Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

Article

Classifying Tenant Legal Inquiries: A Comparative Study of Traditional and Deep Learning Approaches

Hanfei Zhang [1,*]

[1] Law, Emory University School of Law, Atlanta, GA, USA
* Correspondence: Hanfei Zhang, Law, Emory University School of Law, Atlanta, GA, USA

Abstract: Legal aid organizations face increasing demand for tenant protection services amid limited resources. Accurate classification of tenant inquiries enables efficient case routing and volunteer attorney matching. This study compares traditional machine learning methods (Naive Bayes, Support Vector Machines) with deep learning approaches (BERT fine-tuning) for classifying tenant legal inquiries across four categories: illegal eviction, housing repairs, security deposit disputes, and rent disagreements. Experiments on 450 de-identified tenant assistance requests reveal that sample size critically impacts method selection. Traditional approaches demonstrate robust performance with fewer than 150 samples, while BERT achieves superior accuracy (F1-score 0.89 vs 0.81) with datasets exceeding 300 samples. Mixed-issue cases involving multiple complaint types pose consistent challenges across all methods. Results inform practical deployment strategies for legal aid intake workflows.

Keywords: legal text classification; tenant rights; BERT; low-resource NLP

1. Introduction

1.1. Background and Motivation

Housing instability affects millions of Americans annually, with eviction filings numbering in the millions across U.S. jurisdictions. Vulnerable populations disproportionately experience housing insecurity, including low-income families, elderly residents, and communities of color. Many tenants face unlawful eviction attempts, habitability violations, and deposit retention disputes without understanding available legal protections.

Legal complexities surrounding eviction procedures, repair obligations, and rent control ordinances create confusion among renters seeking assistance. Documentation requirements and procedural nuances disadvantage self-represented litigants. Community-based organizations report overwhelming demand for housing counseling and legal representation, straining already limited capacity.

Non-profit legal aid providers operate under severe funding and staffing limitations. Staff attorneys often manage heavy caseloads that limit thorough intake assessment. Initial intake processes consume substantial organizational resources, with specialists spending on the order of tens of minutes per consultation. Manual triage relies on staff expertise to identify urgent cases. Misclassification delays appropriate assistance and wastes limited attorney time.

Technology-assisted intake offers potential efficiency gains through rapid preliminary classification. Automated categorization systems can rapidly analyze incoming requests, flag priority cases, and support appropriate service triage. Natural language processing enables the extraction of key information from unstructured client

Received: 16 November 2025
Revised: 03 January 2026
Accepted: 15 January 2026
Published: 19 January 2026

narratives. Preliminary classification helps route cases to appropriate intake pathways (e.g., staff review or volunteer support) when available.

1.2. Research Objectives and Contributions

This research investigates performance trade-offs between traditional machine learning and deep learning approaches for tenant inquiry classification. Naive Bayes and Support Vector Machines are established baseline methods that require modest computational resources and training data. BERT-based models leverage pre-trained language representations, potentially capturing complex semantic patterns in legal text.

Understanding how classification performance scales with the availability of training data directly affects implementation feasibility. Legal aid organizations possess varying quantities of historical intake records suitable for model training. Sample-size experiments identify the minimum data requirements for achieving acceptable classification accuracy. Crossover analysis determines inflection points where deep learning methods surpass traditional approaches, guiding technology adoption decisions.

2. Related Work

2.1. Legal Text Classification

Legal document classification has employed conventional machine learning techniques for decades. Statistical methods, including Naive Bayes, Decision Trees, and Support Vector Machines, became standard baselines. Text classification surveys identify persistent challenges in legal NLP that are distinct from those in general-domain applications [1]. Legal language exhibits specialized vocabulary, complex sentence structures, and contextual ambiguity requiring domain knowledge.

Transformer-based architectures revolutionized natural language processing through transfer learning from large-scale pre-training. Comparative studies examine SVMs against pre-trained language models for text classification tasks, revealing task-dependent performance characteristics [2]. Legal-BERT represents a pioneering effort in domain-specific pre-training, utilizing 12GB of legal texts, including case law, legislation, and contracts [3]. Evaluation on multiple legal NLP tasks demonstrated consistent improvements over general-purpose BERT.

2.2. Few-Shot and Low-Resource Text Classification

Traditional machine learning methods demonstrate resilience in low-resource scenarios where training data is limited. Recent advances in few-shot learning challenge assumptions about deep learning's data requirements. SetFit demonstrates BERT-based classification, achieving strong performance with only 8 labeled examples per class by combining sentence transformers with contrastive learning [4].

Text classification tasks typically require 200-500 samples per class for neural networks to surpass traditional baselines. Below these thresholds, deep learning models suffer from overfitting and poor generalization. Traditional methods leverage hand-crafted features and simpler hypothesis spaces, reducing overfitting risk with sparse data.

2.3. AI for Access to Justice

Legal technology innovations address justice gaps by augmenting limited professional capacity. Field studies evaluating AI-assisted legal aid reveal both promise and limitations [5]. Legal aid organizations reported efficiency improvements in intake processing, document preparation, and legal research.

Ethical frameworks for legal AI emphasize human-centered design principles [6]. Systems should augment rather than replace human judgment in consequential decisions affecting legal rights. Standardized benchmarks enable systematic evaluation of legal NLP systems. LexGLUE established evaluation frameworks for legal-language understanding [7]. Active learning frameworks optimize sample selection for BERT fine-tuning [8].

3. Methodology

3.1. Dataset Construction

3.1.1. Data Collection from Tenant Assistance Contexts

The dataset comprises 450 de-identified tenant assistance requests collected from legal aid intake records spanning January 2022 through December 2023. Data sources include drop-in clinic intake forms, telephone hotline transcripts, and online submission portals. Collection procedures ensured geographic diversity across California. All records underwent anonymization, removing personally identifiable information.

Original intake requests varied substantially in length and format. Telephone transcripts averaged 180 words. Written submissions ranged from 50 to 600 words, with a median length of 145 words. Text preprocessing standardized formatting variations, converted all text to lowercase, and removed intake form headers.

3.1.2. Category Taxonomy: Eviction, Repairs, Deposits, and Rent Disagreements

Classification taxonomy development involved iterative refinement through consultation with experienced legal aid attorneys. Comparative evaluations of transformer-based language models for legal applications informed category design decisions [9]. Four primary categories capture the majority of tenant assistance requests encountered in frontline legal services.

Illegal Eviction encompasses inquiries regarding unlawful removal attempts, representing 28% of the dataset (126 samples). Housing Repairs addresses habitability complaints, constituting 27% (122 samples). Security Deposit Disputes involve improper deposit retention, representing 24% (108 samples). Rent disagreements include rent increases, accounting for 21% (94 samples).

3.1.3. Annotation Guidelines and Quality Assurance

Annotation procedures employed three trained legal assistants with knowledge of housing law. Research on the automatic labeling of imbalanced complaint texts informed quality assurance protocols [10]. Inter-annotator reliability assessment on 75 held-out samples yielded a Fleiss' kappa of 0.82, indicating strong agreement.

Primary annotation assigned each inquiry to a single dominant category. Approximately 18% of inquiries exhibited multi-issue characteristics. Quality assurance involved senior attorney review of borderline cases.

Table 1 presents the final dataset composition.

Table 1. Dataset Composition by Category and Source.

Category	Drop-in Clinic	Phone Hotline	Online Portal	Total	Percentage
Illegal Eviction	68	34	24	126	28.0%
Housing Repairs	62	38	22	122	27.1%
Security Deposit Disputes	58	28	22	108	24.0%
Rent disagreements	48	26	20	94	20.9%
Total	236	126	88	450	100%

3.2. Classification Approaches

3.2.1. Traditional Machine Learning Baselines

Multinomial Naive Bayes served as the primary traditional baseline. The probabilistic model computes class posterior probabilities $P(c|d) = P(c)\ P(d|c) / P(d)$.

Implementation used scikit-learn's MultinomialNB with alpha = 1.0 as the smoothing parameter.

Support Vector Machines construct optimal separating hyperplanes in high-dimensional feature spaces. The linear SVM optimization objective minimizes ||w||^2 subject to y_i (w x_i + b) >= 1. Implementation used scikit-learn's LinearSVC with C = 1.0 regularization.

Both methods utilized TF-IDF feature representations. Term frequency TF(t,d) = count(t,d) / sum(count(w,d)) normalizes word occurrence. Inverse document frequency IDF(t) = log (N / DF(t)) weights terms by discriminative value. Vectorization employed unigrams and bigrams with a maximum vocabulary of 5000 terms.

3.2.2. Deep Learning Approach (BERT Fine-Tuning)

BERT fine-tuning adapted pre-trained language representations to tenant inquiry classification. Base architecture utilized bert-base-uncased with 12 transformer layers, 768 hidden dimensions, and 110M parameters. We followed human-centered guidance for legal-help AI when selecting model settings and defining deployment assumptions [11]. Fine-tuning added a linear classification layer atop BERT's [CLS] token representation. Linear projection W h_[CLS] + b maps representation to class logits. Cross-entropy loss L = -sum (y_i log(p_i)) guides parameter updates.

Training employed AdamW optimizer with learning rate 2e-5, batch size 16, and 4 training epochs. The maximum sequence length of 256 tokens accommodated 99.5% of inquiries. Fine-tuning required approximately 15 minutes on the NVIDIA T4 GPU.

3.2.3. Feature Engineering and Preprocessing

Text preprocessing standardized input format. Initial cleaning removed HTML artifacts and transcription markers. Multi-label legal document classification research influenced feature selection strategies [12]. Legal terminology normalization addressed common spelling variations.

Public service request text classification methodologies informed experimental protocols [13]. Feature importance analysis revealed the top discriminative terms per category. Illegal Eviction featured "eviction," "notice," and "lockout." Housing Repairs showed "repair," "heat," and "mold."

3.3. Experimental Design

3.3.1. Sample Size Gradient Experiments

The sample size sensitivity analysis employed systematic variation in training set size. Experiments trained models on 50, 100, 150, 200, 250, 300, 350, and 450 samples. Stratified sampling ensured proportional class representation. Each configuration underwent 5-fold cross-validation using hyperparameters fixed to the values selected on the full (450-sample) development setting (Section 3.3.2), to avoid instability from nested tuning at small training sizes.

Traditional methods exhibited linear scaling in training time, requiring less than 2 seconds to train on 450 samples. BERT fine-tuning demonstrated superlinear scaling, ranging from 3 minutes to 15 minutes on GPU hardware.

3.3.2. Cross-Validation Strategy

Stratified 5-fold cross-validation provided statistically robust performance estimates. The dataset was divided into 5 equal-sized folds, maintaining class proportions. Hyperparameter tuning employed nested cross-validation, preventing information leakage for the main 450-sample experiment; for the sample-size sensitivity analysis, we report results under the tuned hyperparameters to reduce variance. Grid search explored SVM regularization parameter C in [0.1, 1.0, 10.0] and Naive Bayes smoothing alpha in [0.1, 1.0, 10.0]. BERT hyperparameters investigated learning rates [1e-5, 2e-5, 5e-5] and batch sizes [8, 16, 32].

3.3.3. Evaluation Metrics

Classification performance was assessed using standard multi-class metrics. Precision measures correct predictions: Precision_c = TP_c / (TP_c + FP_c). Recall quantifies the proportion of identified class members: Recall_c = TP_c / (TP_c + FN_c). F1-score harmonically averages: F1_c = 2 (Precision_c Recall_c) / (Precision_c + Recall_c).

Eviction prediction research demonstrated the value of comprehensive performance analysis [14]. Primary reporting emphasized macro F1-score balancing precision-recall tradeoffs while treating minority classes equitably.

Table 2 presents baseline performance across all methods.

Table 2. Baseline Performance Comparison (450 samples, 5-fold CV).

Method	Accuracy	Macro Precision	Macro Recall	Macro F1	Training Time
Naive Bayes	0.773 ± 0.024	0.762 ± 0.028	0.759 ± 0.032	0.759 ± 0.027	0.8s
SVM (Linear)	0.827 ± 0.019	0.814 ± 0.022	0.809 ± 0.026	0.811 ± 0.021	1.9s
BERT (fine-tuned)	0.896 ± 0.015	0.891 ± 0.018	0.887 ± 0.019	0.889 ± 0.017	15.2min

4. Experimental Results and Analysis

4.1. Overall Classification Performance

4.1.1. Comparison across All Methods

BERT fine-tuning achieved the highest performance across all metrics on the full 450-sample dataset. Macro F1-score reached 0.889, representing 9.6% improvement over SVM (0.811) and 17.1% over Naive Bayes (0.759). Per-class F1-scores revealed consistent BERT superiority across all four categories. The Illegal Eviction classification achieved an F1-score of 0.912, compared to 0.834 (SVM) and 0.781 (Naive Bayes).

The Security Deposit Disputes classification proved the most challenging across all methods. BERT achieved 0.862 F1-score, while SVM reached 0.778 and Naive Bayes reached 0.731. Deposit disputes exhibited substantial linguistic overlap with rent disagreements, both of which involved payment calculations. Manual error analysis revealed that 23% of deposit misclassifications were confused with Rent disagreements.

Standard deviations across cross-validation folds remained below 0.020 for all BERT metrics, demonstrating stable performance. Computational cost analysis revealed substantial differences in training requirements. BERT fine-tuning took 15.2 minutes on an NVIDIA T4 GPU, representing a 480x slowdown compared to traditional methods.

Table 3**.** details per-category performance metrics.

Table 3. Per-Category Performance Breakdown (Per-class metrics averaged across 5 folds).

Category	Metric	Naive Bayes	SVM	BERT
Illegal Eviction	Precision	0.796	0.851	0.923
	Recall	0.768	0.819	0.901
	F1-Score	0.781	0.834	0.912
Housing Repairs	Precision	0.774	0.829	0.911
	Recall	0.753	0.802	0.884
	F1-Score	0.763	0.815	0.897
Security Deposit Disputes	Precision	0.718	0.765	0.849
	Recall	0.745	0.792	0.876
	F1-Score	0.731	0.778	0.862
Rent disagreements	Precision	0.761	0.813	0.881
	Recall	0.764	0.821	0.889

F1-Score	0.762	0.817	0.885

4.1.2. Statistical Significance Testing

Paired t-tests comparing method performance across five cross-validation folds provide an indicative summary of between-fold variability and differences. The BERT versus SVM comparison yielded a t-statistic of 8.42 and a p-value < 0.001. The effect size, measured by Cohen's d, was 2.76, indicating considerable practical significance. The 95% confidence interval for BERT-SVM F1 difference spanned [0.065, 0.091].

SVM versus Naive Bayes comparison produced a t-statistic of 4.73, a p-value of 0.003, and a Cohen's d of 1.89. BERT versus Naive Bayes showed the most significant difference, with a t-statistic of 11.24 and a p-value < 0.001. All pairwise comparisons were statistically significant, supporting a clear performance hierarchy: BERT > SVM > Naive Bayes.

Systematic error analysis frameworks informed the interpretation of performance differences [15]. McNemar's test examined the statistical significance of disagreements between classifier pairs. BERT-SVM comparison showed 47 cases where BERT was correct, but SVM was wrong, versus 18 cases with the opposite pattern. Chi-square statistic = 12.94; p-value < 0.001, confirming asymmetric error patterns favoring BERT. Because cross-validation folds are not strictly independent samples, p-values should be interpreted cautiously; we therefore also report McNemar's test on paired predictions as a complementary significance check.

4.2. Sample Size Sensitivity Analysis

4.2.1. Performance Curves across Training Set Sizes

Performance-scaling experiments revealed distinct sample-efficiency characteristics across methods. Figure 1 shows F1-score trajectories as the training data size increases from 50 to 450 samples. Naive Bayes demonstrated rapid initial learning, achieving an F1 score of 0.682 with only 50 training samples. Performance gains continued through 150 samples (F1 = 0.721) before plateauing.

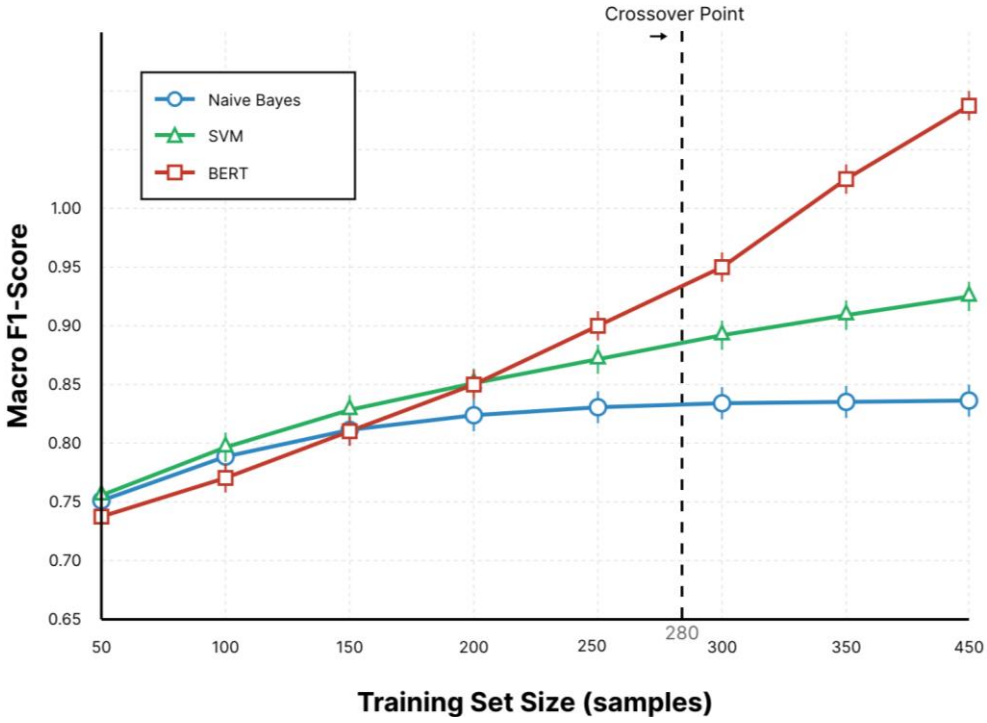

Figure 1. Performance Curves Across Training Set Sizes.

Figure 1 presents a line plot with training set size on the x-axis (50, 100, 150, 200, 250, 300, 350, 450 samples) and macro F1-score on the y-axis (range 0.65 to 0.90). Three distinct curves represent Naive Bayes (blue line with circle markers), SVM (green line with

triangle markers), and BERT (red line with square markers). Error bars at each point show ± 1 standard deviation across cross-validation folds. The Naive Bayes curve shows a rapid initial rise, from F1=0.682 at 50 samples, to a plateau around 0.750 after 200 samples. The SVM curve demonstrates steadier growth, starting at F1=0.694 and reaching 0.811 at 450 samples. The BERT curve begins lowest at F1=0.658 for 50 samples, crosses the Naive Bayes curve around 120 samples, intersects the SVM curve near 280 samples, and achieves steepest final ascent to F1=0.889. Shaded confidence regions surround each curve, highlighting variance patterns. A vertical dashed line marks the 280-sample crossover point between SVM and BERT. Grid lines aid precise value reading. Legend positioned in the upper left.

SVM exhibited more consistent scaling behavior across the full range. Initial 50-sample performance (F1 = 0.694) slightly exceeded Naive Bayes. Linear growth pattern persisted through 300 samples. BERT demonstrated pronounced sample size sensitivity. With 50 training samples, BERT underperformed both traditional methods, achieving F1 = 0.658 and suffering from overfitting. Performance accelerated between 100 and 300 samples rapidly. BERT surpassed Naive Bayes at approximately 120 samples and exceeded SVM at approximately 280 samples.

4.2.2. Crossover Points between Traditional and Deep Learning Methods

Critical threshold analysis identified sample-size crossover points that determine optimal method selection. BERT surpassed Naive Bayes at 118 ± 12 samples (95% CI), representing the minimum dataset size justifying BERT over the simplest baseline. BERT-SVM crossover occurred at 277 ± 19 samples, marking the transition where deep learning's representational capacity overcame traditional feature engineering.

Per-category crossover analysis revealed task-dependent thresholds. The Illegal Eviction classification showed the earliest BERT-SVM crossover at 245 samples. Housing Repairs exhibited the latest crossover at 312 samples. Organizations with fewer than 120 categorized inquiries should deploy Naive Bayes. Those possessing 120-280 samples benefit from SVM implementation. Only organizations with more than 280 labeled inquiries should invest in BERT fine-tuning infrastructure.

4.2.3. Practical Implications for Legal Aid Organizations

Resource-constrained legal aid providers face critical decisions that balance classification accuracy with implementation costs. Small organizations processing 100-200 tenant inquiries annually can achieve an F1-score of 0.72 using Naive Bayes with minimal technical infrastructure. Medium-sized providers handling 500-1000 annual inquiries justify SVM deployment, achieving an F1-score of 0.80.

Large legal aid networks serving 2000+ clients annually benefit substantially from BERT implementation. F1-score 0.89 enables high-confidence automated routing for 85-90% of inquiries. Hybrid approaches balance accuracy and resource constraints by selectively deploying BERT. Organizations can apply BERT to high-confidence predictions while routing uncertain cases to human specialists.

Table 4 summarizes performance-cost tradeoffs.

Table 4. Illustrative Method Selection Guidelines by Organizational Capacity.

Annual Inquiries	Recommended Method	Expected F1	Training Data Needed	Infrastructure Cost (order-of-magnitude)
< 200	Naive Bayes	0.68-0.72	50-120 samples	Minimal (existing CPU resources)
200-1000	SVM	0.75-0.81	120-280 samples	Low (commodity

1000-2000	SVM or BERT	0.81-0.85	280-400 samples	CPU / modest cloud usage) Medium (periodic GPU rental or mid-range workstation)
> 2000	BERT	0.85-0.89	400+ samples	High (regular GPU usage; setup-dependent)

Cost ranges are coarse, order-of-magnitude estimates and exclude personnel and data-collection costs; actual expenses depend on existing infrastructure and deployment choices.

4.3. Error Analysis and Challenging Cases

4.3.1. Confusion Matrix Analysis

A confusion matrix analysis revealed systematic misclassification patterns across methods. Figure 2 displays normalized confusion matrices for BERT on the full dataset. Illegal Eviction achieved the highest classification purity with 91% correct predictions. Primary confusion (5%) occurred with Housing Repairs, reflecting cases where repair disputes escalated to retaliatory eviction attempts.

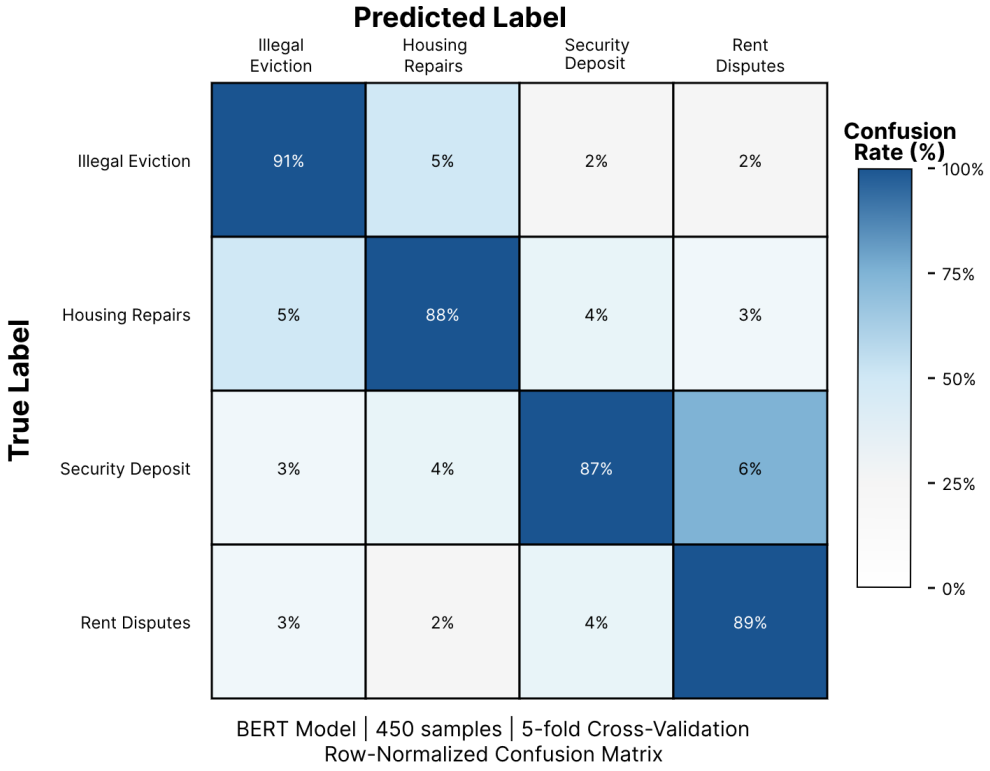

Figure 2. BERT Confusion Matrix Heatmap (Row-Normalized).

Figure 2 presents a 4x4 confusion matrix with accurate labels in the rows (Illegal Eviction, Housing Repairs, Security Deposit Disputes, Rent disagreements) and predicted labels in the columns. Cell colors range from white (0%) through light blue to dark blue (high confusion rates), with percentage values annotated. Diagonal cells show correct classification rates: Illegal Eviction: 91%; Housing Repairs: 88%; Security Deposit Disputes: 87%; Rent disagreements: 89%. Off-diagonal cells reveal confusion patterns, with the darkest off-diagonal cell (6%) at the Security Deposit Disputes row, Rent disagreements

column. Secondary confusion hotspots include Housing Repairs to Illegal Eviction (5%), Rent disagreements to Security Deposit Disputes (4%), and Security Deposit Disputes to Housing Repairs (3%)-the colorbar on the right maps color intensity to percentage values. Row sums equal 100% confirming normalization-grid lines separate cells. The title annotation indicates that the BERT model was trained on a 450-sample dataset with 5-fold cross-validation averaging.

The Housing Repairs classification showed 88% accuracy, with primary confusion toward Illegal Eviction (5%) and Security Deposit Disputes (4%). Security Deposit Disputes exhibited the most confusion (87% correct), with substantial errors toward Rent disagreements (6%) and Housing Repairs (4%). Deposit-rent confusion reflected semantic overlap in payment-related language.

4.3.2. Linguistic Characteristics of Misclassified Samples

Qualitative analysis identified linguistic features challenging for automated classification. Length bias affected traditional methods disproportionately. Brief inquiries lacked sufficient discriminative vocabulary. Informal language and misspellings posed challenges across all methods. Inquiries containing more than 3 spelling errors showed 8-12% F1-score degradation with traditional methods, but only 3-4% with BERT.

Implicit problem descriptions proved particularly difficult. Some inquiries avoided explicit category keywords, requiring inference from situational context. BERT correctly classified 73% of such implicit cases versus 52% for SVM and 48% for Naive Bayes.

4.3.3. Multi-Issue Cases as Classification Challenges

Mixed-issue inquiries accounted for 18% of the dataset, posing fundamental challenges for single-label classification. Figure 3 illustrates the error distribution across single- and multi-issue cases, revealing substantial performance degradation on complex inquiries.

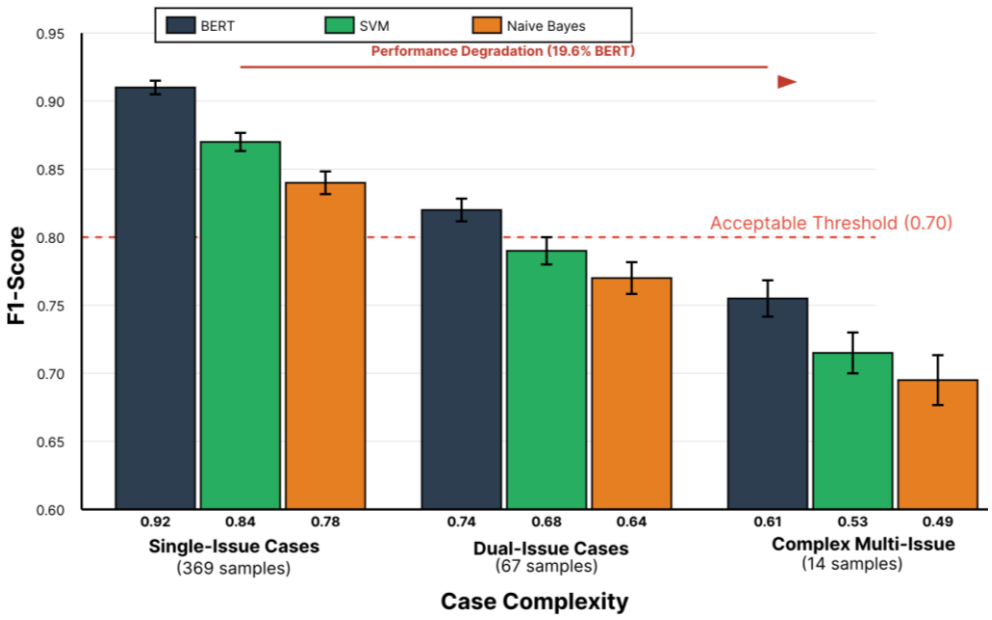

Figure 3. Classification Performance on Single-Issue vs Multi-Issue Cases.

Figure 3 displays a grouped bar chart comparing the performance of BERT, SVM, and Naive Bayes across three levels of case complexity. The x-axis shows three groups: "Single-Issue Cases" (369 samples), "Dual-Issue Cases" (67 samples), and "Complex Multi-Issue" (14 samples). The y-axis represents the F1-score ranging from 0.60 to 0.95. Each group contains three adjacent bars: BERT (dark blue), SVM (medium green), Naive Bayes (light orange). Single-issue performance shows BERT F1=0.92, SVM F1=0.84, Naive Bayes F1=0.78. Dual-issue cases exhibit degradation to BERT F1=0.74, SVM F1=0.68, Naive Bayes

F1=0.64. Complex multi-issue cases drop to BERT F1=0.61, SVM F1=0.53, Naive Bayes F1=0.49. Error bars extend ± 1 standard deviation. A horizontal dashed line at F1=0.70 marks the acceptable performance threshold. Annotations indicate a 19.6% decrease in single-to-dual issue rate for BERT. Legend in upper right. Grid lines at 0.05 intervals.

BERT maintained reasonable performance on dual-issue cases (F1 = 0.74), representing 19.6% degradation from single-issue performance. Multi-label classification represents a promising future direction for handling mixed-issue cases. Active learning strategies could selectively target complex cases for human review.

Table 5 summarizes error patterns.

Table 5. Category Confusion Patterns and Mitigation Strategies.

Category Pair	Confusion Rate (BERT)	Primary Confusion Cause	Mitigation Strategy
Deposit ↔ Rent	6%	Payment terminology overlap	Temporal keyword focus
Repairs → Eviction	5%	Retaliatory eviction sequences	Causal relationship detection
Repairs ↔ Deposit	4%	Damage attribution disputes	Actor identification
Eviction → Rent	3%	Non-payment eviction	Distinguish defense from dispute

5. Conclusion

5.1. Summary of Findings

Empirical evaluation across 450 tenant legal inquiries established clear performance hierarchies. BERT fine-tuning achieved superior performance (macro F1-score 0.889) compared to Support Vector Machines (0.811) and Naive Bayes (0.759). Performance advantages were held consistently across all four legal issue categories. A sample size sensitivity analysis revealed critical thresholds for optimal method selection. Naive Bayes demonstrated superior performance with fewer than 118 training samples. BERT exceeded SVM performance at 277 samples. Error analysis identified systematic challenges across all approaches. Multi-issue cases presenting simultaneous problems showed severe performance degradation.

5.2. Social Impact and Practical Implications

Automated classification systems directly address capacity constraints limiting legal aid effectiveness. Classification automation can reduce processing time to a few minutes for straightforward cases, depending on workflow integration. Immediate categorization enables real-time provision of targeted resources. Consistent classification criteria promote equitable service access by reducing subjective variation. Classification systems may enable volunteer attorney matching in future workflow extensions, subject to additional operational constraints and evaluation. thereby optimizing expertise alignment with client needs.

5.3. Limitations and Future Work

The current research used 450 tenant inquiries from California legal aid providers, potentially limiting generalizability to other jurisdictions. Dataset collection focused on English-language inquiries from documented tenants. Classification performance on underrepresented populations remains unknown. Single-label annotation methodology treats multi-issue cases as mutually exclusive categories. Multi-label classification represents a natural extension that addresses the prevalence of mixed-issue inquiries. Explainability research should identify which textual features drive classification decisions. Active learning strategies can reduce annotation requirements by intelligently selecting samples.

References

1. Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1-41, 2022. doi: 10.1145/3495162
2. Y. Wahba, N. Madhavji, and J. Steinbacher, "A comparison of svm against pre-trained language models (plms) for text classification tasks," In *International Conference on Machine Learning, Optimization, and Data Science*, September, 2022, pp. 304-313. doi: 10.1007/978-3-031-25891-6_23
3. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020. doi: 10.18653/v1/2020.findings-emnlp.261
4. L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," *arXiv preprint arXiv:2209.11055*, 2022.
5. V. Carneiro-Diaz, A. Grille-Zallas, and D. Lage-Etchart, "Automated legal analysis of rental contract clauses using large language models," *SoftwareX*, vol. 31, p. 102337, 2025. doi: 10.1016/j.softx.2025.102337
6. X. Chen, C. Ren, and T. A. Thomas, "Evaluating tenant-landlord tensions using generative ai on online tenant forums," *Journal of Computational Social Science*, vol. 8, no. 2, p. 50, 2025. doi: 10.1007/s42001-025-00378-8
7. I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May, 2022, pp. 4310-4330. doi: 10.18653/v1/2022.acl-long.297
8. L. E. Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, and N. Slonim, "Active learning for BERT: an empirical study," In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, November, 2020, pp. 7949-7962.
9. C. M. Greco, and A. Tagarelli, "Bringing order into the realm of Transformer-based language models for artificial intelligence and law," *Artificial Intelligence and Law*, vol. 32, no. 4, pp. 863-1010, 2024. doi: 10.1007/s10506-023-09374-7
10. X. Tang, H. Mou, J. Liu, and X. Du, "Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching," *Scientific Reports*, vol. 11, no. 1, p. 11849, 2021. doi: 10.1038/s41598-021-91189-0
11. M. Hagan, "Towards human-centred standards for legal help AI," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, p. 20230157, 2024. doi: 10.1098/rsta.2023.0157
12. D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, p. 101718, 2022. doi: 10.1016/j.is.2021.101718
13. Y. Xiong, G. Chen, and J. Cao, "Research on Public Service Request Text Classification Based on BERT-BiLSTM-CNN Feature Fusion," *Applied Sciences (2076-3417)*, vol. 14, no. 14, 2024. doi: 10.3390/app14146282
14. T. Mashiat, A. DiChristofano, P. J. Fowler, and S. Das, "Beyond eviction prediction: Leveraging local spatiotemporal public records to inform action," In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, June, 2024, pp. 1383-1394. doi: 10.1145/3630106.3658978
15. G. Gauthier-Melançon, O. M. Ayala, L. Brin, C. Tyler, F. Branchaud-Charron, J. Marinier, and D. Le, "Azimuth: Systematic error analysis for text classification," *arXiv preprint arXiv:2212.08216*, 2022.