).
4. **Pricing:** competitive and transparent pricing strategies (Rita et al., 2019).
5. **Customer Service:** the helpfulness and responsiveness of store representatives (Mero, 2018).
6. **Product Features:** detailed product descriptions, accurate specifications, and clear images (Holloway & Beatty, 2008).
7. **Delivered Product Status:** the condition in which products arrive upon delivery (Holloway & Beatty, 2008).
8. **Refund Process:** efficiency and transparency in handling refunds (Martínez-López et al., 2022).
9. **Return Process:** the convenience and speed of returning products (Martínez-López et al., 2022).
10. **Information:** accuracy and availability of critical information for purchase decisions (Griva, 2022).
11. **Product Availability:** variety and availability of products or brands (Griva, 2022).
12. **Packaging:** the quality and attractiveness of product packaging (Arora, 2016).
13. **Payment:** a smooth, secure, and diverse payment process (Blut, 2016).
14. **App Experience:** usability and functionality of the platform's website and mobile app (Blut, 2016).

3.1.2. Aspect-Based sentiment analysis dataset

To create a dataset for modeling user-generated content and performing ABSA tasks, Davoodi et al. (2025) collected 12,000 English reviews from Trustpilot between 2013 and 2021. A random selection of 3500 messages was used for aspect extraction and manual annotation. After filtering out reviews that did not meet specific criteria, the final dataset consisted of 2782 data points from five online-only stores: Zalando (35 %), Wish (24 %), Boozt (20 %), SHEIN (16 %), and Nelly (5 %). The reviews were labelled for the 14 aspects, and each aspect was assigned a polarity (positive or negative). Two annotators manually labelled the data, resolving disagreements through discussion. The aspects and their associated sentiments are presented in Fig. 4.

Additionally, each review was labelled with an overall sentiment (positive, negative, or mixed), providing a more holistic understanding of user feedback(shown in Fig. 5).

The cleaned data was pre-processed through several steps, including lowercasing, removing non-English text, and performing lemmatization, resulting in a vocabulary of 4580 unique words. This dataset serves as the foundation for further ABSA model training and testing.

3.1.3. Aspect term extraction dataset

To generate synthesized ATE training data, we used the manually annotated dataset introduced by Davoodi (2023), which consists of 2782 data points designed for the Aspect Detection task. For performing manual annotation, the position of each aspect in the text is determined, along with its associated positive or negative sentiment (Davoodi, 2023). The dataset used in this study is the same as the one employed for manual sentiment tagging of aspects. However, this time, the annotation process was extended to include the exact positions of the aspects within the text.

Prior to the manual annotation, Davoodi (2023) took several steps. First, the reviews were standardized using Python libraries such as contextualSpellCheck and Caribe to correct spelling and grammar errors. This preprocessing step ensured consistency in the data. Second, several rules were established to convert implicit aspects into explicit ones, aiming to preserve the original wording as much as possible. Finally, following this standardization, two annotators independently worked to label the data. The focus was on identifying nouns and noun phrases that expressed sentiment towards aspects, specifically excluding any neutral sentiments (Samha et al., 2014).

To enhance the dataset's suitability for aspect detection tasks even more, Davoodi (2023) expanded the scope of the annotations. This involved considering a broader range of sentence components as potential aspects or opinions, with a particular focus on identifying positive or negative sentiments within individual sentences. For this purpose, the

[1] The data is available at https://doi.org/10.6084/m9.figshare.30411193
[2] The data is available at https://doi.org/10.6084/m9.figshare.24980994.v1

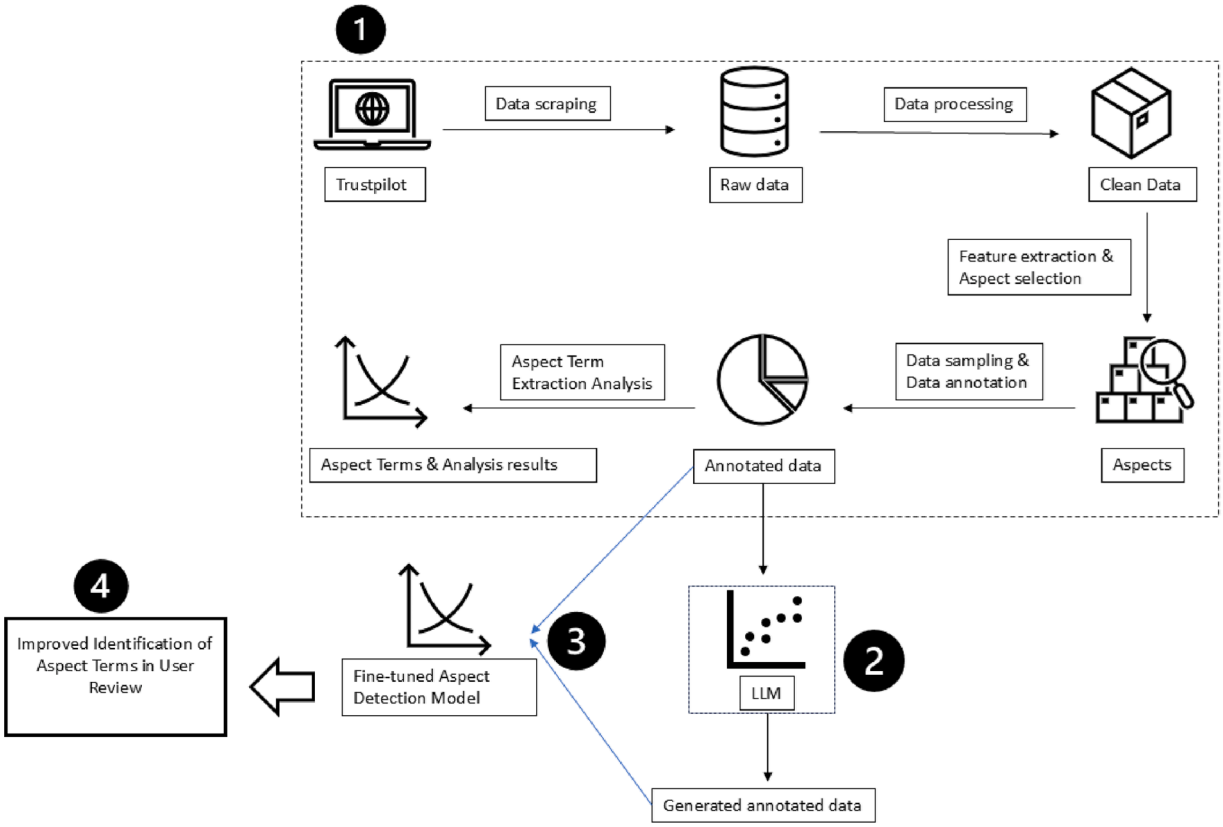

Fig. 2. Generating synthesized data.

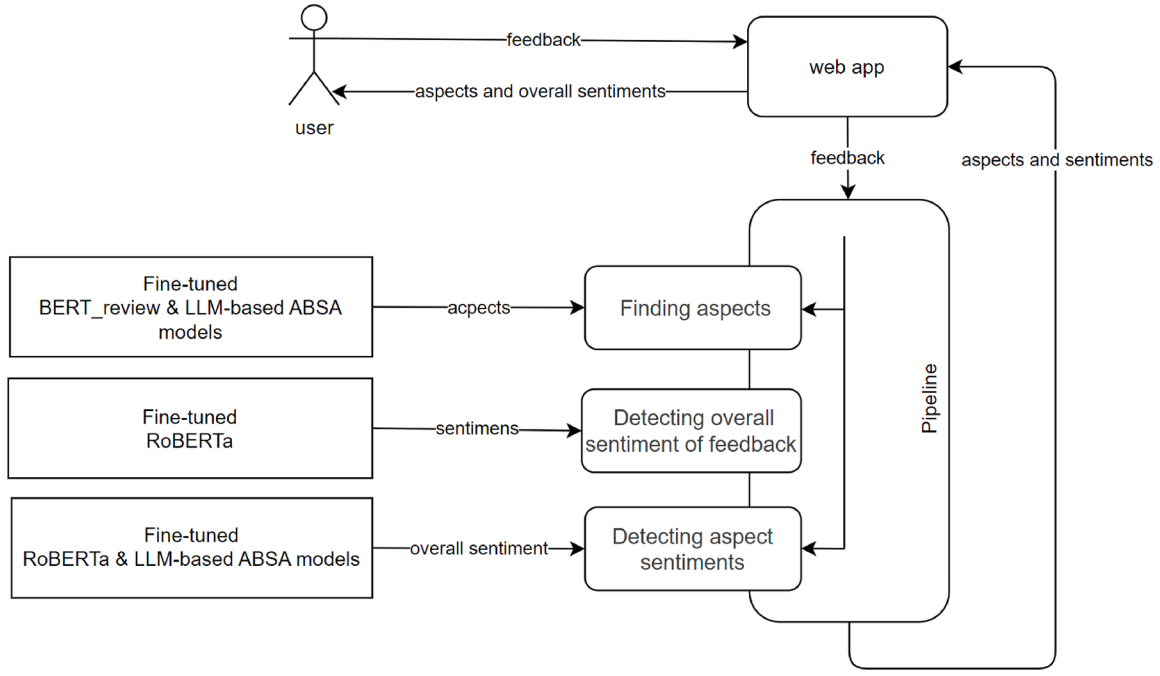

Fig. 3. App work flow.

IOB (Inside, Outside, Beginning) tagging system can be adopted. During the final annotation phase, specific labels were assigned to mark the beginning or continuation of an aspect mentioned, such as App_experience_B, App_experience_I, Trust_B, Trust_I, Customer_service_B, and Customer_service_I. This approach allowed for a thorough extraction of the aspect terms. The dataset ultimately includes 29 labels, consisting of two labels per aspect and an 'O' label for words not associated with any aspect. For instance, in the sentence "slow customer service. I do not recommend this company," the text was split into individual words, and each word was labelled accordingly. The resulting sequence of labels was ['O', 'Shipping_B', 'Shipping_I', 'O', 'O', 'O', 'O', 'O', 'Company_B']. Any discrepancies in the annotations were resolved through discussion, ensuring a consensus between the annotators.

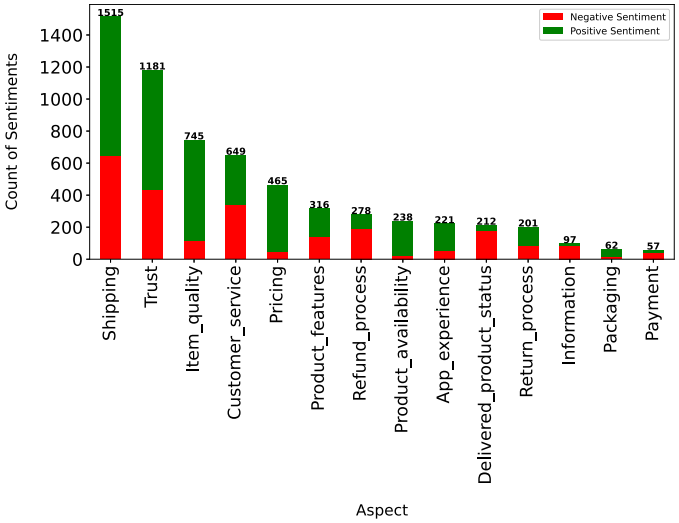

Fig. 4. Aspects with their associated sentiments.

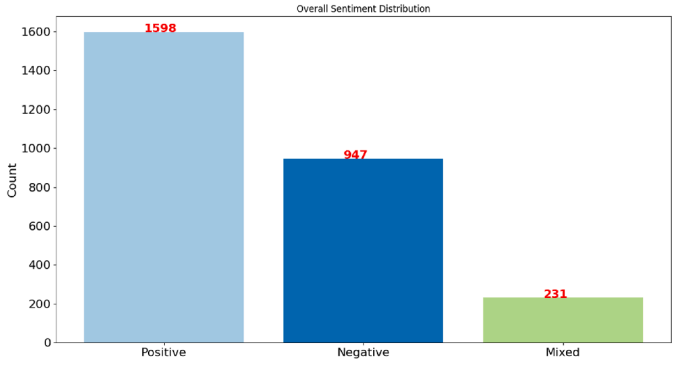

Fig. 5. The distribution of review sentiments.

Afterwards, to evaluate the aspect detection dataset, we selected three transformer-based models: BERT, RoBERTa, and BERT_Review. We employed cross-validation to provide a more robust estimate of the models' performance. Among these, the BERT_Review model showed slightly better results, achieving an F1 score of 84 %. Given that the dataset involved 29 different categories, this level of performance is encouraging. Further details about the modelling process can be found in Davoodi (2023).

4. Pipeline stage 1: Aspect detection

This section explores methods for extracting aspects from customer feedback, comparing traditional approaches with transformer-based models like BERT and RoBERTa. It introduces our hybrid approach, combining BERT_Review and a fine-tuned LLM, and evaluates its performance, including the impact of synthetic data augmentation.

In the context of e-commerce, annotated data for ABSA is often limited and costly to produce due to the need for manual labelling. Manually annotating customer feedback with specific aspects and their corresponding sentiment polarity is time-consuming. As a result, there is a growing need for data augmentation techniques to generate synthetic annotated data, which can expand the training dataset and improve model performance while minimizing the reliance on expensive and labour-intensive manual annotation processes.

To address the challenge of ABSA in scenarios with limited labelled data, we propose to utilize prompt fine-tuning, first to extract aspects (typically nouns) from an unlabelled corpus and categorize them according to sentiment. Subsequently, the LLM is used to generate synthetic data based on these extracted aspects, producing both single-aspect and mixed-aspect pseudodata. Each generated instance was paired with a corresponding sentiment label (positive, negative, or neutral). Finally, a discriminator was employed to assess the generated data for domain and sentiment relevance (Li et al., 2024).

The authors utilized GPT-2 for synthetic data generation by fine-tuning the model on task-specific datasets, focusing on input data without class labels. After fine-tuning, GPT-2 was prompted with a beginning-of-sequence token to generate domain-specific text, ensuring diversity through top-k random sampling and random seed variation. This process produced a large and diverse set of synthetic text, sometimes up to 40 times larger than the original dataset. The generated text was subsequently annotated with pseudo-labels and used in a self-training and knowledge distillation framework, enhancing model performance across multiple NLP tasks (He et al., 2022). In contrast to this approach, we utilized an LLM to simultaneously annotate and tag the text, eliminating the need for a separate model for the Named Entity Recognition (NER) task.

To the best of our knowledge, no prior work in the e-commerce and customer feedback domain has focused on generating synthesized, annotated customer feedback. Previous research primarily concentrates on using LLMs to generate automated responses to customer reviews (Azov et al., 2024). As indicated in Fig. 5, the presence of infrequent aspects, such as payment, negatively affects the performance of aspect detection models. Therefore, we opted to generate additional synthesized data rather than rely on manual annotation to expand our training dataset. Therefore, we decided to augment more training data without going through manual annotation by fine-tuning the TinyPixel/Llama-2-7B model. For this purpose, to generate synthesized training data, we implemented prompts as follows:

- `### Instruction:` Take entity types as input and generate customer feedback text containing entities. Place each entity inside entity type tags.
- `### Input:` Return_process, App_experience
- `### Response:` Very easy <Return_process >return process </Return_process>and <App_experience>ordering</App_experience>.

The rationale for creating this prompt was to exert control over the generated customer feedback by specifying the desired aspects within the text. By doing so, we could direct the model to focus on particular aspects, including those that may be less frequently mentioned in the original data. This approach allowed us to ensure the presence of specific entities in the output, thereby enabling the generation of feedback that reflects under-represented or infrequent aspects, ultimately improving the model's ability to capture a broader range of customer experiences.

In a similar approach to our method for generating synthesized annotated data, Frei and Kramer (2023) employed prompt engineering within the medical domain rather than fine-tuning LLMs. Their method involved tagging sentences for Named Entity Recognition (NER) tasks by enclosing each word associated with a specific label *l* within the tags `<class=''l''>` and `</class>`. In contrast, our method tags entity names in the sentence and fine-tunes the LLM using the same tagging system. The key advantage of our approach lies in the ability to control the entities generated and tagged by the model by manipulating the input, thus offering greater flexibility in entity generation.

The models used in this experiment were optimized using Google Colab Pro, and training took about 6 hours. We generated 1400 synthesized annotated data points, predominantly featuring infrequent aspects, of which 1206 were shown to be of good quality after manual inspection. This inspection took around two days. Compared to the five months required for manual annotation, this approach was a significant success.

For model training, we divided the dataset into three subsets as follows: training (2,482 samples), test (200 samples), and validation (100 samples). We used the test set for error analysis. We conducted an error

Table 1
Examples of synthesized data.

Input	True Response	Generated Response
Pricing, Trust	Wish has great <Pricing>deals</Pricing>on anything you can think of. I love this <Company>site</Company>.	Excellent <Pricing>price</Pricing>. 100 % I am sure. I have no doubt. I am a loyal customer and I would never leave <Trust>Zalando</Trust>.
Shipping, Product_features	Quick <Shipping>delivery</Shipping>, <Product_features>product</Product_features>as described.	I have received the <Shipping>item</Shipping>on the date provided. <Product_features>Item</Product_features>is just as described.
Item_quality	Great service and <Item_quality>products</Item_quality>.	Everything is good, especially <Item_quality>sneakers</Item_quality>.
Shipping, Product_ availability	Super fast <Shipping>delivery</Shipping>and wide <Product_availability>range of products</Product_availability>.	Fast <Shipping>delivery</Shipping>and a big <Product_availability>selection of clothes</Product_availability>.
Customer _service	Really good <Customer_service>customer service</Customer_service>.	Always willing to help. <Customer_service>Zalando</Customer_service>always helps when it comes to ordering and getting things sorted.
Information	Can someone please share the <Information>customer service Email address</Information>with me? I placed an order on 19 March and still have not received anything.	<Information>Information</Information>on the website is confusing.

Table 2
BERTScore evaluation on test set.

Precision	Recall	F1 Score
0.8954	0.8959	0.8955

Table 3
ROUGE score evaluation on test set.

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0.38	0.15	0.31	0.31

Table 4
Comparison between reference text and synthesized text.

Input	True Response	Generated Response
Product _features	Some things need to be at least 3 <Product_features>sizes</Product_features>bigger.	<Product_features>Size</Product_features>is very good and it was cheaper than other places I looked at for the same.

analysis on the test set to identify common areas where the model's generated responses deviated from the true responses. Examples of model output are presented in Table 1, illustrating the degree of variability in the accuracy of aspect inclusion and general fluency. While the model generally succeeded in embedding the correct aspects, some nuances, such as phrasing and detail level, were occasionally missed.

The performance of the model was evaluated using two key metrics: BERTScore and ROUGE score. BERTScore was calculated based on the precision, recall, and F1 scores to measure the semantic similarity between the generated text and the true feedback responses. The BERTScore results are presented in Table 2. These results indicate that the fine-tuned model demonstrates strong alignment with the reference data, effectively capturing the specified entities in the generated feedback.

In addition to the BERTScore, the model's performance was further assessed using the ROUGE score. The ROUGE-1, ROUGE-2, and ROUGE-L scores reflect the overlap between the generated and reference text in terms of n-grams and longest common subsequences. The results are shown in Table 3. The low ROUGE score indicates that the generated text is not a direct replication of the original text but is instead synthetically generated with the expected aspects as input. This diversity in the output is crucial when the goal is to produce synthesized data, particularly for sensitive data tagging tasks, where the generated text should maintain semantic relevance without being identical to the original. This controlled divergence is a strength of the method, as it suggests that the model can create diverse yet relevant outputs without overfitting the original data, which is beneficial in scenarios requiring privacy or sensitive data generation.

In conclusion, the fine-tuned TinyPixel/Llama-2-7B model successfully generated synthesized customer feedback based on specified aspects. It achieved strong semantic similarity to the reference data yet was different from the original text. As shown in Table 4, although both the reference and generated texts are semantically similar and focus on the "Product Features" aspect, the generated text is still notably different from the reference. While the reference text discusses the need for larger sizes, the generated text praises the product size and compares its price to other places. This difference in content makes the generated text not completely identical to the reference, highlighting its potential as a valuable resource for data augmentation. Moreover, the BERT Score and ROUGE metrics highlight the model's ability to capture relevant entities in the feedback accurately.

For improving the activebus BERT_Review model performance, we merged the manually annotated dataset with the synthesized data, resulting in a total of 3988 annotated data points as presented in Fig. 6.

BERT_Review is specifically designed to address Review Reading Comprehension (RRC) based on a well-known benchmark for aspect-based sentiment analysis (Xu et al., 2019). We combined our manually annotated dataset with the synthesized dataset consisting of 3430 data points. The Fig. 6 presents the final dataset after adding the synthesized data. To obtain a realistic assessment of the model's performance, we conducted a 5-fold cross-validation and averaged the results of the five-fold to report the final outcome. As shown in Table 5, performance improved with the addition of more training data. Subsequently, we used a fixed training-test evaluation split and fine-tuned the model again to integrate it into the final pipeline for aspect detection. As demonstrated

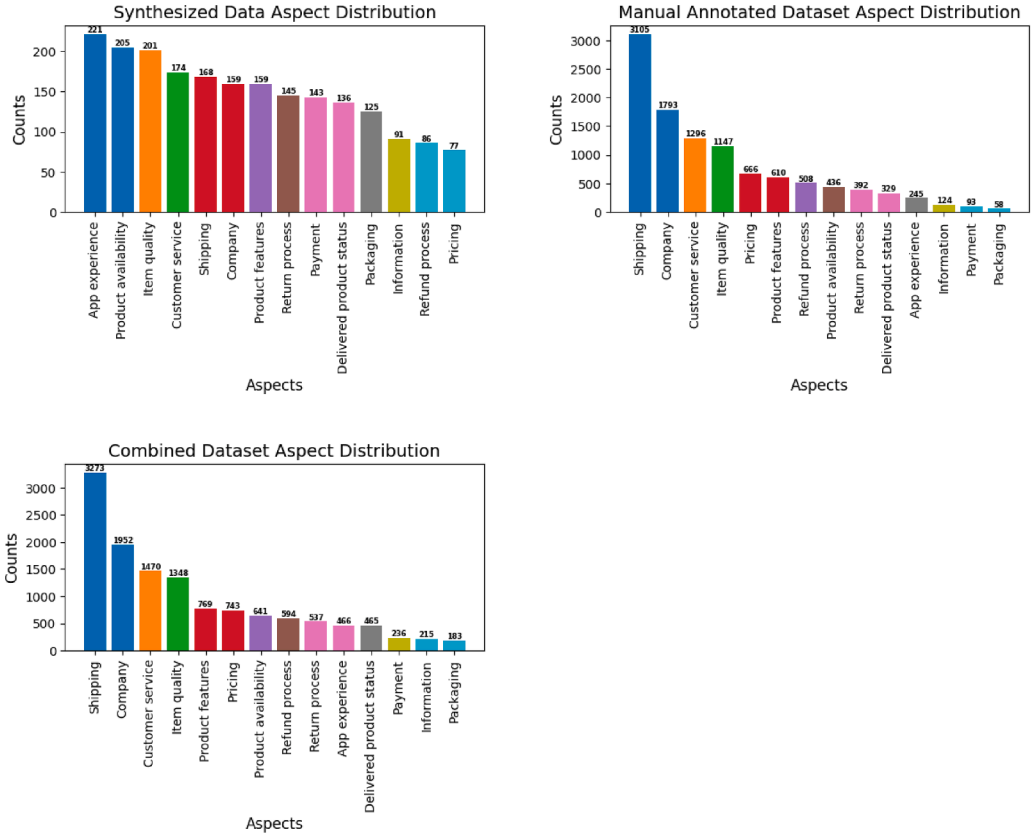

Fig. 6. Comparison of Aspect Frequencies in Synthesized, Manual, and Combined Datasets.

Table 5
Performance comparison of BERT_Review models.

Method	Loss-validation set	Accuracy-validation set	F1-validation set	Execution time (in seconds)
BERT_Review_annotated_data	0.104	0.972	0.841	3077
BERT_Review_combined_data	**0.1**	**0.974**	**0.86**	**2,733**

in Table 5, these efforts led to an improvement in the aspect detection model's performance, achieving an F1 score of 86 %.

However, the weak point of this model is that it cannot identify aspects when they are not directly mentioned in the text in the form of nouns or noun phrases, such as "not refunded," due to the method used for data annotation and fine-tuning the model. Therefore, we used the union output of the fine-tuned LLM-based ABSA model, described in Section 6, and the BERT_Review model to return the final aspects of the pipeline.

5. Pipeline stage 2: Sentiment analysis

For this task, we used our annotated SA dataset with sentiments as follows: Positive, Negative, and Mixed (Davoodi & Mezei, 2022). We employed a comprehensive evaluation of various machine learning models. We found that RoBERTa significantly outperformed the traditional models and also BERT. RoBERTa achieved an accuracy of 98.8 % on the test set, which was higher than the 92.1 % accuracy of BERT. Naïve Bayes and SVM models also showed strong performance, with Naïve Bayes using BOW achieving an accuracy of 91.6 % and SVM with TF-IDF achieving 90.5 % accuracy. However, the transformer models consistently outperformed traditional models, both in terms of accuracy and F1 score.

One important observation made during the evaluation is that the transformer-based models, particularly RoBERTa, demonstrated robustness across different e-commerce platforms, achieving similarly high performance on all datasets. This highlights the potential of transformer models for real-world applications in sentiment analysis. Therefore, we chose the best model, RoBERTa, for detecting the overall sentiment of the feedback in the final pipeline because RoBERTa proved to be highly effective for sentiment classification tasks. The superior performance of RoBERTa, with an accuracy exceeding 98 %, demonstrates the advancement in language models for natural language processing tasks. The study also suggests that companies should consider using such advanced models for better customer review analysis, as manual sentiment classification can be prone to error when relying solely on user ratings.

6. Pipeline stage 3: Aspect-Based sentiment analysis

For the task, we fine-tuned two different models: RoBERTa and TinyPixel/Llama-2-7B. The rationale behind this approach was that while RoBERTa is faster at running inference, it struggled to detect sentiments in sentences with mixed sentiments accurately. On the other hand, Llama-2-7B excels at detecting mixed sentiment aspects but operates more slowly. Therefore, in the final pipeline, we only used the fine-tuned Llama-2-7B model to detect the sentiment of each aspect when the overall sentiment was mixed. Otherwise, we used the RoBERTa model for aspect sentiment detection.

6.1. RoBERTa-Based aspect-based sentiment analysis model

To test our ABSA dataset, a range of machine learning models were employed and evaluated for their performance in aspect-based

sentiment classification. These models included a Long Short-Term Memory (LSTM) network and four transformer-based models: BERT, RoBERTa, DistilBERT, and XLNet. Given the relatively small size of the dataset, a 4-fold cross-validation approach was adopted, with each fold undergoing five epochs of training. Performance metrics such as accuracy, F1 score, and Area Under the Curve (AUC) were tracked to assess model effectiveness.

The LSTM model was implemented using TensorFlow and Keras, while the transformer-based models leveraged respective tokenizers, including the AutoTokenizer for BERT and the RoBERTaTokenizer for RoBERTa. The Adam optimizer and categorical cross-entropy loss function were applied across models, with accuracy serving as the primary performance metric during training and validation. Among the evaluated models, RoBERTa demonstrated the highest performance, achieving an accuracy of 89.56 % and an F1 score of 0.89 during cross-validation. BERT also performed well, obtaining an F1 score of 0.84, though its results were slightly lower in comparison to RoBERTa, particularly in the context of the dataset's 29 distinct sentiment categories.

Although the LSTM model achieved a higher F1 score of 0.93 and a validation accuracy of 90.89 % in one specific fold, RoBERTa was identified as the most consistent and effective model overall for fine-grained sentiment classification. Its ability to generalize across various aspects and sentiment polarities rendered it the most suitable model for ABSA within the context of this study. Finally, for implementing our pipeline to detect the sentiment of each aspect, we used the best model with our annotated ABSA dataset and fine-tuned the RoBERTa model with a fixed splitting. The RoBERTa model was trained again using optimal hyperparameters, including a learning rate of 1e-07, seven epochs, and a dropout rate of 0.5. The dataset was divided into three subsets: training (80 %), validation (12 %), and test (8 %). The training set was used to train the RoBERTa model, while the validation set was utilized to fine-tune and assess the model's performance during training. The test set was reserved solely for conducting error analysis after the model had been trained. We employed the final model to detect the sentiment of each aspect when the overall sentiment was not mixed. The model achieved an F1 score of 92 % on the validation data.

6.2. LLM-Based aspect-based sentiment analysis model

The fine-tuned LLM model processes raw customer text to output detected aspects, associated sentiments, and overall sentiment, combining aspect detection and sentiment analysis into a single step. We used this model for identifying mixed sentiment in ABSA as it performs better than RoBERTa-based model. Moreover, the ABSA LLM model's capability extends to detecting aspects that were not even explicitly introduced during training, showcasing the robustness and adaptability of the LLM-based models.

The training process employed for this project utilized the TinyPixel/Llama-2-7B-bf16-sharded model with parameter-efficient fine-tuning (PEFT). The model was trained on a dataset containing 2782 customer reviews. The dataset was split into three subsets as follows: Training (2482), Validation (200), and Test (100) sets. Fine-tuning was performed using the Soft Prompt Tuning (SFT) technique with quantization, making the model more computationally efficient while preserving its ability to learn complex patterns. The hyperparameters used for the training include Learning rate, Batch size, Epochs, Gradient accumulation, LoRA dropout, and LoRA parameters.

Based on our manually annotated ABSA dataset, we used the following prompt to fine-tune a Llama model:

- `### Instruction:` Take customer feedback as input and identify key aspects discussed in the feedback. Categorize these aspects into predefined tags as follows 'Information', 'Product_features', 'Item_quality', 'App_experience', 'Pricing', 'Product_availability', 'Payment', 'Packaging', 'Delivered_product_status', 'Shipping', 'Return_process', 'Refund_process', 'Customer_service', and 'Trust'
- `### Input:` Excellent and well organized. I will recommend this site to everyone.
- `### Response:` Trust: Positive

Table 6
ROUGE score evaluation on test set.

ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0.88	0.84	0.88	0.88

Table 7
BERT score evaluation on test set.

Precision	Recall	F1 Score
0.978	0.9762	0.9772

Table 8
The generated sentiments and aspects by the model on the test set.

Input	True Response	Generated Response
They sent a totally different item, not what I ordered. Had to return everything.	Delivered_product _status: Negative	Delivered_product _status: Negative
Oh so late!	Shipping: Negative	Shipping: Negative
You do not always get what you thought	Product_features: Negative	Delivered_product _status: Negative, Product_features: Negative

After training, the model was evaluated on the test set (100 examples). The performance of the model was evaluated using two key metrics: ROUGE and BERT scores.

As shown in Table 6, the ROUGE scores represent the performance of the model in terms of n-gram overlap between the generated responses and the true responses. ROUGE-1 measures the overlap of unigrams (single words), ROUGE-2 measures the overlap of bigrams (pairs of consecutive words), and ROUGE-L focuses on the longest common subsequence between the two texts. The high values across all ROUGE scores indicate that the model produces highly accurate summaries that are close to the ground truth. Moreover, the high ROUGE-1, ROUGE-2, and ROUGE-L scores demonstrate that the model performs exceptionally well in generating aspect-based sentiment annotations that overlap significantly with the ground truth. The near-perfect scores (above 0.88 for all ROUGE metrics) indicate the model's robust ability to extract the correct aspects and sentiments from customer feedback.

As presented in Table 7, the BERT score evaluates the semantic similarity between the generated text of the model and the true reference. The high BERT scores demonstrate that the model captures not only the surface-level similarities but also the underlying meaning of the feedback and its corresponding aspect-based sentiments. This is particularly valuable for ensuring that the sentiment analysis task is not just syntactically but also semantically accurate. In other words, with a near-perfect F1 score of 0.977, the model accurately understands the exact meanings in customer feedback and responds with highly relevant aspect-based sentiment annotations. High precision and recall further suggest that the model outputs highly relevant predictions and captures a wide range of correct aspect-sentiment pairs.

Table 8 shows some examples generated by the model and the True answer annotated by the author. Inference results on the test set further corroborate the model's strong performance. For instance, when given input feedback like "I am so pissed right now at SHEIN" the model correctly identifies the 'Trust' as Negative, matching the true response.

As presented in Table 9, alongside BERT and ROUGE, we computed the F1 score and accuracy at both the overall level on the test set and separately for each aspect.

The performance of the fine-tuned model for the ABSA task shows a strong overall result. This indicates that the model is relatively effective

Table 9
Overall performance for Aspect and Sentiment analysis.

Metric	Aspect	Sentiment
F1 Score	0.89	0.91
Accuracy	0.88	0.90

at identifying the correct aspects from user feedback, although there is room for improvement in distinguishing some categories. On the sentiment side, the model performs slightly better, with an overall sentiment F1 score of 0.91 and an accuracy of 0.90. This suggests that the model is more adept at correctly classifying sentiment polarity (positive, negative) within feedback, which is essential for strong analysis in ABSA tasks.

In conclusion, the model, trained on customer feedback for aspect-based sentiment analysis, performed exceptionally well, as indicated by the high ROUGE and BERT scores. Its ability to generate responses with strong semantic and syntactic overlap with true annotations demonstrates its suitability for real-world applications in sentiment analysis tasks, particularly in categorizing and analysing customer feedback. Moreover, the model demonstrates an impressive ability to detect aspects that are not explicitly mentioned in the text through nouns and noun phrases, such as in the sentence “easy to use but takes long to come.” This capability significantly enhances the model’s performance in ABSA and aspect detection tasks, making it more robust and versatile in identifying implicit aspects.

7. Final pipeline for feedback analysis automation

Despite the increasing importance of customer feedback analysis in e-commerce, few comprehensive pipelines have been proposed to address the specific challenges of ABSA in the e-commerce domain. While some studies focus on individual components like aspect detection or sentiment analysis, the development of a fully integrated system that captures both explicit and implicit aspects and handles mixed sentiments remains limited in the literature. Our proposed pipeline could empower businesses to understand customer opinions better, enhance product offerings, and improve overall customer satisfaction, making it a critical advancement in the field of e-commerce analytics.

7.1. Existing solutions and their drawbacks

The PyABSA framework is an open-source, modular tool designed for ABSA, supporting tasks such as ATE and ASC (Yang et al., 2023). It integrates a variety of pre-built datasets, provides data augmentation tools, and supports multilingual modelling. The framework is highly customizable, allowing users to add models and datasets, and it simplifies training and inference with minimal code. It also includes a Metric Visualizer to track and visualize model performance metrics automatically. However, a key weakness of PyABSA lies in its dependence on pre-existing datasets, which may lead to performance volatility in domain-specific applications. Additionally, while multilingual, its performance may vary across languages, especially those with less training data. Further, complex sentiment nuances can still pose challenges for the framework (Yang & Li, 2022).

Lakatos et al. (2024) presents a machine learning-driven pipeline designed to extract insights from customer reviews in e-commerce. The pipeline integrates various natural language processing techniques, including vector embedding-based keyword extraction, clustering, and sentiment analysis. Initially, the text data undergoes preprocessing to remove irrelevant elements, followed by keyword/keyphrase extraction using a combination of N-gram, dependency parsing, and embedding-based methods like SBERT (Sentence-BERT). The extracted key phrases are then grouped using recursive hierarchical clustering to identify relevant topics. Sentiment analysis is conducted using a modified BERT + R model with a regression layer to predict sentiment on a granular scale. The recursive clustering approach further enhances topic detection by reapplying clustering methods until optimal results are achieved. However, the model’s performance heavily relies on the quality of embeddings and is computationally demanding when applied to large datasets.

Maalej et al. (2025) outlines a machine learning and NLP-based pipeline to process large volumes of user feedback efficiently. The pipeline involves multiple stages, including feedback collection, preprocessing (sentiment analysis, quality assessment), classification (e.g., bug reports, feature requests), and clustering for topic identification. It also includes summarization techniques to extract actionable insights and match feedback to development artifacts, like issue trackers. The model utilizes transformers such as BERT for classification in combination with clustering, with the summarization improved using LLMs. Despite advancements, challenges remain in accurately clustering and summarizing feedback due to its informal and complex nature, often requiring human intervention.

7.2. The proposed pipeline

To address the need for an automated system for customer feedback analysis in the e-commerce sector, we propose a comprehensive machine-learning pipeline. While the mentioned previous works in e-commerce often focus on either aspect extraction or sentiment analysis, we integrated LLMs, BERT, and RoBERTa models into a unified pipeline. Moreover, earlier studies, such as Modi et al. (2011), proposed a socially inspired framework for combining multiple inference algorithms or ’experts’ through opinion aggregation principles to improve decision accuracy in human state inference. The idea that each model contributes partial expertise parallels our current study’s design philosophy. However, unlike the rule-based expert integration in Modi et al.’s framework, our pipeline integrates transformer-based models in a sequential and data-driven manner, where each model’s output feeds into the next to enable automated, large-scale text understanding across multiple e-commerce dimensions. The purpose of this pipeline is to provide businesses with granular and macro-level insights into customer opinions, thus enabling more informed decision-making and targeted improvements. This pipeline covers multiple layers of feedback analysis, from aspect detection to granular sentiment analysis, and even includes models for handling mixed sentiments. This level of integration offers a more nuanced understanding of feedback compared to pipelines that focus on either aspect detection or sentiment analysis in isolation. The pipeline comprises several key components:

1. LLM Model for Generating Augmented Annotated Feedback: The fine-tuned LLM generates annotated customer feedback, including the identification of various aspects within the text. The output serves as a synthetic dataset for training and testing purposes, improving the performance of downstream models. This approach addresses the data scarcity challenge, especially in less frequent aspects such as “Packaging” or “Information,” which harm the performance of the machine learning models.
2. BERT-Based Aspect Detection Model: The BERT_Review model is utilized for identifying specific aspects of customer feedback. Using a deep contextual understanding of BERT, this model accurately detects various dimensions of customer experiences, such as product quality, delivery, or service-related aspects. Even though this model performs fast and has the ability to detect a wide range of aspects in the text, but has difficulty detecting implicit aspects. Therefore, we combined the aspects detected by this model with aspects detected by the fine-tuned ABSA LLM model, and we reported the union of the two models as the final set of aspects. The combination of BERT for explicit aspect detection and the fine-tuned LLM for implicit aspect detection helps overcome the limitations of previous models that struggled with implicit aspects. By leveraging both models and reporting their union, our pipeline ensures a more comprehensive

Table 10
Number of misclassification in final pipeline.

	Aspect	Total number	Number of misclassification
1	Payment	251	99
2	Return_process	594	24
3	Refund_process	628	32
4	Delivered_product_status	701	89
5	Information	710	137
6	Packaging	804	103
7	Product_availability	954	175
8	Customer_service	1014	45
9	Product_features	1046	80
10	Pricing	1215	73
11	Item_quality	1510	79
12	App_experience	2187	237
13	Trust	3431	79

aspect detection, addressing a common gap in the literature where implicit aspects are often overlooked.

3. RoBERTa-Based Aspect-Based Sentiment Analysis (ABSA) Model: This model focuses on identifying the sentiment associated with each aspect of customer feedback. It offers an enhanced sentiment classification that allows for differentiating between positive, negative, or neutral sentiments tied to specific aspects. However, this model has difficulty detecting the mixed sentiment correctly, there for when the overall sentiment of the sentence is mixed, we used the fine-tuned LLM ABSA model for identifying the sentiments. A key advancement is the inclusion of the fine-tuned LLM-based ABSA model, which surpasses traditional models like RoBERTa in handling mixed sentiments. This is crucial in e-commerce feedback, where customer reviews often express both positive and negative sentiments in the same entry.
4. LLM-Based ABSA Model: This model integrates aspect detection and sentiment analysis in one step. The LLM is fine-tuned to process raw customer feedback and extract aspects, their associated sentiments, as well as the overall sentiment of the feedback. Notably, this model performs well in detecting mixed sentiments, surpassing the performance of the RoBERTa-based ABSA model in handling complex feedback.
5. RoBERTa-Based Model for Overall Sentiment Detection: Designed to assess the overall sentiment of customer feedback, this model aggregates the sentiment data from individual aspects and generates a holistic sentiment label, helping businesses to gauge customer satisfaction at a glance.

The final pipeline aims to facilitate the identification of dissatisfaction or churn factors, ultimately accelerating managerial decision-making. For instance, in a feedback like, “Even though I like their variety of brands and prices, I won’t shop here again because of the late shipping”, the pipeline would analyze the feedback and generate the following aspects and associated sentiments: Trust - Negative; Shipping - Negative; Product Availability - Positive; Pricing - Positive; Overall Sentiment - Mixed. This pipeline ensures a detailed and automated understanding of customer feedback, providing actionable insights that businesses can use to refine their offerings, address pain points, and enhance overall customer satisfaction.

For a final test of the pipeline, we scraped 100,745 reviews from Trustpilot. We used the pipeline to predict the aspects and their associated aspects. We deleted the rows whose aspects were not detected, the final size of the dataset after cleaning is 100,650. To check the performance of the pipeline, we randomly selected 3431 data points for manual inspection, assuring that the infrequent aspects such as Information and Packaging are present in the pipeline. In Table 10, we present the number of misclassifications in each aspect. We have also implemented a pipeline as an application for testing purposes; a screenshot of the basic interface is presented in Fig. 7.

When comparing the pipeline’s performance against the original annotated dataset, it becomes evident that the model performs well on frequent aspects such as “Trust,” “App Experience,” and “Customer Service,” showing low misclassification rates (e.g., Trust: 2.3 %, App Experience: 10.8 %). Infrequent aspects like “Packaging” and “Product Availability,” while more challenging for the model, still displayed acceptable performance, though with higher misclassification rates (e.g., Packaging: 12.8 %, Product Availability: 18.3 %). This indicates that while the pipeline can detect less frequent aspects, there is room for further refinement, particularly in improving accuracy for these underrepresented categories. Overall, the pipeline demonstrates robust performance across the dataset, effectively balancing the detection of frequent and infrequent aspects and providing reliable insights for both granular and macro-level customer feedback analysis.

Moreover, in addition to the 14 original aspects presented to the model during the fine-tuning process, the pipeline also demonstrated

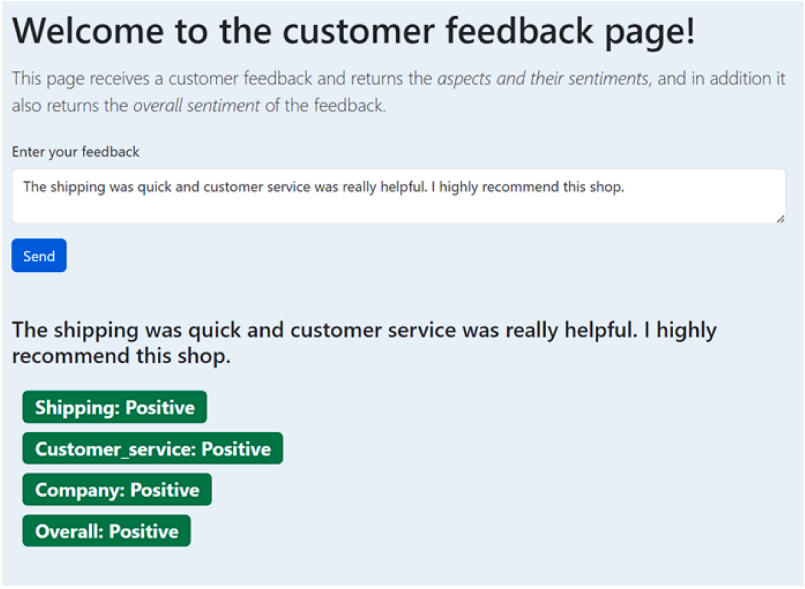

Fig. 7. A screenshot of the implementation of the pipeline with an example review.

Table 11
Comparison between PyABSA and the proposed pipeline.

Sentence	PyABSA output	Proposed pipeline output
The price was high but the location is great.	*price: Negative; location: Positive*	*Pricing: Positive; App_experience/Location: Positive; Overall: Mixed*
easy checkout, but bad customer service.	*checkout: Positive; customer service: Negative*	*App_experience: Positive; Customer_service: Negative; Overall: Mixed*
not refunded.	*refunded: Negative*	*Refund_process: Negative; Overall: Negative*
terrible waiting time but good quality.	*waiting time: Negative; quality: Positive*	*Shipping: Negative; Item_quality: Positive; Overall: Mixed*
came late, but well packed.	*(no output or missing aspect)*	*Packaging: Positive; Shipping: Negative; Overall: Mixed*
can find everything, but late dispatch.	*dispatch: Negative*	*Product_availability: Positive; Shipping: Negative; Overall: Mixed*
good collection but bad shipping.	*shipping: Positive*	*Product_availability: Positive; Shipping: Negative; Overall: Mixed*
No answer.	*answer: Negative*	*Customer_service: Negative; Overall: Negative*
I ordered a pair of boots online and they arrived well within the timescale . I was very pleased with them and would definitely use Zalando again . The price was very reasonable and the packaging was very smart and discreet .	*price: Positive; packaging: Positive*	*Shipping: Positive; Item_quality: Positive; Pricing: Positive; Trust: Positive; Packaging: Positive; Overall: Positive*

the ability to detect various random aspects that were not explicitly part of the training set. These aspects included terms such as 'Communication,' 'App,' 'Search,' 'Email,' 'Filters,' 'Deal,' 'Inventory,' 'Claim_handling,' 'Marketing,' 'Input_error,' 'Quality_control,' 'Discount,' 'Tracking,' 'Guarantee,' 'Advertisement,' 'Privacy,' and others. While this showcases the versatility and powerful aspect-detection capabilities of the LLM models, these additional aspects appeared sporadically throughout the dataset without showing any discernible pattern.

The appearance of these random aspects suggests that the model is highly sensitive to subtle and specific feedback, even detecting aspects that were not intended to be a focus during the fine-tuning process. However, the lack of consistency in these emergent aspects highlights the need for further model refinement to better filter or consolidate less relevant or one-off detections, ensuring that the pipeline focuses on actionable insights rather than generating noise in the form of isolated, infrequent aspects.

Moreover, to assess the practical value of the proposed pipeline beyond architectural description, we conducted a side-by-side comparison against a widely used open-source baseline PyABSA (Yang et al., 2023). Both systems were executed on a curated set of short review-like sentences designed to stress common ABSA phenomena. We observed that the PyABSA baseline returns aspect terms with a single polarity per term; our pipeline (i) jointly identifies multiple aspects, (ii) assigns aspect-level polarities while also producing an overall judgment when the sentence contains conflicting cues, and (iii) maps raw aspects to a managerial taxonomy (e.g., Shipping, Customer_service, Pricing, Packaging), which allows immediate aggregation for decision support. Across the examples, our proposed pipeline yields higher aspect coverage, demonstrates greater polarity robustness under conjunctions and negations, and handles very short complaints (with implicit aspects) such as "not refunded" or "not responded", for which PyABSA often returns a single aspect without an overall assessment. Representative outputs are reported in Table 11. Taken together, these observations support the design choice of a multi-aspect, taxonomy-aware pipeline aimed at managerial use, rather than a single-model ABSA baseline optimised primarily for token-level extraction. We also acknowledge that our proposed pipeline is designed and tuned for customer-service contexts in online retail (e.g., shipping, delivery, returns, customer service, pricing, packaging). In this domain, the ontology and training data yield high aspect coverage and stable polarity assignment. Outside this context, such as restaurant or venue reviews, performance may degrade because certain aspects are out of distribution relative to the training set. For example, in the sentence "The food menu was very diverse but the parking was tight," the system incorrectly mapped parking to App_experience rather than a logistics-style access/venue category. This error reflects a domain shift rather than a systematic flaw in the labeling logic, as the training corpus contains many e-commerce app and website mentions. In practice, we therefore recommend deploying the pipeline within online shopping customer service-related settings.

7.3. Interactions among pipeline components and comparative evaluation

To clarify how the different components of the proposed pipeline interact, Table 12 summarizes their connections, roles, and comparative advantages. Rather than multiple independent pipelines, the system consists of one unified end-to-end pipeline composed of several complementary sub-modules. The LLM-based synthetic data generator supports both the Aspect Detection and ABSA components by increasing data diversity. The Aspect Detection module (BERT-based) identifies explicit aspects, while the ABSA module (RoBERTa + LLM) links those aspects to sentiment polarity, and the final Sentiment module aggregates all results into a holistic trust-satisfaction measure.

These components operate sequentially and complementarily: synthetic data generation improves training robustness; BERT ensures high-precision explicit aspect tagging; the LLM handles implicit aspects and mixed emotions; and RoBERTa enables scalable sentiment classification. Their integration reduces error propagation between subtasks and balances accuracy, interpretability, and computational efficiency, providing a coherent, explainable workflow rather than several isolated pipelines.

Table 12
Interactions among pipeline components and comparative evaluation.

Component	Input	Output	Role / Connection	Key Advantages	Limitations
Synthetic Data Generation (LLM)	Raw customer feedback	Synthesized, annotated customer feedback	Augments training data for Aspect Detection and ABSA; improves coverage of rare aspects and phrasing variants	Reduces manual labeling effort and increases data diversity	Requires quality control; risk of distribution drift for low-frequency aspects
Aspect Detection (BERT)	Cleaned raw customer reviews	Explicit aspect terms or phrases	Feeds high-precision aspect spans to ABSA modules (RoBERTa/LLM)	High precision for explicit aspects; efficient inference	May miss implicit or context-dependent aspects
Aspect Detection (LLM)	Raw customer reviews	Implicit and inferred aspects	Complements BERT; union of BERT and LLM outputs maximizes recall	Captures non-lexical or implicit aspects and complex phrasing	Higher computational cost; slower inference; variable outputs
ABSA (RoBERTa)	Aspect-review pairs or sentences	Sentiment per aspect (polarity)	Default ABSA stage when text is not highly mixed; consumes BERT/LLM aspect spans	Fast, scalable, and stable sentiment classification	Degrades in mixed or complex reviews
ABSA (LLM)	Raw text (with or without aspects)	Aspect-sentiment tuples	Fallback or parallel ABSA for mixed or ambiguous reviews; reconciled with RoBERTa outputs	Handles contradictions and long-range context; recovers missed pairs	Slower and more resource-intensive; requires prompt optimization
Overall Sentiment (RoBERTa)	Aspect-level sentiments or full review	Aggregated overall sentiment	Aggregates sentiment for dashboard KPIs and trust indicators.	Provides a robust managerial signal.	problem with mixed/contradicted sentiments

8. Discussion and conclusions

This study presents a comprehensive machine learning pipeline for automated customer feedback analysis in e-commerce, addressing key challenges in aspect detection and sentiment classification. By integrating pre-trained transformer models with fine-tuned large language models and synthetic data augmentation, the proposed approach improves both accuracy and scalability. The results demonstrate that a multi-model approach, rather than relying on a single model, provides a more robust method for analysing customer feedback, particularly in detecting implicit aspects and mixed sentiments, which traditional models often struggle to handle.

One of the most important findings is the superior performance of the fine-tuned LLM-based model in ABSA tasks. The model effectively identifies aspects that are not explicitly mentioned in the text, such as issues related to refunds, service reliability, and delivery experiences. This suggests that LLMs offer a deeper contextual understanding of customer sentiment compared to traditional transformer-based models, which tend to perform better in explicit aspect detection but fail to capture nuanced sentiment shifts within a single review. The hybrid approach improves coverage and minimizes misclassification of aspects, particularly in under-represented categories.

Another notable contribution is the ability to generate synthetic annotated feedback to supplement manually labelled datasets. The use of fine-tuned LLMs for data augmentation addresses the challenge of imbalanced aspect representation, which was previously difficult for models to classify due to insufficient training data. The inclusion of this synthetic data resulted in a more balanced dataset, reducing bias in aspect detection and improving the pipeline's generalization to real-world customer feedback.

Specifically, regarding our first RQ concerning the integration of transformer-based and large language models in a unified pipeline, we demonstrated that integrating transformer-based models (BERT and RoBERTa) with fine-tuned large language models (Llama-2-7B) significantly improves the automation and depth of customer feedback analysis. The combined architecture allows explicit aspects to be detected efficiently through BERT, while implicit and mixed sentiments are captured through the LLM-based ABSA model. By merging their outputs and incorporating synthetic, LLM-generated annotated data, the pipeline achieves high accuracy and broad coverage of both frequent and infrequent aspects. The modular design also enhances explainability, as each component's role, aspect detection, aspect-level sentiment analysis, and overall sentiment aggregation, remains transparent and interpretable.

As for the second RQ, regarding the effect of multi-model architecture, our results confirm that the proposed multi-model approach outperforms single-model and traditional data-mining methods in accuracy, interpretability, and scalability. Compared with single models, the pipeline shows improved robustness in detecting implicit aspects, mixed sentiments, and low-frequency categories. It also generalizes well to unseen data, correctly identifying relevant aspects beyond those explicitly included in the training set. Moreover, the layered structure enables efficient scaling and provides interpretable outputs that can be directly aggregated for managerial decision-making. The side-by-side comparison with the PyABSA baseline further supports the pipeline's superiority in aspect coverage and polarity consistency, validating the effectiveness of a multi-model architecture for large-scale, domain-specific feedback analysis in e-commerce.

8.1. Contributions

From an academic perspective, this study contributes to the ongoing research in aspect-based sentiment analysis by demonstrating how a multi-model framework can overcome common limitations in existing approaches. Previous research has primarily relied on single-model architectures, which often struggle with mixed sentiment detection and fail to generalize well across different e-commerce domains. By implementing a pipeline that integrates multiple models and synthetic data, this study provides a scalable and adaptable solution applicable beyond the e-commerce sector to other domains where feedback analysis is crucial, such as healthcare and finance. Additionally, the study demonstrates how fine-tuning LLMs with structured prompts can significantly enhance aspect extraction and sentiment classification, which remains an area of active research in natural language processing.

The findings also have practical implications for businesses and e-commerce platforms that rely on customer feedback for decision-making and service optimization. The automated feedback analysis pipeline enables businesses to extract actionable insights at both granular and macro levels, providing a data-driven approach to improving customer

experience. Companies can quickly identify common pain points, such as delayed shipping or product quality concerns, and take proactive measures to enhance service offerings. By incorporating real-time feedback processing, businesses can improve customer retention by addressing issues before they escalate into widespread dissatisfaction. Moreover, the ability to detect mixed sentiments allows for a more nuanced understanding of customer experiences rather than simply classifying feedback as positive or negative.

8.2. Limitations

While the pipeline demonstrates strong performance across multiple aspects, there are certain limitations. The misclassification rates for infrequent aspects indicate that even with synthetic data augmentation, some categories require further refinement. The detection of rare or ambiguous aspects, such as "Privacy", suggests that LLMs can identify new aspects beyond the predefined categories, but these aspects appear inconsistently, making it challenging to derive structured insights. Future work could explore reinforcement learning techniques to refine aspect classification and improve the interpretability of aspect detection models.

Another challenge lies in computational efficiency. While fine-tuned LLMs outperform other models in detecting implicit aspects and mixed sentiments, their inference speed can be somewhat slower. Optimizing the trade-off between computational cost and performance remains a key area for improvement, particularly for deploying this pipeline in real-time applications. Future research could investigate the use of parameter-efficient fine-tuning methods to maintain high accuracy while reducing model complexity.

CRediT authorship contribution statement

Laleh Davoodi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization; **József Mezei:** Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision; **Shahrokh Nikou:** Conceptualization, Writing – original draft, Writing – review & editing; **Leonardo Espinosa-Leal:** Conceptualization, Validation, Writing – original draft, Writing – review & editing, Supervision.

Data availability

The link to the datasets is provided in the article.

Acknowledgment

The first author received financial support for conducting this research from Jenny ja Antti Wihurin rahasto, Liikesivistysrahasto, and Marcus Wallenbergin Foundation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akter, S., & Wamba, S. F. (2016). Big data analytics in e-commerce: A systematic review and agenda for future research. *Electronic Markets*, *26*, 173–194.

Ali, L., Haq, M. Z. U., Asadullah, M. A., & Haider, G. (2025). Leveraging emotional intelligence to enhance supply chain visibility and risk management: Evidence from empirical research. *International Journal of Engineering Business Management*, *17*, 18479790251368324.

Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*, *20*(1/2), 142–161.

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Computers in Human Behavior*, *114*, 106548.

Ansari, G., Saxena, C., Ahmad, T., & Doja, M. N. (2020). Aspect term extraction using graph-based semi-supervised learning. *Procedia Computer Science*, *167*, 2080–2090.

Arora, M. (2016). Selection of parameters of e-commerce websites using AHP. In *Proceeding of 2nd conference on innovative practices in information technology & operations management* (pp. 22–27).

Arthamevia, N. P., Purbolaksono, M. D. et al. (2021). Aspect-based sentiment analysis in beauty product reviews using TF-IDF and SVM algorithm. In *2021 9Th international conference on information and communication technology (ICoICT)* (pp. 197–201). IEEE.

Aydin, C. R., & Güngör, T. (2020). Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations. *IEEE Access*, *8*, 77820–77832.

Azov, G., Pelc, T., Alon, A. F., & Kamhi, G. (2024). Self-improving customer review response generation based on LLMs. arXiv preprint arXiv:2405.03845.

Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, *52*, 201–213.

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, *55*(7), 1–39.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment analysis about e-commerce from tweets using decision tree, k-nearest neighbor, and naïve bayes. In *2018 International conference on orange technologies (ICOT)* (pp. 1–6). https://doi.org/10.1109/ICOT.2018.8705796

Biana, J., Zhai, W., Huang, X., Zheng, J., & Zhu, S. (2024). Vaner: Leveraging large language model for versatile and adaptive biomedical named entity recognition. arXiv preprint arXiv:2404.17835.

Biswas, S., Wardat, M., & Rajan, H. (2022). The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In *Proceedings of the 44th international conference on software engineering* (pp. 2091–2103).

Blut, M. (2016). E-Service quality: Development of a hierarchical model. *Journal of Retailing*, *92*(4), 500–517.

Brinkmann, A., Shraga, R., Der, R. C., & Bizer, C. (2023). Product information extraction using chatGPT. arXiv preprint arXiv:2306.14921.

Chen, Q., Lu, Y., Gong, Y., & Xiong, J. (2023). Can AI chatbots help retain customers? impact of AI service quality on customer loyalty. *Internet Research*, *33*(6), 2205–2243.

Coulombe, C. (2018). Text data augmentation made simple by leveraging nlp cloud apis. arXiv preprint arXiv:1812.04718.

Davoodi, L. (2023). Enhancing the understanding of e-commerce reviews through aspect extraction techniques: A BERT-based approach. In *36Th bled econference digital economy and society: The balancing act for digital innovation in times of instability* (p. 233).

Davoodi, L., & Mezei, J. (2022). A comparative study of machine learning models for sentiment analysis: Customer reviews of e-commerce platforms. In *Proceedings of the 35th bled econference digital restructuring and human (re) action* (pp. 1–14). https://doi.org/10.18690/um.fov.4.2022

Davoodi, L., Mezei, J., & Heikkilä, M. (2025). Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms. *Electronic Commerce Research*, (pp. 1–43).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, *118*, 272–299.

Dogra, V. et al. (2024). Aspect-based approaches for measuring customer feedback in the e-commerce industry. In *2024 2Nd international conference on sustainable computing and smart systems (ICSCSS)* (pp. 479–484). IEEE.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*, 1–14.

Frei, J., & Kramer, F. (2023). Annotated dataset creation through large language models for non-english medical NLP. *Journal of Biomedical Informatics*, *145*, 104478.

Gajewska, T., Zimon, D., Kaczor, G., & Madzík, P. (2020). The impact of the level of customer satisfaction on the quality of e-commerce services. *International Journal of Productivity and Performance Management*, *69*(4), 666–684.

Giannakopoulos, A., Musat, C., Hossmann, A., & Baeriswyl, M. (2017). Unsupervised aspect term extraction with b-LSTM & CRF using automatically labelled datasets. arXiv preprint arXiv:1709.05094.

Grewal, D., Levy, M., & Kumar, V. (2009). Customer experience management in retailing: An organizing framework. *Journal of Retailing*, *85*(1), 1–14.

Griva, A. (2022). "I can get no e-satisfaction". what analytics say? evidence using satisfaction data from e-commerce. *Journal of Retailing and Consumer Services*, *66*, 102954.

Gupta, R. (2019). Data augmentation for low resource sentiment analysis using generative adversarial networks. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7380–7384). IEEE.

Hammi, S., Hammami, S. M., & Belguith, L. H. (2023). Advancing aspect-based sentiment analysis with a novel architecture combining deep learning models CNN and bi-RNN with the machine learning model SVM. *Social Network Analysis and Mining*, *13*(1), 117.

Han, H., & Ryu, K. (2012). The theory of repurchase decision-making (TRD): identifying the critical factors in the post-purchase decision-making process. *International Journal of Hospitality Management*, *31*(3), 786–797.

He, X., Nassar, I., Kiros, J., Haffari, G., & Norouzi, M. (2022). Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, *10*, 826–842.

Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, *52*(3), 1495–1545.

Holloway, B. B., & Beatty, S. E. (2008). Satisfiers and dissatisfiers in the online environment: A critical incident assessment. *Journal of Service Research*, *10*(4), 347–364.
Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177).
Huang, W.-H., Shen, G. C., & Liang, C.-L. (2019). The effect of threshold free shipping policies on online shoppers' willingness to pay for shipping. *Journal of Retailing and Consumer Services*, *48*, 105–112.
Huyghe, E., Verstraeten, J., Geuens, M., & Van Kerckhove, A. (2017). Clicks as a healthy alternative to bricks: How online grocery shopping reduces vice purchases. *Journal of Marketing Research*, *54*(1), 61–74.
Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201.
Koufaris, M., Kambil, A., & LaBarbera, P. A. (2001). Consumer behavior in web-based commerce: An empirical study. *International Journal of Electronic Commerce*, *6*(2), 115–138.
Kumar, A., & Anjaly, B. (2017). How to measure post-purchase customer experience in online retailing? a scale development study. *International Journal of Retail & Distribution Management*, *45*(12), 1277–1297.
Kumar, V. V., Raghunath, K. M. K., Muthukumaran, V., Joseph, R. B., Beschi, I. S., & Uday, A. K. (2022). Aspect based sentiment analysis and smart classification in uncertain feedback pool. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), 252–262.
Kyaw, K. S., Tepsongkroh, P., Thongkamkaew, C., & Sasha, F. (2023). Business intelligent framework using sentiment analysis for smart digital marketing in the e-commerce era. *Asia Social Issues*, *16*(3), e252965.
Lakatos, R., Bogacsovics, G., Harangi, B., Lakatos, I., Tiba, A., Tóth, J., Szabó, M., & Hajdu, A. (2024). A machine learning-based pipeline for the extraction of insights from customer reviews. *Big Data and Cognitive Computing*, *8*(3), 20.
Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894.
Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69–96.
Li, H., Zhong, Q., Zhu, K., Liu, J., Du, B., & Tao, D. (2024). Iterative data augmentation with large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2407.00341.
Lindecrantz, E., Gi, M. T. P., & Zerbi, S. (2020). Personalizing the customer experience: Driving differentiation in retail. *Retrieved January*, *9*, 2022.
Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-based big data analytics for customer insight-driven design innovation in SMEs. *International Journal of Information Management*, *51*, 102034.
Ma, S., Huang, S., Huang, S., Wang, X., Li, Y., Zheng, H.-T., Xie, P., Huang, F., & Jiang, Y. (2023). Ecomgpt-ct: Continual pre-training of e-commerce large language models with semi-structured data. arXiv preprint arXiv:2312.15696.
Maalej, W., Biryuk, V., Wei, J., & Panse, F. (2025). On the automated processing of user feedback. Ferrari, A., Ginde, G. (Eds.). *Handbook on Natural Language Processing for Requirements Engineering*. (pp. 279–308). https://link.springer.com/chapter/10.1007/978-3-031-73143-3_10.
Maheshwari, G., Ivanov, D., & Haddad, K. E. (2024). Efficacy of synthetic data as a benchmark. arXiv preprint arXiv:2409.11968.
Martínez-López, F. J., Feng, C., Li, Y., & López-López, D. (2022). Using instant refunds to improve online return experiences. *Journal of Retailing and Consumer Services*, *68*, 103067.
McColl-Kennedy, J. R., Zaki, M., Lemon, K. N., Urmetzer, F., & Neely, A. (2019). Gaining customer experience insights that matter. *Journal of Service Research*, *22*(1), 8–26.
Mero, J. (2018). The effects of two-way communication and chat service usage on consumer attitudes in the e-commerce retailing sector. *Electronic Markets*, *28*, 205–217.
Meyer, C., & Schwager, A. (2007). Understanding customer experience. *Harvard Business Review*, *85*, 116–26, 157.
Modi, S., Lin, Y., Cheng, L., Yang, G., Liu, L., & Zhang, W. J. (2011). A socially inspired framework for human state inference using expert opinion integration. *IEEE/ASME Transactions on Mechatronics*, *16*(5), 874–878.
Mowlaei, M. E., Abadeh, M. S., & Keshavarz, H. (2020). Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, *148*, 113234.
Mudambi, S. M., & Schuff, D. (2010). Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS Quarterly*, (pp. 185–200).
Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, *12*, 60943–60959.
Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845–863.
Newman, M. E. J. (2013). Network data. http://www-personal.umich.edu/~mejn/netdata/.
Niranjanamurthy, M., Kavyashree, N., Jagannath, S., & Chahar, D. (2013). Analysis of e-commerce and m-commerce: Advantages, limitations and security issues. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(6), 2360–2370.
Obaidat, I., Mohawesh, R., Al-Ayyoub, M., Mohammad, A.-S., & Jararweh, Y. (2015). Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches. In *2015 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)* (pp. 1–6). IEEE.
Ogbeyemi, A., Ogbeyemi, A., & Zhang, W. (2024). Integrating human factors into the distribution model of goods and fast-moving consumer goods for effective inventory control. *International Journal of Engineering Business Management*, *16*, 18479790241266352.
Otim, S., & Grover, V. (2006). An empirical study on web-based services and customer loyalty. *European Journal of Information Systems*, *15*(6), 527–541.
Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., & Amice, C. (2001). Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In *Proceedings of the 1st international conference on knowledge capture* (pp. 116–122).
Palen-Michel, C., Wang, R., Zhang, Y., Yu, D., Xu, C., & Wu, Z. (2024). Investigating LLM applications in e-commerce. arXiv preprint arXiv:2408.12779.
Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, *14*(5), 2074.
Pham, T. S. H., & Ahammad, M. F. (2017). Antecedents and consequences of online customer satisfaction: A holistic process perspective. *Technological Forecasting and Social Change*, *124*, 332–342.
Policarpo, L. M., da Silveira, D. E., da Rosa Righi, R., Stoffel, R. A., da Costa, C. A., Barbosa, J. L. V., Scorsatto, R., & Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, *41*, 100414.
Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., & Farri, O. (2016). Neural paraphrase generation with stacked residual LSTM networks. arXiv preprint arXiv:1610.03098.
Psorakis, I., Roberts, S., Ebden, M., & Sheldon, B. (2011). Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, *83*, 066114.
Qian, T. Y., Li, W., Gong, H., Seifried, C., & Xu, C. (2024). Experience is all you need: A large language model application of fine-tuned GPT-3.5 and roBERTa for aspect-based sentiment analysis of college football stadium reviews. *Sport Management Review*, (pp. 1–25).
Rachman, A., Zhang, T., & Ratnayake, R. M. C. (2021). Applications of machine learning in pipeline integrity management: A state-of-the-art review. *International Journal of Pressure Vessels and Piping*, *193*, 104471.
Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: Comparative analysis and survey. *Artificial Intelligence Review*, *46*, 459–483.
Rane, N., Paramesha, M., Choudhary, S., & Rane, J. (2024). Business intelligence and business analytics with artificial intelligence and machine learning: Trends, techniques, and opportunities. *Techniques, and Opportunities (May 17, 2024)*.
Rane, N. L., Achari, A., & Choudhary, S. P. (2023). Enhancing customer loyalty through quality of service: Effective strategies to improve customer satisfaction, experience, relationship, and engagement. *International Research Journal of Modernization in Engineering Technology and Science*, *5*(5), 427–452.
Rita, P., Oliveira, T., & Farisa, A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. *Heliyon*, *5*(10).
Rosario, I. D., Ghosh, A., Huang, B., Yan, Y., Zhang, W., & Lin, W. (2024). Enhancing telehealth patient experience with emotion-sensitive large language models. In *International congress on information and communication technology* (pp. 541–550). Springer.
Rose, S., Hair, N., & Clark, M. (2011). Online customer experience: A review of the business-to-consumer online purchase context. *International Journal of Management Reviews*, *13*(1), 24–39.
Rubtsova, Y., & Koshelnikov, S. (2015). Aspect extraction from reviews using conditional random fields. In *Knowledge engineering and semantic web: 6th international conference, KESW 2015, moscow, russia, september 30-october 2, 2015, proceedings 6* (pp. 158–167). Springer.
Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: tutorials* (pp. 15–18).
Samha, A. K., Li, Y., & Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. arXiv preprint arXiv:1404.1982.
Sanders, M. S., & McCormick, E. J. (1998). Human factors in engineering and design. *Industrial Robot: An International Journal*, *25*(2), 153.
Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 86–96).
Sheth, J. N., Jain, V., & Ambika, A. (2023). The growing importance of customer-centric support services for improving customer experience. *Journal of Business Research*, *164*, 113943.
Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, *8*(1), 101.
Shu, L., Xu, H., & Liu, B. (2017). Lifelong learning CRF for supervised aspect extraction. arXiv preprint arXiv:1705.00251.
Siddiqua, A., Bindumathi, V., Raghu, G., & Bhargav, Y. S. V. (2024). Aspect-based sentiment analysis (ABSA) using machine learning algorithms. In *2024 Third international conference on distributed computing and electrical circuits and electronics (ICDCECE)* (pp. 1–6). IEEE.
Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025.
Singh, U., Saraswat, A., Azad, H. K., Abhishek, K., & Shitharth, S. (2022). Towards improving e-commerce customer review analysis for sentiment detection. *Scientific Reports*, *12*(1), 21983.
Sirisha, U., & Bolem, S. C. (2022). Aspect based sentiment & emotion analysis with ROBERTa, LSTM. *International Journal of Advanced Computer Science and Applications*, *13*(11).
Šmíd, J., Přibáň, P., & Král, P. (2024). Llama-based models for aspect-based sentiment analysis. In *Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis* (pp. 63–70).
Strawderman, L., & Koubek, R. (2008). Human factors and usability in service quality measurement. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *18*(4), 454–463.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv preprint arXiv:1903.09588.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Troya, A., Gopalakrishna Pillai, R., Rodriguez Rivero, D. C., Genc, D. Z., Kayal, D. S., & Araci, D. (2021). Aspect-based sentiment analysis of social media data with pre-trained language models. In *Proceedings of the 2021 5th international conference on natural language processing and information retrieval* (pp. 8–17).

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*, 478–514.

Tulkens, S., & Van Cranenburgh, A. (2020). Embarrassingly simple unsupervised aspect extraction. arXiv preprint arXiv:2004.13580.

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G. et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133.

Vanaja, S., & Belwal, M. (2018). Aspect-level sentiment analysis on e-commerce data. In *2018 International conference on inventive research in computing applications (ICIRCA)* (pp. 1275–1279). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. (pp. 1–11). (*vol. 30*). Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). https://doi.org/10.48550/arXiv.1706.03762

Villaneau, J., Pecore, S., & Saïd, F. (2018). Aspect detection in book reviews: Experimentations. In *Nl4ai*.

Wagner, G., Schramm-Klein, H., & Steinmann, S. (2020). Online retailing across e-channels and e-channel touchpoints: Empirical studies of consumer behavior in the multichannel e-commerce environment. *Journal of Business Research*, *107*, 256–270.

Wang, Y.-S. (2008). Assessing e-commerce systems success: A respecification and validation of the delone and mclean model of IS success. *Information Systems Journal*, *18*(5), 529–557.

Wankhade, M., Kulkarni, C., & Rao, A. C. S. (2024). A survey on aspect base sentiment analysis methods and challenges. *Applied Soft Computing*, *167*, 112249.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.

Wu, Z. F., Li, J., Cai, M. Y., Lin, Y., & Zhang, W. J. (2016). On membership of black-box or white-box of artificial neural network models. In *2016 IEEE 11Th conference on industrial electronics and applications (ICIEA)* (pp. 1400–1404). IEEE.

Xia, M., Huang, B., Yan, Y., Zhang, W., & Lin, W. (2024). Transforming patient experience in underserved areas with innovative voice-based healthcare solutions. In *International congress on information and communication technology* (pp. 643–653). Springer.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

Yang, H., & Li, K. (2022). PyABSA: open framework for aspect-based sentiment analysis, *10*. arXiv preprint arXiv:2208.01368.

Yang, H., Zhang, C., & Li, K. (2023). Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 5117–5122).

Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems*, *6*(3), 621–634.

Yom-Tov, G. B., Ashtar, S., Altman, D., Natapov, M., Barkay, N., Westphal, M., & Rafaeli, A. (2018). Customer sentiment in web-based service interactions: automated analyses and new insights. In *Companion proceedings of the web conference 2018* (pp. 1689–1697).

Yu, X., Guo, S., Guo, J., & Huang, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, *38*(3), 1425–1430.

Zalutska, O., Molchanova, M., Sobko, O., Mazurets, O., Pasichnyk, O., Barmak, O., & Krak, I. (2023). Method for sentiment analysis of ukrainian-language reviews in e-commerce using roBERTa neural network. In *Colins (1)* (pp. 344–356).

Zhang, S., Peng, B., Zhao, X., Hu, B., Zhu, Y., Zeng, Y., & Hu, X. (2024a). Llasa: Large language and e-commerce shopping assistant. arXiv preprint arXiv:2408.02006.

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024b). Sentiment analysis in the era of large language models: A reality check. In *Findings of the association for computational linguistics: NAACL 2024* (pp. 3881–3906).

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, *28*.

Zhang, Y., Zhong, M., Ouyang, S., Jiao, Y., Zhou, S., Ding, L., & Han, J. (2024c). Automated mining of structured knowledge from text in the era of large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 6644–6654).

Zhao, Y., Song, J., Montazeri, A., Gupta, M. M., Lin, Y., Wang, C., & Zhang, W. J. (2018). Mining affective words to capture customer's affective response to apparel products. *Textile Research Journal*, *88*(12), 1426–1436.

Zheng, L., Jin, P., Zhao, J., & Yue, L. (2014). Multi-dimensional sentiment analysis for large-scale e-commerce reviews. In *Database and expert systems applications: 25th international conference, DEXA 2014, munich, germany, september 1–4, 2014. proceedings, part II 25* (pp. 449–463). Springer.

Zhou, J., Du, W., Rokon, M. O. F., Wang, Z., Xu, J., Shah, I., Lee, K.-c., & Wen, M. (2023). Enhanced e-commerce attribute extraction: Innovating with decorative relation correction and LLAMA 2.0-based annotation. arXiv preprint arXiv:2312.06684.

Zineb, E. L. F., Najat, R., & Jaafar, A. (2021). An intelligent approach for data analysis and decision making in big data: A case study on e-commerce industry. *International Journal of Advanced Computer Science and Applications*, *12*(7).