# Sentiment Classification on Noisy Customer Reviews: A Comparative Analysis of Classical and Neural Network Approaches

Elif Şevval Akdeniz
*Universiteit Antwerpen, Faculty of Arts*

January 2026

**Abstract**

This study compares classical machine learning and neural network approaches for sentiment classification on customer reviews containing textual noise. Using 33,396 McDonald's reviews (55.6% noisy), four models were evaluated: Logistic Regression, SVM, Random Forest, and RNN-LSTM with class-weighted loss to address class imbalance. Minimal preprocessing preserved authentic noise patterns to assess robustness. Results show classical models outperform neural networks, with SVM achieving macro F1-score 0.73 and 81.1% accuracy compared to RNN-LSTM's macro F1-score 0.66 and 71.0% accuracy. Despite class weighting enabling neutral sentiment prediction (F1-score 0.39), the neural network underperforms classical models by 7 points in macro F1-score, revealing fundamental limitations in neural network optimization for imbalanced noisy text with limited training data. Performance analysis across four noise levels confirms classical models maintain better robustness.

## 1 Introduction

Sentiment analysis has become essential for understanding customer feedback, enabling businesses to automatically extract insights from large review volumes. However, online platforms present unique challenges. Communication on social media and review platforms differs fundamentally from formal text, characterized by nonstandard language variations, informal expressions, and creative orthography (Barbosa and Feng, 2010). Customer reviews exemplify these patterns through spelling errors, slang, abbreviations, and emoticons—features that substantially affect machine learning performance.

Classical approaches (Logistic Regression, SVM, Random Forest) learn from carefully designed features like TF-IDF, showing strong results on structured datasets (Sarker, 2021). Neural networks, particularly RNN-LSTM architectures, learn directly from raw text, capturing contextual relationships in sequences. While deep learning achieves state-of-the-art results on benchmarks (Wankhade et al., 2022), performance on noisy real-world data with class imbalance remains underexplored. Recent work demonstrates that preprocessing choices significantly impact performance on noisy social media text (Pota et al., 2023), yet systematic comparisons between classical and neural approaches on customer reviews with natural noise remain limited. This study addresses this gap through comparative analysis on McDonald's reviews, examining: How do classical algorithms compare to neural networks in overall accuracy, and how does noise severity differentially affect these algorithmic families?

## 2 Methods

### 2.1 Data and System Architecture

The dataset comprises 33,396 customer reviews from McDonald's locations across the United States, collected from Google Reviews with associated star ratings, timestamps, and location metadata. Reviews exhibit realistic language patterns with mean length of 22.1 words and median of 11 words, reflecting typical customer feedback brevity. Star ratings were converted to three sentiment classes: negative (1-2 stars, 37.5%), neutral (3 stars, 14.4%), and positive (4-5 stars, 48.1%). This distribution reflects realistic customer feedback patterns where positive reviews dominate while neutral constitutes a challenging minority class.

Table 1: Dataset Overview and Properties

| Property | Value | Noise | Count |
|---|---|---|---|
| Total | 33,396 | Typos | 1,830 |
| Noisy | 18,584 | Abbrev. | 14,495 |
| Clean | 14,812 | Slang | 3,033 |
| Mean | 22.1 w | CAPS | 2,188 |
| Median | 11 w | Repeat | 2,821 |
| *Sentiment: Neg 37.5% \| Neu 14.4% \| Pos 48.1%* | | | |

The sentiment classification system follows a supervised learning pipeline specifically designed to preserve textual noise for robustness evaluation, as illustrated in Figure 1. Raw reviews undergo automated noise detection using five indicators: typos identified via dictionary comparison (5.5% of reviews), abbreviations like "thru" and "ur" (43.4%, most prevalent type), slang expressions (9.1%), ALL CAPS words indicating emphasis (6.6%), and repeated characters like "sooooo" (8.4%). Each review receives a total noise score enabling categorization into four noise levels: clean (0 indicators, 44.4%), low (1-2 indicators, 25.0%), medium (3-5 indicators, 15.7%), and high (6+ indicators, 14.9%). This automated annotation ensures objectivity and reproducibility while eliminating inter-annotator agreement concerns. Noise level serves as an analysis dimension rather than prediction target—models predict sentiment, with predictions grouped by pre-labeled noise levels to assess robustness.

### 2.2 Preprocessing and Feature Engineering

Minimal preprocessing was deliberately chosen to preserve authentic noise patterns for robustness testing. The core research question concerns model robustness to naturally occurring textual noise, yet

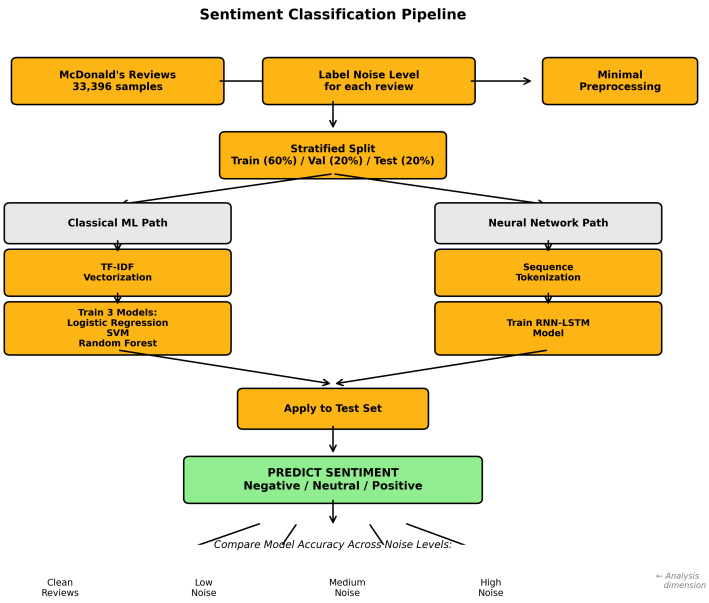


Figure 1: Six-stage sentiment classification pipeline. Models predict sentiment (negative/neutral/positive), with noise levels used as an analysis dimension to evaluate robustness across text quality conditions. Parallel processing paths ensure fair comparison between classical and neural approaches.

standard techniques—stop word removal, stemming, lemmatization, spell correction, slang normalization, punctuation removal—would eliminate the phenomena under investigation. Applying intensive cleaning would invalidate robustness claims, as models would be tested on artificially sanitized text rather than authentic customer reviews. Production sentiment analysis systems process unedited reviews in real-time without human intervention, making a model's ability to handle raw input crucial for operational deployment. To validate this methodological choice, supplementary analysis compared minimal versus aggressive preprocessing strategies, revealing that aggressive cleaning degrades rather than improves performance, particularly for neural networks (Appendix B). This counter-intuitive finding confirms that informal language patterns carry crucial sentiment signals and strengthens the minimal preprocessing approach adopted in this study.

Beyond research alignment, informal language often carries strong sentiment signals that preprocessing would eliminate. The expression "AMAZING!!!" conveys greater intensity than "amazing", while "sooooo good" communicates enthusiasm more effectively than "so good". Emoticons, repeated characters, and capitalization patterns function as sentiment amplifiers in digital communication. Stop words were specifically retained because TF-IDF naturally downweights common words through inverse document frequency, and contextual words like "not" prove crucial for sentiment in short reviews where negation significantly impacts meaning. The preprocessing pipeline consists of lowercase conversion for model compatibility, URL and email removal as non-informative noise, and basic tokenization for neural networks, while critically preserving punctuation, repeated characters, emoticons, slang, typos, abbreviations, and stop words.

Classical models employ TF-IDF vectorization with maximum vocabulary size 5,000 features and bigrams (1,2-grams), measuring term importance by combining local frequency with global rarity. This representation proves particularly robust to orthographic variation because similar sentiment terms receive independent high weights regardless of surrounding noise. Neural networks use sequence tokenization with maximum length 100 tokens, converting text to integer sequences (vocabulary 10,000) that feed into an embedding layer learning dense 128-dimensional vector representations capturing semantic relationships.

### 2.3 Model Selection and Experimental Design

Four models were selected to represent distinct learning paradigms, each chosen for specific theoretical strengths. Logistic Regression serves as the linear baseline with L2 regularization (C=1.0), learning separating hyperplanes in TF-IDF space through maximum likelihood estimation with L-BFGS optimization. Support Vector Machine (LinearSVC) with linear kernel employs margin maximization, finding optimal decision boundaries that maximize class separation while providing strong generalization even with limited training data. Random Forest represents non-linear ensemble methods, combining 100 decision trees (maximum depth 30) trained on bootstrap samples to capture complex feature interactions while reducing overfitting through aggregation.

RNN-LSTM represents deep learning approaches capable of learning sequential dependencies directly from raw text. The architecture consists of an embedding layer (128 dimensions) learning dense vector representations, an LSTM layer (64 hidden units) with memory gates that selectively remember or forget information across sequence positions, and a dense output layer (3 units) with softmax activation for multi-class probability distribution. LSTMs theoretically offer advantages for sentiment analysis by capturing negations, intensifiers, and context-dependent meanings. The architecture employs dropout (0.3) for regularization and trains with categorical cross-entropy loss via Adam optimizer (learning rate 0.001, batch size 32).

To address the class imbalance inherent in the dataset (neutral class 14.4%), the RNN-LSTM employed class-weighted loss function with automatically computed balanced weights inversely proportional to class frequencies. This resulted in weights of 0.889 for

Table 2: Overall Test Set Performance (N=6,680)

| Model | Macro F1 | F1 (wtd) | Acc. | Prec. |
|---|---|---|---|---|
| Baseline | — | — | .481 | — |
| LR | .72 | .799 | .813 | .803 |
| SVM | **.73** | **.801** | .811 | **.800** |
| RF | .67 | .755 | .780 | .807 |
| RNN-LSTM | .66 | .733 | .710 | .772 |

negative, 2.310 for neutral, and 0.693 for positive, ensuring minority class errors receive higher penalties during optimization. Training incorporated early stopping (patience 5 epochs based on validation loss), learning rate scheduling (ReduceLROnPlateau), and gradient clipping (maximum norm 1.0) for stability. Hyperparameter tuning evaluated three configurations across embedding dimensions (128), hidden dimensions (64-96), dropout rates (0.3-0.4), and learning rates (0.0005-0.001), selecting optimal architecture based on validation macro-averaged F1-score to ensure balanced performance across all sentiment classes.

A majority class baseline predicting positive sentiment achieves 48.1% accuracy, establishing the minimum performance threshold. Models are evaluated using macro-averaged F1-score as primary metric to account for class imbalance, treating all sentiment classes equally regardless of their frequency in the dataset. This ensures minority class performance (neutral: 14.4%) substantially influences overall evaluation rather than being masked by majority class accuracy. Complementary metrics include weighted F1-score, accuracy, precision, and recall to provide comprehensive performance assessment. Data splitting employs stratified sampling maintaining class balance: training (20,037 samples, 60%), validation (6,679 samples, 20%), and test (6,680 samples, 20%). Stratification ensures each split reflects original sentiment distribution, preventing class imbalance issues during evaluation.

# 3 Results

Classical approaches substantially outperform neural networks across all evaluation metrics, as shown in Table 2. SVM achieves highest macro F1-score (0.73) with 81.1% accuracy and weighted F1-score 0.801, closely followed by Logistic Regression (macro F1: 0.72, accuracy: 81.3%, weighted F1: 0.799). Random Forest achieves macro F1-score 0.67 with notably high precision (0.807), indicating conservative predictions with few false positives. Despite class-weighted training designed to address minority class handling, RNN-LSTM achieves substantially lower performance with macro F1-score 0.66, weighted F1-score 0.733, and 71.0% accuracy. The performance gap between best classical model (SVM) and neural network spans 7 points in macro F1-score (0.73 versus 0.66), with SVM demonstrating superior precision (0.800 versus 0.772), recall (0.811 versus 0.710), and accuracy (10.1 percentage points higher). All models significantly exceed the baseline, confirming effective learning beyond naive class frequency.

Detailed per-class analysis in Table 3 reveals the mechanisms behind overall performance differences. SVM demonstrates balanced performance across all sentiment classes with F1-scores of 0.85 (negative), 0.47 (neutral), and 0.87 (positive). The neutral class proves challenging even for SVM, reflecting inherent difficulty of identifying reviews with mixed or ambiguous sentiment in 3-star ratings, yet SVM maintains moderate capability across

Table 3: Detailed Per-Class Performance

| Model | Class | Prec. | Rec. | F1 | N |
|---|---|---|---|---|---|
| SVM | Negative | .82 | .88 | .85 | 2,504 |
| | Neutral | .60 | .38 | **.47** | 963 |
| | Positive | .85 | .89 | .87 | 3,213 |
| | Macro | .76 | .72 | **.73** | 6,680 |
| | Weighted | .80 | .81 | .80 | 6,680 |
| RNN-LSTM | Negative | .85 | .74 | .79 | 2,504 |
| | Neutral | .30 | .54 | **.39** | 963 |
| | Positive | .85 | .74 | .79 | 3,213 |
| | Macro | .67 | .67 | **.66** | 6,680 |
| | Weighted | .77 | .71 | .73 | 6,680 |

all classes. RNN-LSTM shows competitive performance on negative (F1-score 0.79) and positive (F1-score 0.79) classes, but exhibits substantially weaker neutral class performance with F1-score of 0.39 despite class-weighted training specifically designed to address minority class handling. The model achieves 54% neutral recall but only 30% neutral precision, indicating aggressive neutral prediction that generates many false positives.

The macro-averaged F1 comparison (SVM: 0.73 versus RNN-LSTM: 0.66) quantifies this 7-point gap, demonstrating that classical models achieve superior balanced performance across all classes. Critically, RNN-LSTM neutral F1-score (0.39) remains substantially below SVM neutral F1-score (0.47), showing that classical models achieve better minority class performance without requiring specialized balancing techniques. The class-weighted RNN-LSTM exhibits balanced recall across classes (negative 0.74, neutral 0.54, positive 0.74) but poor neutral precision (0.30), suggesting the model overcorrects by aggressively predicting neutral, generating many false positive neutral predictions that degrade precision on other classes and contribute to lower macro F1-score.

Figure 2 presents confusion matrices comparing SVM and RNN-LSTM. SVM's confusion matrix shows balanced errors with primary confusions between neutral and positive sentiments (270 and 324 misclassifications), representing reasonable uncertainty on ambiguous 3-star reviews. The strong diagonal (2199, 369, 2852 correct predictions) demonstrates effective classification across all three categories. RNN-LSTM's confusion matrix reveals successful neutral prediction (518 correct predictions, 53.8% recall) but substantial performance degradation on majority classes. Negative class correct predictions total 1852 (74.0% recall), and positive class correct predictions total 2370 (73.7% recall), both substantially lower than SVM's performance. The model exhibits scattered predictions across all classes with substantial off-diagonal errors, indicating confusion between categories. This pattern demonstrates that class-weighted training enables minority class prediction but introduces performance tradeoffs that degrade overall accuracy and macro F1-score.

Performance patterns stratified by noise condition reveal how models maintain accuracy as text quality degrades, directly addressing the second research question. Table 4 shows all models achieve peak performance on clean reviews, with classical models reaching approximately 86% accuracy. Classical models maintain superior robustness across all noise conditions, with the SVM-RNN performance gap ranging from 7.4 percentage points on clean text to 6.0 points on high-noise text. The RNN-LSTM shows substantial performance degradation on low-noise (61.3%) and medium-noise (61.5%) conditions compared to clean text (78.4%), suggesting the class-weighted model struggles particularly with subtle mixed noise patterns. Counter-intuitively, most models demon-

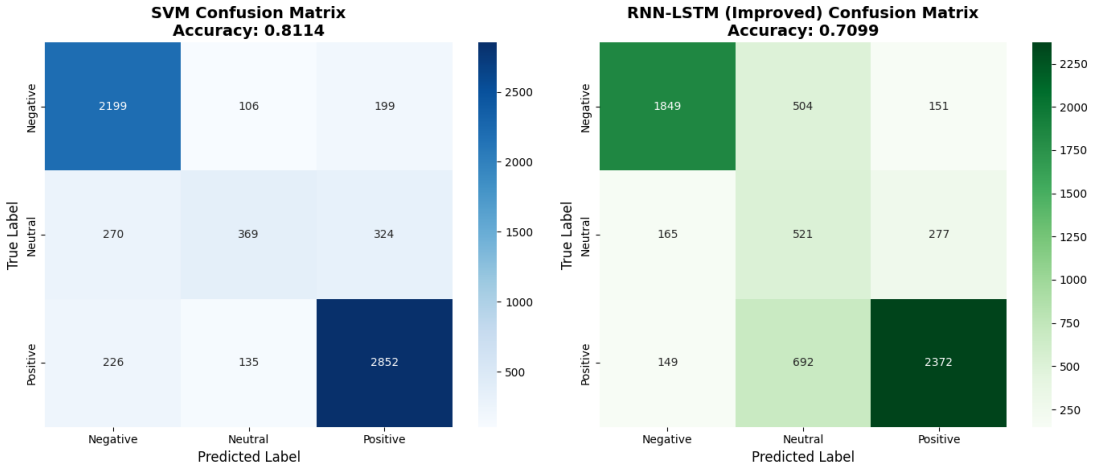


Figure 2: Confusion matrices comparing SVM (left, 81.1% accuracy, 0.73 macro F1) and RNN-LSTM (right, 71.0% accuracy, 0.66 macro F1). SVM shows balanced predictions with strong diagonal and reasonable off-diagonal errors across all classes. RNN-LSTM successfully predicts neutral sentiment (518/963 correct, middle diagonal) through class-weighted training but exhibits scattered predictions and substantial off-diagonal errors, resulting in 7-point lower macro F1-score compared to SVM.

Table 4: Model Accuracy Stratified by Noise Level

| Level | N | LR | SVM | RF | RNN |
|---|---|---|---|---|---|
| Clean | 2,938 | .858 | .858 | .813 | .784 |
| Low | 1,635 | .755 | .755 | .728 | .613 |
| Medium | 1,062 | .783 | .772 | .744 | .615 |
| High | 1,045 | .810 | .809 | .805 | .749 |

strate improved accuracy on high-noise versus medium-noise reviews (SVM: 80.9% versus 77.2%; RNN-LSTM: 74.9% versus 61.5%), suggesting that emphatic noise amplifies rather than obscures sentiment signals.

# 4 Discussion

Several factors explain classical model superiority over neural networks on this dataset. TF-IDF representations effectively capture sentiment-bearing terms despite orthographic variation because each term receives independent importance weighting—even with surrounding typos or irregular capitalization, words like "terrible", "amazing", and "worst" maintain high scores. The relatively short review length (median 11 words) provides insufficient context for LSTM's sequential learning advantages to materialize, as classical bag-of-words approaches work effectively for short texts where global word presence matters more than sequential dependencies. The dataset size (20,037 training samples) proves insufficient for optimal neural network training, particularly for the minority neutral class (2,888 training samples), while classical models efficiently learn from this scale due to simpler parameter spaces.

The RNN-LSTM results reveal fundamental limitations in neural network optimization for imbalanced data with limited samples. Despite class-weighted training with neutral errors receiving 3.33× higher penalties, the model achieves neutral F1-score of only 0.39 compared to SVM's 0.47 without specialized techniques, contributing to a 7-point gap in macro F1-score (0.66 versus 0.73). The class weighting introduces a severe accuracy-balance tradeoff, with the model achieving balanced recall across classes (negative 0.74, neutral 0.54, positive 0.74) but poor neutral precision (0.30) and overall accuracy of 71.0%—10.1 percentage points below SVM. This demonstrates that the network lacks sufficient capacity or training data to simultaneously optimize for all classes when minority class errors receive higher penalties. The model overcorrects by aggressively predicting neutral, generating false positive neutral predictions that degrade precision on other classes. The 7-point macro F1 gap between class-weighted RNN-LSTM (0.66) and SVM (0.73) confirms that even with specialized balancing techniques, neural networks substantially underperform classical models on this dataset. The counter-intuitive finding that high-noise text proves easier to classify than medium-noise text reveals that noise type matters more than noise quantity. Emphatic noise through repeated characters, ALL CAPS, and multiple punctuation amplifies rather than obscures sentiment signals because exaggerated expressions carry clearer emotional content. In contrast, subtle mixed noise through casual abbreviations, mild typos, and informal language creates greater classification ambiguity because it doesn't systematically amplify sentiment. Supplementary preprocessing comparison (Appendix B) confirms this interpretation, demonstrating that aggressive preprocessing removing these emphatic patterns degrades neural network performance substantially (macro F1: 0.38 to 0.26) while minimally affecting classical models (macro F1: 0.73 to 0.72). This validates the minimal preprocessing approach and demonstrates that classical robustness advantage is fundamental rather than preprocessing-dependent.

For operational customer feedback systems, these findings offer actionable guidance. Classical machine learning, specifically SVM or Logistic Regression, provides optimal macro F1-score (0.72-0.73) and accuracy (81%) without requiring specialized balancing techniques or sensitive preprocessing decisions. Neural networks require not only substantially larger datasets (typically 100k+ samples) but also careful hyperparameter tuning and class-balancing strategies that introduce performance tradeoffs, yet still achieve lower macro F1-score (0.66) compared to classical approaches. The practical implication is clear: for datasets with limited samples (20k-50k) and class imbalance, classical ML offers better balanced performance requiring minimal tuning, while neural networks demand extensive optimization that yields inferior results. This study examines a single domain (fast food reviews) and geographic re-

gion (United States), suggesting cross-domain validation on other review types and multilingual contexts would strengthen generalizability. The RNN architecture is relatively simple compared to state-of-the-art models; more sophisticated architectures like Transformers or BERT might narrow the performance gap but require even larger datasets and computational resources. Future research should explore larger datasets (100k+ samples) to determine if neural network advantages emerge at scale, hybrid approaches combining TF-IDF's robustness with neural sequence modeling, alternative class-balancing techniques including focal loss and SMOTE, and noise-type-aware preprocessing strategies that distinguish emphatic from obscuring noise.

# 5 Conclusion

This study demonstrates that classical machine learning substantially outperforms neural networks for sentiment classification on noisy customer reviews. SVM achieves 0.73 macro F1-score and 81.1% accuracy compared to RNN-LSTM's 0.66 macro F1-score and 71.0% accuracy. The neural network particularly struggles with neutral sentiment (F1: 0.39 vs SVM: 0.47), showing poor precision (0.30) despite class-weighted training. Classical models maintain superior robustness across all noise levels, while RNN-LSTM shows particular weakness on low and medium-noise text (61% accuracy). These findings challenge assumptions that neural networks handle informal text better. For this customer feedback dataset with 55.6% noisy text, 20,037 training samples, and class imbalance (14.4% minority class), classical approaches provide optimal performance. Even with class-weighted training, early stopping, and hyperparameter tuning, neural networks underperform by 7 points in macro F1-score, revealing fundamental limitations in this data regime. The results offer actionable guidance: prioritize classical ML for customer review analysis with limited training data and class imbalance. Neural networks might require substantially larger datasets and extensive tuning yet yield inferior balanced performance. This research contributes to understanding how different learning paradigms handle linguistic irregularities and class imbalance in user-generated content, informing deployment decisions for operational sentiment analysis systems.

**Note on AI Assistance**: LLM were used for code debugging and grammatical refinement during paper preparation.

# References


Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010. URL https://dl.acm.org/doi/pdf/10.1145/2065023.2065035.

Marco Pota, Fiammetta Marulli, Massimo Esposito, Giuseppe De Pietro, and Hamido Fujita. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Systems with Applications*, 215:119352, 2023. doi: 10.1016/j.eswa.2022.119352. URL https://arxiv.org/pdf/2304.06934.

Iqbal H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(160), 2021. doi: 10.1007/s42979-021-00592-x. URL https://www.researchgate.net/publication/342890321_Machine_Learning_A_Review_of_Learning_Types.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55:5731–5780, 2022. doi: 10.1007/s10462-022-10144-1.

# Appendix A: Detailed Noise Distribution Analysis

Figure 3 presents comprehensive noise distribution patterns across the dataset. The distribution across noise levels shows 44.4% clean reviews and 55.6% containing various noise types, validating the dataset's ecological validity for robustness testing. Abbreviations dominate at 43.4% of reviews, far exceeding other noise types (typos 5.5%, slang 9.1%, ALL CAPS 6.6%, repeated characters 8.4%), indicating this is the most common form of informal language in customer reviews and explaining why medium-noise reviews prove challenging to classify.

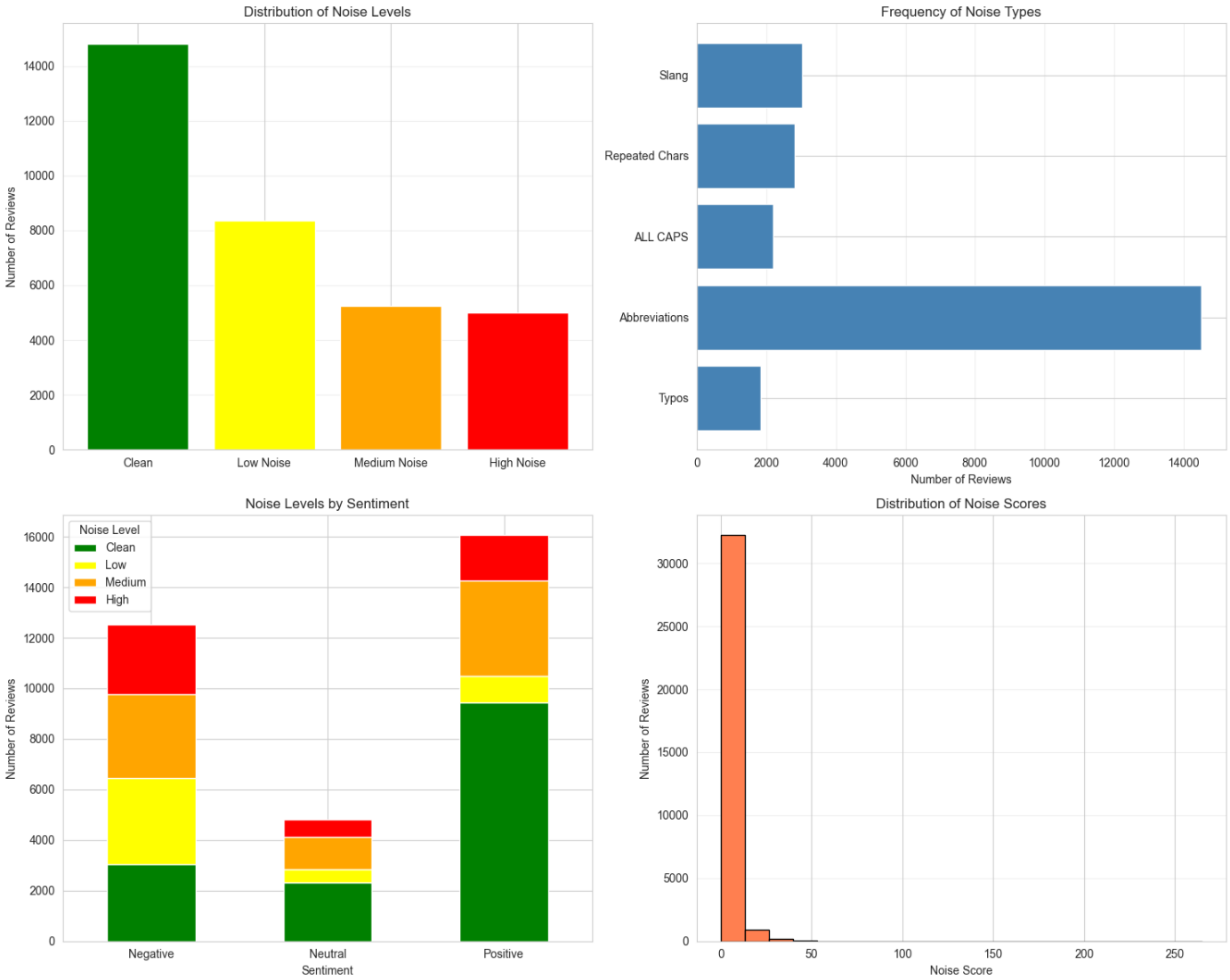


Figure 3: Four-panel noise distribution analysis: (top-left) distribution across four noise levels showing 44.4% clean and 55.6% noisy reviews, (top-right) frequency of five noise types with abbreviations most prevalent at 43.4%, (bottom-left) noise level distribution by sentiment class showing positive reviews contain more varied noise patterns, (bottom-right) histogram of continuous noise scores with peak at zero and gradual decline.

The sentiment-stratified noise analysis reveals that positive reviews demonstrate greater noise diversity with more high-noise examples compared to negative reviews. This pattern may reflect that satisfied customers write enthusiastically using informal, expressive language while dissatisfied customers write more formally to be taken seriously. The continuous noise score distribution shows a clear peak at zero (clean reviews) followed by gradual decline rather than abrupt separation, confirming that natural text contains a spectrum from pristine to extremely noisy rather than discrete categories. Multiple noise types frequently co-occur, as evidenced by many reviews with 5+ indicators, creating the high-noise category where emphatic features combine to amplify sentiment signals.

# Appendix B: Preprocessing Strategy Comparison

To validate the minimal preprocessing approach and assess whether performance differences arise from preprocessing choices rather than inherent algorithmic characteristics, supplementary analysis compared minimal versus aggressive preprocessing strategies. The aggressive preprocessing pipeline applied stopword removal, punctuation removal, repeated character normalization (e.g., "sooooo" → "so"), and short word removal (<3 characters), while minimal preprocessing preserved these elements as described in the main text.

All models were retrained using identical hyperparameters and class-balancing strategies on both preprocessing versions. The aggressive preprocessing removed approximately 40% of tokens per review through stopword elimination and normalization, while minimal preprocessing preserved authentic noise patterns including emoticons, emphatic punctuation, and repeated characters that function as sentiment amplifiers.

Table 5: Preprocessing Strategy Impact on Test Set Performance

| **Model** | **Minimal** | **Aggressive** | $\Delta$ |
|---|---|---|---|
| Logistic Regression | .814 | .809 | −.005 |
| SVM | .811 | .807 | −.004 |
| Random Forest | .781 | .779 | −.002 |
| RNN-LSTM | .705 | .646 | **−.059** |
| *Performance Gap (Best Classical - RNN)* | | | |
| Minimal | 10.8 points | | |
| Aggressive | 16.2 points | | (+5.4) |
| *Neutral Class F1-Score* | | | |
| SVM | .467 | .457 | −.010 |
| RNN-LSTM | .377 | .256 | **−.121** |

Results in Table 5 reveal counter-intuitive patterns. Aggressive preprocessing degrades performance for all models, with neural networks showing substantially greater sensitivity. Classical models demonstrate minimal performance change (SVM: 81.1% → 80.7%, −0.4%), while RNN-LSTM exhibits severe degradation (70.5% → 64.6%, −5.9%). The performance gap between best classical model and neural network widens from 10.8 to 16.2 percentage points under aggressive preprocessing, confirming that classical robustness advantage is fundamental rather than preprocessing-dependent.

Neutral class performance reveals particularly striking differences. SVM neutral F1-score shows minimal change (0.467 → 0.457, −0.010), while RNN-LSTM exhibits catastrophic degradation (0.377 → 0.256, −0.121). Aggressive preprocessing removes emphatic noise patterns (repeated characters, capitalization, punctuation) that serve as sentiment amplifiers, disproportionately affecting the minority neutral class where such signals are already sparse. This 12-point neutral F1 collapse demonstrates that neural networks not only underperform on minority classes but also show extreme sensitivity to preprocessing choices that eliminate subtle contextual cues.

These findings strengthen three core conclusions: First, classical models maintain robust performance regardless of preprocessing strategy, requiring minimal tuning for production deployment. Second, neural networks exhibit fragility to preprocessing choices, with aggressive cleaning degrading rather than improving performance, particularly on minority classes. Third, the performance gap between classical and neural approaches is fundamental rather than preprocessing-dependent, with aggressive preprocessing widening rather than narrowing the gap. This counter-intuitive result—that "cleaner" data yields worse neural network performance—demonstrates that informal language patterns, emoticons, and emphatic punctuation carry crucial sentiment signals that standard NLP preprocessing practices inadvertently eliminate.