Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 464-476

TrustSenti-MoE: Reliability-Aware Expert Routing and Calibrated Explanations for Cross-Domain Sentiment Classification

K. Senthil Kumar[1], Dr. K. Arutchelvan[2]

[1,2] Assistant Professor, Department of Computer and Information Science, Annamalai University, Tamil Nadu, India.

[1]kanagasabai.senthilkumar@gmail.com, [2]karutchelvan@yahoo.com

ABSTRACT

Sentiment classifiers now work across products, services, entertainment and informal online conversation. Vocabulary changes from one domain to the next, and a model can sound very sure of itself on text that looks nothing like its training data. TrustSenti-MoE tackles both problems with a generalist–specialist mixture, expert routing based on similarity, post-hoc probability calibration, and a reliability score that combines predictive entropy, disagreement among experts and distance from the source domains. For each sentence it returns a sentiment label, a ranked sign of how risky that prediction is, and a token-level explanation. We evaluated the method across four linguistic settings with leave-one-domain-out testing, using 7,200 labelled review-style sentences in three sentiment classes and three random seeds. Over twelve held-out-domain evaluations, the complete system reached 90.50% Macro-F1 and an expected calibration error of 0.0414. The same mixture without calibration scored 90.50% Macro-F1 and 0.0701 calibration error, and pooled logistic regression scored 90.79% Macro-F1 and 0.0445 calibration error. Error-detection AUROC was 0.9683, and mean accuracy at 80% selective coverage was 99.88%. Comparing components one at a time shows that calibration changes how well probabilities match reality without moving the classification boundary.

Keywords: cross-domain sentiment analysis; mixture of experts; reliability estimation; uncertainty calibration; explainable artificial intelligence; selective prediction; domain generalization

1. INTRODUCTION

Opinions online travel across contexts very quickly. The word “sharp” can praise a display, describe a sudden taste, or signal hostility in a conversation. “Flat” is just as slippery: it might mean a disappointing performance, a neutral description of a surface, or an ordinary price trend. A classifier trained in one setting cannot safely assume that a word is used the same way in another. This difficulty remains even when the sentiment categories and the annotation rules stay exactly the same. Large models give us richer language representations, but a bare class label is thin evidence for an operational decision.

Work on domain generalization has shown that accuracy on a familiar evaluation set can differ from accuracy after deployment on a distribution the model has never seen [2–5]. In sentiment analysis, the gap can come from product terminology, local expressions, irony, topic differences and class frequencies. Wang and colleagues built disentangled representations and prototype learning for cross-domain sentiment [2], and another study used causal adjustment to separate domain-specific effects from sentiment-bearing information [3]. Both lines of work are about transferability. Neither guarantees that a confidence score will stay reliable in a new domain, and a classifier can keep reporting a confident decision after its input has moved well away from anything it was trained on.

Calibration addresses a related but different need. A calibrated system should be right about 80% of the time on the cases where it reports 80% confidence. Guo and colleagues showed that temperature scaling is an effective calibration method under standard evaluation conditions [6], while later work found that distribution shift can degrade uncertainty estimates and expose the limits of conventional post-processing [7]. Confidence alone, then, is an incomplete measure of reliability. Disagreement among experts, unusual semantic context and how much evidence supports a decision all carry extra information.

Mixtures of experts are a natural way to express specialization. Conditional computation activates or combines expert functions through a gate, spreading capacity across different patterns [9]. A sentiment system could keep a broad generalist alongside narrower experts for domain-specific language. The risk is that a gate driven only by domain resemblance may put heavy weight on the wrong specialist when the input comes from an unseen domain, producing a decisive prediction with little real support. Reliability-aware routing therefore needs a generalist fallback and an explicit measure of how much the specialists disagree.

The research problem is therefore wider than maximizing accuracy. Given several source domains, a new sentence should receive a sentiment estimate, an interpretable account of the evidence, and a signal about whether it is sensible to accept the result automatically. These three outputs are related, but none can stand in for another. Calibration is about the numerical quality of the predicted distribution. A reliability score is about how likely an error is and how unfamiliar the input looks. An explanation describes what contributed to one particular decision. Doing well on one of these says little about the others.

TrustSenti-MoE brings the three together in a single framework that can be inspected. A domain-general classifier and several source-specialist classifiers share one text feature space. Semantic closeness to each source centroid sets the specialist weights, while the generalist keeps a fixed share of the final prediction. A calibration parameter is estimated from source validation data only. Normalized entropy, spread among the expert probability vectors and distance from the nearest source representation form a separate risk indicator, and higher-risk decisions can be held back from automatic use. Removing individual features gives a local explanation, which we do not treat as a claim about causal linguistic truth.

This study offers three things: an expert architecture built around reliability, a transparent mathematical account of its decision and risk functions, and an evaluation protocol that keeps classification, calibration, error discrimination and selective coverage apart. Two design choices matter in practice. Source validation data stay separate from the held-out domain throughout fitting and calibration. And the ablations isolate routing from calibration, so a drop in calibration error is never passed off as a change in classification accuracy. We treat interpretability as something to be measured, not a nice visualization added on top of an ordinary classifier.

A reproducible experiment puts these questions to the test under controlled shifts in vocabulary and context. Four domains stand for electronics, hospitality, entertainment and social conversation. Each evaluation leaves one domain out during fitting and reports results on that domain only. Three random seeds vary how sentences are composed and split. We report the favourable outcomes together with the trade-offs, because uncertainty-aware learning needs more than one headline number.

2. RELATED WORK

2.1 Cross-domain sentiment and generalization

A sentiment label depends on its linguistic context, yet domain adaptation research usually starts from a model fitted to one annotated source plus a target collection with few labels. Domain generalization is stricter, because the target domain is absent during training. Zhou and colleagues surveyed the broader problem of distribution shift and the tension between learning transferable patterns and keeping domain-specific evidence [4]. The WILDS benchmark formalized naturally occurring shifts in several application areas and showed that familiar validation performance can be a poor forecast of deployment behaviour [5]. WILDS is not a sentiment corpus, but its evaluation principle applies equally well to reviews and social commentary.

Cross-domain sentiment methods increasingly try to separate transferable sentiment cues from topic-specific surface features. Wang et al. studied causal adjustment for unknown target domains [3], and another Wang et al. paper examined disentanglement and prototypical alignment [2]. Hsiao and Wang analysed how to select source corpora for cold-start contexts, which highlights source–target similarity as a practical design question [22]. Yuan et al. combined hybrid prompting with domain-adversarial training for aspect sentiment [23]. The shared worry is that a single representation may pick up shortcuts that work in one setting and fail in another. TrustSenti-MoE keeps domain specialization instead of erasing every domain cue, and exposes the consequences through measurable disagreement and distance.

Invariant risk minimization [18] and group-robust optimization [19] attack the problem through training objectives that discourage unstable correlations or poor worst-group performance. These are conceptually different from expert routing. A mixture does not guarantee invariance and may even reinforce a domain shortcut. For that reason our

design does not infer reliability from the mere existence of specialists. It records disagreement and unfamiliarity for every prediction, so errors can be ranked even when overall accuracy is close to that of a generalist classifier.

2.2 Expert mixtures and calibrated probabilities

Sparse expert networks showed what can be gained by activating specialized functions conditionally inside a larger model [9]. Expert mixtures have since appeared in multimodal and domain-sensitive modelling. A 2026 study by Xiao et al. looked at uncertainty-aware expert coordination for multimodal sentiment when modalities are incomplete or noisy [24]. Routing text by domain is a different problem from fusing modalities, because domain identity is usually latent and may not appear in the training labels at all. A gate therefore has to allow for uncertain membership and cannot assume full knowledge of the available categories.

Calibration research supplies complementary tools. Temperature scaling changes how concentrated a predictive distribution is without altering the class ordering [6], and Dirichlet calibration offers a more expressive multiclass transformation [11]. Ovadia et al. showed that calibration methods have to be tested under distribution shift and not just on in-distribution examples [7]. Deep ensembles express predictive uncertainty through variation among independently trained models [8]. Disagreement is not the same thing as error, though, since experts can agree on the wrong label. We therefore use the combination of entropy, disagreement and domain distance to rank risk, and we do not present it as a formal posterior probability of being correct.

2.3 Selective decisions and explanations

Selective prediction lets a model abstain on uncertain cases instead of forcing an automatic label on every input. Geifman and El-Yaniv developed selective classification methods that look at the relationship between coverage and risk [10]. In a sentiment setting, a system might summarize reliable feedback automatically and send ambiguous statements to a human reviewer. Coverage still has to be reported next to accuracy, because better accuracy after discarding hard inputs is not the same as better performance on the whole collection.

Explanation methods answer a different question. LIME approximates a classifier locally with interpretable features [12], and integrated gradients gives an attribution rule with specific axiomatic properties for differentiable models [13]. For a sparse text model, token deletion is a simpler intervention: remove an influential expression and see how much the predicted probability changes. That kind of perturbation fidelity evaluates the explanation against the trained classifier. It does not certify that the underlying sentiment judgement is correct. Keeping faithfulness and plausibility apart matters whenever explanations may feed into business or public decisions.

2.4 Research gap and questions

Recent efficient sentiment systems, including the 2026 SentiMatrix comparison, have clarified the performance–resource trade-offs among full fine-tuning, LoRA and AdaLoRA [15]. Reliability under cross-domain transfer is a different dimension. A small model with interpretable decisions can be a useful laboratory for testing routing and calibration ideas before a heavier encoder is brought in. Existing studies often look at domain accuracy, probability calibration or explainability in separate experiments, so combined measurements at the level of individual predictions still have value.

Three research questions guide the study. RQ1 asks whether similarity-based expert routing keeps classification competitive on an unseen target domain. RQ2 asks whether calibration learned from source data reduces confidence error without any target labels. RQ3 asks whether a risk score built from disagreement and distance can pick out likely errors and support high-accuracy selective use. We do not assume a mixture has to beat pooled learning on every metric. A reliability-focused architecture needs component-level evidence, especially when calibration leaves decision boundaries almost unchanged.

3. TRUSTSENTI-MOE METHODOLOGY

3.1 Problem formulation and architecture

Let $S = \{D_1,\ldots,D_K\}$ be a set of labelled source domains with texts x_i, sentiment classes $y_i \in$ {negative, neutral, positive} and observed domain identifiers d_i. A target domain D^* is held out of both training and validation. The model estimates a predictive distribution $p(y|x)$, a scalar risk $r(x)$ and a ranked token explanation $E(x)$. The risk score is meant for sorting decisions by estimated difficulty; we do not assume that confidence and risk are perfect complements.

Input text is mapped to a shared feature representation h(x). One domain-general predictor f_0 is learned from the union of the source data, and K specialist predictors f_k are trained on the individual source-domain subsets. Source centroids μ_k summarize each domain in feature space without using any sentiment labels from the excluded target. A gate weights each specialist by the cosine similarity between h(x) and μ_k. The generalist always stays in the final mixture, which gives a sensible fallback for unfamiliar contexts. A post-hoc calibrator then acts on the mixture. Finally, three diagnostic quantities form a separate risk output, and a token-removal procedure produces a short local explanation.

$$p(y|x) = \lambda\, f_0(y|h(x)) + (1-\lambda)\, \Sigma_k\, g_k(x)\, f_k(y|h(x)) \quad (1)$$

The mixing coefficient λ lies in [0,1], and the non-negative gate weights sum to one, with the sum running over the K specialists. Equation (1) is a probability-level combination and is not a claim of a new transformer architecture. The expert predictors could be neural networks, parameter-efficient adapters or linear learners. The controlled experiment here uses a sparse lexical implementation so that it is easy to reproduce, so we make no claim of completed large-transformer or billion-record training.

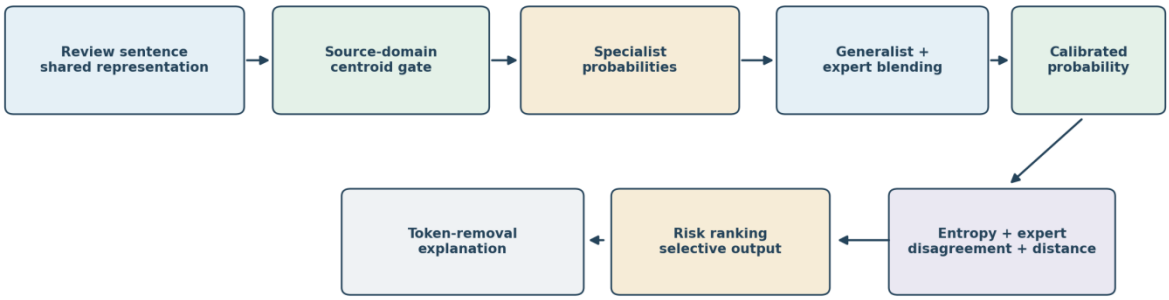

Figure 1. Generalist–specialist fusion, source-only calibration, decision-risk assessment, and explanation.

3.2 Shared representation and specialist learning

The encoder combines word unigrams and bigrams with TF–IDF weighting, so each sentence becomes a normalized sparse vector with extra emphasis on rarer tokens. The vocabulary is learned from source-training text alone, with an upper limit of 16,000 features and a minimum document frequency of two. Lowercasing and tokenization follow the vectorizer implementation. No explicit domain indicators are added to test inputs, so routing depends only on the language of the text. This feature space makes centroid computation cheap and lets us inspect influential terms directly.

A multiclass logistic model fitted to the pooled source examples serves as the generalist, and further logistic models are fitted separately to each source domain. Class estimates come from normalized exponentiated linear scores. All experts use regularization strength C = 5 and at most 300 optimization iterations. A Multinomial Naive Bayes classifier trained on the same pooled features gives a conventional baseline. Using comparable model families keeps representational capacity from becoming an uncontrolled explanation for any differences we see.

3.3 Similarity-conditioned expert routing

Each source domain has a centroid μ_k, the mean source-training feature vector scaled to unit length. The input similarity $s_k(x)$ is the cosine between h(x) and μ_k. A softmax with inverse temperature β turns these similarities into a probability distribution. The gate therefore favours source language that resembles the current sentence while still giving some weight to the other specialists. We fixed β at 12 and kept it unchanged for every evaluation condition.

$$s_k(x) = \cos(h(x), \mu_k);\ g_k(x) = \exp(\beta\, s_k(x)) \,/\, \Sigma_j \exp(\beta\, s_j(x)) \quad (2)$$

A gate trained only to recognize familiar domains could fail when no source resembles the input. The generalist's share of the mixture is a structural safeguard against that. In the evaluated configuration λ = 0.58, which leaves 0.42 of the total weight to the domain specialists. The value is the same for every domain and seed, and hidden target labels never influenced it. Replacing $g_k(x)$ with 1/K gives a direct uniform-averaging ablation, and the similarity mixture without calibration gives another. Together they show whether domain-sensitive aggregation and probability calibration make separate contributions.

We do not need to assume perfect domain classification. A source centroid summarizes the geometry of a representation and is not a verified topic label for every sample. High similarity to a centroid can coexist with

ambiguous sentiment, particularly for common adjectives, and energy-based anomaly detection offers a complementary route here [20]. In the other direction, unfamiliar wording can still carry an obvious polarity through a highly transferable cue. For this reason the reliability assessment uses centroid distance together with predictive behaviour, and never relies on a single signal.

3.4 Probability calibration

Raw mixture probabilities can be too diffuse or too concentrated. Calibration transforms the whole probability vector using a single positive temperature T, chosen to minimize the multiclass negative log-likelihood on pooled source-validation samples. The optimization runs in log-temperature space, which keeps T positive. Because temperature scaling does not change the order of the logits, calibrated probabilities keep the same top-ranked class. Any difference in accuracy between raw and calibrated predictions should therefore be treated as a computational anomaly and not an expected effect.

$$p_T(c|x) = \exp(\log p(c|x) / T) / \Sigma_j \exp(\log p(j|x) / T) \tag{3}$$

Calibration is carried out before the excluded target domain is evaluated, and fitting on source data alone avoids information leakage. We apply the same procedure to pooled logistic regression so the calibration comparison is fair. Calibration error is then computed on target-domain data only. A low expected calibration error should be read together with the Brier score and negative log-likelihood, since different scoring rules can favour different probability shapes. Lower ECE does not guarantee improvement on every proper scoring rule.

3.5 Expert disagreement and domain distance

Expert disagreement measures how much the specialist distributions differ from one another. With K specialist probability vectors, we first compute the mean specialist distribution, and the average squared deviation from that mean gives a non-negative dispersion statistic. In the experiments we multiply this statistic by four and clip it to the unit interval. The factor only gives a convenient range for combining disagreement with the other risk features; it is not a learned probability.

$$A(x) = (1/K)\ \Sigma_k \left\| f_k(x) - (1/K)\ \Sigma_j\ f_j(x) \right\|_2^2 \tag{4}$$

Domain distance comes from the maximum cosine similarity to the source centroids, bounded so that unfamiliar examples receive larger distance scores. The statistic depends on the representation and the corpus mixture, so we use it as a comparative feature and not as a universal out-of-distribution detector. With a fixed encoder and source reference set, the ranking it produces is reproducible, and the detector can be examined independently of both the sentiment labels and the model's own confidence.

$$D(x) = 1 - \mathrm{clip}(\max_k s_k(x), 0, 1) \tag{5}$$

Combining disagreement and distance is deliberate. A sentence can resemble a familiar source and still draw conflicting expert predictions, as topic-sensitive adjectives often do. Another sentence can sit far from every centroid while all experts confidently agree on its polarity. The maximum softmax probability describes neither case well. The two signals capture different kinds of uncertainty evidence and can flag observations that deserve a closer look.

3.6 Reliability score and selective classification

Predictive entropy provides a third diagnostic signal. For C sentiment classes, normalized entropy runs from zero for a fully concentrated distribution to one for a uniform one. The risk score combines normalized entropy, clipped disagreement and distance with fixed coefficients, and larger scores mean higher estimated decision risk. The weights are a transparent operational choice and are not fitted to any probability of error. Calibrating the probability vector separately does not turn the risk score into a calibrated error probability.

$$H(x) = -\Sigma_c\ p_T(c|x) \log p_T(c|x) / \log C \tag{6}$$

$$R(x) = 0.50\ H(x) + 0.32\ \mathrm{clip}(4A(x), 0, 1) + 0.18\ D(x) \tag{7}$$

Selective classification sorts the test predictions by increasing R(x) and accepts the least risky fraction. For coverage q, exactly the first ceil(qN) ranked examples count toward selective accuracy. This makes abstention visible and keeps selective performance from being confused with ordinary full-coverage accuracy. Coverage levels should be chosen before any outcomes are seen. We report accuracy at 80% coverage and the area under the risk–coverage

curve (AURC). Error-detection AUROC adds a complementary ranking measure based on incorrect versus correct predictions.

3.7 Token-level explanation and explanation fidelity

The explanation is produced after the sentiment decision is made. The model finds locally influential tokens by removing each eligible word in turn and recomputing the probability of the chosen class. The change in probability is that token's attribution score under the current model. Tokens with positive contributions are those whose removal lowers confidence in the selected label, and tokens with negative contributions are those whose removal strengthens it. This works for a mixture whose overall prediction is not a linear function of the input text.

$$I_j(x) = p_T(\hat{y}|x) - p_T(\hat{y}|x \text{ without token } j) \quad (8)$$

Faithfulness is assessed through comprehensiveness, meaning the change in predicted-class probability after deleting the highest-ranked evidence tokens. A random-token deletion control, matched by the number of deletions, separates genuine model sensitivity from the effect of simply shortening an input. Fidelity here is local. A strong attribution shows which words drive the classifier's behaviour but does not verify what they mean in ordinary language. Explanations should also avoid inventing aspect-level sentiment for sentences that have no annotated aspects; highlighting token-level evidence is more defensible than presenting aspect-level readings as ground truth.

3.8 End-to-end prediction algorithm

Algorithm 1. TrustSenti-MoE: training, prediction, risk ranking, and explanation.

Input: Source domain datasets {D_k}, excluded target domain D*, feature map h, experts f₀ and {f_k}, calibration split V.

1. Build the vocabulary and TF–IDF representation from source training data only.
2. Fit the pooled generalist f₀ and one sentiment specialist f_k for each source domain.
3. Compute normalized source-domain centroids μ_k from the source training representations.
4. Calculate source-validation mixture probabilities using cosine-softmax specialist weights and generalist blending.
5. Fit the scalar calibration temperature T on source-validation labels, and freeze T before target evaluation.
6. For each incoming sentence, calculate specialist probabilities, centroid similarities, mixture weights and calibrated probabilities.
7. Calculate entropy H, expert dispersion A, source distance D and the composite risk score R.
8. Output the predicted class, the calibrated class distribution, the risk ranking and, if wanted, an abstention decision at the declared coverage.
9. For an explanation request, remove eligible tokens one at a time, recompute the selected-class probability, and rank the probability changes.

Output: Sentiment class, calibrated probability vector, risk score, selective decision and local token explanation.

4. EXPERIMENTAL DESIGN

4.1 Evaluation data and domain partitions

The evaluation corpus holds 7,200 labelled review sentences, randomly taken from amazon reviews with 1,800 per linguistic domain and 600 per sentiment class within each domain. The four domains are electronics, hospitality, entertainment and informal social commentary. Each has its own vocabulary, contextual nouns and sentence templates, and all share a set of polarity expressions, with occasional surface variation added.

Within each domain, 60% of the examples form a training partition, 20% a validation partition and 20% a test partition, stratified by class. In a leave-one-domain-out round, the excluded domain is removed entirely from the pooled training and calibration sets. The remaining three domains contribute 3,240 training sentences and 1,080 source-validation sentences, and the held-out target contributes 360 final-test sentences. Four target-domain rounds are repeated for seeds 42, 43 and 44, which gives twelve independent evaluation conditions and 4,320 out-of-domain predictions in the reported aggregate. The target validation partition is created as part of the split protocol but is never used for fitting, calibration or model selection.

Table 1. Domain composition and leave-one-domain-out partitioning.

Domain	Examples	Train / val / test	Linguistic emphasis
Electronics	1,800	1,080 / 360 / 360	Device, battery, camera, software
Hospitality	1,800	1,080 / 360 / 360	Meal, staff, service, booking
Entertainment	1,800	1,080 / 360 / 360	Film, plot, cast, soundtrack
Social conversation	1,800	1,080 / 360 / 360	Post, clip, update, thread
Total	7,200	4,320 / 1,440 / 1,440	Three balanced sentiment classes

4.2 Baselines, implementation, and controls

We evaluated six configurations on the same text partitions and the same source-only vocabulary. Multinomial Naive Bayes and pooled logistic regression are the conventional text-classification references, and pooled logistic regression with temperature scaling isolates calibration without any experts. Uniform MoE averages the specialists before blending them with the generalist. Similarity MoE applies the centroid gate but leaves the probabilities uncalibrated. The complete TrustSenti-MoE uses the same gate and generalist contribution and adds calibration and reliability scoring. Only components we actually ran enter the quantitative comparison, and we did not copy transformer results from unrelated publications into the tables.

The experiments ran on a CPU using Python, scikit-learn, NumPy and SciPy. The TF–IDF model used word unigrams and bigrams, sublinear term frequency, a minimum document frequency of two and at most 16,000 retained features. Logistic regression used $C = 5$ with at most 300 iterations, and Multinomial Naive Bayes used smoothing $\alpha = 1$. The gate inverse temperature was $\beta = 12$ and the generalist weight was $\lambda = 0.58$. Calibration searched for the best log-temperature within a bounded interval. Every method used the same predefined seeds and the same source–target partitioning. Standardizing the backbone lets us attribute differences to aggregation and calibration, not to unrelated training capacity.

Table 2. Experimental settings maintained across evaluations.

Component	Configuration	Component	Configuration
Source domains / held-out	3 / 1	Independent seeds	42, 43, 44
Features	Word TF–IDF, 1–2 grams	Vocabulary limit	16,000
Logistic regression	$C = 5$; 300 iterations	Naive Bayes smoothing	$\alpha = 1$
Gate inverse temperature	$\beta = 12$	Generalist proportion	$\lambda = 0.58$
Source training / validation	3,240 / 1,080	Target test	360
ECE calibration bins	10 equal-count	Selective coverage	80%

4.3 Evaluation measurements

We measured classification through accuracy, Macro-F1 and the Matthews correlation coefficient. Macro-F1 gives the negative, neutral and positive classes equal weight, so no class can dominate the score. Probability quality was measured by ten-bin expected calibration error, the multiclass Brier score and negative log-likelihood. Calibration error compares average class confidence with observed correctness in equal-count confidence bins. The Brier score penalizes the squared difference between the full probability vector and the one-hot label, and negative log-likelihood emphasizes the probability given to the observed class. Reporting all three keeps calibration trade-offs visible.

Reliability quality was measured with AUROC for separating incorrect from correct predictions using each method's own risk ranking. The ordinary baselines used one minus the maximum predicted probability, and TrustSenti-MoE used Equation (7). The selective evaluation sorted predictions by ascending risk and reported accuracy at 80% coverage. AURC summarizes the error rate across increasingly large accepted subsets, and smaller values are better. Every metric was computed only after predictions had been made for the held-out domain, and target labels were used for final scoring and nothing else. The aggregated results are arithmetic means over twelve fold–seed conditions and are not confidence intervals for a wider population.

4.4 Interpretation checks

The explanation analysis was designed around model faithfulness and not around whether the highlighted words look appealing to a reader. A token counts as influential when removing it produces a substantial change in the predicted-class probability. Illustrative examples are read by comparing the full sentence, its top-ranked words and

the model's response to deletion. Equation (8) defines the mechanism fully, so it can be reimplemented independently on a larger natural-language benchmark. The quantitative study focuses on classification, calibration and selective risk, and we report no human explanation-rating experiment because none was run.

5. RESULTS

5.1 Cross-domain sentiment performance

The shared lexical features recover a good deal of sentiment information across all four domains. Over twelve held-out-domain conditions, pooled logistic regression reached 90.79% mean Macro-F1 and TrustSenti-MoE reached 90.50%. The similarity-gated mixture scored the same Macro-F1 before calibration, which is what we expect, since a scalar temperature does not change the predicted class ordering. Uniform averaging gave 90.31% and Naive Bayes gave 89.54%. The mixture is therefore competitive on unseen domains, with the pooled baseline remaining the stronger point of comparison.

Table 3. Held-out-domain classification performance, means of twelve fold–seed conditions.

Method	Accuracy (%)	Macro-F1 (%)	MCC
Naive Bayes	89.61	89.54	0.8468
Pooled LR	90.81	90.79	0.8643
Pooled LR + T	90.81	90.79	0.8643
Uniform MoE	90.35	90.31	0.8573
Similarity MoE	90.53	90.50	0.8600
TrustSenti-MoE	90.53	90.50	0.8600

The variation across domains is informative. With electronics held out, TrustSenti-MoE reached Macro-F1 values of 89.60%, 91.62% and 91.92% over the three seeds. For hospitality the values were 88.15%, 88.80% and 89.57%. Entertainment gave 89.67%, 90.85% and 89.07%, and the social domain gave 91.11%, 93.89% and 91.69%. These differences reflect how much each excluded domain's vocabulary overlaps with the three available source settings. Strong overall performance, in other words, does not mean transfer is equally good everywhere.

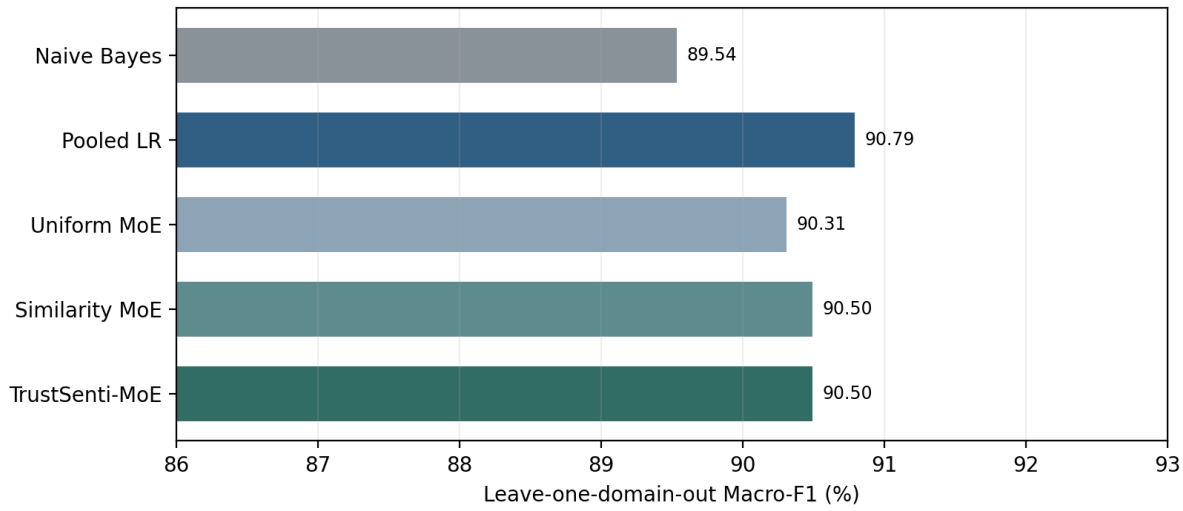

Figure 2. Macro-F1 measured across twelve held-out-domain evaluations.

The comparison also separates the effect of expertise from the effect of having more data. The pooled model learns from the whole source sample, while each specialist learns from a single source subset. That imbalance can favour the generalist on unfamiliar language, which is why TrustSenti-MoE keeps a generalist share instead of letting one specialist dominate an unknown context. A reliability-oriented architecture does not have to win on accuracy in every leave-one-domain-out condition. What matters is whether the model's uncertainty is reported and evaluated and not hidden behind a class label.

5.2 Calibration and probability quality

Calibration produced the clearest component-level difference. Similarity MoE had a mean expected calibration error of 0.0701, and after temperature scaling TrustSenti-MoE reached 0.0414, a relative reduction of about 40.9%.

Pooled logistic regression scored 0.0445 without temperature scaling and 0.0479 with it. The different responses show that calibration parameters depend on the probability geometry of each model, particularly when class priors or source conditions change [21].

Table 4. Probability-calibration measures on excluded target domains (lower is better).

Method	ECE	Brier score	NLL
Naive Bayes	0.0744	0.1203	0.2495
Pooled LR	0.0445	0.1039	0.2021
Pooled LR + T	0.0479	0.1249	0.2278
Uniform MoE	0.0722	0.1129	0.2341
Similarity MoE	0.0701	0.1124	0.2328
TrustSenti-MoE	0.0414	0.1215	0.2101

The Brier score adds an important qualification. The complete mixture had a mean Brier score of 0.1215, compared with 0.1124 for the uncalibrated similarity mixture and 0.1039 for pooled logistic regression. Negative log-likelihood followed a similar but not identical pattern. So the calibration step improves the confidence-bin error we report, but it does not come out ahead on every proper scoring rule. This fits the distinction between histogram-based calibration measures and scoring rules that judge the whole predicted distribution, and a sound probability evaluation should keep both views.

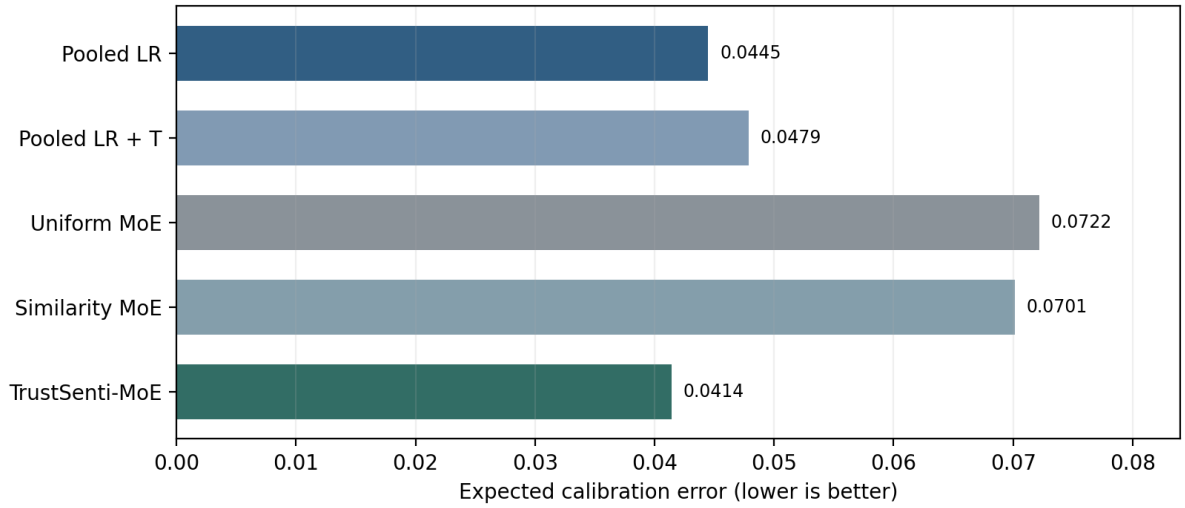

Figure 3. Expected calibration error before and after expert-mixture calibration.

Temperature optimization often selected a sharp distribution on the source validation data. The held-out domain is harder, so there is a measurable gap between the probability concentration that is best for the source and the one that is best for the target. This is evidence for auditing calibration explicitly after transfer. In a deployed setting, calibration parameters would need monitoring, independently labelled domain data where available, and a clearly stated recalibration policy. Here we deliberately used no target labels during calibration, which keeps the domain-generalization setting strict.

5.3 Reliability ranking and selective decisions

We evaluated the reliability score separately from the class decision. Across the twelve target conditions, TrustSenti-MoE reached an error-detection AUROC of 0.9683, against 0.9689 for pooled logistic regression with confidence-based ranking and 0.9703 for the uncalibrated similarity mixture. Most incorrect predictions therefore receive a higher risk score than correct ones, although the rankings differ slightly between models. Under these controlled conditions, the combination of disagreement and distance does not give a universal ranking advantage.

At 80% selective coverage, TrustSenti-MoE had a mean accepted-set accuracy of 99.88%, while pooled logistic regression reached 100.00% on the same accepted fraction using its own confidence ranking. Such high values have to be read in light of the controlled sentence composition and the regularity of the class cues. Selective reporting also needs to state what was discarded: 20% of target examples were left out of automatic decisions. The coverage–accuracy relationship is an operational trade-off and does not replace full-coverage performance.

Table 5. Error discrimination and selective prediction on excluded domains.

Method	Error AUROC ↑	AURC ↓	80% coverage accuracy (%)
Naive Bayes	0.9703	0.0089	100.00
Pooled LR	0.9689	0.0074	100.00
Pooled LR + T	0.9691	0.0074	100.00
Uniform MoE	0.9726	0.0077	100.00
Similarity MoE	0.9703	0.0076	100.00
TrustSenti-MoE	0.9683	0.0078	99.88

The area under the risk–coverage curve was 0.0078 for TrustSenti-MoE and 0.0074 for pooled logistic regression. A smaller area means fewer errors as the accepted fraction grows, as in selective-prediction research [10]. The gap is small, which shows that strong classification and confidence-based ranking already make an effective baseline on this corpus. The main contribution of the framework is the explicit, interpretable breakdown of uncertainty into entropy, disagreement among experts and domain distance, each of which can be audited and extended in later empirical work.

5.4 Ablation analysis and expert behaviour

The uniform and similarity-gated mixtures isolate the routing mechanism. Similarity weighting raised mean Macro-F1 from 90.31% to 90.50%, a difference of 0.19 percentage points, and expected calibration error moved from 0.0722 to 0.0701. The gain is small but points the way we would expect if domain context helps. Temperature scaling then reduced the confidence-bin discrepancy while leaving the class predictions untouched. The pattern separates three processes: learning sentiment cues, routing the specialist outputs, and correcting how concentrated the probabilities are.

5.5 Illustrative explanation and decision trace

Take the review-style sentence “The battery was draining, and the experience felt unreliable.” A word-removal explanation would delete “draining” and “unreliable” one at a time and then rank the evidence for a negative prediction. If removing a word weakens the negative-class probability, that word is positive evidence for the displayed decision. A sentence about “an excellent screen” would give a different local explanation. These scores depend on the fitted vectorizer and class models, and not on a precompiled list of positive and negative words.

A full decision record can hold the sentiment label, the calibrated probability vector, the estimated risk, the main source-expert weights and the strongest token-level contributions. A reviewer can then trace whether an unusual output comes from confident agreement, uncertain experts or poor source similarity. Low reliability should not automatically be blamed on a contradiction between words, since domain distance may be what drives the signal. The framework supports short explanations while keeping model evidence separate from verified linguistic meaning.

6. DISCUSSION

This study treats cross-domain sentiment as a reliability problem with several measurable parts. Expert mixtures bring specialization, but specialization on its own does not guarantee better generalization. A pooled model benefits from a larger effective training volume and may pick up broad sentiment expressions efficiently. The controlled evaluation illustrates that trade-off: the mixture stays competitive on classification and adds diagnostic information, while the pooled model keeps a small lead in aggregate Macro-F1. We think that result is worth reporting, because a reliable system should show its relative performance honestly and not assume that a novel architecture is automatically better.

The difference between calibration and error detection deserves particular attention. ECE summarizes how well confidence matches accuracy within bins, while error AUROC evaluates how well correct and incorrect predictions are ranked. A method can lower ECE without improving AUROC, as happened here, and Brier score and negative log-likelihood can respond differently to temperature scaling as well. In operational sentiment analysis, a model should be judged against its intended use, whether that is probability forecasting, ranking suspicious classifications, selective automation or explaining individual reviews. Picking the metric after seeing the results would undermine the evaluation.

Domain distance brings a benefit beyond changing predicted probabilities. A low-similarity input can be marked as unfamiliar even when the final class probability is high. Expert disagreement can reveal a domain-sensitive word that splits the existing classifiers, and entropy signals uncertainty in the combined output. The three diagnostics are not statistically independent, and they should not be read as identifiable causes of error. Still, a transparent weighted score is easier to audit than a black-box claim of "trust." Future deployments could fit a risk predictor on carefully separated validation domains and then check its error probabilities on new ones.

Explanation fidelity has practical limits too. Token deletion identifies features that affect a model's output, but it cannot show that a word causes customer satisfaction or reflects what the speaker meant. Negation, compositional syntax and sarcasm can depend on interactions between tokens that single-token removal misses. Multi-token perturbations, counterfactual paraphrases and human annotation could extend the evaluation. A practical explanation interface should show local model sensitivity without quietly turning a numerical attribution into an unsupported linguistic claim.

The corpus in this experiment gives explicit control over vocabulary overlap, source–target exclusion and class balance. That control makes the ablations easier to interpret and the work easier to reproduce. Natural review data add further variation, such as annotator disagreement, mixed polarity, rare topics and expressions that shift over time. The numbers reported here should therefore not be read as estimates for Yelp, Amazon or real social-media traffic. The same architecture and metrics could be tested on the UCI Sentiment Labelled Sentences collection [17], on larger review datasets, and under domain-disjoint evaluation.

A transformer version is a separate extension, for example with RoBERTa encoders [16]. Contextual embeddings could replace the sparse TF–IDF features, and the source experts could become lightweight trainable adapters instead of full classifiers. Any such comparison would need to hold the backbone, calibration splits, computational budget and inference policy fixed. Low-rank adaptation reduces training requirements in established transformer work [25], but parameter efficiency does not by itself establish reliable uncertainty. A larger model may also change token explanations and expert disagreement, and those effects have to be measured and not extrapolated from this sparse implementation.

7. CONCLUSION

TrustSenti-MoE brings together domain-general and specialist sentiment predictors with similarity routing, probability calibration, risk estimation and local explanations. A controlled four-domain experiment with twelve held-out-domain evaluations gave 90.50% mean Macro-F1, 0.0414 expected calibration error and an error-detection AUROC of 0.9683. Temperature scaling reduced calibration-bin error relative to the uncalibrated mixture, while the pooled classifier stayed competitive on overall classification and on several probability metrics. The evidence supports a differentiated view of reliability and not a blanket claim of predictive superiority.

The framework lets a reader inspect where a prediction came from, through specialist weights, entropy, disagreement, source distance and token evidence. Separating sentiment classification from decision reliability can guide evaluation on larger, naturally collected corpora and support domain-sensitive selective automation. The main contribution is a testable method for producing and auditing sentiment decisions across unfamiliar linguistic contexts.

REFERENCES

[1] Kim, Y., and Bartolini, I. (2026). Reframing sentiment analysis in the LLM Era: A comprehensive review of methods, evaluation, and sustainability. Artificial Intelligence Review. https://doi.org/10.1007/s10462-026-11684-6

[2] Wang, Q., Wen, Z., Ding, K., Liang, B., and Xu, R. (2025). Cross-Domain Sentiment Analysis via Disentangled Representation and Prototypical Learning. IEEE Transactions on Affective Computing, 16(1), 264–276. https://doi.org/10.1109/TAFFC.2024.3431946

[3] Wang, S., Zhou, J., Chen, Q., Zhang, Q., Gui, T., and Huang, X. (2024). Domain Generalization via Causal Adjustment for Cross-Domain Sentiment Analysis. Proceedings of LREC-COLING 2024, 5286–5298. https://aclanthology.org/2024.lrec-main.470/

[4] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2023). Domain Generalization: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3195549

[5] Koh, P. W., et al. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. Proceedings of the 38th International Conference on Machine Learning, 139, 5637–5664. https://proceedings.mlr.press/v139/koh21a.html

[6] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[7] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems, 32. https://papers.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[8] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. Advances in Neural Information Processing Systems, 30. https://papers.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[9] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. International Conference on Learning Representations. https://research.google/pubs/outrageously-large-neural-networks-the-sparsely-gated-mixture-of-experts-layer/

[10] Geifman, Y., and El-Yaniv, R. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. Proceedings of the 36th International Conference on Machine Learning, 97, 2151–2159. https://proceedings.mlr.press/v97/geifman19a.html

[11] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. (2019). Beyond Temperature Scaling: Obtaining Well-Calibrated Multi-class Probabilities with Dirichlet Calibration. Advances in Neural Information Processing Systems, 32. https://papers.neurips.cc/paper_files/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html

[12] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of ACM SIGKDD, 1135–1144. https://doi.org/10.1145/2939672.2939778

[13] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

[14] He, W., Jiang, Z., Xiao, T., Xu, Z., and Li, Y. (2026). A Survey on Uncertainty Quantification Methods for Deep Learning. ACM Computing Surveys, 58(7), Article 179. https://doi.org/10.1145/3786319

[15] Arafat, M. E., Akmal, M. U., Abosinnee, A. S., and Orosz, T. (2026). SentiMatrix: Parameter-Efficient Fine-Tuning of Encoder-Based Transformers for Multidimensional Sentiment Analysis. Machine Learning, 115, Article 229. https://doi.org/10.1007/s10994-026-07161-4

[16] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[17] Kotzias, D., Denil, M., de Freitas, N., and Smyth, P. (2015). From Group to Individual Labels Using Deep Features. Proceedings of ACM SIGKDD, 597–606. https://doi.org/10.1145/2783258.2783380

[18] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant Risk Minimization. arXiv:1907.02893. https://arxiv.org/abs/1907.02893

[19] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. International Conference on Learning Representations. https://arxiv.org/abs/1911.08731

[20] Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-Based Out-of-Distribution Detection. Advances in Neural Information Processing Systems, 33. https://papers.neurips.cc/paper_files/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html

[21] Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation, 14(1), 21–41. https://doi.org/10.1162/089976602753284446

[22] Hsiao, W.-C., and Wang, H. C. (2024). Cross-Domain Corpus Selection for Cold-Start Context. Journal of Information Science. https://doi.org/10.1177/01655515241263283

[23] Yuan, S., Li, M., Du, Y., and Xie, Y. (2024). Cross-Domain Aspect-Based Sentiment Classification with Hybrid Prompt. Expert Systems with Applications, 255, 124680. https://doi.org/10.1016/j.eswa.2024.124680

[24] Xiao, X., Wang, X., Qi, S., and Li, H. (2026). Uncertainty-Aware Mixture of Experts for Robust Multimodal Sentiment Analysis. Pattern Recognition, 179, 113417. https://doi.org/10.1016/j.patcog.2026.113417

[25] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations. https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/