

An Uncertainty-Aware Hybrid Symbolic-Lexicon Framework for Multi-Domain Sentiment Classification

Saleh Yaakub¹, Mustafa Mat Deris², Ika Safitri Windiarti³

¹Faculty of Science, Technology, and Health, Universitas Muhammadiyah Jambi, Jambi, Indonesia

²Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia

³Computer Science, Universitas Muhammadiyah Palangkaraya, Indonesia

Received:

May 29, 2026

Revised:

August 8, 2026

Accepted:

September 15, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Saleh Yaakub

Email*:

saleh@umjambi.ac.id

DOI:

10.63158/journalisi.v8i5.1872

© 2026 Journal of
Information Systems and
Informatics. This open
access article is distributed
under a (CC-BY License)



Abstract. Purpose: Multi-domain sentiment classifiers may return a polarity label even when the evidence behind that decision becomes fragile under domain shift. This study evaluates an uncertainty-aware hybrid symbolic-lexicon framework designed to pair each polarity prediction with a traceable accept-or-defer status. Methods: The framework combines a lexical branch adapted only on training data with training-only fitted symbolic feature states, variable-precision rough evidence, and nested score fusion. Evaluation uses the UCI Sentiment Labelled Sentences dataset, comprising 3,000 balanced English sentences from Amazon Cells, IMDb, and Yelp, under repeated within-source, pooled multi-domain, and leave-one-source-out protocols. Results: In pooled evaluation, the framework achieved 83.73% accuracy and 83.76% F1-score, improving F1 by 0.50 points over the adaptive lexical branch. The boundary mechanism deferred 53.55% of cases and retained 92.54% accuracy among accepted predictions; however, matched-coverage logistic regression and SVM confidence rankings produced lower selective risk. Novelty: The contribution is therefore a reproducible, algorithmically traceable polarity-and-deferment mechanism rather than a claim of universal predictive or uncertainty-ranking superiority.

Keywords: multi-domain sentiment classification, selective prediction, deferment, domain shift, lexicon adaptation, rough-set evidence, algorithmic traceability

1. INTRODUCTION

Sentiment indicators derived from online reviews, service evaluations, and platform comments are increasingly used in customer-service triage, product monitoring, and information-system dashboards. Their apparent simplicity hides a persistent generalisation problem: vocabulary, review length, rating conventions, and evaluative language change across sources, so a classifier that is reliable in one domain may lose discrimination in another. Recent reviews continue to identify domain shift, ambiguity, uncertainty, explainability, and reproducibility as unresolved issues in sentiment analysis [1], [2]. Cross-domain research has therefore moved toward representations and augmentation strategies intended to separate domain-specific cues from sentiment-bearing information [3], [4].

Modern sentiment classification is now dominated by pretrained transformers and, increasingly, large language models. These models can capture contextual dependencies and compositional meaning that sparse lexical representations often miss, and recent work has also explored distillation to reduce the deployment burden of larger models [5], [6]. At the same time, higher model capacity does not remove uncertainty: LLM-based sentiment outputs can vary with prompts, sampling conditions, and distribution shift, while explanation quality remains a separate concern [7]. The framework studied here is therefore positioned as a complementary design rather than a substitute for contextual encoders. Its aim is to make the evidence path behind a polarity decision inspectable and to identify cases that should not be accepted automatically.

Lexicon-based methods remain relevant to this design space because their decisions can be traced to explicit lexical contributions. VADER combines a human-rated sentiment lexicon with rules for intensification, negation, punctuation, capitalisation, and contrast [8]. That portability is useful, but a fixed prior cannot fully represent polarity shifts that emerge within a particular review domain. Barik and Misra showed that domain-specific lexical adaptation can improve review classification [9]. A polarity label alone, however, does not reveal whether the supporting evidence is coherent, contradictory, or too weak to justify an unqualified automated decision.

Rough-set reasoning offers a symbolic language for representing this type of indeterminacy. Classical rough sets distinguish lower approximations, upper approximations, and boundary regions [10], while the variable-precision model allows a controlled level of inconsistency [11]. Direct equivalence over raw TF-IDF vectors is unsuitable for sparse text because continuous values fragment observations into highly specific groups. The proposed framework therefore learns a fold-local three-state transformation—absent, low, or high—from non-zero training values and estimates variable-precision evidence over those states. The purpose is not to recast continuous TF-IDF as inherently categorical, but to create an auditable symbolic layer whose rules can be traced back to training evidence.

Four terms are kept distinct throughout the manuscript. Uncertainty denotes the broader situation in which the available evidence does not justify an unqualified automated decision. Low confidence refers to a small numerical margin or weak class score. Boundary status is the operational deferment flag defined in this study from fused-margin proximity and the amount of consistent symbolic evidence. Neutral sentiment, in contrast, would be a third semantic polarity class; it is not used because the dataset is binary. A boundary case still receives a positive or negative label, but the label is flagged for secondary processing. This accept-or-defer logic is consistent with selective prediction [12], while calibration work shows why ranking scores should not automatically be interpreted as probabilities [13], [14]. Conformal prediction provides a separate family of methods for coverage-aware uncertainty under explicit assumptions [15]; the present boundary is not claimed to provide conformal or probabilistic guarantees.

The intended operational use is a two-stage review pipeline. An observation with sufficiently separated polarity evidence and adequate lower-region support can be accepted for automated aggregation or routine monitoring. A boundary observation can instead be routed to a stronger contextual model, reviewed by a human analyst, or temporarily excluded from a high-stakes aggregate. The correct action depends on the cost of a false automated decision, the capacity and cost of the secondary queue, and the value placed on coverage. This distinction is important because deferment is an operational decision, not an additional sentiment class.

The research gap is therefore more specific than the search for another high-accuracy classifier. Domain-adaptive lexicons provide explicit lexical evidence but usually do not attach a separate evidential status; conventional linear models provide useful margins but comparatively limited symbolic traces; and symbolic approaches are not always evaluated under strict leakage control, multi-scenario transfer tests, and matched-coverage selective baselines. This study addresses that gap by integrating training-only lexical adaptation, explicit symbolic granulation, variable-precision feature-state evidence, nested score fusion, and a separate deferment status. The investigation asks whether this integration remains competitive across known, pooled, and unseen sources, whether the boundary concentrates errors sufficiently to support routing, and whether the resulting evidence path is auditable without overstating predictive or uncertainty-ranking superiority.

2. METHODS

2.1. Research Design and Evidence Boundary

The study used an artefact-oriented experimental design centred on a reproducible multi-source benchmark. All inferential results reported in this article come from the independently executable UCI Sentiment Labelled Sentences suite. Archived values from an earlier Amazon Books, DVD, Electronics, and Kitchen methodological precursor are retained only as historical motivation because their observation-level predictions were not available for independent regeneration. Keeping this evidence sources separate prevents historical benchmark figures from being presented as if they were outputs of the verified experiment reported here.

2.2. Dataset and Experimental Scenarios

The verified dataset is the UCI Sentiment Labelled Sentences collection [16], associated with the work of Kotzias et al. [17]. It contains 3,000 English sentences: 1,000 Amazon mobile-phone review sentences, 1,000 IMDb movie-review sentences, and 1,000 Yelp restaurant-review sentences. Each source is balanced with 500 positive and 500 negative labels. Within-source and pooled experiments used stratified fivefold cross-validation repeated twice with seed 17. Thus, the within-source setting contains 30 outer folds (3 sources $\times$ 5 folds $\times$ 2 repeats), the pooled setting contains 10 outer folds, and leave-one-

source-out evaluation contains three folds, each holding out one complete source. Table 1 summarises the data sources and the three evaluation scenarios.

Table 1. Dataset and evaluation design

Evidence source / scenario	Observations or folds	Labels / split	Role
Amazon Cells	1,000	500 positive; 500 negative	Product-review source
IMDb	1,000	500 positive; 500 negative	Movie-review source
Yelp	1,000	500 positive; 500 negative	Restaurant-review source
Repeated within- source	30 outer folds	5-fold x 2 repeats x 3 sources	Known-source evaluation
Pooled multi-domain	10 outer folds	5-fold x 2 repeats	Mixed-source evaluation
Leave-one-source- out	3 outer folds	One source held out	Unseen-source evaluation

2.3. Leakage-Controlled Processing

Every transformation learned from data was fit only inside the relevant training partition. This restriction applies to the TF-IDF vocabulary and document frequencies, adaptive lexical statistics, non-zero medians used for symbolic cut points, candidate ranking, feature-state granules, fusion weight, and the hyperparameters of the linear baselines. Any inner validation search was therefore nested inside the outer training fold; the outer test fold was never used to construct a vocabulary, rank features, define discretisation thresholds, choose the 600-feature subset, select a fusion weight, or tune C. This protocol addresses train-test contamination, a well-documented threat in machine-learning experiments [18], and follows current reproducibility guidance for empirical ML research [19].

2.4. Adaptive Lexical Scoring

The adaptive lexical branch first estimates class-conditional term evidence from the training fold and combines it with a general VADER prior. Let $n_+(t)$ and $n_-(t)$ denote the numbers of positive and negative training documents containing term t , let N_+ and N_- denote the corresponding class totals, and let $\delta=1$ be the additive-smoothing constant. Equation (1) defines the smoothed training log-odds contribution assigned to term t :

$$\ell(t) = \log \frac{n_+(t)+\delta}{N_++2\delta} - \log \frac{n_-(t)+\delta}{N_-+2\delta}. \quad (1)$$

Equation (2) then combines the fold-specific log-odds with the VADER prior and maps the resulting document score to a bounded interval. Here $v(t)$ is the VADER valence prior, $\kappa=0.35$ is its predeclared scale, and $w(t,x)$ is the sparse term weight for sentence x obtained from a TF-IDF representation fit only on the training data. Adapted lexical weights are clipped to $[-6, 6]$ before aggregation:

$$s_{\text{lex}}(x) = \tanh \left\{ \frac{\sum_t w(t,x) [\kappa v(t) + \ell(t)]}{\epsilon + \sum_t w(t,x)} \right\}. \quad (2)$$

This branch operationalises domain-adaptive lexical classification while preserving a clear evidence trail. It should not be interpreted as an exact reproduction of the implementation in [9], because the complete machine-readable settings of that study are not available. The present implementation instead specifies its own training-only adaptation so that the reported evaluation can be regenerated.

2.5. Symbolic Discretisation, Dependency Ranking, and Granules

The symbolic branch starts from the TF-IDF matrix fitted inside the outer training fold. At most 1,500 candidate features enter symbolic screening. For each feature j , the median m_j is computed from its non-zero values in the training partition only. Equation (3) converts each continuous feature value into one of three states—absent, low, or high—without consulting validation or test observations:

$$q_j(x_j) = 0 \text{ if } x_j = 0; \ 1 \text{ if } 0 < x_j \leq m_j; \ 2 \text{ if } x_j > m_j. \quad (3)$$

For each state r in $\{0,1,2\}$, the feature-state granule is $G_{jr}=\{i:q_j(x_{ij})=r\}$. If $n_{jr,c}$ is the number of training observations in G_{jr} assigned to class c , Equation (4) defines the majority-class confidence and its corresponding inconsistency:

$$\text{conf}_{jr} = \frac{\max_c n_{jr,c}}{|G_{jr}|}, \quad \text{err}_{jr} = 1 - \text{conf}_{jr}. \quad (4)$$

The variable-precision thresholds are fixed before outer-fold testing. A state contributes lower-region evidence only when its inconsistency is at most 0.20 and its majority-class confidence is at least 0.80. Equation (5) defines the single-feature dependency score as the proportion of training observations covered by states satisfying both criteria:

$$D_j = \frac{1}{n} \sum_r |G_{jr}| I(\text{err}_{jr} \leq 0.20 \wedge \text{conf}_{jr} \geq 0.80). \quad (5)$$

Candidates are ranked by D_j , with chi-square association used only as a deterministic tie-breaker inside the same training partition. The principal configuration retains at most 600 ranked features. This cap was predeclared as a controlled compression setting rather than selected by outer-test performance; Section 3.4 reports sensitivity results showing that it is not a unique optimum. For a matched feature-state, the archived implementation assigns a signed contribution $a_{jr}=\text{sigma}_{jr} h(\text{conf}_{jr},|G_{jr}|)$, where sigma_{jr} is the majority-class sign and h is the non-negative confidence-support weighting implemented in the reproducibility archive. Equation (6) defines the normalised symbolic score, and Equation (7) defines $L(x)$, the share of matched symbolic evidence originating from lower-region states:

$$s_{\text{rough}}(x) = \frac{\sum_j a_{j,q_j(x_j)}}{\epsilon + \sum_j |a_{j,q_j(x_j)}|}. \quad (6)$$

$$L(x) = \frac{\text{lower-region symbolic evidence mass}}{\text{total matched symbolic evidence mass}}. \quad (7)$$

Equations (6) and (7) separate direction from evidential status: $s_{\text{rough}}(x)$ contributes to polarity, whereas $L(x)$ quantifies how much matched symbolic evidence comes from sufficiently consistent training granules. The generic $h(\cdot)$ notation is retained because the archived source code, rather than an independently documented closed-form coefficient,

is the authoritative specification of the confidence-support weighting. No undocumented coefficient is introduced.

2.6. Fusion and Operational Boundary

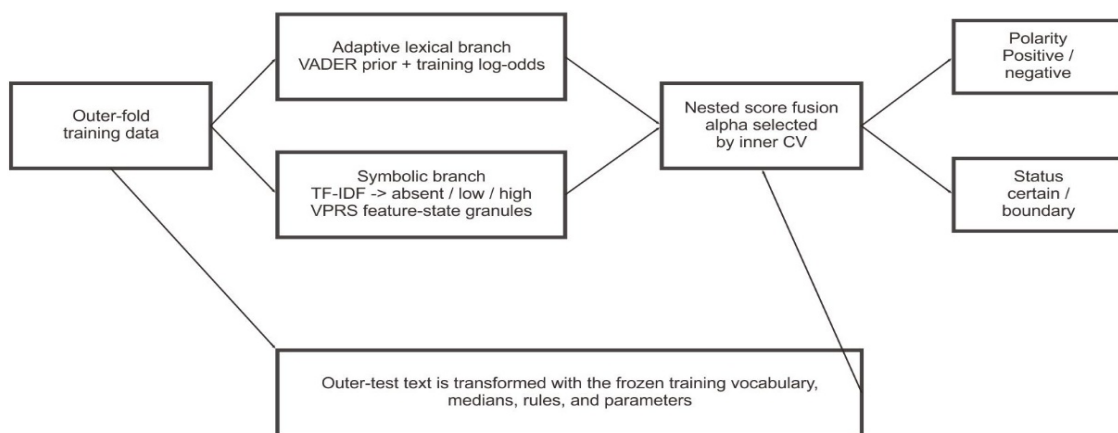
The two branches are fused only after their scores have been placed on bounded scales. Within each outer training fold, alpha is selected by stratified threefold inner validation from the fixed grid {0.0, 0.1, ..., 1.0}; the outer test fold does not participate in this search. Equation (8) defines the convex fused score, while $\alpha=0.50$ is retained as the predeclared equal-weight ablation:

$$s_{\text{fuse}}(x) = \alpha s_{\text{lex}}(x) + (1 - \alpha) s_{\text{rough}}(x). \quad (8)$$

Polarity is assigned from the sign of $s_{\text{fuse}}(x)$: non-negative scores are positive and negative scores are negative. The deferment status is computed only after polarity has been assigned. Two fixed operational cut points are used: a lower-evidence share threshold of 0.15 and an absolute fused-margin threshold of 0.10. Equation (9) marks an observation as boundary when either source of evidence is insufficient:

$$u(x) = \text{boundary if } L(x) < 0.15 \text{ or } |s_{\text{fuse}}(x)| < 0.10; \text{ certain otherwise.} \quad (9)$$

Both cut points were predeclared before outer-fold testing. They do not create a neutral class and they do not alter the predicted polarity. Figure 1 illustrates the full training-only workflow, including the nested tuning stage and the separate placement of the operational boundary after score fusion.



Boundary is an operational deferment flag: it does not replace or alter the binary polarity label

Figure 1. Training-only workflow of the hybrid symbolic-lexicon framework

2.7. Baselines, Hyperparameters, and Score Interpretation

Seven models use identical outer partitions: VADER, the adaptive lexical branch, rough-only classification, equal-weight static fusion, the proposed nested fusion, logistic regression, and a linear support vector machine. The sparse representation uses word unigrams and bigrams, a maximum vocabulary of 5,000, sublinear term frequency, smoothed inverse document frequency, and L2 normalisation. Logistic regression and linear SVM search C over $\{0.01, 0.1, 1, 10, 100\}$ using only inner folds constructed from the outer training partition. AUC is computed from each model's continuous ranking score: fused score for hybrid models, branch score for component models, logistic-regression probability or decision score for logistic regression, and decision margin for linear SVM. These values are used for ranking and are not described as calibrated probabilities. Table 2 consolidates the principal model, tuning, threshold, and evaluation settings.

Table 2. Principal model and evaluation configuration

Component	Configuration
Sparse representation	Word unigrams+bigrams; max vocabulary 5,000; sublinear TF; smoothed IDF; L2 norm
Adaptive lexicon	VADER prior scale 0.35; additive smoothing 1.0; adapted weights clipped to $[-6, 6]$
Symbolic branch	1,500 candidates; absent/low/high states; error tolerance 0.20; confidence threshold 0.80; cap 600
Fusion	alpha 0.0-1.0 by 0.1; inner 3-fold selection; static alpha=0.50
Linear baselines	Logistic regression and linear SVM; C in $\{0.01, 0.1, 1, 10, 100\}$
Operational boundary	$L(x) < 0.15$ OR $ fused\ margin < 0.10$; thresholds predeclared
Outer evaluation	Seed 17; within and pooled 5-fold x 2 repeats; LOSO 3 held-out sources

2.8. Metrics, Statistical Analysis, and Reproducibility

The primary verified predictive metrics are accuracy, F1-score, area under the ROC curve (AUC), and Matthew's correlation coefficient (MCC). The stored aggregate evidence available for this revision does not support defensible reconstruction of precision, recall, specificity, source-disaggregated confusion matrices, expected calibration error,

calibration slope, runtime, or peak memory. Those values are therefore not back-calculated from summary statistics. For pooled accuracy and F1-score, descriptive 95% t intervals are reported over the ten stored outer-fold scores; because repeated cross-validation folds are correlated, these intervals are descriptive rather than independent-sample confidence statements.

Paired fold-level model comparisons use two-sided Wilcoxon signed-rank tests [20]. P-values are adjusted within scenario-metric families by Holm's sequential procedure [21], and rank-biserial effect sizes are retained. Leave-one-source-out comparisons are interpreted descriptively because only three held-out-source folds are available. Selective evaluation reports deferred coverage, accepted-case accuracy, boundary accuracy, error capture, lift, and area under the risk-coverage curve (AURC). The combined rule is evaluated against margin-only and lower-evidence-only ablations, random deferment at matched coverage, and matched-coverage rankings from logistic regression and linear SVM.

The reproducibility archive used for verification contains 105,000 observation-model prediction rows across 43 outer folds and 301 model-fold groups, together with 77,400 feature-state rules, fold parameters, uncertainty ablations, risk-coverage values, source data, configurations, and executable code. Independent recalculation reproduced the stored accuracy, F1, AUC, and MCC values for all 301 model-fold groups and found no duplicate prediction key.

3. RESULTS AND DISCUSSION

3.1. Predictive Performance Across Evaluation Scenarios

Table 3 reports the verified results for all seven models under within-source, pooled multi-domain, and leave-one-source-out evaluation. In the pooled setting, the proposed fusion achieved 83.73% accuracy, 83.76% F1-score, 91.42% AUC, and 67.48% MCC. The descriptive 95% interval across the ten pooled outer folds was 82.74%-84.72% for accuracy and 82.79%-84.73% for F1-score. Static fusion produced a marginally higher pooled AUC of 91.44%, while the proposed model improved F1 by only 0.50 percentage points over the adaptive lexical branch. This modest difference is important: the empirical

contribution should be read as complementary, traceable evidence integration rather than as a large predictive advance.

Table 3. Verified predictive results across evaluation scenarios

Scenario	Model	Accuracy (%)	F1 (%)	AUC (%)	MCC (%)
Within-source	VADER	74.97	79.26	89.26	54.90
Within-source	Adaptive lexicon	81.43	81.79	89.50	63.26
Within-source	Rough-only	74.62	74.49	84.01	49.98
Within-source	Static fusion	81.52	81.73	89.76	63.22
Within-source	Proposed fusion	81.57	81.94	89.62	63.47
Within-source	Logistic regression	80.62	80.58	88.43	61.35
Within-source	Linear SVM	80.33	80.40	88.20	60.80
Pooled	VADER	74.97	79.22	89.12	54.73
Pooled	Adaptive lexicon	83.17	83.26	91.04	66.34
Pooled	Rough-only	78.30	78.28	87.02	56.79
Pooled	Static fusion	83.52	83.52	91.44	67.04
Pooled	Proposed fusion	83.73	83.76	91.42	67.48
Pooled	Logistic regression	82.75	82.67	90.52	65.52
Pooled	Linear SVM	82.35	82.26	90.33	64.73
LOSO	VADER	74.97	79.24	89.25	54.85
LOSO	Adaptive lexicon	76.03	75.22	84.35	52.40
LOSO	Rough-only	73.10	72.86	82.03	46.79
LOSO	Static fusion	77.00	76.01	85.75	54.33
LOSO	Proposed fusion	76.47	75.59	84.83	53.27
LOSO	Logistic regression	75.20	73.79	83.36	50.77
LOSO	Linear SVM	74.90	73.22	83.30	50.26

Known-source F1 (81.94%) and pooled F1 (83.76%) were clearly higher than leave-one-source-out F1 (75.59%). This decline is consistent with evidence that feature-label relationships can shift across sentiment domains and that generalisation to an unseen source remains difficult [3], [22], [23]. Equal-weight fusion reached 76.01% F1 in the unseen-source setting, slightly above the proposed nested fusion at 75.59%, so no universal cross-domain advantage is claimed. Figure 2 visualises the proposed framework's F1-score across the known-source, pooled, and unseen-source conditions.

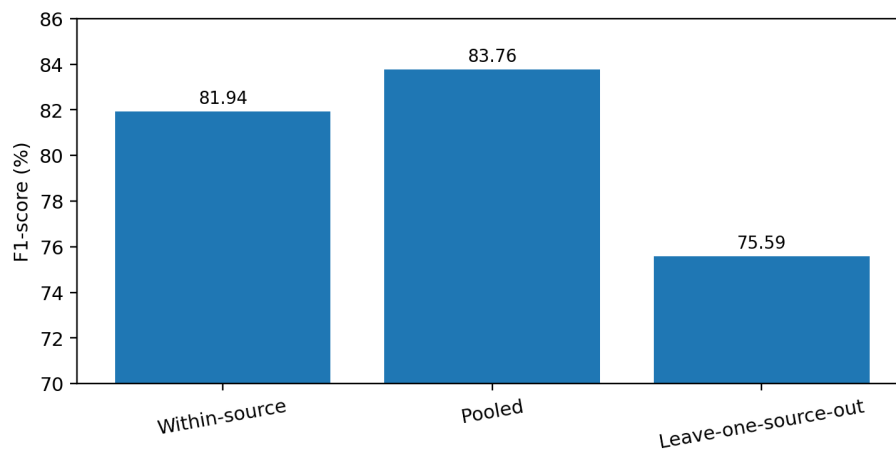


Figure 2. Proposed-framework F1-score across known-source, pooled, and unseen-source evaluation

3.2. Paired Statistical Evidence

In pooled folds, the proposed model exceeded the adaptive lexical branch, logistic regression, linear SVM, VADER, and rough-only classification after Holm adjustment, but its difference from static equal-weight fusion was not significant. Table 4 reports the paired F1 comparisons, adjusted p-values, and rank-biserial effects. The result narrows the central claim: nested weighting can provide a small pooled improvement, yet a simpler equal-weight combination remains a strong alternative and should be retained as a benchmark.

Table 4. Pooled F1-score comparisons against the proposed fusion

Comparator	Mean difference (points)	Holm-adjusted p	Rank-biserial	Decision
Adaptive lexicon	0.50	0.03125	0.891	Significant

Comparator	Mean difference (points)	Holm-adjusted p	Rank- biseiral	Decision
Linear SVM	1.50	0.03125	0.956	Significant
Logistic regression	1.09	0.03125	0.891	Significant
Static fusion	0.24	0.31250	0.524	Not significant
VADER	4.54	0.01172	1.000	Significant
Rough-only	5.48	0.01172	1.000	Significant

3.3. Operational Uncertainty, Deferment Cost, and Ablation

The combined boundary deferred 53.55% of pooled observations and accepted 46.45%. Accuracy among accepted predictions rose to 92.54%, and 78.72% of all classification errors were concentrated in the deferred subset. This demonstrates meaningful error enrichment, but the operational burden is substantial: more than half of incoming reviews would require a second-stage decision. At a scale of one million reviews, the fixed reference thresholds would route approximately 535,500 cases away from automatic acceptance. A deployment policy must therefore calibrate the deferment threshold against the cost of false automation, the capacity and cost of the secondary model or human queue, and the coverage required by the application. Table 5 quantifies this trade-off and compares the combined boundary with matched-coverage and ablation controls.

Table 5. Pooled uncertainty ablation and matched-coverage controls

Mechanism	Deferred (%)	Accepted acc. (%)	Boundary acc. (%)	Errors captured (%)	Lift	AURC
Combined	53.55	92.54	76.06	78.72	1.47	0.0848
Margin only	16.47	88.52	59.30	41.09	2.50	0.0612
Lower evidence only	49.15	89.83	77.44	68.12	1.39	0.1097
Logistic matched	53.55	94.58	72.36	85.36	1.60	0.0712
SVM matched	53.55	94.56	71.66	85.58	1.60	0.0728
Random matched	53.55	83.64	83.75	53.27	1.00	-

The selective-risk comparison is a central limitation of the proposed mechanism. At the same 53.55% deferred coverage, logistic-regression confidence retained 94.58% accepted accuracy and SVM confidence retained 94.56%, compared with 92.54% for the combined symbolic boundary; both also captured about 85.5% of errors. Margin-only ranking produced a smaller AURC than the combined rule. The boundary should therefore not be presented as the best uncertainty-ranking mechanism. Its distinct contribution is that the deferment decision can be traced to an explicit fused-margin condition and/or insufficient lower-region symbolic support. Figure 3 complements Table 5 by visualising accepted-case accuracy and error capture for the principal deferment mechanisms.

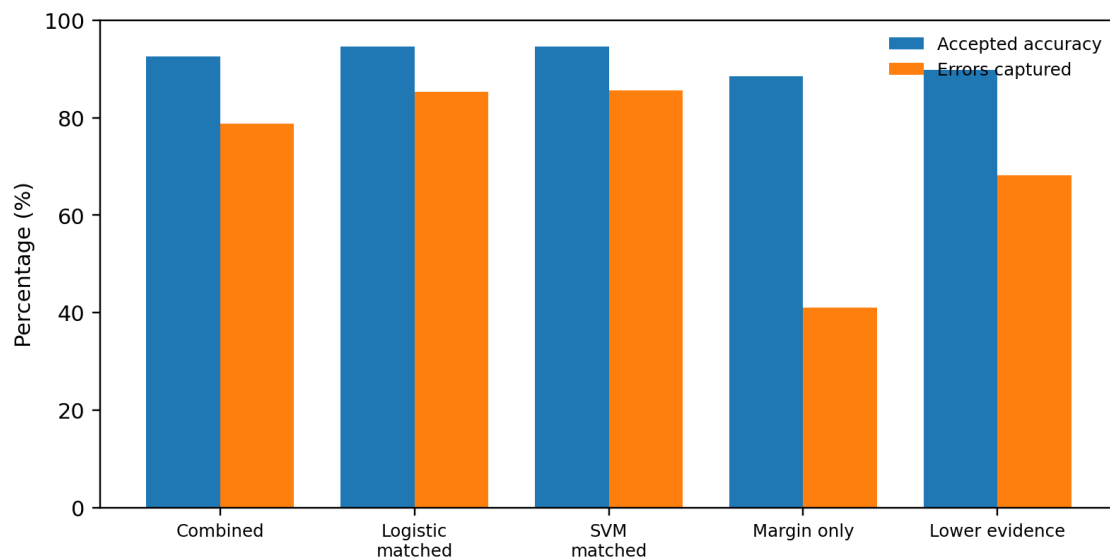


Figure 3. Accepted-case accuracy and error capture for principal uncertainty mechanisms

3.4. Feature Sensitivity and Algorithmic Traceability

The 600-feature setting was examined rather than treated as a hidden optimum. A one-repeat pooled sensitivity analysis evaluated dependency-ranked caps from 100 to 1,500 features together with chi-square and random 600-feature controls. The highest point estimate occurred at 300 dependency-ranked features (84.29% F1), whereas the principal 600-feature configuration reached 83.93% F1. A chi-square 600-feature control reached 84.15%, and the random 600-feature control reached 83.77%. Table 6 presents the complete sensitivity results. The narrow performance range supports controlled feature

compression, but not a claim that 600 features or dependency ranking is uniquely optimal.

Table 6. One-repeat pooled feature-selection sensitivity

Selection	Features	Accuracy (%)	F1 (%)	AUC (%)	MCC (%)
Dependency	100	83.27	82.95	91.28	66.73
Dependency	300	84.27	84.29	91.51	68.55
Dependency	600	83.87	83.93	91.27	67.75
Dependency	900	84.00	84.08	91.31	68.02
Dependency	1,200	84.07	84.15	91.30	68.15
Dependency	1,500	83.73	83.79	91.31	67.48
Chi-square	600	84.07	84.15	91.33	68.15
Random	600	83.80	83.77	91.16	67.62

The symbolic-state audit shows another interpretive constraint. Mean absolute symbolic evidence mass was dominated by absent states (71.19%), compared with 8.48% from low states and 20.33% from high states, as summarised in Table 7. Term absence can be discriminative in sparse text, but an explanation dominated by features that are not present may be less intuitive to a human analyst than one driven by observed evaluative phrases. For this reason, interpretability is used here in the restricted sense of algorithmic traceability: a prediction can be linked to branch scores, matched feature states, state support, majority class, confidence, fusion weight, and boundary trigger. No claim of human comprehensibility is made because no expert or user study was conducted, consistent with the distinction between faithful computational traces and human-centred explanation [24], [25].

Table 7. Mean symbolic evidence mass in pooled evaluation

Symbolic state	Mean absolute evidence mass (%)
Absent	71.19
Low	8.48
High	20.33

Case-level trace examples would provide an additional test of this traceability claim, particularly if an accepted case and a boundary case were exported directly from the observation-level archive with sentence text, lexical score, matched feature states, $L(x)$, fused score, α , and the triggering condition. Those case-level values are not recoverable from the aggregate tables alone. Rather than inventing illustrative numbers, such examples should be generated directly from the archived prediction rows and included in the supplementary package or a subsequent expanded evaluation.

3.5. Position Relative to Transformer Models and Practical Efficiency

Contextual encoders and LLM-based sentiment systems capture semantic dependencies that the sparse symbolic-lexicon framework cannot represent. Recent cross-domain methods use causal representation learning and counterfactual augmentation [3], [4], while sentiment-specific distillation demonstrates that smaller models can retain strong performance with reduced data or model scale [6]. Compact transformer encoders such as DistilBERT [26] and MiniLM [27] are therefore especially relevant future baselines: they provide contextual modelling while remaining closer to the resource envelope of practical information systems than full-scale LLMs. A defensible comparison should place these models on the identical outer folds, use the same label protocol, and report both predictive and selective-risk metrics rather than importing headline scores from unrelated experiments.

The same matched-design principle applies to efficiency. The present verification archive does not contain fold-matched wall-clock runtime or peak-memory measurements for the seven evaluated models, so this article makes no claim of proven computational superiority. A future benchmark should record training time, inference throughput, peak memory, hardware, software versions, and, where feasible, energy or compute budget for the hybrid framework and compact transformer baselines under identical partitions. Such measurements are necessary because the additional symbolic discretisation and rule construction have their own cost, while compact contextual models incur different memory and inference trade-offs.

3.6. Calibration and Additional Evidence for External Validity

Uncertainty-aware terminology also requires calibration discipline. A confidence score can rank observations without being a calibrated probability [13], [14]. The verified suite

provides selective-risk and risk-coverage diagnostics but does not contain numerical expected calibration error, calibration slope, reliability diagrams, or a fully documented Brier analysis. The manuscript therefore does not claim probability calibration. An expanded experiment should add fold-local calibration diagnostics and compare the operational boundary with probability-calibrated and conformal alternatives where their assumptions are appropriate [15].

Several additional analyses would strengthen external validity but cannot be reconstructed responsibly from aggregate statistics: fully source-disaggregated model tables, separate results for each held-out source, pooled and leave-one-source-out confusion matrices, precision/recall/specificity tables, case-level trace examples, runtime and memory measurements, and compact-transformer baselines. The present article therefore reports only analyses supported by the verified archive and treats these items as explicit extensions. This evidence boundary avoids filling unresolved analytical gaps with synthetic or post-hoc numerical claims.

3.7. Practical Implications for Information Systems

The framework supports a two-output architecture for information systems that process heterogeneous textual feedback. The first output is binary polarity for monitoring or aggregation; the second is a routing status. Accepted cases may enter automated summaries, whereas boundary cases may be re-analysed by a contextual model, reviewed by an analyst, or withheld from high-stakes aggregates until resolved. This design is useful when an organisation needs an auditable reason for escalation rather than a confidence number alone. The reference thresholds should not be transferred unchanged. Deployment should define the relative costs of false positive, false negative, and deferment outcomes, impose a capacity constraint on the secondary queue, tune the policy on representative validation data, and monitor both coverage and selective risk as the input distribution changes.

A staged deployment roadmap follows from these findings. First, the current pipeline and observation-level archive should be released with a persistent identifier. Second, compact contextual and efficiency baselines should be evaluated on the same folds. Third, deferment thresholds should be tuned against an explicit application-specific cost function and calibration should be assessed separately. Fourth, analyst-facing trace

quality should be evaluated, especially where absent-state evidence dominates. Finally, validation should extend beyond short English binary sentences to multilingual reviews, longer documents, temporal drift, sarcasm, and aspect-level sentiment before the framework is used in broader operational settings.

3.8. Discussion

The findings indicate that the main value of the proposed framework lies in combining competitive multi-domain sentiment classification with an explicit mechanism for identifying predictions that should not be accepted automatically. The pooled results show that nested fusion can improve on the adaptive lexical branch and conventional linear baselines, although the gain over static equal-weight fusion is small and not statistically significant. The decline from pooled and within-source performance to leave-one-source-out performance further confirms that cross-domain sentiment remains sensitive to shifts in feature-label relationships, consistent with prior work on domain adaptation and cross-domain representation learning [3], [4], [22], [23]. Accordingly, the framework should be interpreted as a traceable hybrid decision architecture rather than as a universally superior classifier.

The deferment analysis provides a second important result because it separates predictive accuracy from the operational question of when a prediction should be trusted. The combined boundary increased accepted-case accuracy to 92.54% and concentrated 78.72% of errors in the deferred subset, demonstrating that the boundary can identify a substantial proportion of difficult observations. However, logistic-regression and SVM confidence rankings achieved lower selective risk at matched coverage, so the proposed boundary does not establish superiority as an uncertainty-ranking method. Its contribution is instead aligned with selective-prediction principles [12]: a deferment decision is linked to observable conditions in the fused margin and symbolic lower-region evidence rather than to an opaque confidence score alone.

The symbolic component also clarifies the scope of the framework's interpretability claim. Because the evidence path can be reconstructed from feature states, class support, state confidence, branch scores, fusion weight, and the boundary trigger, the method offers algorithmic traceability that can support audit and debugging. At the same time, the dominance of absent-state evidence shows that a faithful computational trace is not

necessarily intuitive to a human analyst, which agrees with the distinction between computational faithfulness and human-centred explanation discussed in prior work [24], [25]. This limitation means that future evaluation should include case-level trace studies and analyst-facing assessments before the framework is described as providing human-comprehensible explanations.

From an information-systems perspective, the results support a staged architecture in which automatic sentiment aggregation is reserved for accepted cases and boundary cases are routed to a stronger contextual model or human review. Such a design is particularly relevant when organisations require an auditable reason for escalation, but the current fixed thresholds should be treated only as reference settings because the observed 53.55% deferment rate may be too costly in many deployments. Future work should therefore tune coverage against application-specific error and review costs, add fold-local calibration and conformal comparisons [13]- [15], and evaluate compact transformer baselines, runtime, memory, multilingual data, longer documents, temporal drift, sarcasm, and aspect-level sentiment. These extensions are necessary to determine whether the framework's traceability advantages remain useful when predictive performance, selective risk, computational cost, and human review capacity are evaluated jointly.

4. CONCLUSION

This study presented an uncertainty-aware hybrid symbolic-lexicon framework for multi-domain binary sentiment classification that combines training-only lexical adaptation, sparse TF-IDF representation, three-state symbolic discretisation, variable-precision feature-state evidence, dependency ranking, nested score fusion, and a deferment status that remains separate from polarity. Evaluation on 3,000 UCI Sentiment Labelled Sentences covered repeated within-source, pooled multi-domain, and leave-one-source-out conditions. The strongest result occurred in pooled evaluation, where the framework achieved 83.73% accuracy and 83.76% F1-score; its 0.50-point F1 improvement over the adaptive lexical branch was statistically significant after correction, whereas the 0.24-point difference from equal-weight static fusion was not. The combined boundary concentrated 78.72% of errors and retained 92.54% accuracy among accepted cases, but it deferred 53.55% of observations and was outperformed in selective risk by matched-

coverage logistic-regression and SVM confidence rankings. The contribution is therefore a reproducible and algorithmically traceable integration of complementary evidence, not universal superiority in classification or uncertainty ranking. Practical use requires task-specific cost calibration for deferment, capacity-aware secondary processing, and ongoing monitoring of coverage and selective risk. The evidence is currently limited to short English binary text; future work should add compact transformer baselines, runtime and memory benchmarking, calibration analysis, human evaluation of decision traces, source-disaggregated diagnostics, multilingual and longer documents, temporal drift, sarcasm, and aspect-level sentiment.

ACKNOWLEDGMENT

The authors express their appreciation to Universitas Muhammadiyah Jambi, Universiti Muhammadiyah Malaysia, and Universitas Muhammadiyah Palangkaraya for their support of the academic environment in which this research was developed.

REFERENCES

- [1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
- [2] K. Alahmadi, S. Alharbi, J. Chen, *et al.*, "Generalizing sentiment analysis: A review of progress, challenges, and emerging directions," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, 2025, doi: 10.1007/s13278-025-01461-8.
- [3] S. Wang, J. Zhou, Q. Chen, Q. Zhang, T. Gui, and X. Huang, "Domain generalization via causal adjustment for cross-domain sentiment analysis," pp. 5286–5298, 2024, arXiv: 2402.14536.
- [4] D. Xin, K. Zhao, J. Sun, and Y. Li, "CDA²: Counterfactual diffusion augmentation for cross-domain adaptation in low-resource sentiment analysis," in *Proc. 31st Int. Conf. Comput. Linguistics*, Jan. 2025, pp. 61–72.
- [5] W. Zhang, Y. Deng, B. Liu, *et al.*, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3881–3906, 2024, doi: 10.18653/v1/2024.findings-naacl.246.

- [6] G. Xie, Y. Zhang, J. Bao, *et al*, "Comprehensive and efficient distillation for lightweight sentiment analysis models," in *Proc. 2025 Conf. Empirical Methods in Natural Language Processing*, pp. 22070–22091, 2025, doi: 10.18653/v1/2025.emnlp-main.1122.
- [7] D. Herrera-Poyatos, C. Peláez-González, C. Zuheros, A. Herrera-Poyatos, V. Tejedor, F. Herrera, and R. Montes, "An overview of model uncertainty and variability in LLM-based sentiment analysis: Challenges, mitigation strategies, and the role of explainability," *Front. Artif. Intell.*, 2025, doi: 10.3389/frai.2025.1609097.
- [8] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Web and Social Media*, vol. 8, no. 1, pp. 216–225, 2014, doi: 10.1609/icwsm.v8i1.14550.
- [9] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," *J. Big Data*, vol. 11, no. 1, art. no. 10, 2024, doi: 10.1186/s40537-023-00861-x.
- [10] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.
- [11] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, no. 1, pp. 39–59, 1993.
- [12] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proc. 36th Int. Conf. Machine Learning*, PMLR, vol. 97, 2019, arXiv: 1901.09192.
- [13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning*, PMLR, vol. 70, pp. 1321–1330, 2017.
- [14] Y. Ovadia, E. Fertig, J. Ren, *et al*, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. 33rd Conf. Neural Information Processing Systems*, 2019, arXiv: 1906.02530.
- [15] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, 2023, doi: 10.1561/22000000101.
- [16] K. Gunaratne, S. Kobourov, and J. A. Miller, "Class angular distortion index for dimensionality reduction," *Comput. Graph. Forum*, 2026, doi: 10.1111/cgf.70465.
- [17] D. Kotzias, M. Denil, N. de Freitas, *et al*, "From group to individual labels using deep features," in *Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 597–606, 2015, doi: 10.1145/2783258.2783380.
- [18] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

- [19] J. Pineau, P. Vincent-Lamarre, K. Sinha, *et al*, "Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program)," *J. Mach. Learn. Res.*, vol. 22, pp. 1–20, 2021.
- [20] M. Ock, N. Yi, J. Ahn, and M. Jo, "How many alternatives can be ranked? A comparison of the paired comparison and ranking methods," *Value Health*, vol. 19, no. 5, pp. 655–660, 2016, doi: 10.1016/j.jval.2016.03.1836.
- [21] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
- [22] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP — A survey," in *Proc. 28th Int. Conf. Computational Linguistics*, pp. 6838–6855, 2020, doi: 10.18653/v1/2020.coling-main.603.
- [23] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, doi: 10.1109/TPAMI.2022.3195549.
- [24] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?," in *Proc. 58th Annu. Meeting Association for Computational Linguistics*, pp. 4198–4205, 2020, doi: 10.18653/v1/2020.acl-main.386.
- [25] Q. Lyu, M. Apidianaki, and C. Callison-Burch, "Towards faithful model explanation in NLP: A survey," *Comput. Linguist.*, 2024, doi: 10.1162/coli_a_00511.
- [26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv: 1910.01108, 2019.
- [27] W. Wang, F. Wei, L. Dong, *et al*, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. Advances in Neural Information Processing Systems*, 2020.