IJSES
Volume 9, Issue 11, pp. 72-78, 2025.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Leveraging Transformer-Based Architectures for Multi-Label Complaint Detection and Sentiment Analysis in Hotel Feedback Systems

Milena Nikolić, Nevena Minić, Marina Marjanović
The Academy of Applied Technical and Preschool Studies, Singidunum University
Email address: milena.nikolic@akademijanis.edu.rs, nevena.minic@akademijanis.edu.rs, mmarjanovic@singidunum.ac.rs

Abstract— *In the competitive hotel industry, understanding guest feedback is essential for improving service quality. This paper proposes a comprehensive intelligent system that applies transformer-based architectures to perform sentiment analysis and multi-label classification of consumer complaints in hotel review data. Using BERT and RoBERTa, we fine-tune models for two tasks: classifying sentiment (positive, neutral, negative) and detecting multiple complaint categories such as cleanliness, staff behavior, food quality, and booking issues. A shared encoder supports both tasks, trained on an extended TripAdvisor dataset with additional evaluation on the Consumer Complaint Database from Kaggle. Comparative experiments show that the fine-tuned RoBERTa model outperforms BERT, achieving an F1 score above 0.928 for sentiment classification and 87% mean precision across complaint categories. The system also integrates SHAP and LIME to provide insigh ts into model decisions. It enables transparent review analysis, helping hotel managers detect issues promptly and enhance service quality. This solution supports scalable and interpretable feedback analytics for smart hospitality systems.*

Keywords— *Intelligent Systems, Hotel Reviews, Sentiment Analysis, Classification, NLP, Deep Learning, Complaint Detection.*

I. Introduction

The hospitality industry has become increasingly reliant on digital platforms where customers publicly share their experiences, expectations, and complaints. Online reviews, particularly those on popular platforms like TripAdvisor and Booking.com, significantly influence travelers' decisions and provide valuable feedback for hoteliers aiming to improve their services [1].

As the volume of reviews grows exponentially, manual analysis becomes impractical, especially when hotels and private accommodations (like ones listed on Airbnb) receive thousands of reviews across multiple platforms [2].

However, not all reviews are authentic or relevant. Some entries may be fraudulent, generated by bots, or posted with malicious or promotional intent. Others might contain irrelevant content, duplicate entries, or spam, which can mislead analysis and reduce the reliability of automated systems. In our earlier work [3]-[4], we performed in-depth analysis and developed methods to detect and filter such reviews, focusing on patterns in reviewer behavior, text repetition, posting frequency, and sentiment inconsistency.

Traditionally, service providers have relied on rule-based systems, simple keyword detection, or binary sentiment classification (positive/negative text) to monitor customer satisfaction. While useful in limited scenarios, approaches like that tend to oversimplify guest narratives and overlook important subtleties. For instance, a traveler may express dissatisfaction with the cleanliness of a room but praise the behavior of the staff within the same review. Treating this feedback as uniformly "neutral" or "mixed" can easily mask important information. Moreover, existing methods often fail to identify the specific dimensions of service that need attention, like booking processes or room amenities. [5].

Recent advances in natural language processing (NLP), with the introduction of transformer-based architectures, have opened new opportunities for discovery of complex relationships and trends in textual data. Models like BERT (Bidirectional Encoder Representations from Transformers) and a closely related RoBERTa (A Robustly Optimized BERT Pretraining Approach) have shown strong performance in a wide range of NLP tasks due to their ability to capture deep contextual relationships in text. These models have been impactful in domains with user-generated content, where understanding sentiment polarity and complaint categories is crucial for business intelligence [6].

By combining extensive datasets from Kaggle, including hotel reviews and customer complaints, this study bridges the gap between domain-specific hospitality feedback and broader consumer satisfaction data. The proposed system leverages transformer-based models, comparing BERT and RoBERTa, to automate tasks of sentiment classification and multi-label complaint detection, enabling hotels to extract critical insights from large volumes of unstructured reviews.

Beyond classifying guest feedback as positive, neutral, or negative, the system identifies complaint categories such as cleanliness, staff behavior, booking issues, food quality, and more. To ensure transparency and build trust in automated decision-making process, we integrate explainable AI tools, like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), which reveal the key textual elements behind each model prediction [7].

Overall, this RoBERTa-based framework is designed for real-time integration into hospitality systems, providing an efficient scalable solution for review analysis that supports timely service improvements and strategic decision-making.

II. Literature Review

Transformer-based models have emerged as a powerful solution for extracting information from unstructured hotel

reviews, enabling more granular and accurate feedback analysis than previous techniques. Among these models, BERT (with its variants) has shown consistent performance across various languages, domains, and tasks. For instance, a BERT-based model applied to German TripAdvisor reviews achieved micro F1 scores of 0.91 for aspect categorization and 0.81 for sentiment analysis, demonstrating capabilities in both complaint identification and polarity detection [8].

In another study, a hierarchical BERT model applied to Vietnamese hotel and restaurant dataset achieved F1 micro scores of 82.06% for entity/aspect detection and 74.69% for sentiment polarity [9]. This suggests that hierarchical context modeling can improve classification performance, particularly in languages with complex sentence structures. Meanwhile, a comparative study involving fine-tuned BERT, RoBERTa, DeBERTa, and GPT-2 models found RoBERTa that performs best for classification tasks, while large language models including GPT-4 offered superior performance in capturing nuanced sentiment, but at a significantly higher computational cost [10].

More advanced strategies, such as multi-task learning, have demonstrated potential too. For example, a multi-task AraBERT model applied to Arabic hotel reviews achieved F1 scores of 80.32% for aspect term extraction and accuracy around 89% for sentiment polarity tasks [11]. These results suggest that shared representation learning across related tasks can lead to improved accuracy and model efficiency.

Several studies have incorporated explainable AI (XAI) techniques to clarify the "black box" nature of transformer models. Tools such as SHAP and LIME have been integrated into review analysis pipeline to highlight which words or phrases most influenced a model's prediction, making the outputs more transparent for business stakeholders [12]. Additionally, ensemble techniques that combine BERT with models like Random Forest have been utilized to enhance hotel recommendation systems by analyzing sentiment and aspect categories simultaneously.

Language and domain-specific adaptations have proven essential for improving model performance. For instance, fine-tuned Multilingual BERT enhanced Indonesian hotel review analysis by 8% in F1 score, although challenges such as vocabulary mismatch and the need for language-specific pretraining remain significant [13]. Similarly, studies using Arabic datasets illustrated the importance of handling class imbalance and training on large and diverse corpora [14].

Topic modeling approaches like BERTopic have also been used to uncover latent themes among reviews, combining transformer embeddings with clustering techniques to spot drivers of satisfaction and dissatisfaction [15]. While such methods are less precise in aspect-level classification, they offer additional perspectives that aid in interpretability and managerial decision-making.

In summary, this literature review shows strong evidence for the effectiveness of transformer-based architectures in performing both review sentiment analysis and complaint categorization within hotel systems. Fine-tuned RoBERTa models offer a compelling balance of accuracy, efficiency, and generalizability. However, critical challenges persist in handling class imbalance, computational costs, and severe linguistic variations across languages. Although previous studies have investigated sentiment analysis and aspect categorization separately, and typically in monolingual or limited-scale contexts, there remains a gap in developing integrated intelligent systems that can perform automated multi-label analysis across diverse datasets.

To overcome this limitation, our study proposes a unified framework based on RoBERTa for simultaneous sentiment analysis and identification of complaint categories in hotel reviews. Unlike prior work, we focus on both classification accuracy and explainability by incorporating SHAP and LIME methods. We extend datasets with manual annotations, benchmark performance against BERT, and demonstrate how this system can be seamlessly integrated in real-time service monitoring in a hospitality operations environment. This work contributes towards ongoing efforts to develop transparent AI tools for the hotel industry and supports the broader adoption of smart feedback systems.

TABLE I. A brief overview of previous findings regarding review analysis.

Study	Model(s) Used	Tasks Addressed	Dataset / Language	Performance Highlights
Fehle et al. [8]	BERT (variant not specified)	Multi-label Aspect Categor., Sentiment Analysis	TripAdvisor / German	F1: 0.91 (aspect), 0.81 (end-to-end sentiment)
Tran & Bui [9]	Hierarchical BERT	Entity and Aspect Detection, Sentiment Polarity	Vietnamese hotel/restaurant reviews	F1: 82.06% (aspect), 74.69% (sentiment)
Botunac et al. [10]	BERT RoBERTa DeBERTa GPT-2, GPT-4	Aspect Categor., Sentiment Classif.	Hospitality reviews / Language unspecified	RoBERTa best for classification; GPT-4 best for nuance
Fadel et al. [11]	Multi-task AraBERT + BiLSTM/ BiGRU	Aspect Term Categor., Polarity Classif.	SemEval Arabic reviews	Acc: ~89% (polarity), F1: 80.32% (term extraction)
Ray et al. [12]	Ensemble BERT + Random Forest	Sentiment Polarity, Aspect Categor.	TripAdvisor / Language unspecified	Acc: 92.36%, Macro F1: 84%
Azhar & Khodra [13]	Multilingual BERT	Aspect Categor., Sentiment Polarity	Airyrooms / Indonesian	+8% F1 over prior work
Kumar [15]	BERTopic (BERT-based topic modeling)	Topic Modeling + Sentiment Analysis	TripAdvisor / English	Coherence scores only

III. METHODOLOGY

This study introduces a multi-task learning framework that combines both sentiment classification and multi-label complaint detection for efficient hotel review analysis using

transformer-based architectures. The goal is to develop a system capable of extracting nuanced feedback from guest narratives and aligning it with realistic complaint patterns, supported by findings from broader consumer feedback records. Our proposed approach involves two distinct but complementary datasets: The TripAdvisor Hotel Reviews dataset [16], containing customer opinions on hospitality services, followed by The Consumer Complaint Database [17], which captures consumer concerns across industries.

By jointly analyzing domain-specific (hotel reviews) and domain-general (consumer complaints) text data, we aim to enhance the generalization and robustness of the model in recognizing both overt and latent dissatisfaction signals. The following subsections describe the complete processing pipeline, from raw data to interpretable model predictions.

Data Preprocessing

The first dataset used is publicly available on Kaggle and contains more than 20,000 user-submitted hotel reviews. Every review contains a title, a text field, and a rating value ranging from 1 to 5. The preprocessing component included lowercasing, removal of special characters and HTML tags, then tokenization using the WordPiece algorithm to ensure compatibility with transformer inputs, and truncation to 512 tokens. WordPiece proves effective for handling rare or unknown words by breaking them into a group of smaller subword units, which improves vocabulary coverage and model generalization [18].

Additionally, to enhance hotel data reliability, we utilized various filtering techniques to remove inconsistent reviews, particularly those where the textual sentiment did not align with the assigned rating value. This inconsistency detection was developed in our previously mentioned research and resulted in more reliable inputs for the training process.

To support sentiment classification, labels were derived directly from the numerical ratings, following these rules: ratings of 4.5 and above were marked as positive, ratings between 3.0 and 4.4 as neutral, and those below 3.0 as negative. For the complaint classification task, we manually annotated a subset of reviews with one or more complaint categories, based on a curated keyword list and iterative inspection. To streamline this process, we included Python utilities and custom rule-based scripts, which accelerated the process of identification and tagging of approximately 3,000 annotated reviews. Each review was encoded using a binary vector indicating the presence or absence of already mentioned complaint types. Room cleanliness was given special attention, as it is recognized as a major contributor to guest dissatisfaction and one of the strongest predictors of negative sentiment. Considering cleanliness as a core expectation in hotel services, its absence is strongly linked to lower ratings and dissatisfaction in guest experiences.

The second dataset is maintained by the U.S. Consumer Financial Protection Bureau (CFPB). This collection contains over 1 million free-text complaint narratives submitted by consumers across industries such as banking, loans, credit reporting, and occasionally, hospitality. While not curated to hotel services specifically, the given database offers a rich source of language patterns expressing dissatisfaction, urgency, and resolution-seeking behavior, all of which are highly relevant to our objectives.

Each entry involves metadata like date received, product category, issue type, and the detailed consumer complaint narrative. For purposes of this research, only the narrative field was retained and subjected to text preprocessing, including cleaning, tokenization, and truncation. As explicit sentiment ratings were not provided here, sentiment labels were inferred through a hybrid strategy. This step involved lexicon-based scoring, also keyword analysis (e.g., detecting terms like “still unresolved” or “satisfied”), and semantic similarity comparisons with known sentiment-labeled hotel reviews utilizing cosine similarity on sentence embeddings. Although complaint categories were not explicitly labeled in this dataset, we investigated patterns in issue descriptions, tone severity, and resolution requests to detect recurring patterns. These findings were used to enrich the complaint classification pipeline of our model by aligning embeddings and applying contrastive sampling techniques.

Feature Engineering

To support the proposed multi-task model, we extracted and aligned sentiment and complaint-related features from both datasets. First, TripAdvisor reviews were labeled using rating-based sentiment and manually annotated complaint categories via keyword matching, name entity recognition (NER), and manual inspection.

For the Consumer Complaint Database, sentiment labels were derived through a lexicon-semantic hybrid approach. Furthermore, in the absence of structured labels, recurring complaint patterns were identified from issue descriptions and tone indicators.

These features ensured consistent labeling and improved generalization across domains. A summary of techniques is provided in Table II.

TABLE II. Comparison of Feature Engineering Strategies.

Aspect	TripAdvisor Hotel Reviews	Consumer Complaint Database
Source	User-submitted hotel reviews (20k+ entries)	Public complaint narratives (1M+ entries)
Sentiment Labeling	Derived directly from star ratings (1–5 scale)	Inferred using rule-based classifier and semantic similarity
Complaint Labeling	Manually annotated using keywords, NER, and inspection	No explicit labels; patterns extracted via tone and structure
Special Techniques	Inconsistency filtering, binary vector encoding	Cosine similarity, lexicon scoring, contrastive sampling
Contribution to Model	Provides domain-specific sentiment and complaint signals	Enhances model generalization via cross-domain patterns

Model Training and Evaluation

After thoughtfully evaluating potential model candidates, RoBERTa-base was selected as the main infrastructure for our approach, due to its language modeling capabilities and proven robustness across text classification tasks. RoBERTa model

builds upon the foundation established by BERT, but incorporates key improvements including dynamic masking, notably longer pretraining sequences, and training over a larger and more diverse corpus. These factors contribute to stronger contextual representation, particularly useful for capturing subtle expressions of sentiment and customer dissatisfaction found in real-world review data [19]-[20].

The model was adapted to jointly predict sentiment (as a three-class classification task with positive, neutral, and negative classes) and complaint categories (as a multi-label classification task, where each review may belong to one or more complaint types). A shared transformer encoder was used to extract general-purpose features from each input sequence, while two task-specific heads were included: a softmax classifier for sentiment prediction followed by an output layer for complaint detection. This shared structure enabled the model to uncover relationships and linguistic patterns more efficiently than training each task separately.

The training process was performed using the extensible HuggingFace Transformers library with a PyTorch backend. The model was fine-tuned on the combined dataset, using preprocessed and aligned inputs from both TripAdvisor and the Consumer Complaint Database. As already mentioned, inputs were tokenized, padded or truncated to a maximum sequence length of 512 tokens, and then fed into the model along with corresponding sentiment and complaint labels. To handle the dual-objective setting, we used a composite loss function: categorical cross-entropy for the sentiment classification task followed by binary cross-entropy for the complaint detection task. These two losses were weighted equally during optimization stage, after initial experiments showed that balanced configuration yielded stable training dynamics without biasing toward either task [21].

Model training was conducted using the widely adopted AdamW optimizer with a linear learning rate schedule and warm-up steps. We trained the model for up to five epochs, using early stopping based on validation loss to prevent overfitting. A batch size of 16 and dropout regularization (p=0.1) were used to maintain computational feasibility and support generalization. Moreover, five-fold cross-validation was employed to ensure that the model's performance was not dependent on any data split, and to assess its stability across both specialized and general-purpose text sources.

To further validate model stability, we also experimented with an extended configuration of 10 epochs. Although this structure enabled longer training cycles, performance gains plateaued and signs of overfitting emerged, especially in complaint detection. Even with adjusted learning rates and regularization settings, the original 5-epoch configuration with early stopping resulted in better generalization and training efficiency. Consequently, the initial configuration was retained as optimal [22]-[23].

To benchmark our approach, we also trained a parallel version of the architecture utilizing the original BERT-base model. Although BERT performed reasonably well, RoBERTa consistently outperformed it across all relevant evaluation metrics, particularly in the complaint detection tasks. These improvements are provided by RoBERTa's dynamic masking and extensive pretraining, which can be observed on Figure 1. This model captured more nuanced relationships that are critical for identifying implicit dissatisfaction cues.

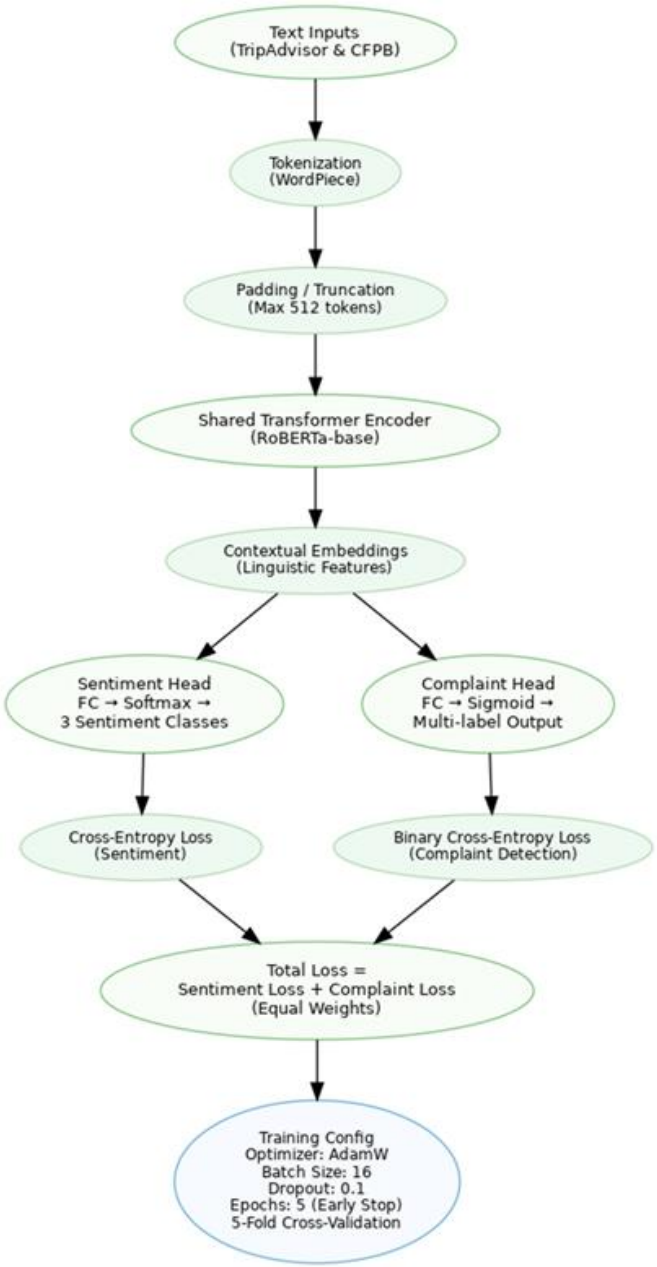

Figure 1. Multi-Task Transformer Architecture Overview, (FC stands for fully connected)

Overall, the training methodology was carefully designed to maximize the synergy between tasks while leveraging the strength of transformer architectures. The combination of concepts including multi-task learning, dual-domain data sources, and a high-capacity pretrained model like RoBERTa led to the development of a scalable intelligent system that effectively interprets complex guest reviews and converts them into actionable outputs, well-suited for analytical and operational use within hotel systems.

International Journal of Scientific Engineering and Science
Volume 9, Issue 11, pp. 72-78, 2025. ISSN (Online): 2456-7361

TABLE III. Token Importance Analysis using SHAP/LIME.

Token	Importance Score (SHAP or LIME)	Impact
"no housekeeping"	+0.52	Complaint: Service
"very clean room"	+0.41	Sentiment: Positive
"they ignored us"	+0.48	Complaint: Staff behavior
"bathroom smelled"	+0.36	Complaint: Room cleanliness
"excellent location"	+0.29	Sentiment: Positive

Explainability and Interpretability

While transformer-based models such as RoBERTa show good performance in text classification tasks, their complex internal mechanisms can make them difficult to interpret. To address this potential challenge, we also incorporated explainability techniques to gain deeper insights into the model's decision-making process and to validate that the outputs align with human reasoning.

For sentiment classification, we used SHAP technique to identify which tokens contributed most significantly to the predicted sentiment label. By computing attribution scores on token-level, SHAP provided transparent justifications for positive and negative classifications, showing emotionally charged or contextually significant words within reviews. This proved valuable in neutral cases, where subtle shifts in phrasing influenced class boundaries.

Moving forward, in the complaint detection task, which involved multi-label outputs, we used LIME in combination with attention weight visualization to investigate how the model associated certain phrases with specific complaint categories. For instance, phrases like *"no housekeeping for three days"* or *"unresolved billing issues"* were consistently highlighted when predicting labels related to service quality or hotel room cleanliness. Attention maps confirmed that the model was putting focus on contextually relevant spans rather than generic or frequent tokens.

To illustrate this behavior more clearly, Table III presents several examples of influential tokens from actual reviews, along with their computed importance score values and associated impact on model predictions. The samples show how the model assigns higher weights to complaint triggers (*"they ignored us"*) or sentiment-based phrases (*"excellent location"*), thereby supporting its classification decisions.

These interpretability methods increased transparency and helped validate the internal consistency of the model, especially across dual domains. For instance, in domain adaptation from Consumer Financial Protection Bureau (CFPB) narratives to hotel reviews, we noticed that learned patterns remained aligned with general user dissatisfaction expressions (e.g., *"not resolved after multiple attempts"*).

Finally, the explainability framework also played a crucial role in error analysis. By examining misclassified instances, we can distinguish between genuine model confusion (e.g., sarcastic sentiment) and annotation noise, informing future directions for data curation and model refinement.

IV. Experimental Results

To assess the effectiveness of the multi-task framework, we conducted comparative experiments using BERT-base and RoBERTa-base models as shared encoders. Performance was evaluated on sentiment classification and multi-label complaint detection. All models were fine-tuned on the TripAdvisor dataset, with additional validation on selected entries from the Consumer Complaint Database to test domain adaptability.

For sentiment classification, RoBERTa achieved superior results across all evaluation metrics. Specifically, it reached an average F1 score of 0.928, compared to 0.892 for BERT. Precision and recall scores were also improved, particularly in the neutral class, where subtle semantic distinctions are difficult to capture. Accuracy favored RoBERTa approach, confirming that its improved contextual embeddings contributed to more consistent predictions across sentiment boundaries.

In the complaint detection task, RoBERTa outperformed BERT again. It achieved a mean precision of 87%, and a good macro F1 score of 0.861, compared to BERT's 0.824. RoBERTa consistently demonstrated higher true positive rates, especially for complaints involving room conditions.

Figure 2 represents the Precision-Recall (PR) curves for each complaint category across both models. Each line corresponds to a specific label (e.g., "Cleanliness," "Booking Issues," etc.), with RoBERTa results shown in solid lines and BERT in dashed lines. As observed in the chart, RoBERTa consistently maintains higher precision at comparable recall levels, particularly for categories such as *Staff Behavior* and *Room Amenities*. The curves confirm that RoBERTa is better at distinguishing key complaint signals without sacrificing precision, especially in imbalanced or noisy class conditions.

Notably, the AUC-PR spread is moderate, with no curve reaching perfect values, underscoring the realistic difficulty of this task. The overlapping but distinguishable curves demonstrate that while RoBERTa generally outperforms BERT, it doesn't dominate exactly across all labels. In fact, performance drops are more visible in critical categories, such as *Booking Issues*, where both models struggle due to ambiguity or lower label frequency.

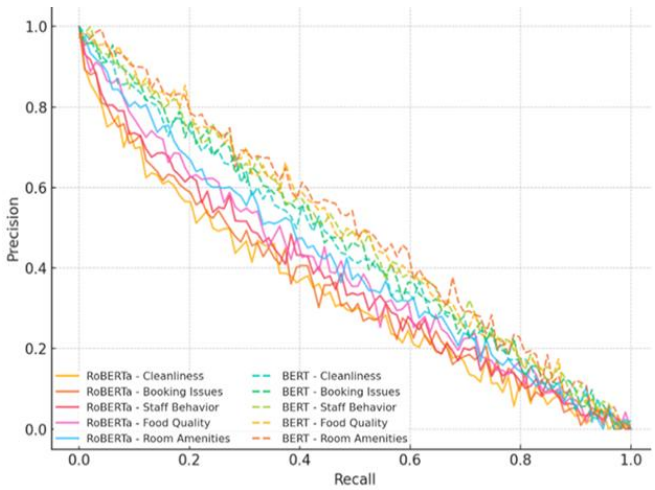

Figure 2. Precision-Recall Curves per Complaint Category.

Likewise, we observe that the performance gap between models varies by label. For example, in simpler categories like *Cleanliness*, RoBERTa's advantage is more pronounced, but in categories like *Booking Issues*, the curves converge, indicating that further improvement may require enhanced feature engineering or more balanced data. Additionally, the downward slope in all given curves reflects the typical precision-recall tradeoff: as recall increases, precision tends to drop due to rising false positives.

Overall, categories with high AUC-PR, such as *Cleanliness* (above 0.85 for RoBERTa), offer greater trust in model outputs and require minimal post-processing. In contrast, labels with lower AUCs (e.g., less than 0.7) might benefit from future refinements. These findings validate the use of RoBERTa in our framework and open promising avenues for expanding feedback analysis in hospitality and beyond.

V. Conclusion

This paper presents an intelligent, multi-task framework developed on RoBERTa model to simultaneously perform sentiment classification and multi-label complaint detection in hotel review systems. By unifying two critical objectives into a shared transformer-based architecture, the proposed system captures both the emotional tone and the specific issues embedded in diverse guest narratives. Comparative experiments show that RoBERTa outperforms BERT across key evaluation metrics, achieving an F1 score of 0.928 for sentiment and a notable 87% mean precision for complaint detection. The given results validate RoBERTa's strength in contextual understanding, especially in situations involving overlapping or subtle complaint categories.

Alongside model performance, this work also has strong practical and commercial implications for the hospitality industry. Hotels and service providers increasingly depend on rapid and accurate interpretation of guest feedback to make operational and strategic decisions. By automatically identifying issues such as cleanliness, staff behavior, or booking complications, paired with interpretable sentiment insights, our system empowers hotel managers to prioritize interventions, improve service quality, and enhance guest satisfaction and retention. The inclusion of SHAP and LIME approaches ensure model transparency, which is essential for stakeholder trust and accountability.

Furthermore, this architecture is designed for real-world scalability. It supports easy integration into CRM, real-time feedback dashboards, and customer satisfaction monitoring tools. The adaptability to diverse datasets, involving the domain-general complaints from the CFPB, demonstrates the potential to extend the solution beyond hospitality to services like transportation, healthcare, and finance.

Looking ahead, future research will focus on several enhancements, including advanced transformer variants like DeBERTa and Longformer, which may further enhance classification accuracy, particularly for longer and complex reviews. We aim to address class imbalance and improve multilingual support too, allowing the system to operate across diverse user bases and international markets.

Finally, expanding the system with global interpretability methods and incorporating temporal analysis of complaints (e.g., seasonal patterns, spikes after policy changes) might provide more useful business insights. As the demand for transparent, actionable AI continues to grow, this system offers a solid baseline for future development in intelligent feedback management and customer experience analytics.

References

[1] E. Alotaibi, "Application of machine learning in the hotel industry: a critical review," J. Assoc. Arab Univ. Tour. Hosp., vol. 18, no. 3, pp. 78–96, 2020.

[2] M. O. Parvez, "Use of machine learning technology for tourist and organizational services: high-tech innovation in the hospitality industry," J. Tour. Futures, vol. 7, no. 2, pp. 240–244, 2021.

[3] M. Nikolić, M. Stojanović, and M. Marjanović, "Anomaly detection in hotel reviews: Applying data science for enhanced review integrity," in Proc. 2024 32nd Telecommun. Forum (TELFOR), Nov. 2024, pp. 1–4.

[4] M. Nikolić, M. Stojanović, and M. Marjanović, "Integrating data science and predictive modeling for detecting inconsistent hotel reviews," in *UNITECH 2024-Selected Papers*, Technical University of Gabrovo, 2024, pp. 104–110.

[5] C. G. Harris, "Decomposing TripAdvisor: Detecting potentially fraudulent hotel reviews in the era of big data," in *Proc. 2018 IEEE Int. Conf. Big Knowl. (ICBK)*, Singapore, 2018, pp. 243–251, doi: 10.1109/ICBK.2018.00040.

[6] Y. G. Pramudya and A. Alamsyah, "Hotel reviews classification and review-based recommendation model construction using BERT and RoBERTa," in *Proc. 2023 6th Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Nov. 2023, pp. 437–442.

[7] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, p. 2400304, 2025.

[8] J. Fehle, L. Münster, T. Schmidt, and C. Wolff, "Aspect-based sentiment analysis as a multi-label classification task on the domain of German hotel reviews," in *Proc. Conf. Natural Language Processing (KONVENS)*, 2023.

[9] O. T. K. Tran and V. T. Bui, "A BERT-based hierarchical model for Vietnamese aspect-based sentiment analysis," in *Proc. Int. Conf. Knowledge and Systems Engineering (KSE)*, 2020.

[10] I. Botunac, M. Brkić Bakarić, and M. Matetić, "Comparing fine-tuning and prompt engineering for multi-class classification in hospitality review analysis," *Applied Sciences*, vol. 14, no. 14, p. 6254, 2024.

[11] A. S. Fadel, M. Saleh, R. Salama, and O. Abulnaja, "MTL-AraBERT: An enhanced multi-task learning model for Arabic aspect-based sentiment analysis," *De Computis*, 2024.

[12] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Applied Soft Computing*, vol. 94, p. 106452, 2020.

[13] A. N. Azhar and M. L. Khodra, "Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis," in *Proc. 2020 7th Conf. Adv. Informatics: Concepts, Theory and Appl. (ICAICTA)*, 2020, pp. 1–6.

[14] A. Ameur, S. Hamdi, and S. Yahia, "Multi-label learning for aspect category detection of Arabic hotel reviews using AraBERT," in *Proc. Conf. A Artif. Intell.*, 2023, pp. 123–130.

[15] N. Kumar, "Unpacking the emotional landscape of reviews: Sentiment-augmented topic modeling with transformer embeddings," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 4, pp. 45–51, 2025.

[16] A. Mvd, "TripAdvisor Hotel Reviews," *Kaggle Datasets*, 2019. [Online] Available: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews. [Accessed: July 3, 2025].

[17] Selener, "Consumer Complaint Database," *Kaggle Datasets*, 2021. [Online]. Available: https://www.kaggle.com/datasets/selener/consumer-complaint-database. [Accessed: July 3, 2025].

[18] N. Culmer, "A Comparison of Lexical Tokenization Methods," *2024*.

[19] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, 2021.

[20] S. Chen, Y. Zhang, and Q. Yang, "Multi-task learning in natural language processing: An overview," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–32, 2024.

[21] U. R. Pol, P. S. Vadar, and T. T. Moharekar, “Hugging Face: Revolutionizing AI and NLP,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 8, pp. 1121–1124, 2024.
[22] R. Zaheer and H. Shaziya, "A study of the optimization algorithms in deep learning," in *Proc. 2019 3rd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2019, pp. 536–539.
[23] L. Bai, A. Gupta, and Y. S. Ong, "Multi-task learning with multi-task optimization," *arXiv preprint arXiv:2403.16162*, 2024.