*Journal of Science, Innovation & Social Impact*
***EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)***

*Article*

# A Comparative Study of Machine Learning Methods for Automated Customer Service Dialogue Quality Assessment

**Yajing Zhang [1,*]**

[1] UCD Smurfit Graduate Business School, University College Dublin, Dublin, Ireland
* Correspondence: Yajing Zhang, UCD Smurfit Graduate Business School, University College Dublin, Dublin, Ireland

**Abstract:** The rapid expansion of digital customer service channels has created an urgent need for automated quality assessment methods capable of evaluating dialogue interactions at scale. This paper presents a comprehensive comparative study of machine learning approaches for automated assessment of customer service dialogue quality, examining traditional machine learning algorithms, deep learning architectures, and transformer-based models. A multi-dimensional quality assessment framework is proposed, incorporating three primary evaluation categories: information accuracy, communication appropriateness, and process compliance. An experimental evaluation on a customer service dialogue dataset demonstrates that BERT-based models achieve superior overall classification accuracy (94.2%), while traditional methods offer advantages in computational efficiency and interpretability. The analysis reveals significant performance differences across service defect categories, with transformer models excelling at detecting subtle compliance violations and attitude-related issues. These findings provide practical guidance for enterprises seeking to implement standardized quality-monitoring systems that align with consumer protection regulations.

**Keywords:** Customer Service Quality Assessment; Natural Language Processing; Text Classification; Machine Learning Comparison

## 1. Introduction

### *1.1. Background and Motivation*

The transformation of customer service operations through digital channels has fundamentally altered how organizations interact with consumers. Contact centers now process millions of text-based interactions daily through chatbots, email systems, and social media platforms. Maintaining consistent service quality across these channels directly impacts customer satisfaction, brand reputation, and regulatory compliance. Traditional quality assurance approaches that rely on manual review by human supervisors can evaluate only 1-3% of total interactions, leaving the vast majority of customer communications unmonitored [1].

Artificial intelligence and natural language processing technologies have emerged as promising solutions to address this quality monitoring gap. Machine learning algorithms can analyze complete conversation transcripts to identify service defects, compliance violations, and opportunities for improvement. The application of these technologies enables organizations to achieve comprehensive quality coverage while reducing operational costs associated with manual review processes [2].

### *1.2. Problem Statement and Research Gap*

Despite growing interest in automated quality assessment, significant challenges remain in selecting appropriate machine learning methods for specific evaluation tasks.

Received: 01 February 2026
Revised: 11 March 2026
Accepted: 25 March 2026
Published: 31 March 2026

Customer service dialogues present unique characteristics that distinguish them from general text classification problems. These conversations contain domain-specific terminology, implicit sentiment, and contextual dependencies that require sophisticated language understanding [3].

The research community has produced numerous studies examining individual algorithms for sentiment analysis and text classification tasks. A systematic examination comparing multiple algorithmic approaches specifically for service quality assessment remains insufficient [4]. Organizations implementing automated quality monitoring systems lack empirical guidance on selecting methods that align with their specific evaluation requirements and operational constraints.

### *1.3. Research Objectives and Contributions*

This study addresses the identified research gap by conducting a comprehensive comparative analysis of machine learning methods for assessing customer service dialogue quality. The primary objectives include establishing a multi-dimensional quality evaluation framework aligned with industry standards and regulatory requirements, implementing and evaluating representative algorithms from traditional machine learning, deep learning, and transformer-based categories, and analyzing performance characteristics across different service defect types.

The contributions of this research extend to both academic and practical domains. The proposed quality assessment framework provides a structured approach for operationalizing service quality metrics. The comparative experimental results provide empirical evidence to guide method selection. The analysis of interpretability characteristics addresses the growing emphasis on explainable AI in customer-facing applications.

## 2. Related Work

### *2.1. Natural Language Processing in Customer Service*

Natural language processing applications in customer service environments have evolved substantially over the past decade. Early implementations focused on keyword matching and rule-based routing of customer inquiries to appropriate service agents [5]. The development of statistical machine learning methods enabled more sophisticated intent recognition and entity extraction capabilities. Recent advances in neural network architectures have further expanded the scope of automated language understanding in service contexts.

Chatbot systems are among the most visible applications of NLP in customer service. These conversational agents utilize intent classification, slot filling, and dialogue management components to conduct automated interactions with customers [6]. The integration of sentiment analysis capabilities allows chatbots to detect customer frustration and escalate conversations to human agents when appropriate. Research has demonstrated that effective sentiment detection can improve customer satisfaction scores by enabling timely intervention in problematic interactions.

### *2.2. Service Quality Assessment Methods*

Service quality assessment has traditionally relied on frameworks developed through customer satisfaction research. The SERVQUAL model and its derivatives provide structured approaches for measuring service quality across dimensions, including reliability, responsiveness, assurance, empathy, and tangibles. Adapting these conceptual frameworks to automated evaluation requires operationalizing abstract quality dimensions into measurable textual features [7].

Contact center operations have developed specific key performance indicators for monitoring agent performance. Metrics, including average handle time, first contact resolution rate, customer satisfaction scores, and compliance adherence rates, provide quantitative measures of service quality. Automated speech and text analytics platforms have been deployed to extract these metrics from recorded interactions [8]. The

application of machine learning algorithms enables more nuanced quality assessment beyond simple metric calculation.

*2.3. Machine Learning for Text Classification*

Text classification forms the technical foundation for automated quality assessment systems. Traditional machine learning approaches, including Support Vector Machines, Naive Bayes classifiers, and Random Forest algorithms, have demonstrated effectiveness across various text classification tasks. These methods rely on hand-crafted features such as bag-of-words representations, TF-IDF weightings, and n-gram statistics [9].

Deep learning architectures have achieved state-of-the-art performance on numerous NLP benchmarks. Recurrent neural networks, particularly Long Short-Term Memory networks and their bidirectional variants, capture sequential dependencies in text data [10]. The introduction of attention mechanisms and transformer architectures has further advanced language understanding capabilities. Pre-trained language models, including BERT and its variants, leverage transfer learning to achieve strong performance with limited task-specific training data.

## 3. Methodology

*3.1. Quality Assessment Framework*

The quality assessment framework developed for this study incorporates multiple evaluation dimensions derived from industry standards and regulatory requirements. Consumer protection principles and common contact-center quality assurance policies emphasize the importance of accurate information provision, transparent communication, and respectful customer treatment. These regulatory principles inform the selection of quality dimensions for automated assessment.

The framework defines three primary quality categories for evaluation. The first category addresses information accuracy, examining whether agent responses contain factual information about products, services, policies, and procedures. Errors in this category include providing incorrect pricing information, misrepresenting product features, or conveying inaccurate policy details. The second category encompasses communication appropriateness, evaluating the tone, politeness, and professionalism of agent responses. Defects in this dimension include dismissive language, inappropriate informality, or failure to acknowledge customer concerns. The third category focuses on process compliance, assessing adherence to required procedures, including identity verification, disclosure statements, and escalation protocols [11].

Table 1 presents the quality dimension taxonomy with associated defect types and example indicators.

**Table 1.** Quality Assessment Dimension Taxonomy.

| Quality Dimension | Defect Category | Example Indicators |
|---|---|---|
| Information Accuracy | Factual Error | Incorrect pricing, wrong policy details |
| Information Accuracy | Incomplete Information | Missing disclosure, partial response |
| Communication | Tone Violation | Dismissive language, unprofessional remarks |
| Communication | Empathy Deficit | Failure to acknowledge concerns |
| Process Compliance | Verification Failure | Skipped identity confirmation |
| Process Compliance | Disclosure Omission | Missing required statements |

The quantification of quality metrics requires mapping model outputs to numerical scores. Each quality dimension receives a score in [0,1] derived from predicted class probabilities. Information accuracy is computed as

$$1 - P(\text{Information Error})$$

. Communication appropriateness scores are derived from sentiment analysis of agent utterances, combined with the detection of specific courtesy phrases and expressions. Process compliance is computed as

$$1 - P(\text{Compliance Violation})$$

The mathematical formulation for the overall quality score Q combines individual dimension scores through weighted aggregation:

Q = w1 * S_accuracy + w2 * S_communication + w3 * S_compliance

The weights w1, w2, and w3 are configurable parameters allowing organizations to emphasize specific quality dimensions based on their operational priorities. Default weights of 0.4, 0.35, and 0.25 are applied, reflecting the relative importance typically assigned to these dimensions in contact center operations.

*3.2. Dataset and Preprocessing*

The experimental evaluation uses a customer service dialogue dataset containing 45,000 conversation transcripts collected across multiple service channels, including live chat, email, and social media. Each conversation has been annotated by trained quality analysts according to the defined quality taxonomy. The annotation process involved two independent reviewers, with disagreements resolved through adjudication by a senior analyst. Inter-annotator agreement measured by Cohen's kappa coefficient reached 0.82, indicating substantial agreement [12].

Table 2 summarizes the dataset characteristics, including distribution across quality categories.

**Table 2.** Dataset Statistics and Distribution.

| Characteristic | Value |
|---|---|
| Total Conversations | 45,000 |
| Average Turns per Conversation | 8.4 |
| Average Words per Turn | 42.3 |
| Vocabulary Size | 28,456 |
| Information Accuracy Defects | 3,240 (7.2%) |
| Communication Defects | 4,950 (11.0%) |
| Compliance Defects | 2,160 (4.8%) |
| No Defect Identified | 34,650 (77.0%) |

The preprocessing pipeline applies standard text normalization procedures adapted for customer service dialogue characteristics. Text is converted to lowercase, and punctuation is standardized. Customer personally identifiable information, including names, account numbers, and contact details are replaced with placeholder tokens to protect privacy while preserving conversational structure. Spelling correction is applied to address common typographical errors present in chat-based interactions.

Feature extraction procedures vary according to the requirements of different classification algorithms. Traditional machine learning methods utilize TF-IDF weighted unigram and bigram features with a vocabulary limited to terms appearing in at least five conversations. Word embedding representations are generated using pre-trained GloVe vectors with 300 dimensions. Transformer-based models employ subword tokenization using the WordPiece algorithm with a vocabulary size of 30,000 tokens. The dataset is partitioned into training, validation, and test sets using stratified sampling to maintain consistent class distributions. The training set contains 31,500 conversations (70%), the validation set contains 6,750 conversations (15%), and the test set contains 6,750 conversations (15%).

*3.3. Comparative Algorithms*

The comparative study examines representative algorithms from three major categories of machine learning approaches. Traditional machine learning methods

include Support Vector Machine with radial basis function kernel, Multinomial Naive Bayes, Random Forest with 100 estimators, and Logistic Regression with L2 regularization. These algorithms represent well-established approaches with proven effectiveness on text classification tasks and favorable interpretability characteristics [13].

Deep learning sequence models include Long Short-Term Memory networks and Bidirectional LSTM architectures. The LSTM implementation uses 128 hidden units with a dropout rate of 0.3 applied between layers. The BiLSTM architecture processes input sequences in both forward and backward directions, concatenating hidden states to capture bidirectional context. An attention mechanism is incorporated to enable the model to focus on salient portions of the input sequence. Table 3 details the hyperparameter configurations for deep learning models.

**Table 3.** Deep Learning Model Hyperparameters.

| Model | Hidden Units | Layers | Dropout | Learning Rate | Batch Size |
|---|---|---|---|---|---|
| LSTM | 128 | 2 | 0.3 | 0.001 | 32 |
| BiLSTM | 128 | 2 | 0.3 | 0.001 | 32 |
| BiLSTM-Attention | 128 | 2 | 0.3 | 0.001 | 32 |

Transformer-based models include BERT-base, RoBERTa, and DistilBERT. BERT-base consists of 12 transformer encoder layers with 768 hidden dimensions and 12 attention heads, containing approximately 110 million parameters. RoBERTa applies optimized pre-training procedures, including dynamic masking and larger batch sizes. DistilBERT provides a compressed alternative with 6 layers and 66 million parameters, achieving faster inference while retaining substantial performance [14].

Fine-tuning procedures for transformer models follow established practices. A classification head consisting of a dropout layer followed by a linear projection is added to the pre-trained encoder. Training proceeds for 4 epochs with a learning rate of 2e-5 and a linear warmup over 10% of training steps. Early stopping based on validation loss prevents overfitting (As shown in Figure 1).

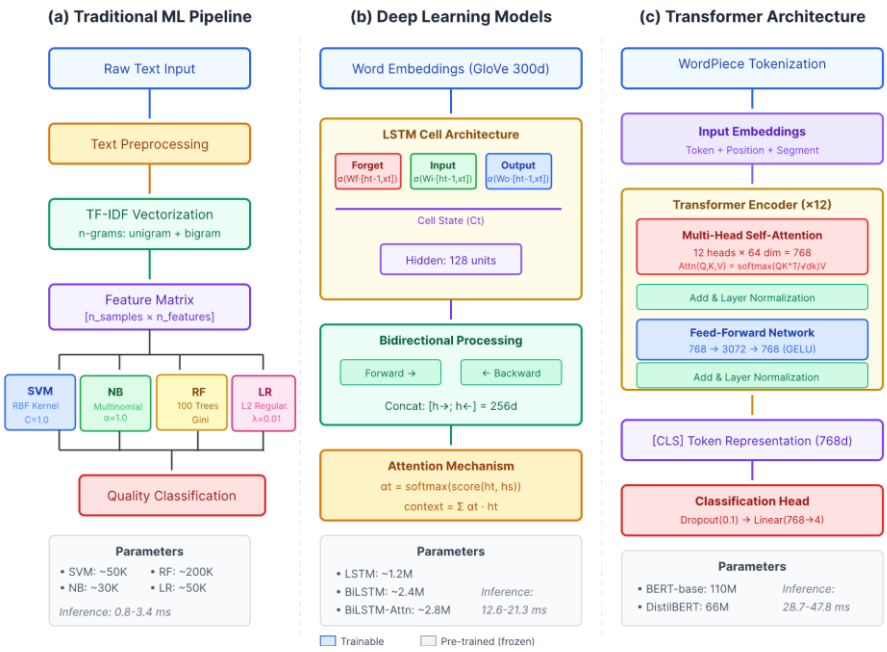


**Figure 1.** Comparative Model Architecture Diagram.

This figure presents a comprehensive architectural comparison of the three model categories examined in this study. The diagram is organized in three vertical panels. The left panel illustrates the traditional machine learning pipeline, showing the flow from raw text through TF-IDF vectorization to the classification algorithms (SVM, Naive Bayes, Random Forest, Logistic Regression) with their respective kernel functions and decision boundaries visualized. The center panel depicts the deep learning sequence models, displaying the LSTM cell structure with input, forget, and output gates, the bidirectional processing flow for BiLSTM, and the attention weight distribution across sequence positions. The right panel shows the transformer architecture, illustrating the multi-head self-attention mechanism, layer normalization components, and the fine-tuning classification head. Arrows indicate data flow between components, and parameter counts are annotated for each model variant. Color coding distinguishes trainable parameters (blue) from frozen pre-trained weights (gray).

The implementation utilizes PyTorch framework for deep learning models and scikit-learn library for traditional machine learning algorithms. All experiments are conducted on a computing environment with NVIDIA V100 GPU with 32GB memory. Training time and inference latency are recorded to assess computational efficiency across methods.

## 4. Results and Analysis

### *4.1. Performance Comparison of Classification Methods*

The classification performance evaluation applies standard metrics including accuracy, precision, recall, and F1-score computed on the held-out test set. Macro-averaged scores provide equal weighting across quality categories regardless of class imbalance. The experimental results reveal substantial performance differentiation across algorithmic approaches. Table 4 presents comprehensive performance metrics for all evaluated methods on the multi-class quality classification task.

**Table 4.** Classification Performance Comparison.

| Method | Accuracy | Precision | Recall | F1-Score | Time(ms) |
|---|---|---|---|---|---|
| Naive Bayes | 0.763 | 0.712 | 0.698 | 0.705 | 0.8 |
| Logistic Regression | 0.801 | 0.756 | 0.742 | 0.749 | 1.2 |
| SVM (RBF) | 0.824 | 0.784 | 0.771 | 0.777 | 3.4 |
| Random Forest | 0.812 | 0.768 | 0.759 | 0.763 | 2.1 |
| LSTM | 0.856 | 0.821 | 0.814 | 0.817 | 12.6 |
| BiLSTM | 0.873 | 0.842 | 0.836 | 0.839 | 18.4 |
| BiLSTM-Attention | 0.891 | 0.863 | 0.857 | 0.860 | 21.3 |
| DistilBERT | 0.912 | 0.889 | 0.884 | 0.886 | 28.7 |
| BERT-base | 0.942 | 0.924 | 0.918 | 0.921 | 45.2 |
| RoBERTa | 0.938 | 0.919 | 0.912 | 0.915 | 47.8 |

BERT-base achieves the highest overall accuracy at 94.2%, representing a significant improvement over traditional machine learning baselines. The performance gap between transformer models and deep learning sequence models (approximately 5-7% absolute improvement) demonstrates the value of pre-trained language representations for this task. Traditional machine learning methods achieve moderate performance levels, with

SVM obtaining 82.4% accuracy, confirming their continued viability for resource-constrained deployment scenarios.

Inference-time measurements reveal the computational trade-offs associated with model complexity. Naive Bayes completes inference in 0.8 milliseconds per sample, compared to 45.2 milliseconds for BERT-base, resulting in a 56-fold difference in processing speed. These efficiency considerations become significant when processing high-volume interaction streams in real-time monitoring applications [15] (As shown in Figure 2).

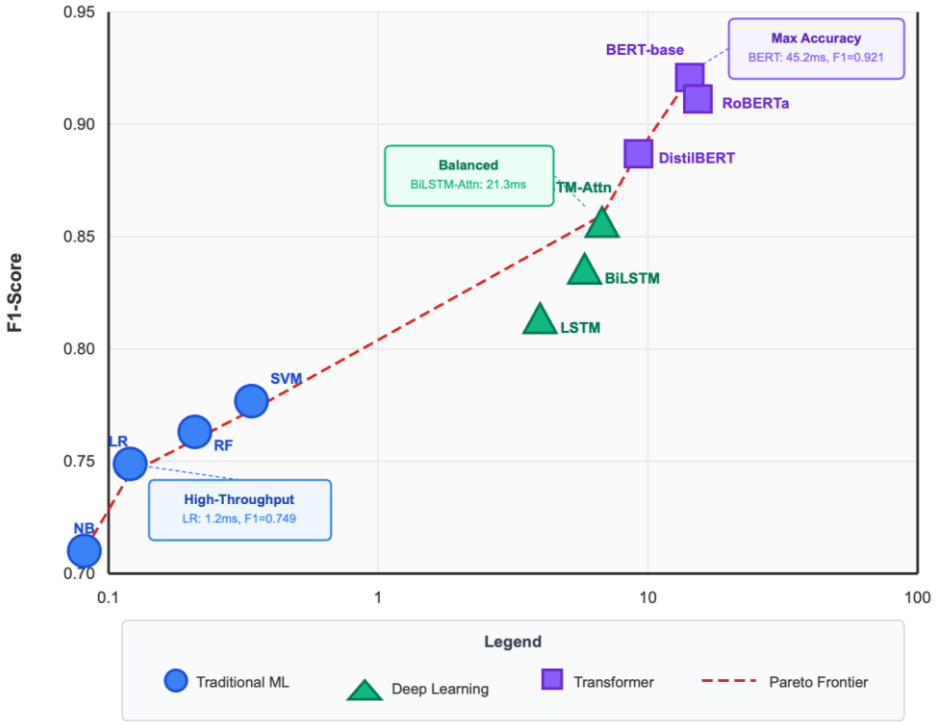


**Figure 2.** Performance-Efficiency Trade-off Visualization.

This figure shows a scatter plot of classification accuracy versus computational efficiency for all evaluated methods. The x-axis represents per-sample inference time on a logarithmic scale (milliseconds; see Table 4). The y-axis shows F1-score ranging from 0.70 to 0.95. Each method is represented by a distinct marker shape (circles for traditional ML, triangles for deep learning, squares for transformers) with color intensity indicating model parameter count. Pareto-optimal methods forming the efficiency frontier are connected by a dashed line. Annotation boxes highlight the three recommended operating points: high-throughput deployment (Logistic Regression), balanced performance (BiLSTM-Attention), and maximum accuracy (BERT-base). Results correspond to the held-out test set; no cross-validation confidence intervals are reported. Performance differences are discussed based on the held-out test set metrics reported in Table 4. BERT-base achieves the highest overall performance among the evaluated methods. RoBERTa performs comparably to BERT on this task.

*4.2. Sensitivity and Specificity Analysis*

The aggregate performance metrics mask important variations in classification effectiveness across different quality defect categories. Analysis of category-specific performance reveals distinct patterns that inform method selection for targeted quality assessment applications. Table 5 presents F1-scores disaggregated by quality defect category for selected high-performing methods.

**Table 5.** Category-Specific F1-Scores by Method.

| Method | Info Error | Comm Defect | Compliance | No Defect |
|---|---|---|---|---|
| SVM | 0.692 | 0.724 | 0.658 | 0.834 |
| BiLSTM-Attn | 0.812 | 0.847 | 0.789 | 0.892 |
| DistilBERT | 0.856 | 0.878 | 0.834 | 0.924 |
| BERT-base | 0.897 | 0.912 | 0.891 | 0.948 |

The detection of compliance violations presents the greatest classification challenge across all methods. Compliance defects often involve subtle omissions rather than explicit erroneous statements, requiring models to recognize the absence of required disclosure language. BERT-base achieves 89.1% F1-score on compliance detection, compared to 65.8% for SVM, representing a 23.3 percentage-point improvement. This performance gap underscores the importance of contextual language understanding for identifying procedural compliance issues.

Communication defects involving inappropriate tone or empathy failures are of intermediate difficulty. The BiLSTM-Attention model demonstrates notable effectiveness in this category, achieving an 84.7% F1-score by identifying sentiment-bearing phrases and attending to emotionally charged portions of conversations. The attention mechanism provides interpretable evidence for classification decisions by highlighting specific utterances contributing to defect identification.

Information accuracy errors present distinct detection characteristics depending on error type. Factual errors involving numerical values (e.g., incorrect pricing, wrong dates) achieve higher detection rates than nuanced policy misrepresentations. The feature importance analysis for traditional machine learning models reveals that specific lexical patterns, including hedge words ("approximately," "around," "usually"), are associated with higher error rates.

The confusion matrix analysis identifies systematic misclassification patterns. The most frequent error involves the false negative classification of communication defects as no-defect cases. Manual examination of misclassified samples reveals that borderline cases involving mildly dismissive language without explicit rudeness challenge all evaluated methods. The distinction between acceptable brevity and inappropriate curtness requires nuanced judgment that current algorithms struggle to replicate consistently (As shown in Figure 3).

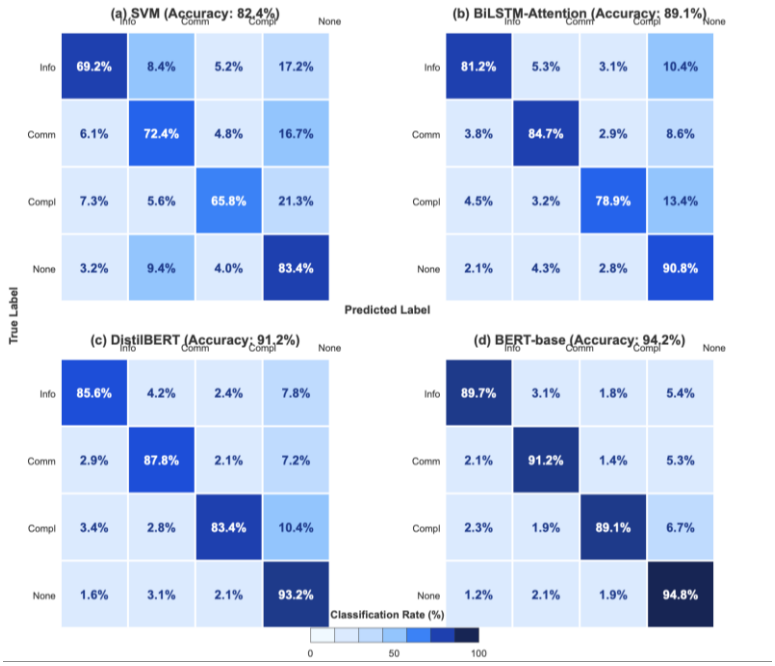


**Figure 3.** Confusion Matrix Heatmap Comparison.

This figure presents a 2x2 confusion matrix showing classification patterns across four methods (SVM, BiLSTM-Attention, DistilBERT, BERT-base). Each heatmap displays a 4x4 matrix, with rows representing the true labels (Information Error, Communication Defect, Compliance Violation, No Defect) and columns representing the predicted labels. Cell colors use a sequential blue colormap with intensity proportional to the percentage of predictions, ranging from white (0%) to dark blue (100%). Diagonal cells representing correct classifications are outlined in bold. Off-diagonal cells with a misclassification rate exceeding 5% are annotated with the exact percentage value. The figure caption notes that BERT-base achieves the most concentrated diagonal pattern, while SVM shows notable confusion between Communication Defect and No Defect categories (16.7% misclassification rate).

### *4.3. Interpretability and Explainability Assessment*

The deployment of automated quality assessment systems in customer service environments requires consideration of interpretability requirements. Quality managers and supervisors need to understand the basis for automated quality scores to validate system outputs and provide meaningful feedback to agents. Regulatory frameworks increasingly emphasize the importance of explainable AI in consumer-facing applications.

Traditional machine learning methods offer inherent interpretability advantages by relying on explicit feature representations. Logistic Regression coefficients directly indicate the contribution of individual terms to classification decisions. The top positive and negative coefficients reveal lexical indicators associated with quality defects. Terms including "unfortunately," "cannot," and "policy prohibits" carry negative weights indicating correlation with information accuracy issues. Positive indicators for high-quality interactions include "happy to help," "certainly," and acknowledgment phrases.

Feature importance rankings derived from Random Forest models provide complementary insights into interpretability. Permutation importance analysis identifies the most influential features in classification decisions. Bigram features capturing negation patterns ("not able," "cannot provide") rank among the most important predictors for defect detection.

The attention mechanisms in BiLSTM-Attention and transformer models provide post-hoc interpretability through attention weight visualization. Close attention weights indicate that tokens have a significant influence on classification decisions. Analysis of attention patterns reveals that models attend strongly to utterance boundaries and turn-taking transitions in multi-turn conversations. This behavior aligns with quality assessment practices where initial greeting and closing statements carry particular importance for overall impression formation.

The evaluation of explainability quality employs human assessment of explanation utility. A sample of 200 classified conversations with associated explanations (feature importance or attention weights) was presented to quality analysts, who rated the helpfulness of the explanations on a 5-point scale. BERT-base attention explanations received a mean helpfulness rating of 3.4, while Logistic Regression coefficient-based explanations achieved a 3.8 rating. This result suggests that simpler explanation formats may provide greater practical utility despite the superior classification accuracy of complex models.

The trade-off between classification performance and interpretability emerges as a critical consideration for deployment decisions. Organizations prioritizing automated decision-making efficiency may favor BERT-base despite limited interpretability. Applications requiring human review and validation of automated assessments may benefit from the transparency of traditional machine learning approaches or the application of post-hoc explanation methods to neural network outputs.

## 5. Conclusion

### *5.1. Summary of Findings*

This comparative study has examined machine learning methods for automated quality assessment of customer service dialogue across traditional, deep learning, and

transformer-based approaches. The experimental evaluation on a substantial dataset of 45,000 annotated conversations demonstrates clear performance stratification across algorithmic categories.

BERT-base achieves the highest classification accuracy at 94.2% with an F1-score of 0.921, representing state-of-the-art performance for this task. The pre-trained language representations enable effective detection of subtle quality defects, including compliance violations and inappropriate communication patterns. RoBERTa and DistilBERT offer competitive alternatives with distinct computational efficiency profiles, suitable for varied deployment requirements.

Deep learning sequence models, including BiLSTM with attention, achieve intermediate performance (89.1% accuracy) while offering improved interpretability through attention-weight visualization. Traditional machine learning methods remain viable in resource-constrained scenarios, with SVM achieving 82.4% accuracy at a dramatically reduced computational cost. The category-specific analysis reveals that compliance violation detection benefits most substantially from advanced language models, with BERT-base improving the F1-score by 23.3 percentage points over the SVM baseline.

### *5.2. Practical Implications*

The findings provide actionable guidance for organizations implementing automated quality monitoring systems. The method selection decision should consider operational priorities, including accuracy requirements, computational budget, and interpretability needs. High-volume monitoring applications processing millions of daily interactions may favor DistilBERT as a balanced solution, achieving 91.2% accuracy with moderate computational requirements. Organizations emphasizing real-time feedback and low latency may deploy traditional machine learning methods for initial screening, with transformer models applied selectively to flagged conversations. The quality assessment framework proposed in this study aligns with consumer protection principles and common compliance requirements, providing a structured approach to operationalizing quality assurance practices. The multidimensional evaluation, encompassing information accuracy, communication appropriateness, and process compliance, addresses the full range of service quality considerations.

### *5.3. Limitations and Future Work*

Several limitations constrain the generalizability of these findings. The dataset derives from English-language interactions within specific industry sectors, and performance may vary for other languages or domains. The binary annotation scheme for each quality dimension simplifies the nuanced spectrum of service quality into discrete categories.

Future research directions include extending multilingual quality assessment to leverage the cross-lingual transfer learning capabilities of multilingual transformer models. The development of joint models that simultaneously perform quality classification and generate explanations represents a promising approach to addressing interpretability requirements. Investigation of few-shot and zero-shot learning paradigms may enable rapid adaptation to new quality dimensions without extensive annotation effort. The integration of automated quality assessment with agent coaching and training systems offers opportunities for closed-loop quality improvement.

## References

1. A. O. Afolabi and J. K. Alhassan, "NLP techniques for automating responses to customer queries: A systematic review," Discover Artificial Intelligence, vol. 3, no. 1, p. 65, 2023. https://doi.org/10.1007/s44163-023-00065-5
2. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: A comprehensive review," ACM Computing Surveys, vol. 54, no. 3, pp. 1-40, 2021. https://doi.org/10.1145/3439726
3. S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," Expert Systems with Applications, vol. 184, p. 115461, 2021. https://doi.org/10.1016/j.eswa.2021.115461
4. H. Chen, X. Liu, D. Yin, and J. Tang, "Recent advances in deep learning based dialogue systems: A systematic survey," Artificial Intelligence Review, vol. 55, no. 3, pp. 2387-2444, 2022.

5. M. Nuruzzaman and O. K. Hussain, "A survey on chatbot implementation in customer service industry through deep neural networks," in 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), pp. 54-61, 2018. https://doi.org/10.1109/ICEBE.2018.00019
6. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1-42, 2018. https://doi.org/10.1145/3236009
7. Z. Li, C. Yang, and C. Huang, "A comparative sentiment analysis of airline customer reviews using BERT and its variants," Mathematics, vol. 12, no. 1, p. 53, 2024. https://doi.org/10.3390/math12010053
8. T. A. Al-Qablan, M. H. Mohd Noor, M. Al-Betar, and A. Khader, "A survey on sentiment analysis and its applications," Neural Computing and Applications, vol. 35, no. 11, pp. 8013-8034, 2023. https://doi.org/10.1007/s00521-023-08334-5
9. F. Wei et al., "Empirical study of LLM fine-tuning for text classification in legal document review," in 2023 IEEE International Conference on Big Data, pp. 2786-2792, 2023. https://doi.org/10.1109/BigData59044.2023.10386911
10. M. Durairaj and A. Chinnalagu, "Transformer based contextual model for sentiment analysis of customer reviews: A fine-tuned BERT," International Journal of Advanced Computer Science and Applications, vol. 12, no. 11, pp. 423-432, 2021.
11. B. Galitsky and D. Ilvovsky, "A review of natural language processing in contact centre automation," Pattern Analysis and Applications, vol. 26, no. 3, pp. 823-846, 2023. https://doi.org/10.1007/s10044-023-01182-8
12. S. Bharati, M. R. H. Mondal, and P. Podder, "A review on explainable artificial intelligence for healthcare: Why, how, and when?" IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1429-1442, 2023. https://doi.org/10.1109/TAI.2023.3266418
13. A. Zangari, M. Marcuzzo, M. Schiavinato, A. Gasparetto, and A. Albarelli, "Are we really making much progress in text classification? A comparative review," ACM Computing Surveys, vol. 56, no. 8, pp. 1-38, 2024.
14. A. Patel, P. Oza, and S. Agrawal, "Sentiment analysis of customer feedback and reviews for airline services using language representation model," Procedia Computer Science, vol. 218, pp. 2459-2467, 2023.
15. H. Mohammadi, A. Bagheri, A. Giachanou, and D. L. Oberski, "Explainability in practice: A survey of explainable NLP across various domains," arXiv preprint arXiv:2502.00837, 2025.