TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 24, No. 6, December 2026, pp. 2114~2125
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i6.28180

Uncertainty-aware multi-domain sentiment classification using adaptive lexicons and rough-set reasoning

Saleh Yaakub[1], Mustafa Mat Deris[1], Ika Safitri Windiarti[1,2]
[1]Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia
[2]Computer Science, Universitas Muhammadiyah Palangkaraya, Indonesia

Article Info

Article history:

Received Jul 29, 2026
Revised Sep 18, 2026
Accepted Oct 11, 2026

Keywords:

Domain shift
Evidence granulation
Interpretable natural language processing
Risk-coverage analysis
Selective prediction
Symbolic artificial intelligence
Variable-precision rough sets

ABSTRACT

Sentiment classifiers deployed across heterogeneous review domains face vocabulary shift and often return polarity labels without an auditable indication of ambiguous evidence. This study evaluates RS-IVADER, a leakage-controlled hybrid framework that combines a training-adapted valence aware dictionary and sentiment reasoner (VADER) lexicon with variable-precision rough-set evidence and a separate operational boundary status. Experiments use 3,000 balanced sentences from Amazon cells, internet movie database (IMDb), and Yelp under repeated within-source, pooled, and leave-one-source-out protocols. In pooled evaluation, RS-IVADER attained 83.73% accuracy, 83.76% F1-score, 91.42% area under the curve (AUC), and 67.48% Matthews correlation coefficient (MCC). The combined boundary deferred 53.55% of cases, raising accepted-case accuracy to 92.54% and capturing 78.72% of errors. Equal-weight fusion remained statistically competitive, and matched linear-confidence baselines achieved lower selective risk. The contribution is therefore a reproducible integration of adaptive lexical evidence and symbolic uncertainty traces for auditable review triage, rather than a claim of universal predictive or uncertainty superiority.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Saleh Yaakub
Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia
Bangunan Wisma MAIPs, Blok S1, Kompleks Desa Siswa, Uniciti Alam, 02100 Padang Besar
Perlis, Malaysia
Email: p5240052@student.umam.edu.my

1. INTRODUCTION

Digital platforms continuously generate customer reviews, service evaluations, film opinions, and other texts that encode satisfaction and perceived quality. Sentiment analysis converts this evidence into operational labels, yet reliability depends on whether a model remains stable when vocabulary, writing style, and evaluative conventions change across sources. Recent surveys identify domain dependence, linguistic variability, robustness, and generalization as persistent challenges across lexicon, classical machine-learning, deep-learning, and large-language-model approaches [1]–[4]. Words such as light, unpredictable, or cheap may carry different evaluative force in electronics, film, and hospitality. Consequently, performance measured on a familiar source cannot by itself establish dependable transfer to an unseen domain. Recent domain-adaptive lexicon work likewise shows that sentiment evidence can vary materially across review domains [5].

Contextual models improve representation capacity, but they do not eliminate the need for transparent evaluation, calibrated confidence, and reproducible pipelines. Classical machine-learning methods can remain competitive on structured consumer-review tasks [6]–[9], whereas domain-adaptive

pretraining and domain-generalization methods explicitly address distributional differences [10]–[12]. Large-language-model sentiment systems add variability related to prompting, stochastic generation, and explanation faithfulness [13], [14]. Probabilistic uncertainty estimates can also degrade under dataset shift [15]. These observations motivate uncertainty mechanisms that are evaluated independently of predictive accuracy and that expose why a case is considered unreliable.

Lexicon-based systems offer a complementary route because their term scores and linguistic rules can be inspected. The valence aware dictionary and sentiment reasoner (VADER) combines a human-rated lexicon with rules for intensity, negation, punctuation, capitalization, and contrast [16]. Barik and Misra [5] extended this lineage by adapting VADER to domain-specific review evidence and reported strong results under their own four-domain protocol. Their study motivates domain-adaptive lexical scoring, but its reported values are not used as paired evidence in the present experiments because the available description does not expose every split, threshold, seed, and observation-level prediction required for exact independent reproduction. Moreover, a polarity classifier alone does not provide a formally separate, auditable evidence-status output.

Rough-set theory provides a non-probabilistic language for distinguishing lower-region evidence from boundary evidence [17], while variable-precision rough sets permit controlled inconsistency in noisy granules [18]. This choice is deliberate. Bayesian calibration and ensemble confidence estimate uncertainty through probability distributions or repeated predictive models; such methods are valuable when calibrated probabilities are the primary requirement [15], [19], [20]. Rough sets instead represent indiscernibility and evidence consistency directly in a symbolic information table, making the source of ambiguity inspectable at the feature-state level. The trade-off is that rough boundaries are not posterior probabilities and should not be treated as automatically superior uncertainty estimators. Recent rough feature-selection work emphasizes scalability, noise tolerance, and validation [21], [22]. For sparse text, however, exact equality over continuous vectors fragments granules, train-test leakage can occur during feature construction, and symbolic boundaries are rarely benchmarked against ordinary confidence at matched coverage.

This study addresses these gaps through rough set-improved VADER (RS-IVADER), a hybrid symbolic-lexicon sentiment framework. It retains the domain-adaptive intent of Improved VADER while introducing a documented leakage-controlled lexical operationalization, training-fitted symbolic granulation, variable-precision feature-state evidence, nested score fusion, and a polarity-independent review status. The evaluation uses repeated within-source, pooled, and leave-one-source-out settings; lexical, symbolic, linear, and static-fusion baselines; paired significance testing; matched-coverage uncertainty controls; risk-coverage diagnostics; feature-count sensitivity; state-mass auditing; and independent metric regeneration from fold-level predictions. The innovation is therefore the controlled integration of adaptive lexical scoring and rough-set evidence into a traceable decision pipeline, rather than the use of either component in isolation.

The objective is to determine whether this integration offers competitive sentiment classification and a useful error-triage mechanism under multi-domain evaluation. The engineering value lies in separating the polarity decision from an inspectable operational status that can be consumed by a decision-support system. In a deployment pipeline, certain predictions can proceed automatically, whereas boundary cases can be routed to human review or a higher-capacity model. The remainder of the paper presents the framework, experimental design, results, deployment implications, limitations, and conclusions.

2. PROPOSED SYMBOLIC-LEXICON FRAMEWORK

2.1. Methodological lineage and architecture

The proposed framework is methodologically rooted in the domain-adaptive VADER strategy reported by Barik and Misra [5], but the current article separates conceptual inheritance from numerical replication. The earlier study establishes the value of a domain-specific sentiment dictionary, whereas the present implementation, denoted IVADER-R for leakage-controlled operationalization, uses a documented corpus log-odds adjustment rather than claiming to reconstruct every semantic-refinement, entropy, and composite-score detail of the published workflow. Published headline values are therefore used only for contextual discussion and are not entered into paired tests. All statistical comparisons in this study use models fitted and evaluated on identical outer folds within the verified experiment.

Figure 1 shows four connected stages. The first stage applies fixed outer partitions and reserves all adaptive operations for training data. The second stage builds a sparse representation. The third stage contains two evidence branches: an adaptive lexical branch and a symbolic rough-set branch. The fourth stage combines branch scores, predicts polarity, assigns an operational certainty status, and exports a complete decision trace. A reproducibility layer records seeds, parameters, predictions, rules, checksums, and independently recalculated metrics.

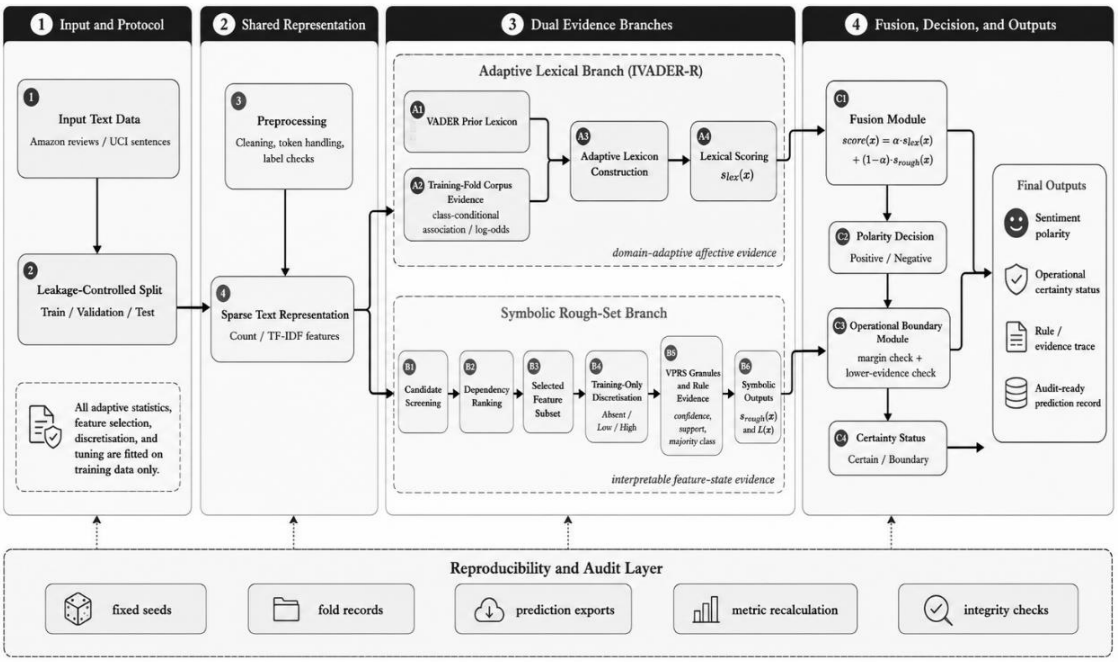


Figure 1. Proposed leakage-controlled symbolic-lexicon sentiment classification architecture

2.2. Shared sparse representation and adaptive lexicon

The shared representation uses word unigrams and bigrams with a maximum vocabulary of 5,000 features, sublinear term frequency, smoothed inverse document frequency, and L2 normalization. The vectorizer is fitted only on the current outer-training partition. Let $tf(t,d)$ be the frequency of term t in document d, N the number of training documents, and $df(t)$ the training-document frequency. The smoothed feature value is defined in (1).

$$\text{tfidf}(t,d) = \text{tf}(t,d)\left[\log\left(\frac{1+N}{1+\text{df}(t)}\right)+1\right] \quad (1)$$

The adaptive lexical branch begins with the VADER prior valence $v(t)$. For each feature, the training corpus supplies class-conditional document counts. With additive smoothing delta, the log-odds adjustment $ell(t)$ measures whether the feature occurs proportionally more often in positive or negative training documents. The adapted document score combines the prior and corpus evidence, weights them by the sparse representation, clips extreme contributions, and applies a hyperbolic tangent to keep the result bounded. This preserves the main paper's domain-adaptive objective while making label dependence and train-only fitting explicit.

$$\ell(t) = \log\left(\frac{n_{\text{pos}}(t)+\delta}{N_{\text{pos}}+2\delta}\right) - \log\left(\frac{n_{\text{neg}}(t)+\delta}{N_{\text{neg}}+2\delta}\right) \quad (2)$$

The bounded lexical score is the weighted average of the VADER prior and corpus log-odds after clipping and hyperbolic-tangent transformation. The adaptation coefficient, smoothing constant, clipping limit, and vocabulary controls are stored in the experiment configuration.

2.3. Symbolic granulation and rough-set evidence

Continuous sparse features cannot be treated as classical rough-set equivalence attributes without transformation. For each retained feature j, the median m_j of its non-zero values is estimated from outer-training data. Each observation is then mapped to one of three states as shown in (3): absent when the value is zero, low when the positive value does not exceed the training median, and high when it exceeds that median. Features without non-zero training observations are removed. This operation creates a categorical information table while preserving a readable connection between a feature and its relative intensity.

The variable precision rough set (VPRS) is an extension of rough set theory used to analyze and identify data patterns which represent statistical trends rather than functional. The main idea of VPRS is to allow objects to be classified with an error smaller than a certain pre-defined level. This introduced threshold relaxes the rough set notion of requiring no information outside the dataset itself.

Definition 2.3: let a set U as a universe and $X, Y \subseteq U$, where $X, Y \neq \emptyset$. The error classification rate of X relative to Y is denoted by $e(X,Y)$, is defined by [17],

$$e(X,Y) = \begin{cases} 1 - \dfrac{|X \cap Y|}{|X|} & ,|X| > 0 \\ 0 & ,X = 0 \end{cases}$$

Definition: let U be a finite set and let β be a real number within the range $0 \leq \beta < 0.5$ and a set $X \subseteq U$. The B_β −lower-approximation of X, denoted by $\underline{B_\beta}(X)$ and B_β −upper-approximation of X, denoted by $\overline{B}_\beta(X)$, respectively, are defined by:

$$\underline{B_\beta}(X) = \{X \in U : e([x]_B, X) \leq \beta\}$$
$$\overline{B}_\beta(X) = \{X \in U : e([x]_B, X) < 1 - \beta\}$$

where $[x]_B$ is the equivalence class for $x \in U/B$, and $\underline{B_\beta}(X) \subseteq \overline{B}_\beta(X)$ if and only if $0 \leq \beta < 0.5$.

Candidate features are screened inside training data and ranked by variable-precision dependency. For each feature-state granule, the majority class, support, and smoothed confidence are estimated. A granule qualifies as lower-region evidence when its majority confidence is at least 0.80, equivalent to a tolerated error of at most 0.20. The signed rule contribution increases with class consistency and logarithmic support. The normalized symbolic score is the sum of matched signed contributions divided by the sum of their absolute magnitudes. A separate lower-evidence proportion $L(x)$ measures how much of the matched evidence comes from lower-region granules. These records are associations in the training decision table; they are not causal linguistic rules.

Candidate screening is capped at 1,500 features and the principal configuration retains at most 600 dependency-ranked features. The term rough-set-informed dependency-ranked subset is used deliberately: the method ranks and caps features but does not claim a unique mathematically minimal reduct. Chi-square selection, random selection, and alternative feature counts are included as controls.

$$q_j(x) = \begin{cases} 0, & x_j = 0, \\ 1, & 0 < x_j \leq m_j, \\ 2, & x_j > m_j. \end{cases} \quad (3)$$

2.4. Fusion, polarity, and operational boundary

The two branch scores are placed on bounded scales and combined by a convex weight alpha selected in inner validation from 0.0 to 1.0 in increments of 0.1. Equal-weight fusion is retained as a static baseline. The sign of the fused score determines polarity, as defined in (4). This arrangement allows validation to favor either evidence branch while ensuring that the outer test set does not influence model selection.

$$s_{\text{fuse}}(x) = \alpha s_{\text{lex}}(x) + (1 - \alpha) s_{\text{rough}}(x) \quad (4)$$

$$\hat{y}(x) = \begin{cases} \text{positive}, & s_{\text{fuse}}(x) \geq 0, \\ \text{negative}, & \text{otherwise}. \end{cases} \quad (5)$$

Operational status is calculated after polarity. A prediction is marked boundary when the lower-evidence proportion is below 0.15 or the absolute fused margin is below 0.10; otherwise, it is marked certain. These constants are predeclared operating points rather than probabilities or globally optimized rejections. The boundary does not replace the sentiment label. A positive-boundary output, for example, means that positive is the current best class but the supporting evidence is weak, conflicting, or near the decision threshold.

$$u(x) = \begin{cases} \text{boundary}, & L(x) < 0.15 \ \vee \ |s_{\text{fuse}}(x)| < 0.10, \\ \text{certain}, & \text{otherwise}. \end{cases} \quad (6)$$

The complete output includes the lexical score, symbolic score, fused score, selected alpha, polarity, lower-evidence proportion, triggered boundary conditions, and the highest-contributing feature-state records. This trace supports faithfulness checks because every displayed explanation can be compared directly with the implemented decision function. Explanation quality nevertheless requires more than readable rules; stability, support, and user comprehension remain separate empirical questions [23], [24].

3. METHOD

3.1. Data and evaluation scenarios

The primary verified data are the sentiment labelled sentences collection [25], [26], containing 1,000 Amazon mobile-phone sentences, 1,000 internet movie database (IMDb) sentences, and 1,000 Yelp sentences. Each source contains 500 positive and 500 negative observations. Labels are those distributed with the benchmark; no relabeling or post-hoc annotation correction was performed, which avoids introducing an additional subjective annotation layer but also means that annotation consistency is inherited from the source dataset. Duplicate handling and split construction were performed before adaptive fitting. Raw punctuation and case were preserved for the VADER sentence branch, while the sparse branch used the training-fitted word unigram/bigram representation defined in section 2.2. Historical Amazon books, digital versatile disc (DVD), Electronics, and kitchen aggregates from the dissertation are not used as direct evidence because their original observation-level prediction archive was unavailable; the University of California, Irvine (UCI) suite is the independently executable evidence tier.

Three evaluation settings were used. Repeated within-source evaluation applied five stratified folds twice to each source, producing 30 outer folds. Pooled evaluation combined all 3,000 sentences and applied five stratified folds twice, producing 10 outer folds. Leave-one-source-out evaluation trained on two sources and tested on the third, producing three outer folds. These settings respectively assess source specialization, shared multi-domain learning, and transfer to an unseen source. Table 1 summarizes the design.

Table 1. Dataset and evaluation design

Setting	Data	Outer protocol	Purpose
Within-source	Amazon cells, IMDb, Yelp; 1,000 each	5-fold stratified, repeated twice per source (30 folds)	Known-source specialization
Pooled	All 3,000 balanced sentences	5-fold stratified, repeated twice (10 folds)	Shared multi-domain learning
Leave-one-source-out	Train on two sources; test on the third	Three held-out-source evaluations	Unseen-source transfer

3.2. Training controls, baselines, and metrics

All adaptive operations were restricted to the current outer-training partition: vectorization, class-conditional lexical statistics, non-zero medians, feature ranking, symbolic rule construction, fusion-weight selection, and baseline hyperparameter selection. Raw sentence case, punctuation, negation, capitalization, and contrast cues were retained for VADER. The sparse branch used word unigrams and bigrams, at most 5,000 features, sublinear term frequency, smoothed inverse-document frequency, and L2 normalization. Candidate symbolic screening was capped at 1,500 features and the principal dependency-ranked subset at 600 features. The lower-region confidence threshold was 0.80, the lower-evidence boundary threshold 0.15, and the absolute fused-margin threshold 0.10. Fusion weight alpha was selected from {0.0, 0.1, ..., 1.0}. Logistic regression and linear support vector machine (SVM) were selected as strong linear text baselines because they are competitive on sparse review representations and provide straightforward confidence or margin controls; C was selected from 0.01 to 100 within inner validation. VADER, IVADER-R, rough-only, and equal-weight static fusion isolate the contribution of each proposed component. The implementation used Python with NumPy, SciPy, and scikit-learn [27]-[29]; exact environment versions, seeds, fold identifiers, parameters, and checksums are recorded in the verified experiment archive.

Performance was evaluated with accuracy, precision, recall, specificity, F1-score, area under the receiver operating characteristic curve / area under the curve (AUC), and the Matthews correlation coefficient (MCC). Accuracy is directly interpretable because every source is class-balanced; F1 summarizes the precision-recall trade-off; MCC provides a balanced correlation-based measure; and AUC evaluates score ranking independently of one decision threshold. A diagnostic Brier mapping is also reported, but it is not interpreted as a calibrated probability score because the lexical, symbolic, and fused outputs were not probability-calibrated [30]. This metric set separates label quality, ranking quality, and operational uncertainty rather than relying on a single headline measure.

Reported within-source and pooled values are means over repeated outer folds, not single experiments. Paired fold comparisons used the Wilcoxon signed-rank test [31], Holm adjustment within each scenario and metric family [32], and rank-biserial correlation as an effect-size measure. Cross-validation folds are correlated, so these tests are interpreted as scenario-specific supporting evidence rather than independent population replications [33]. Leave-one-source-out contains only three paired source evaluations and is therefore treated descriptively. A confidence interval based only on the summary values in the manuscript would be statistically misleading because repeated folds are dependent; interval estimation should be regenerated from the archived observation-level or fold-level outputs using a dependence-aware resampling design. This requirement is included in the reproducibility checklist for the final empirical archive.

3.3. Uncertainty diagnostics, parameter controls, and reproducibility

Uncertainty evaluation used coverage, accepted-case accuracy, selective risk, boundary accuracy, proportion of all errors captured, error-capture lift, complete risk-coverage curves, and area under the risk-coverage curve. The combined symbolic-margin boundary was compared with margin-only, lower-evidence-only, random matched-coverage, logistic-regression matched-coverage, and SVM matched-coverage mechanisms. This design follows the principle that a reject or defer option must be evaluated jointly by risk and coverage rather than by accepted-case accuracy alone [15], [19], [20]. The constants 0.80, 0.15, and 0.10 are treated as predeclared operating parameters, not learned probabilities. A complete threshold-sensitivity rerun should vary lower-region confidence, lower-evidence proportion, and fused-margin thresholds jointly while reporting coverage, selective risk, and error capture; the main text therefore avoids claiming that the selected operating point is globally optimal.

The final archive contains 105,000 observation-model prediction rows across 43 outer folds, 301 model-fold metric records, 77,400 symbolic records, and 43 parameter records. All 301 groups passed independent recalculation from predictions, no duplicate model-fold-observation key was detected, and file integrity was recorded through manifests and checksums. Leakage controls and artifact reporting respond directly to reproducibility concerns in machine-learning-based science [34], [35].

4. RESULTS AND DISCUSSION

4.1. Predictive performance across scenarios

Table 2 reports the independently verified mean performance. In pooled evaluation, RS-IVADER achieved 83.73% accuracy, 83.76% F1-score, 91.42% AUC, and 67.48% MCC. Its pooled F1-score exceeded IVADER-R by 0.50 percentage points, logistic regression by 1.09 points, linear SVM by 1.50 points, VADER by 4.54 points, and the rough-only branch by 5.48 points. The difference from equal-weight static fusion was only 0.24 points, indicating that most of the predictive gain came from combining lexical and symbolic evidence rather than from learning a non-equal fusion weight.

Table 2. Independently verified performance by scenario

Scenario	Model	Accuracy (%)	F1 (%)	AUC (%)	MCC (%)
Within-source	VADER	74.97	79.26	89.26	54.90
Within-source	IVADER-R	81.43	81.79	89.50	63.26
Within-source	VPRS-Rough-only	74.62	74.49	84.01	49.98
Within-source	Static fusion	81.52	81.73	89.76	63.22
Within-source	RS-IVADER	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	IVADER-R	83.17	83.26	91.04	66.34
Pooled	VPRS-Rough-only	78.30	78.28	87.02	56.79
Pooled	Static fusion	83.52	83.52	91.44	67.04
Pooled	RS-IVADER	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
Leave-one-source-out	VADER	74.97	79.24	89.25	54.85
Leave-one-source-out	IVADER-R	76.03	75.22	84.35	52.40
Leave-one-source-out	VPRS-Rough-only	73.10	72.86	82.03	46.79
Leave-one-source-out	Static fusion	77.00	76.01	85.75	54.33
Leave-one-source-out	RS-IVADER	76.47	75.59	84.83	53.27
Leave-one-source-out	Logistic regression	75.20	73.79	83.36	50.77
Leave-one-source-out	Linear SVM	74.90	73.22	83.30	50.26

Paired pooled comparisons support a bounded interpretation. RS-IVADER exceeded IVADER-R, logistic regression, linear SVM, VADER, and the rough-only branch after Holm correction, with adjusted p-values between 0.0117 and 0.0313. The comparison with equal-weight static fusion was not significant for either accuracy or F1-score. This pattern is consistent with complementary evidence rather than a universally superior fusion rule: the adapted lexical branch captures domain-conditioned term polarity, whereas the rough branch contributes class-consistency information from discretized feature states. When both branches agree, fusion can reinforce the decision; when they disagree, the symbolic trace can expose the conflict. The small difference from static fusion also shows that inner-fold weight selection added limited predictive value on this dataset, even though the nested procedure prevents test-set information from entering model selection.

Within-source performance reached 81.57% accuracy and 81.94% F1-score, and the difference from static fusion was not significant. Pooled evaluation was stronger because combining the three balanced sources increased the amount and diversity of training evidence while every test fold still contained examples from the same source mixture. Leave-one-source-out evaluation removed that advantage: the held-out source contributed no lexical statistics, medians, feature-state frequencies, or fusion-tuning data, so source-specific associations learned from the remaining two domains could become less reliable. Performance consequently declined to 76.47% accuracy and 75.59% F1-score, while static fusion reached 77.00% accuracy and 76.01% F1-score. This pattern is consistent with domain-adaptation and domain-generalization research showing that diversity among observed sources does not guarantee invariance to a previously unseen source [10]–[12]. Published results should therefore be compared cautiously across protocols: Barik and Misra [5] reported substantially higher accuracy for an adaptive VADER variant on different Amazon product domains, while other review-classification studies also show strong dependence on representation, dataset, and validation design [6], [36]. The inferential claims here are restricted to the identical-fold comparisons in Table 2.

4.2. Operational boundary and selective risk

Table 3 summarizes the pooled operating points and their practical consequences. The combined boundary deferred 53.55% of pooled observations; the remaining accepted cases achieved 92.54% accuracy, and 78.72% of all errors were concentrated in the deferred set. The resulting error-capture lift of 1.47 relative to boundary coverage indicates that deferment was informative rather than random. By comparison, random matched deferment achieved 83.64% accepted-case accuracy and captured 53.27% of errors. The relevant engineering question is therefore not whether deferment mechanically raises accepted-case accuracy, but whether the deferred set concentrates enough errors to justify the associated review workload.

The same comparison defines the principal uncertainty limitation. At the identical 53.55% deferment rate, logistic-regression and SVM confidence achieved accepted-case accuracies of 94.58% and 94.56%, respectively, with selective risks near 5.4%, compared with 7.46% for the combined symbolic-margin boundary. Margin-only confidence deferred fewer cases and captured fewer errors, yet its complete pooled risk-coverage area was 0.0612, below 0.0848 for the combined mechanism. Thus, the symbolic boundary is not the best pure error-ranking device. The lower-evidence criterion measures structural support consistency rather than a monotonic probability of misclassification, so a case can have weak or conflicting symbolic support and still be correctly separated by a linear margin. This distinction is consistent with selective-prediction and calibration work, which directly targets confidence quality or risk ranking [15], [20], [30], [37], [38], whereas rough-set boundaries encode indiscernibility and evidence consistency [17], [18], [21], [22]. The value of the symbolic status is therefore diagnostic: it can expose which feature-state evidence was absent, conflicting, or weak when deferment was triggered. This distinction matters in engineering design. Let c_h denote the unit cost of human review and c_e the expected cost of an unreviewed classification error. A deferment policy is economically useful when the reduction in expected error cost among accepted cases exceeds the review cost introduced by the deferred fraction. The present results do not estimate organization-specific monetary values, so no cost-saving claim is made; instead, Table 3 supplies the coverage and error-capture quantities needed to instantiate such a calculation. Systems prioritizing minimum selective risk may prefer calibrated linear confidence, whereas decision-support settings that require inspectable reasons may accept a modest risk penalty in exchange for symbolic traces.

Table 3. Pooled uncertainty mechanisms at the verified operating points

Mechanism	Boundary (%)	Accepted accuracy (%)	Selective risk (%)	Errors captured (%)
Combined	53.55	92.54	7.46	78.72
Margin only	16.47	88.52	11.48	41.09
Lower evidence only	49.15	89.83	10.17	68.12
Logistic matched	53.55	94.58	5.42	85.36
SVM matched	53.55	94.56	5.44	85.58
Random matched	53.55	83.64	16.36	53.27

4.3. Feature sensitivity, interpretability, and evidential boundaries

Feature sensitivity further narrows the performance claim. In a one-repeat pooled control, 300 dependency-ranked features produced the highest F1-score, 84.29%, compared with 83.93% for the principal 600-feature configuration. Chi-square selection with 600 features reached 84.15%, while random selection reached 83.77%. Figure 2 shows a relatively flat region between 300 and 1,200 dependency-ranked features, so 600 features should be interpreted as a transparent controlled configuration rather than a unique optimum or mathematically minimal reduct. The modest differences are also consistent with sentiment feature-selection research showing that weighting and subset choice can influence performance without implying a single universally optimal representation [36].

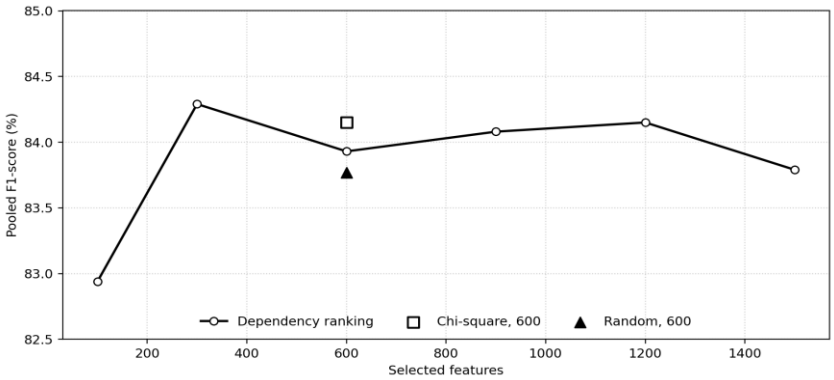


Figure 2. Pooled feature-count sensitivity and selection controls

Table 4 reports the symbolic state-mass composition before its interpretability implications are discussed. In pooled evaluation, 71.19% of matched states were absent, 8.48% were low, and 20.33% were high. Absence is informative in sparse text, but its dominance means that an explanation interface should distinguish non-occurrence from observed positive evidence. Positive-state summaries, absence caps, or conjunction rules may provide more intuitive traces. This audit reinforces recent work arguing that interpretability must be evaluated against the actual decision process rather than assumed from readable outputs [23], [24], [39], [40].

Table 4. Symbolic state-mass composition

Scenario	Absent state (%)	Low state (%)	High state (%)
Within-source	77.74	4.05	18.21
Pooled	71.19	8.48	20.33
Leave-one-source-out	75.36	4.74	19.90

The relationship to the earlier Improved VADER study must remain precise. Barik and Misra [5] supplied the central domain-adaptive lexicon motivation and reported strong results under their own protocol. This article does not claim to reproduce those figures. IVADER-R is a documented comparison model evaluated on the same partitions as RS-IVADER. The resulting pooled improvement of 0.50 F1 points over IVADER-R is modest but supported by paired folds, while static fusion remains statistically competitive. Separating published contextual evidence from independently regenerated fold-level evidence prevents non-equivalent experiments from being combined into a single superiority claim.

The findings clarify how symbolic reasoning differs from probabilistic uncertainty mechanisms, as summarized in Table 5. Calibration methods aim to align predictive scores with empirical outcome frequencies, while confidence and ensemble methods rank or aggregate predictive evidence; these mechanisms are appropriate when probability quality or selective-risk ordering is central [15], [20], [30], [37], [38]. RS-IVADER instead records whether discretized feature states receive sufficiently consistent training support. The rough-set boundary is therefore an explanation-bearing structural signal, not a substitute for Bayesian uncertainty, ensemble variance, or calibrated confidence. In deployment these mechanisms can be complementary: calibrated confidence can rank cases by expected reliability, while symbolic traces can support audit and human review.

Table 5. Comparison of uncertainty mechanisms for deployment

Mechanism	Primary signal	Probability-calibrated	Inspectable reason	Best use in this study
RS-IVADER boundary	Lower-region support + fused margin	No	Yes	Auditable triage
Linear confidence	Decision score/margin	No (unless calibrated)	Limited	Low selective risk
Post-hoc calibration	Empirical score-probability mapping	Yes, by design	Limited	Cost-sensitive probability decisions
Ensemble uncertainty	Predictive dispersion	Potentially	Limited	Model-uncertainty estimation

4.4. Engineering deployment and decision-support implications

A customer-review platform can integrate RS-IVADER as a service between text ingestion and downstream decision support. After language detection and language-specific preprocessing, an English model can produce polarity, fused score, boundary status, and supporting feature-state records. Certain cases may feed dashboards or aggregate service-quality indicators directly; boundary cases can be routed to a reviewer queue, a domain-specific model, or a contextual transformer. In multilingual pipelines, lexicons, tokenization, and symbolic thresholds must be fitted separately or aligned across languages; the present English-only experiment does not establish multilingual transfer. For real-time monitoring, the boundary rate itself can be tracked as an operational signal: a sustained rise may indicate vocabulary drift or a new review source and can trigger retraining or model-governance checks. This architecture links uncertainty-aware sentiment analysis to monitoring and control through a measurable feedback variable, while avoiding the stronger claim that the current study validates a closed-loop control system.

4.5. Threats to validity and future work

Several threats limit generalization. The verified data contain only three small English sources with binary labels; no neutral, ordinal, multilingual, code-mixed, aspect-level, or temporal-drift task was evaluated, and no live customer-review platform was instrumented. Leave-one-source-out evaluation contains only three source-level test pairs, which is insufficient for strong population-level inference. The sparse representation does not model long-range context as directly as transformer architectures, and no domain-adapted transformer or deep-ensemble baseline was included. The diagnostic Brier values are not calibrated probabilities. The boundary thresholds were predeclared rather than optimized under a deployment cost function, and a full joint sensitivity surface was not regenerated from the manuscript summary alone. Finally, feature-state explanations were not evaluated with human users for speed, comprehension, or trust calibration. These constraints are material because recent uncertainty and explainability studies show that confidence quality can degrade under shift and that readable explanations are not automatically faithful or useful [37]–[40].

Future work should therefore evaluate larger and more heterogeneous corpora, include domain-adapted compact transformers and ensemble/calibrated baselines under identical folds, regenerate dependence-aware confidence intervals from the archived predictions, and report a complete sensitivity surface for the 0.80 lower-region confidence, 0.15 lower-evidence, and 0.10 margin operating constants. Cost-sensitive validation or conformal operating points can then be compared with the symbolic boundary. Multilingual and aspect-level extensions should separate language-specific lexical adaptation from shared symbolic logic, while field studies should measure review time, avoided error cost, reviewer agreement, and whether symbolic traces improve human decisions rather than merely adding cognitive load.

5. CONCLUSION

This study evaluated RS-IVADER, a leakage-controlled framework that combines an adaptive sentiment lexicon with variable-precision rough-set evidence and a separate operational boundary status. Across 3,000 balanced Amazon cells, IMDb, and Yelp sentences, all adaptive operations were fitted inside training partitions and assessed under repeated within-source, pooled, and leave-one-source-out protocols. The strongest pooled result was 83.73% accuracy, 83.76% F1-score, 91.42% AUC, and 67.48% MCC. RS-IVADER exceeded several component and linear baselines, but equal-weight static fusion remained statistically competitive. Its combined boundary captured 78.72% of errors and raised accepted-case accuracy to 92.54%; however, matched linear confidence achieved lower selective risk. The contribution is therefore an interpretable, reproducible review-triage mechanism that complements predictive scoring rather than a universally superior classifier or uncertainty estimator. The evidence remains limited to small English benchmark datasets. Future work should evaluate larger and multilingual corpora, stronger contextual and

calibrated baselines, dependence-aware interval estimates, joint threshold sensitivity, deployment cost, and the usefulness of symbolic traces in human review.

ACKNOWLEDGMENTS

The authors acknowledge the Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia, for academic and research support. The authors also acknowledge the maintainers of the UCI sentiment labelled sentences dataset, the VADER lexicon, and the open-source Python ecosystem used in the reproducibility pipeline. No specific research grant or contract from any public, commercial, or not-for-profit funding agency supported this study.

FUNDING INFORMATION

The authors state that no external funding was involved in this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Saleh Yaakub	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Mustafa Mat Deris	✓	✓		✓						✓		✓	✓	✓
Ika Safitri Windiarti				✓	✓		✓			✓	✓			✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The benchmark sentences are publicly available from the UCI Machine Learning Repository under the Sentiment Labelled Sentences dataset [40]. The verified experimental archive contains the code, configurations, seeds, fold identifiers, observation-level predictions, metrics, symbolic records, manifests, and checksums used to regenerate the reported results. Derived experimental materials are available from the corresponding author upon reasonable request.

REFERENCES

[1] K. Alahmadi, S. Alharbi, J. Chen, and X. Wang, "Generalizing sentiment analysis : a review of progress , challenges , and emerging directions," *Social Network Analysis and Mining*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.1007/s13278-025-01461-8.
[2] M. Kumar, L. Khan, and H.-T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," *PeerJ Computer Science*, vol. 11, p. e2592, Jan. 2025, doi: 10.7717/peerj-cs.2592.
[3] M. Birjali, M. Kasri, and A. Beni-hssane, "Knowledge-Based Systems A comprehensive survey on sentiment analysis : Approaches , challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.
[4] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Feb. 2022, doi: 10.1007/s10462-022-10144-1.
[5] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," *Journal of Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-023-00861-x.
[6] H. J. Alantari, I. S. Currim, Y. Deng, and S. Singh, "International Journal of Research in Marketing An empirical comparison of machine learning methods for text- based sentiment analysis of online consumer reviews," *International Journal of Research in Marketing*, vol. 39, no. 1, pp. 1–19, 2022, doi: 10.1016/j.ijresmar.2021.10.011.
[7] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Computer Science Review*, vol. 41, p. 100413, 2021, doi: 10.1016/j.cosrev.2021.100413.
[8] Q. Li *et al*., "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022, doi: 10.1145/3495162.

[9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification," *ACM Computing Surveys*, vol. 54, no. 3, 2021, doi: 10.1145/3439726.

[10] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in NLP—a survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 6838–6855, doi: 10.18653/v1/2020.coling-main.603.

[11] S. Gururangan *et al*., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, pp. 8342–8360, 2020. doi: 10.18653/v1/2020.acl-main.740.

[12] K. Zhou, Z. Liu, Y. Qiao, T. Xiang and C. C. Loy, "Domain Generalization: A Survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396-4415, 1 April 2023, doi: 10.1109/TPAMI.2022.3195549.

[13] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," *Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics*, pp. 3881–3906, 2024. doi: 10.18653/v1/2024.findings-naacl.246.

[14] D. Herrera-Poyatos *et al*., "An overview of model uncertainty and variability in LLM-based sentiment analysis: challenges, mitigation strategies, and the role of explainability," *Frontiers in Artificial Intelligence*, vol. 8, Aug. 2025, doi: 10.3389/frai.2025.1609097.

[15] Y. Ovadia *et al*., "Can You Trust Your Model's Uncertainty ? Evaluating Predictive Uncertainty Under Dataset Shift," *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf

[16] C. J. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014, vol. 8, no. 1, pp. 216–225, doi: 10.1609/icwsm.v8i1.14550.

[17] I. T. R. Yanto, P. Vitasari, T. Herawan, and M. M. Deris, "Applying variable precision rough set model for clustering student suffering study's anxiety," *Expert Systems with Applications*, vol. 39, no. 1, pp. 452–459, Jan. 2012, doi: 10.1016/j.eswa.2011.07.036.

[18] W. Ziarko, "Variable precision rough set model," *Journal of Computer and System Sciences*, vol. 46, no. 1, pp. 39–59, Feb. 1993, doi: 10.1016/0022-0000(93)90048-2.

[19] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," 2021, *arXiv*. doi: 10.48550/ARXIV.2107.07511.

[20] Y Geifman and R El-Yaniv, "SelectiveNet : A Deep Neural Network with an Integrated Reject Option," *International conference on machine learning*, 2019. [Online]. Available: https://proceedings.mlr.press/v97/geifman19a/geifman19a.pdf

[21] K. Liu *et al*., "A survey on rough feature selection: recent advances and challenges," *IEEE/CAA Journal of Automatica Sinica*, vol. 13, no. 3, pp. 521–542, Mar. 2026, doi: 10.1109/JAS.2025.125231.

[22] Y. Chen and Y. Chen, "Feature subset selection based on variable precision neighborhood rough sets," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 572–581, 2021, doi: 10.2991/ijcis.d.210106.003.

[23] Q. Lyu, M. Apidianaki, and C. Callison-Burch, "Towards Faithful Model Explanation in NLP: A Survey," *Computational Linguistics*, vol. 50, no. 2, pp. 657–723, 2024, doi: 10.1162/coli_a_00511.

[24] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386.

[25] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From Group to Individual Labels Using Deep Features," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, pp. 597–606, Aug. 10, 2015. doi: 10.1145/2783258.2783380.

[26] F. Pedregosa *et al*., "Scikit-learn : machine learning in Python," *The Journal of machine Learning* research, vol. 12, pp. 2825–2830, 2011.

[27] C. R. Harris *et al*., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[28] P. Virtanen *et al*., "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," *International conference on machine learning*, 2017. [Online]. Available: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

[30] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.

[31] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[32] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[33] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[34] J. Pineau *et al*., "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)," 2020, *arXiv*. doi: 10.48550/ARXIV.2003.12206.

[35] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Information Processing & Management*, vol. 58, no. 5, p. 102656, Sep. 2021, doi: 10.1016/j.ipm.2021.102656.

[36] A. N. Holm, D. Wright, and I. Augenstein, "Revisiting softmax for uncertainty approximation in text classification," *Information*, vol. 14, no. 7, p. 420, Jul. 2023, doi: 10.3390/info14070420.

[37] T. S. Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.

[38] S. El Mrabti, J. EL-Mekkaoui, A. Hachmoud, and M. Lazaar, "An explainable machine learning model for sentiment analysis of online reviews," *Knowledge-Based Systems*, vol. 302, no. August, p. 112348, Oct. 2024, doi: 10.1016/j.knosys.2024.112348.

[39] Y. Sun, S. He, X. Han, and Y. Luo, "Interpretability in sentiment analysis: a self-supervised approach to sentiment cue extraction," *Applied Sciences*, vol. 14, no. 7, p. 2737, Mar. 2024, doi: 10.3390/app14072737.

[40] D. Kotzias, "Sentiment Labelled Sentences." *UCI Machine Learning Repository*, 2015. doi: 10.24432/C57604.

BIOGRAPHIES OF AUTHORS



Saleh Yaakub is a Doctoral researcher in Information Technology at the Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia. His research focuses on machine learning, natural language processing, text classification, sentiment analysis, symbolic reasoning, and reproducible evaluation. In this study, he led the conceptualization, software implementation, experimentation, formal analysis, data curation, visualization, and original manuscript preparation. He can be contacted at email: p5240052@student.umam.edu.my.



Mustafa Mat Deris is a Professor and research supervisor at the Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia. His scholarly work spans computer science, data management, intelligent systems, rough-set-based methods, text classification, and advanced information technology. In this study, he contributed to conceptualization, methodological review, validation, supervision, project administration, funding acquisition oversight, and manuscript review and editing. He can be contacted at email: mustafa@umam.edu.my.



Ika Safitri Windiarti is affiliated with Computer Science at Universitas Muhammadiyah Palangkaraya and the Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia. Her research interests include computer science, machine learning, natural language processing, information systems, and applied data analysis. In this study, she contributed to validation, formal analysis, resources, visualization, investigation, and manuscript review and editing. She can be contacted at email: ika.windiarti@umpr.ac.id, ika.windiarti@umam.edu.my.