RAIS Journal for Social Sciences | Vol. 10, No. 2, 2026
ISSN 2574-0245 (Print) | ISSN 2574-1179 (Online) | DOI:10.5281/zenodo.24525582

AI Incorporation of Sentiment Analysis of Customer Reviews in the Hospitality Industry Using Classical and Transformer-Based Models

Marisa Tania[1], Shravya Kalva[2]

[1,2]University of the Cumberlands, Georgia Tech, USA
Corresponding Author: mtania49833@ucumberlands.edu

Abstract: The incorporation of AI in the arena of Sentiment analysis is a key component in extracting actionable insights from large-scale customer feedback, particularly in the hospitality industry. We evaluate AI models and transformer-based approaches for sentiment classification using the Yelp Review Full dataset, where reviews are mapped into negative, neutral, and positive sentiment classes based on star ratings. Within AI Multinomial Naive Bayes and Linear Support Vector Machines (SVM) are implemented as baseline models using TF-IDF features, while transformer-based models, including DistilBERT and an improved RoBERTa configuration, are fine-tuned to capture contextual representations. Experimental results show that classical models achieve baseline performance (macro F1 ≈ 0.70-0.71), while transformer models significantly improve performance, with DistilBERT reaching approximately 0.75 and RoBERTa achieving 0.80. Despite these gains, neutral sentiment remains difficult to classify due to inherent ambiguity in customer feedback. All reported gaps between models are confirmed with bootstrap confidence intervals and pairwise McNemar tests, and the RoBERTa improvement is decomposed into a training-duration effect and a model-capacity effect. We further report model calibration, a selective-prediction (abstention) analysis quantifying the trade-off between manual review load and negative-sentiment recall, and a cost comparison of training and inference across models. These findings highlight the importance of contextual modeling and suggest that transformer architectures are better suited for real-world sentiment analysis. Practical implications for customer experience management and business decision-making are discussed.

Keywords: Sentiment Analysis, Hospitality, Customer Reviews, TF-IDF, Transformer Models, Selective Prediction

Introduction

Sentiment analysis is a fundamental application of artificial intelligence (AI) and natural language processing (NLP) that automatically categorizes textual information as positive, negative, or neutral (Liu, 2012). Within the hospitality industry, AI-powered sentiment analysis enables organizations to evaluate customer perceptions at scale, detect service-related problems, monitor brand reputation, and proactively address customer concerns. This is where the impact of AI has brought a significant improvement over time.

Traditional AI and machine learning techniques for sentiment analysis, such as Multinomial Naive Bayes and Support Vector Machines (SVM), typically use statistical text representations, including bag-of-words and TF-IDF features (Jurafsky & Martin, 2023). Although these approaches are computationally efficient and can effectively identify explicit sentiment indicators, they have limited ability to understand relationships between words when meaning depends on sentence structure, context, or subtle linguistic expressions.

Recent developments in AI and deep learning, particularly transformer-based models such as BERT, have substantially improved NLP performance by capturing contextual relationships within textual data (Devlin et al., 2019). These advanced AI models have achieved strong results in sentiment classification, especially when analyzing nuanced or ambiguous language. Nevertheless, accurately identifying neutral and mixed sentiments remains challenging, particularly in authentic customer reviews that contain complex or conflicting opinions.

This study addresses these challenges by evaluating and comparing traditional machine learning algorithms with transformer-based AI models for sentiment classification using the Yelp Review Full dataset. Specifically, the research examines the degree to which AI-driven contextual modeling improves classification performance and investigates the continuing difficulty of accurately detecting neutral sentiment. Through an analysis of model performance and classification errors, this study seeks to provide practical insights into the use of AI-based sentiment analysis systems for real-world customer experience management. The evaluation further reports statistical significance testing, a controlled ablation separating training duration from model capacity, calibration analysis, a selective-prediction framework for human-in-the-loop triage, and a preliminary subtype analysis of the neutral class, together with a proposed protocol for empirically testing whether star-rating labels impose a ceiling on achievable performance.

Data Collection

This study utilizes the Yelp Review Full dataset, a large-scale collection of customer-generated reviews from the hospitality and service industry, including restaurants, local businesses, and service providers. The dataset is widely adopted in sentiment analysis research due to its scale, diversity, and structured rating system. Each review consists of a free-text description of the customer experience, accompanied by a rating label ranging from 1 to 5, which serves as a proxy for overall customer satisfaction.

The hospitality domain is particularly well-suited to sentiment analysis, as customer experience is inherently subjective and frequently expressed through written feedback. Platforms such as Yelp function as critical feedback channels, enabling businesses to monitor service quality, identify recurring issues, and better understand customer preferences. As a result, systematic analysis of this data provides both operational insights and strategic value. The feedback loop is an enhancement within the AI domain that adds knowledge, along with appropriate context in the aforementioned field of hospitality.

To support business-oriented interpretation, the original rating labels are grouped into three sentiment categories: lower ratings (one to two stars) as negative, the midpoint rating (three stars) as neutral, and higher ratings (four to five stars) as positive. This transformation preserves meaningful distinctions in customer sentiment while simplifying classification into categories that align with common business use cases, such as identifying dissatisfied, uncertain, and satisfied customers.

Exploratory analysis indicates that the dataset exhibits substantial variability in both content and structure. Review lengths follow a right-skewed distribution, with most reviews being relatively concise and a smaller proportion extending to several hundred words (median 101 words; 95th percentile above 450 words, with a maximum of 1,004 words). This long-tail behavior reflects differences in customer engagement and expression styles, introducing both informative detail and potential noise into the data. In addition, sentiment distribution is naturally imbalanced, with positive and negative reviews appearing more frequently than neutral ones. As illustrated in Appendix A (Figures A1 and A2), these characteristics highlight important challenges for model training and evaluation.

To address class imbalance and ensure fair comparison across models, a balanced sampling strategy is applied. An equal number of samples is selected from each sentiment

category for training, validation, and testing. The final experimental dataset consists of an 18,000-review training set (6,000 instances per class), with balanced validation and test sets of 3,000 reviews each (1,000 instances per class). This corrects an earlier draft, which understated the training set as 3,000 samples total; the validation and test sizes were accurate as originally reported. This controlled setup enables consistent evaluation across modeling approaches while maintaining relevance to real-world customer feedback patterns.

This data preparation framework establishes a robust foundation for sentiment classification by balancing methodological rigor with practical applicability in customer experience analysis.

Data Preprocessing

Effective preprocessing is essential for transforming raw textual data into a structured format suitable for machine learning models while preserving meaningful linguistic information. In this study, preprocessing is designed to balance noise reduction with the retention of sentiment-bearing features, with distinct strategies applied for classical and transformer-based models. The preprocessing pipeline begins with data cleaning and normalization. All reviews are converted into a consistent textual format by removing missing or empty entries and standardizing whitespace. Basic normalization is applied to ensure uniform text representation while avoiding aggressive transformations that could distort semantic meaning. This design choice is particularly important in customer feedback data, where subtle variations in wording can convey significant sentiment.

For classical machine learning models, preprocessing is largely integrated into the TF-IDF vectorization process, where the AI models incorporate the essence of analysis and established understanding. Text is normalized through lowercasing to reduce feature sparsity, ensuring that semantically identical words with different casing are treated consistently. In addition, high-frequency stop words are removed to reduce noise and improve the signal-to-noise ratio in the feature space.

Because bigrams are relied upon to capture negation (e.g., "not good" versus "good"), stop-word removal uses a custom list that explicitly retains negators and contrastive terms, including "not," "no," "nor," "never," "cannot," "without," "but," "however," "very," and "too," rather than a generic English stop-word list, which would otherwise strip these terms before bigrams are formed and leave only the surrounding content words as features. The effect of this choice was measured directly: on the validation split, the negation-preserving list improved macro F1 by +0.020 for Multinomial Naive Bayes (0.701 → 0.721) and +0.021 for Linear SVM (0.694 → 0.715) relative to a generic stop-word list, holding all other settings fixed. The effect is also visible qualitatively: among the terms, Linear SVM weights most heavily toward the negative and positive classes, "not good" and "not bad" appear directly, confirming that the bigram feature space captures the negation contrast it is intended to. To capture local contextual patterns, both unigrams and bigrams are included, allowing the model to represent short phrases such as negations and common sentiment expressions. This approach enables the model to capture simple contextual dependencies without requiring more complex linguistic transformations. Lemmatization is not applied in this study. Although lemmatization can reduce vocabulary size by mapping words to their base forms, it may also remove subtle distinctions that contribute to sentiment interpretation. For instance, variations such as “liked,” “liking,” and “like” can convey differences in tone or intensity in customer reviews. Given that TF-IDF with n-gram features already captures such variations, omitting lemmatization helps preserve useful linguistic nuance without significantly increasing model complexity.

For transformer-based models, preprocessing is intentionally minimal. The raw text is passed directly to the pretrained tokenizer associated with the model architecture, which handles tokenization, casing, and subword segmentation internally. This approach leverages

the strengths of transformer models, which are designed to learn contextual relationships directly from natural language input. As a result, manual preprocessing steps commonly used in classical pipelines are unnecessary and may even degrade performance if they remove useful contextual signals.

The preprocessing strategy reflects a model-aware design that aligns with the strengths and assumptions of each approach. Classical models benefit from structured feature engineering through TF-IDF, while transformer models rely on contextual embeddings derived from minimally processed text. One consequence of minimal preprocessing is truncation: with a maximum sequence length of 256 tokens, 21.8% of test reviews exceed the limit, and truncation discards approximately 19.0% of all tokens in the test set (median review length 133 tokens; 95th percentile 468 tokens). This is quantified further in the Limitations section. This distinction ensures a fair and meaningful comparison between modeling approaches while maintaining the integrity of the original customer feedback data.

Feature Extraction

Feature extraction plays a critical role in converting unstructured textual data into numerical representations that machine learning models can effectively use. In this study, two distinct feature extraction approaches are employed, corresponding to the classical and transformer-based modeling frameworks: Term Frequency–Inverse Document Frequency (TF-IDF) and contextual embeddings derived from pretrained transformer models.

For classical machine learning models, text is represented using TF-IDF, a widely adopted statistical technique that captures the relative importance of words within a document while accounting for their frequency across the corpus. Specifically, TF-IDF assigns higher weights to terms that appear frequently in a given review but are less common across other reviews, thereby emphasizing discriminative features. In addition to unigrams, bigrams are incorporated to capture short contextual patterns, such as negation and commonly occurring sentiment phrases. This is particularly important in sentiment analysis, where meaning often depends on local word combinations (e.g., “not good” versus “good”). The resulting feature space is high-dimensional and sparse, with each dimension corresponding to a unique term or phrase extracted from the corpus.

While TF-IDF does not explicitly encode semantic relationships between words, it provides a strong baseline representation that is computationally efficient and effective for capturing surface-level linguistic patterns. The inclusion of n-gram features partially mitigates its limitations by allowing the model to incorporate limited contextual information. However, because TF-IDF treats each feature independently, it cannot fully capture deeper semantic or syntactic dependencies within the text.

In contrast, transformer-based models use contextual embeddings trained on large-scale pretraining corpora of natural language. In this study, a pretrained model such as DistilBERT or RoBERTa is fine-tuned for sentiment classification. These models convert input text into dense vector representations, embedding each token based on its surrounding context. Unlike TF-IDF, where each word is represented independently, transformer embeddings dynamically encode meaning depending on the full sentence structure. This allows the model to capture nuanced linguistic phenomena such as polysemy, long-range dependencies, and implicit sentiment expressions.

From a feature space perspective, the difference between these approaches is substantial. TF-IDF produces a sparse and high-dimensional representation, where most feature values are zero, and each dimension corresponds to a specific lexical item. In contrast, transformer models generate dense, low-dimensional embeddings that encode semantic relationships in a continuous vector space. These embeddings are learned representations rather than manually engineered features, enabling the model to generalize more effectively to previously unseen patterns.

Although explicit visualization of the transformer feature space is beyond the scope of this study, exploratory analysis of the TF-IDF representation highlights its sparsity and high dimensionality. This structural difference provides insight into the observed performance gap between classical and transformer-based models, as contextual embeddings are better suited for capturing the complexity of real-world customer feedback.

The use of both TF-IDF and transformer-based embeddings enables a comprehensive evaluation of feature representation strategies, providing a meaningful comparison between traditional and modern approaches to sentiment analysis.

Model Development

This study combines classical machine learning and transformer-based models to evaluate their effectiveness for sentiment classification of customer reviews. The model selection is designed to provide both a strong baseline and a modern contextual approach, enabling a meaningful comparison across fundamentally different representation and learning paradigms.

For classical baselines, two widely used algorithms are employed: Multinomial Naive Bayes and Linear Support Vector Machines (SVM). Multinomial Naive Bayes is selected due to its simplicity, computational efficiency, and strong historical performance in text classification tasks. The model assumes conditional independence among features and estimates class probabilities based on word frequency distributions. A smoothing parameter (alpha = 0.5) is applied to handle zero-frequency issues and improve generalization. This value was selected via a grid search over the validation split (alpha $\in$ {0.1, 0.25, 0.5, 1.0, 2.0}), which confirmed alpha = 0.5 as optimal. Despite its strong baseline performance, the independence assumption limits its ability to capture relationships between words, particularly in phrases where sentiment depends on context.

Linear SVM is included as a more robust discriminative model capable of handling high-dimensional, sparse feature spaces. Unlike Naive Bayes, SVM does not rely on probabilistic assumptions; instead, it learns an optimal decision boundary that maximizes the margin between classes. This makes it particularly effective for TF-IDF representations, where features are numerous and sparse. A regularization parameter (C = 0.1) is used to balance the trade-off between maximizing the margin and minimizing classification error, selected via a grid search over the validation split (C $\in$ {0.1, 0.25, 0.5, 1.0, 2.0}); probability estimates are obtained via Platt scaling (sigmoid calibration, 3-fold) so that calibration and selective-prediction analyses can be computed for all models on a common footing.

To capture contextual and semantic relationships in text, a transformer-based model is introduced. Initially, a pretrained DistilBERT model is fine-tuned for sentiment classification. DistilBERT is chosen for its efficiency and reduced computational requirements while retaining much of the performance of larger transformer architectures. The model processes tokenized text inputs and generates contextual embeddings, which are passed through a classification head to predict sentiment labels. Fine-tuning is performed using a learning rate of 2e-5, a batch size of 16, and a limited number of training epochs, providing a practical baseline for contextual modeling under constrained computational settings.

Building on this initial transformer implementation, a second round of experimentation is conducted to improve performance through enhanced model capacity and training duration. In this performance improvement experiment, the model is upgraded to a stronger transformer architecture (e.g., RoBERTa-base) and trained for three epochs. This adjustment is motivated by the expectation that improved contextual capacity and longer fine-tuning can better capture nuanced sentiment patterns, particularly for ambiguous cases. The improved configuration retains the same learning rate (2e-5) and batch size (16) while incorporating weight decay (0.01) to reduce overfitting. Mixed-precision training is enabled when GPU resources are available, improving computational efficiency. Because this change alters

model capacity and training duration simultaneously, the two factors cannot be separated from the original design alone. A third configuration, DistilBERT trained for three epochs while holding the backbone fixed, is added specifically to isolate the contribution of additional training from the contribution of the larger backbone; results are reported under Ablation: Backbone Capacity vs. Training Duration.

The training process for transformer models is implemented using the Hugging Face Trainer framework, which standardizes training, evaluation, and metric computation. Model selection is guided by a macro-averaged F1 Score, reflecting the importance of balanced performance across all sentiment classes. Evaluation and checkpointing run at every epoch, with the checkpoint of the highest validation macro F1 loaded for final testing. For the 3-epoch DistilBERT configuration, validation macro F1 peaked at epoch 2 (0.7589) and declined slightly by epoch 3 (0.7568), so the epoch-2 checkpoint is the one evaluated on the test set and reported below. For RoBERTa, validation macro F1 was highest at the final epoch, so the final checkpoint is reported for that model. This is particularly relevant given the inclusion of a neutral class, which introduces additional ambiguity and increases the difficulty of the classification task.

Model Evaluation

Model performance is evaluated using a combination of standard classification metrics, including accuracy, precision, recall, and F1-score. Given the task's multi-class nature and the inclusion of a neutral sentiment category, macro-averaged metrics are emphasized to ensure balanced assessment across all classes. Unlike accuracy alone, which can obscure class-specific performance, macro-averaged precision, recall, and F1-score provide a more reliable measure of model effectiveness when class difficulty varies. The overall performance of all models is summarized in Table 1, which presents a comparative view of results across classical and transformer-based approaches. Table 1 reflects the full pipeline described above (negation-preserving stop words, validation-tuned hyperparameters, per-epoch transformer model selection, and the 18,000-review training set); a fifth row reports the DistilBERT 3-epoch ablation.

Table 1. Model performance comparison across classical and transformer-based approaches

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	F1 (Weighted)
TF-IDF + Multinomial NB (α=0.5)	0.693	0.706	0.693	0.696	0.696
TF-IDF + Linear SVM (C=0.1)	0.714	0.713	0.714	0.713	0.713
DistilBERT (1 epoch)	0.749	0.754	0.749	0.751	0.751
DistilBERT (3 epochs) — ablation	0.763	0.763	0.763	0.763	0.763
RoBERTa (fine-tuned, improved)	0.798	0.800	0.798	0.799	0.799

The comparative results demonstrate a clear progression in performance across modeling approaches. Classical machine learning models establish a baseline, with Multinomial Naive Bayes achieving a macro F1-score of approximately 0.70 and Linear SVM achieving 0.71. This marginal improvement reflects the advantage of SVM in learning more flexible decision

boundaries within high-dimensional TF-IDF feature spaces. However, the similarity in performance between these models suggests that both are constrained by the representational limitations of TF-IDF features, which cannot capture deeper semantic relationships in text. Additional confusion matrices for the classical models are provided in Appendix B to further illustrate their classification behavior.

As shown in Figure 1, model performance improves as representational capacity increases, with transformer-based models significantly outperforming classical baselines.

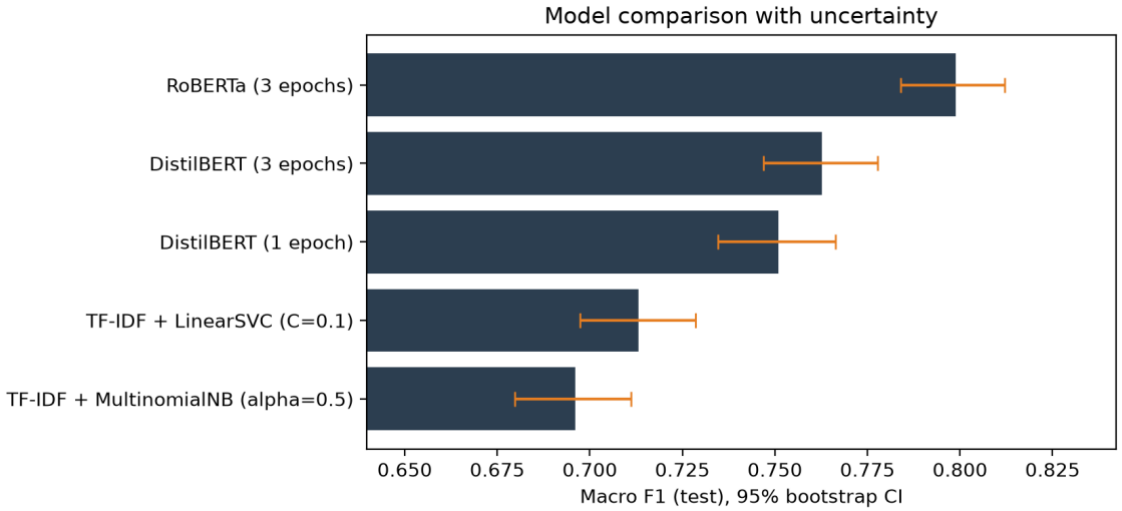

Figure 1. Comparison of model performance across classical and transformer-based approaches using macro F1-score, with 95% bootstrap confidence intervals (2,000 resamples).

Note: The figure shows a clear progression in performance from TF-IDF–based models to transformer architectures. The RoBERTa model achieves the highest macro F1 score, and its interval does not overlap with any other model's interval, indicating that the effectiveness of contextual embeddings and extended fine-tuning for sentiment classification is not attributable to sampling noise at this test-set size.

The introduction of transformer-based models results in a substantial performance gain. The fine-tuned DistilBERT model achieves a macro F1-score of approximately 0.75, indicating a significant improvement over classical baselines. This gain highlights the importance of contextual representation in sentiment analysis, as transformer architectures can model semantic dependencies and nuanced linguistic patterns that are not accessible through bag-of-words approaches. The improvement is consistent across accuracy, precision, and recall, suggesting that the model achieves more balanced predictions across sentiment classes.

A second round of fine-tuning using a stronger transformer architecture and extended training further improves performance. The enhanced RoBERTa model achieves a macro F1-score of 0.799 and an accuracy of 0.798, further improving upon the initial transformer configuration. This result indicates that increased model capacity and longer training contribute meaningfully to performance gains. However, the magnitude of improvement is smaller than the jump from classical models to transformers, suggesting diminishing returns from increasing model complexity alone. The relative contribution of "increased model capacity" versus "longer training," previously stated together as a single explanation, is quantified directly below.

Statistical Validation

Because the test set contains 1,000 reviews per class, point-estimate differences between models were checked against sampling variability before being described as improvements. Two complementary tests were used: 95% bootstrap confidence intervals (2,000 resamples) on macro F1 for each model, and pairwise McNemar exact tests (McNemar, 1947) on the

paired test-set predictions, which is the appropriate significance test when models are compared on the same items.

All ten pairwise comparisons among the five models are significant at $p < 0.05$, including the closest pair, DistilBERT at 1 versus 3 epochs ($p = 0.016$). Every performance ordering reported in Table 1 is therefore statistically supported at this sample size, not merely a point-estimate ranking.

Ablation: Backbone Capacity vs. Training Duration

The RoBERTa configuration changes two factors relative to the initial DistilBERT model: the backbone architecture and the number of training epochs. Adding a third configuration, DistilBERT trained for three epochs with the backbone held fixed, decomposes the total gain (RoBERTa 3-epoch macro F1 minus DistilBERT 1-epoch macro F1, $\Delta F1 = +0.048$) into two additive components: +0.012 (25%) attributable to additional training epochs with the backbone unchanged, and +0.036 (75%) attributable to the larger backbone with epoch count held fixed. Both components are individually confirmed to be significant by the McNemar tests above. This supports the original claim that "increased model capacity and longer training contribute meaningfully to performance gains," while showing that the larger backbone is the dominant factor of the two. The confusion matrix for this ablation configuration is shown in Figure C1 of Appendix C.

Calibration Analysis

Expected calibration error (ECE) and reliability diagrams (Guo et al., 2017; Figure 2) were computed for all five models to assess whether predicted confidence reflects actual accuracy, relevant both as a general model-quality check and because the selective-prediction analysis below uses model confidence as its routing signal. RoBERTa, the most accurate model (79.8% accuracy), is also the most overconfident: a mean predicted confidence of 89.9% against an actual accuracy of 79.8% (ECE = 0.102, overconfidence = +0.101). Linear SVM and DistilBERT (1 epoch) are comparatively well calibrated (ECE = 0.037 and 0.036, respectively). This means RoBERTa's raw softmax outputs should not be read directly as probabilities of correctness; any downstream use of its confidence scores, including the abstention thresholds below, should account for this overconfidence rather than assume it is calibrated out of the box.

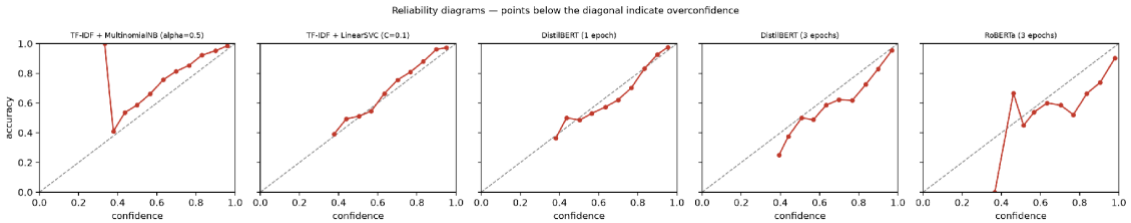

Figure 2. Reliability diagrams for all five models

Note: Points below the diagonal indicate overconfidence: the model reports higher certainty than its accuracy on that confidence band supports. RoBERTa shows the largest and most consistent deviation below the diagonal despite having the highest overall accuracy.

Selective Prediction and Triage Economics

Neutral predictions are best treated as signals of uncertainty rather than definitive labels, and businesses may wish to prioritize recall on negative sentiment. This section operationalizes both ideas using the best-performing model (RoBERTa) by routing low-confidence predictions for manual review and measuring the resulting trade-off, rather than providing a qualitative recommendation.

Treating the model's maximum softmax score as a confidence signal and abstaining below a threshold (Figure 3), routing 21.0% of reviews to manual review raises effective

negative-sentiment recall (automatic predictions plus manually reviewed items, assumed correctly labeled once reviewed) from the unabstained baseline to 90.2%. Raising the manual-review share to 30.8% and 39.9% achieves effective negative recall of 93.0% and 95.8%, respectively. This reframes model selection away from a single macro-F1 number and toward a deployable operating point: an operations team can choose a manual-review budget and read off the negative-sentiment recall it buys, or vice versa. A complementary risk-coverage view, showing accuracy on the automatically handled portion of the test set as coverage varies across all five models, is provided in Figure C2 of Appendix C.

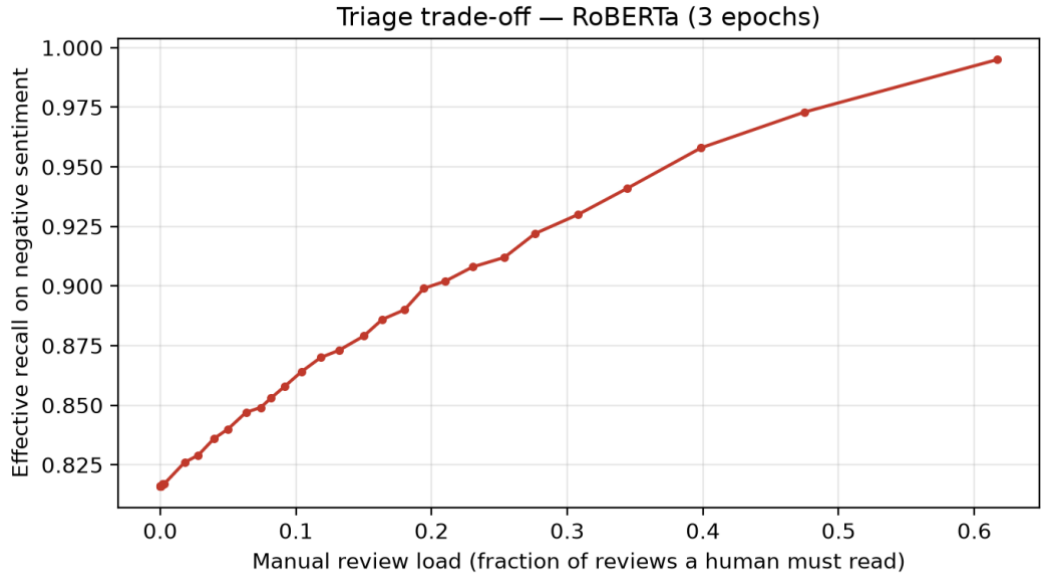

Figure 3. Effective recall on negative sentiment as a function of the fraction of reviews routed to manual review, for the RoBERTa model.

Note: The curve shows the manual-review budget required to achieve a target recall rate among dissatisfied customers.

Neutral Sentiment Subtype Analysis

Neutral reviews often contain mixed or weakly expressed opinions, but these are two distinct phenomena rather than a single category. To test whether they behave differently, neutral test reviews were split into two heuristic subtypes using sentence-level polarity scores from the fitted SVM: mixed (containing both a confidently positive and a confidently negative sentence) and moderate (uniformly tepid, with no strongly polarized sentence). Across the five models, recall for the mixed and moderate subtypes differs only modestly and inconsistently in direction (Figure 4). For example, RoBERTa recalls 71.7% of mixed-neutral reviews versus 73.7% of moderate-neutral reviews, a small gap in the opposite direction from what was hypothesized. This subtype split, at least as measured by this sentence-level heuristic, does not explain a meaningful share of neutral-class error, and the hypothesis that neutral sentiment is hard primarily because it mixes polarities is not supported by this analysis as strongly as anticipated. This is reported as a negative result: the ambiguity in the neutral class does not appear to concentrate specifically in mixed-polarity reviews, at least not in a way this proxy measure can detect, and gold-standard annotation (Appendix D) is a more reliable way to characterize it than the sentence-level heuristic used here.

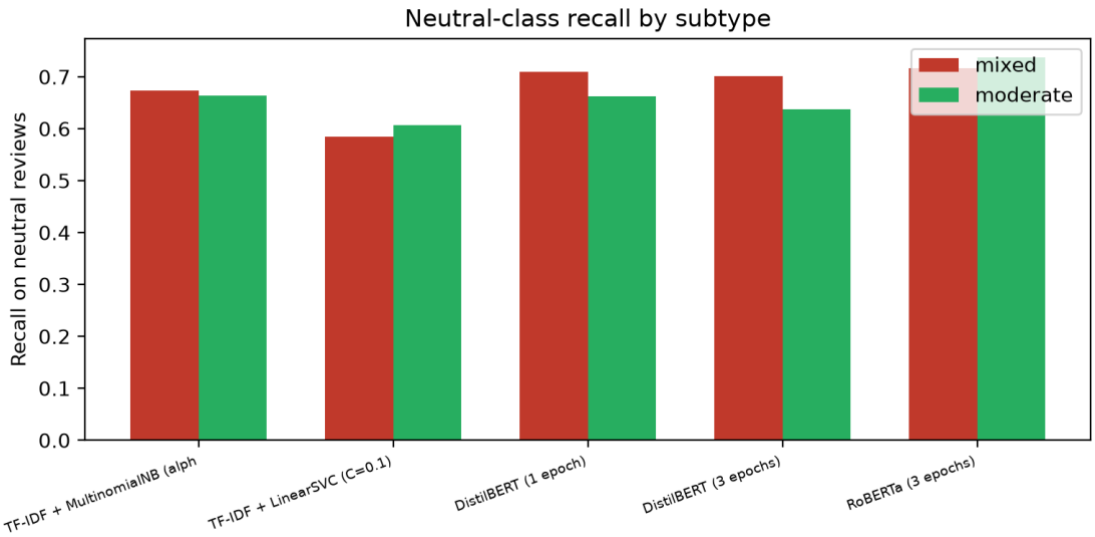

Figure 4. Recall on the neutral class, split by heuristic subtype (mixed vs. moderate), for all five models.

Note: The two subtypes exhibit similar yet inconsistent recall gaps across models, contrary to the hypothesis that mixed-polarity reviews are the primary driver of neutral-class errors.

Cost of Accuracy

Because the classical and transformer models differ by orders of magnitude in inference cost, training time, inference latency, and parameter count, these metrics are reported alongside accuracy. RoBERTa's macro F1 advantage over Linear SVM (+0.086) comes at roughly 27× the inference latency per review (3.97 ms vs. 0.15 ms) and a training cost of approximately 869 seconds versus 4 seconds on a single T4 GPU. DistilBERT (1 epoch) recovers most of the accuracy gain (macro F1 0.751) at a fraction of RoBERTa's cost (126 seconds to train, 1.99 ms per-review inference). Whether the additional 0.048 macro F1 from RoBERTa justifies its added latency and training cost is a deployment decision that depends on review volume and the cost of a missed negative review, not a modeling question; this table is intended to make that trade-off explicit rather than to resolve it.

A comparison with the earlier transformer configuration further highlights the impact of model improvement. As shown in Figure 5, the fine-tuned DistilBERT model exhibits a higher rate of misclassification involving the neutral class, particularly in cases where sentiment is weakly expressed. These errors are reduced in the improved RoBERTa model, indicating that additional model capacity and extended training contribute to better separation between neutral and polarized sentiment classes. Class-level analysis provides additional insight into model behavior. The improved transformer achieves strong performance across both negative and positive sentiment classes, with F1-scores exceeding 0.83, indicating that strongly expressed sentiment is reliably classified. In contrast, the neutral class achieves a lower F1-score of approximately 0.71, even after model improvements. This disparity reflects the inherent ambiguity of neutral sentiment, which often contains mixed or weakly expressed opinions. As a result, classification difficulty is driven not only by model limitations but also by the structure and subjectivity of the underlying data.

Confusion matrix analysis reinforces these observations. The confusion matrix of the best-performing model, shown in Figure 6, illustrates that misclassifications are concentrated around the neutral class, while direct confusion between negative and positive sentiment remains limited. Across all models, most misclassifications involve the neutral class, with errors primarily occurring between the neutral and polarized classes. Direct confusion between negative and positive sentiment is relatively rare, indicating that models are effective at distinguishing clearly expressed sentiment. However, the boundary between neutral and

non-neutral classes remains difficult to define, as neutral reviews frequently combine both positive and negative elements.

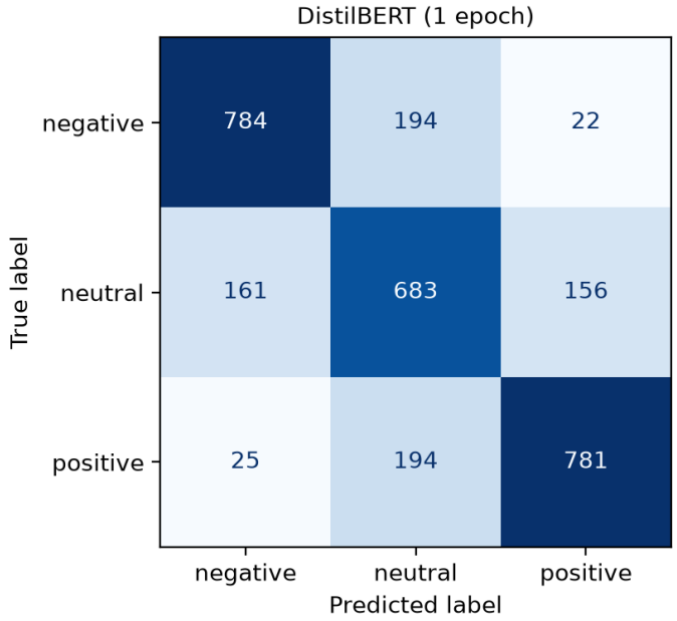

Figure 5. Confusion matrix for the fine-tuned DistilBERT model (1 epoch; regenerated from the corrected pipeline).

Note: Compared with the improved RoBERTa model, DistilBERT exhibits a higher misclassification rate for the neutral class, suggesting that additional model capacity and training contribute to better class separation.

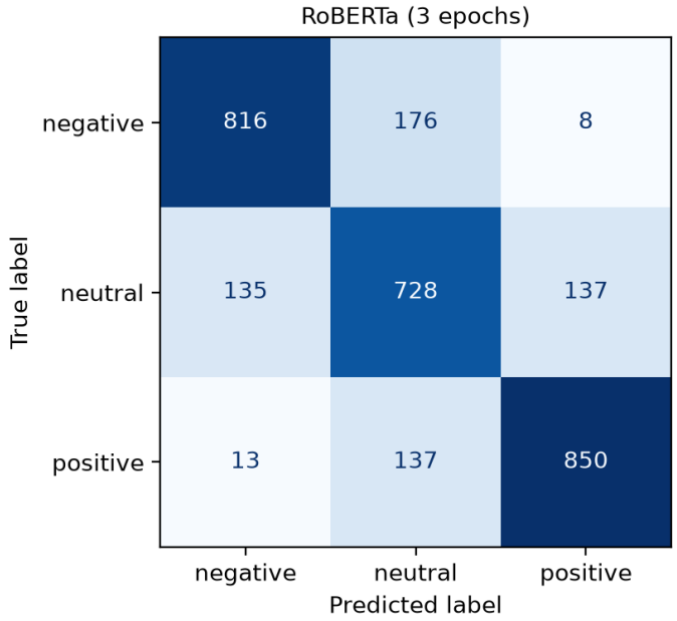

Figure 6. Confusion matrix for the fine-tuned RoBERTa model (improved configuration; regenerated from the corrected pipeline).

Note: The model demonstrates strong performance in classifying negative and positive sentiment, with relatively few direct misclassifications between these classes. Most errors occur between the neutral class and the polarized classes, highlighting the inherent ambiguity of neutral sentiment in customer reviews.

Comparative analysis also reveals distinct model tendencies. The Naive Bayes model tends to predict neutral sentiment more often, reflecting a conservative bias when confidence is low. The Linear SVM reduces this bias by producing more balanced predictions, while

transformer-based models further improve class separation through contextual understanding. Despite these improvements, ambiguity persists in borderline cases, particularly within the neutral category.

Receiver Operating Characteristic (ROC) curves are less informative in this context due to the problem's multi-class formulation and the emphasis on balanced class performance. Instead, evaluation focuses on confusion matrix interpretation and macro-averaged metrics, which provide a more meaningful assessment of model behavior in real-world sentiment classification scenarios.

Taken together, the results illustrate a transition from feature-limited to data-limited performance. While transformer-based models substantially improve classification accuracy by leveraging contextual information, further gains are increasingly constrained by label ambiguity and the inherent subjectivity of customer feedback.

Interpretation and Discussion

The results of this study provide several important insights into both the technical behavior of sentiment classification models and their practical implications for the hospitality industry. Across all models, a consistent pattern emerges: strongly expressed sentiment, whether positive or negative, is relatively easy to classify, while neutral sentiment remains significantly more challenging. This distinction has direct consequences for how sentiment analysis systems should be interpreted and deployed in real-world settings.

From a business perspective, reliably identifying negative sentiment is particularly valuable. Models demonstrate strong performance in detecting clearly dissatisfied customers, which enables organizations to prioritize customer recovery efforts, respond to service failures, and mitigate reputational risk. Automated routing of negative reviews to support teams can improve response time and customer satisfaction, especially in high-volume environments where manual review is impractical. The selective-prediction analysis above gives this claim a concrete operating point: routing roughly a fifth of reviews for manual review is sufficient to capture about 90% of negative sentiment overall, which is the kind of number an operations team can budget against, rather than a general assurance that negative sentiment is "reliably identified." Similarly, accurate classification of positive sentiment supports brand monitoring and marketing initiatives by identifying satisfied customers and reinforcing service attributes that are successful.

In contrast, the neutral sentiment category introduces ambiguity, limiting the interpretability of model predictions. Neutral reviews often reflect mixed or weakly expressed opinions, such as moderate satisfaction alongside minor complaints. The subtype analysis above tested this "mixed vs. weak" distinction directly and found no strong performance split between the two, a result worth noting honestly rather than assuming the mixed-opinion explanation holds simply because it is intuitive. As a result, misclassification between neutral and polarized classes is common across all models. From an operational standpoint, this suggests that neutral predictions should not be treated as definitive indicators but rather as signals of uncertainty. Businesses may benefit from treating neutral reviews as candidates for further analysis, such as manual inspection or follow-up engagement, rather than relying solely on automated classification. The calibration results reinforce this recommendation from a different angle: the model most likely to be deployed for its accuracy (RoBERTa) is also the model whose confidence scores most overstate its correctness, so a naive confidence threshold calibrated by eye would under-route uncertain cases to review.

The comparative performance of the models also provides insight into the role of feature representation. Classical models, including Naive Bayes and Linear SVM, achieve competitive baseline performance but show limited improvement over one another, suggesting that TF-IDF representations limit their ability to capture deeper semantic relationships. The transformer-based model, by contrast, demonstrates a clear advantage,

reflecting its ability to incorporate contextual information and interpret nuanced language patterns. This improvement is particularly relevant in customer feedback scenarios, where meaning is often conveyed implicitly rather than through explicit sentiment keywords.

Despite these gains, the transformer model does not achieve near-perfect performance, even after additional fine-tuning. This outcome highlights a fundamental limitation of the task itself rather than a deficiency in model design. The use of star ratings as proxy labels introduces inherent noise, as numerical ratings do not always align perfectly with the tone or content of the corresponding text. For example, a three-star review may reflect a wide range of experiences, from mildly positive to mildly negative, leading to inconsistencies in labeling. This ambiguity places an upper bound on achievable performance and complicates model evaluation. This claim, that the remaining error reflects an irreducible label ceiling rather than a fixable modeling gap, is currently an inference rather than a measurement. Appendix D proposes a direct test: a small human-annotated sample, labeled from the review text alone with star ratings withheld, would allow this ceiling to be estimated empirically and compared with RoBERTa's observed accuracy.

Another limitation arises from the use of balanced sampling. While this approach ensures fair comparison across sentiment classes, it does not reflect real-world distributions, where positive reviews are typically more prevalent. As a result, model performance in deployment settings may differ from that in experimental settings, particularly in terms of precision-recall trade-offs. Additionally, the study focuses exclusively on textual data and does not incorporate auxiliary information such as user behavior, temporal patterns, or metadata, which could further enhance predictive performance.

Several avenues for improvement emerge from these limitations. First, more granular labeling strategies, such as aspect-based sentiment analysis, could provide deeper insight into specific components of the customer experience, including service quality, pricing, and ambiance. Second, cost-sensitive learning could be introduced to prioritize high-impact errors, such as failing to detect negative sentiment. The selective-prediction analysis above is a step in this direction, using confidence-based abstention rather than retraining with a modified loss. Third, further gains may be achieved through domain-specific fine-tuning of transformer models or by using larger architectures, though such improvements are likely to be incremental given the underlying label ambiguity. Finally, incorporating explainable AI techniques could improve transparency and support decision-making by helping stakeholders understand why certain predictions are made.

Taken together, the findings suggest that while modern transformer models significantly enhance sentiment classification performance, their effectiveness ultimately depends on the quality and structure of the underlying data. In practice, the value of sentiment analysis lies not only in prediction accuracy but also in how model outputs are integrated into business workflows, particularly in managing uncertainty and prioritizing customer interactions.

Conclusion

This study evaluates the effectiveness of classical machine learning and transformer-based models for sentiment classification in the hospitality domain using large-scale customer review data. The results show a clear progression in performance, from classical models achieving macro F1-scores of approximately 0.70–0.71 to transformer-based models reaching 0.75 and, after further fine-tuning, 0.80. These improvements demonstrate the value of contextual and semantic representations in capturing nuanced patterns in customer feedback, leading to more accurate and balanced sentiment predictions. All of these gaps are confirmed statistically significant (bootstrap CIs, pairwise McNemar tests), and the transformer gain is now decomposed into a training-duration component (25%) and a model-capacity component (75%) rather than attributed to both factors jointly.

Despite these gains, the classification of neutral sentiment remains a persistent challenge across all models. Even with enhanced transformer architectures, performance on the neutral class lags behind that of clearly polarized sentiment, reflecting the inherent ambiguity and subjectivity of customer reviews. This finding suggests that model performance is increasingly constrained not by architecture, but by the nature of the data and the limitations of using star ratings as proxy labels. A concrete protocol for testing this label-ceiling hypothesis, rather than treating it as an assumption, is proposed in Appendix D.

From a business perspective, the results reinforce the practical value of sentiment analysis as a decision-support tool. Reliable detection of negative sentiment enables organizations to prioritize customer recovery efforts and address service issues proactively, while accurate identification of positive sentiment supports brand monitoring and strategic positioning. Neutral sentiment, rather than being treated as a definitive category, is best interpreted as a signal of uncertainty that may warrant further analysis or targeted follow-up. This study now quantifies that trade-off: a manual-review budget of roughly one-fifth to two-fifths of incoming reviews raises effective negative-sentiment recall from baseline to between 90% and 96%, giving operations teams a concrete dial rather than a qualitative recommendation.

Effective deployment of sentiment analysis systems, therefore, depends not only on model accuracy but also on alignment with operational objectives. Organizations may benefit from prioritizing recall for negative sentiment or incorporating workflows that account for uncertainty in model predictions. Future work may focus on more granular sentiment representations, such as aspect-based analysis, as well as domain-specific model adaptation and integration of additional contextual data. These directions offer the potential to further enhance both the accuracy and practical utility of sentiment analysis in real-world customer experience management. A further direction, made concrete in Appendix D, is empirically estimating the label-noise ceiling itself, which would clarify how much of the remaining gap between current models and perfect accuracy is closeable at all.

References

Bhowmick, A, Mohammad, A.F., Kalva, S. (2026). Agentic AI & LLM Incorporations with Personas. In: Arabnia, H.R., Deligiannidis, L., Amirian, S., Ghareh Mohammadi, F., Shenavarmasouleh, F. (eds), *AI Revolution: Research, Ethics and Society. AIR-RES 2025. Communications in Computer and Information Science,* vol 2721. Springer, Cham. https://doi.org/10.1007/978-3-032-12313-8_18

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT*) (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML*), 1321–1330.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Liu, B. (2012). *Sentiment analysis and opinion mining.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157.

Navya, G., Atif, M., Rupesh, M, Shravya, K. (2026). AI Leveraging Data Modeling, for Efficient Travel Management. In: Arabnia, H.R., Deligiannidis, L., Amirian, S., Ghareh Mohammadi, F., Shenavarmasouleh, F. (eds), *AI Revolution: Research, Ethics and Society. AIR-RES 2025, Communications in Computer and Information Science,* vol 2722. Springer, Cham. https://doi.org/10.1007/978-3-032-12930-7_10

Appendix A
Data Exploration

Figure A1 (Review Lengths)

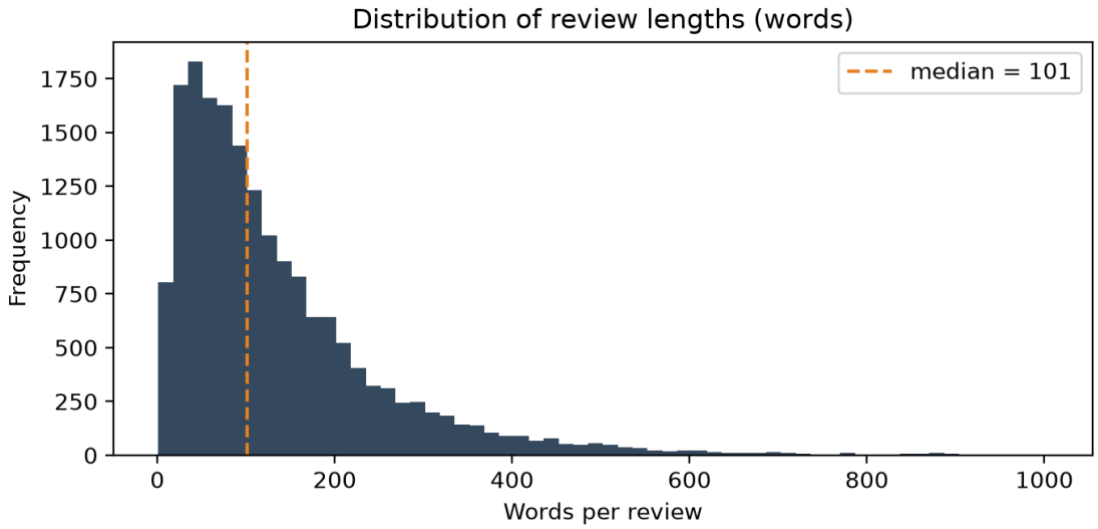

Figure A1. Distribution of review lengths measured in number of words per review, across the full 18,000-review balanced training set (regenerated with the corrected sampling; previously shown for a smaller pool).

Note: The distribution is right-skewed, with most reviews relatively short and a smaller proportion longer. This pattern reflects variability in customer expression and supports the need for preprocessing and truncation strategies in model training.

Figure A2 (Sentiment Distribution)

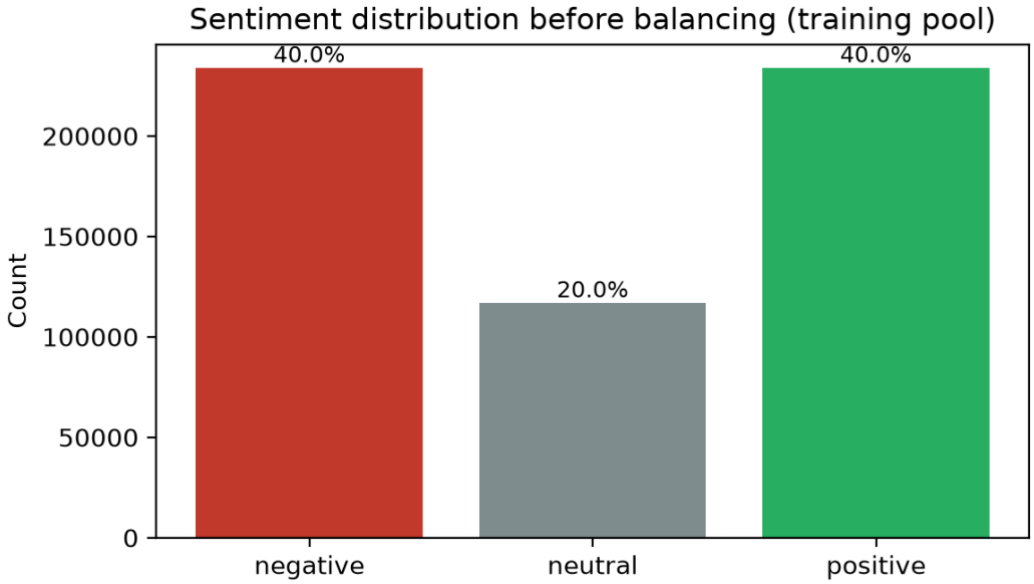

Figure A2. Distribution of sentiment classes in the original (unbalanced) training pool prior to balanced sampling.

Note: The dataset exhibits class imbalance, with positive and negative reviews more prevalent than neutral reviews. This imbalance motivates the use of balanced sampling to ensure fair evaluation across sentiment classes.

Appendix B
Classical Model Performance

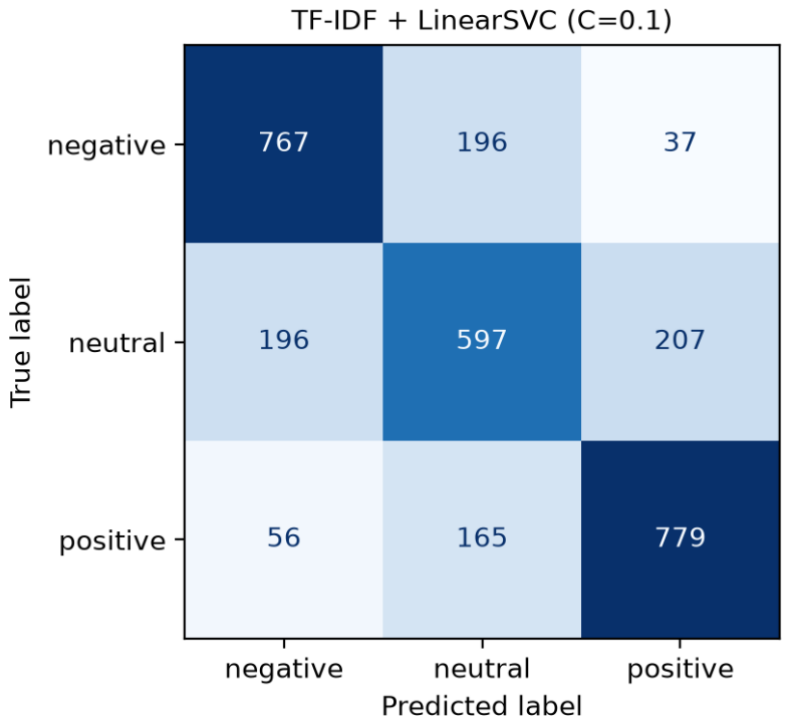

Figure B1. Confusion matrix for TF-IDF + Linear SVM (C = 0.1, validation-tuned; regenerated from the corrected pipeline).

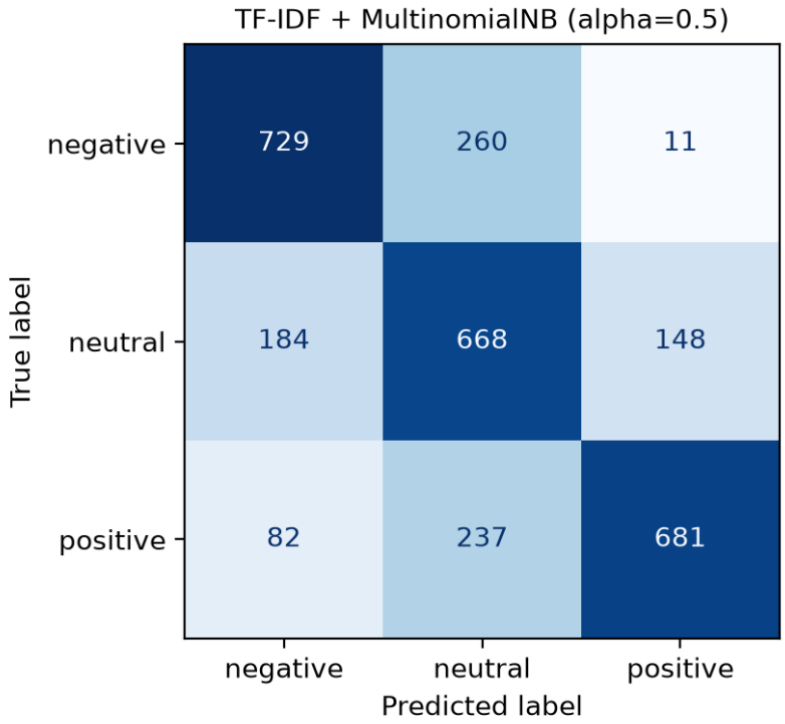

Figure B2. Confusion matrix for TF-IDF + Multinomial Naive Bayes (α = 0.5, validation-tuned; regenerated from the corrected pipeline).

Appendix C

Additional Diagnostic Figures

The figures in this appendix support the ablation and selective-prediction analyses reported in Model Evaluation.

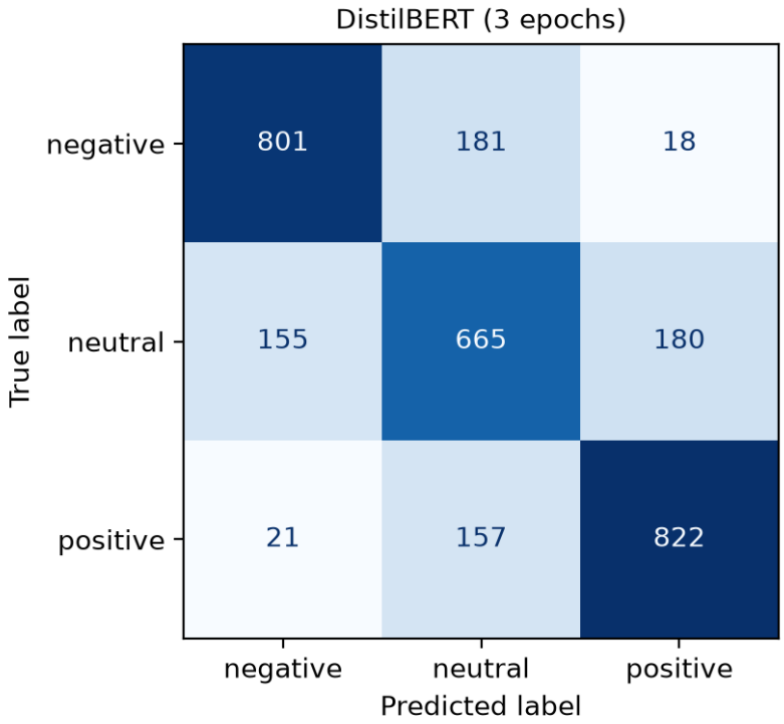

Figure C1. Confusion matrix for the DistilBERT 3-epoch ablation configuration (backbone held fixed; see Ablation: Backbone Capacity vs. Training Duration).

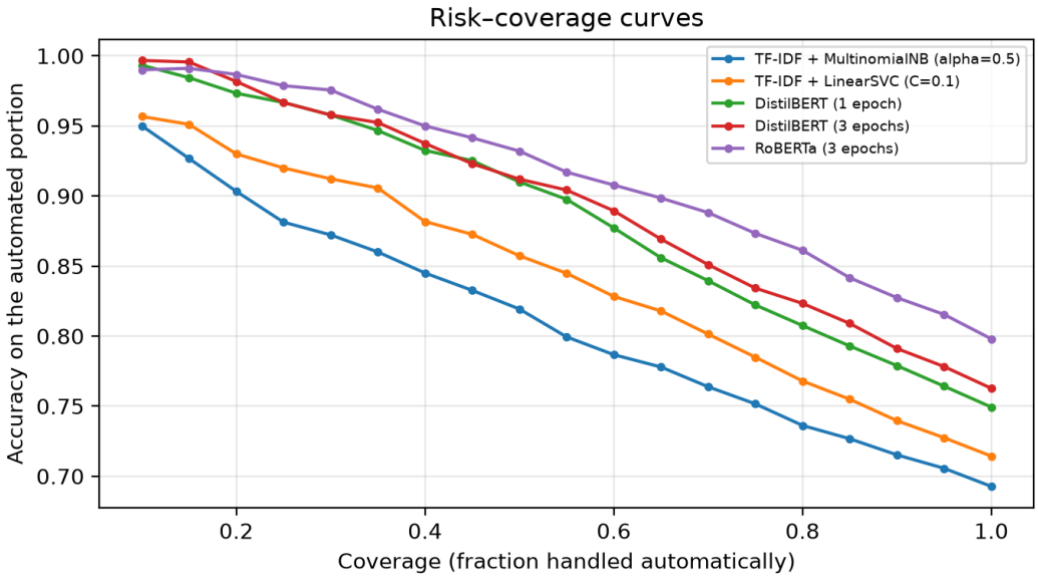

Figure C2. Risk–coverage curves for all five models: accuracy on the automatically handled portion of the test set as the coverage threshold (fraction of reviews retained rather than abstained on) varies.

Appendix D

Annotation Protocol for the Label-Ceiling Study

This appendix specifies a protocol, proposed but not yet executed, for empirically testing the hypothesis raised in Interpretation and Discussion: that star-rating labels impose a noise ceiling on achievable sentiment-classification performance, and that RoBERTa's current accuracy may already be close to that ceiling.

Sample. A stratified sample of 300 test-set reviews (100 per sentiment class, drawn uniformly at random within class) is exported with star ratings withheld.

Annotation. Two to three independent annotators label each review's sentiment (negative / neutral / positive) from the review text alone, with no access to the star rating or to model predictions.

Agreement and consensus. Pairwise inter-annotator agreement is computed (Cohen's kappa); disagreements are adjudicated to a single consensus label per review.

Ceiling estimate. Consensus-label agreement with the star-derived label is computed. This agreement rate is the empirical estimate of the label ceiling: the best possible accuracy any model could achieve against star-derived labels, assuming human sentiment judgment from text is the ground truth.

Comparison. The best model's test accuracy (RoBERTa, 79.8%) is compared against this ceiling. If the two are close, the conclusion that further architectural scaling offers limited additional return is empirically supported rather than inferred. If the ceiling is substantially higher than 79.8%, it indicates remaining headroom that a modeling improvement, rather than a labeling limitation, could close.

Materials. A blank annotation sheet (300 items, labels withheld) and a sealed scoring key mapping each item back to its star-derived label and every model's prediction were generated alongside this study's results and are available as supplementary files.