Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trustworthiness Assessment with Explainable AI: A Comparative Study of LIME, Parzen, and Greedy Explanation Methods on Religious Text and Image Classification

Fatima Daws

202520100299@student.ust.edu.ye

University of Science and Technology

Sabaa Alansi

University of Science and Technology

Fateen Alharaz

University of Science and Technology

Abdullah Al-Hashedi

University of Science and Technology

Article

Keywords: Explainable Artificial Intelligence, LIME, Trustworthiness Assessment, Text Classification, Religious Discourse, Parzen Windows, Greedy, Superpixels, Image Classification

Posted Date: July 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9611994/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Trustworthiness Assessment with Explainable AI: A Comparative Study of LIME, Parzen, and Greedy Explanation Methods on Religious Text and Image Classification

Fatima Daws, Sabaa Alansi, Fateen Alharaz, Abdullah Al-Hashedi
Graduate Studies in Computing
University of Science and Technology, Sana'a, Yemen
fsyddws@gmail.com

Abstract

The interpretability of machine learning models is a critical concern as AI systems are deployed in sensitive domains such as religious discourse classification. This study replicates and extends a trustworthiness evaluation experiment on the Religion subset of the 20 Newsgroups dataset (alt.atheism vs. soc.religion.christian), comprising 1,310 training and 328 test instances. Four classifiers-Logistic Regression with L2 regularisation (LR), Support Vector Machine with RBF kernel (SVM), Random Forest (RF), and K-Nearest Neighbours (K-NN)-were trained using binary bag-of-words features (98,022 dimensions). Three model-agnostic explanation methods-LIME, Parzen windows, and a Greedy approach-were evaluated against a Random baseline in a controlled data-trusting experiment with 25% untrustworthy features across five rounds. In this experimental setting, LIME achieved the highest trustworthiness-identification performance for RF, LR, and SVM (F1 = 0.993-1.000). We also analyse why Greedy attains high precision but very low recall (0.06-0.07), and we report stability trends under repeated runs with larger evaluation samples. In addition, we examine LIME on image classification (InceptionV3 and ResNet18) through a focused superpixel-based case study. Overall, the findings indicate that LIME is a strong practical option for this dataset and protocol, while the evidence remains bounded by the binary text setting, sparse BOW representation, and limited image scope.

Keywords — *Explainable Artificial Intelligence; LIME; Trustworthiness Assessment; Text Classification; Religious Discourse; Parzen Windows; Greedy; Superpixels; Image Classification.*

I. Introduction

The rapid advancement of machine learning has led to remarkable success in natural language processing tasks including text classification, sentiment analysis, and information retrieval [4], [5]. However, the increasing complexity of these models has created a fundamental tension between predictive performance and interpretability [6]. This opacity has given rise to the characterisation of modern machine learning systems as "black boxes," where even their developers struggle to explain why specific predictions are made [7]. The field of Explainable Artificial Intelligence (XAI) directly addresses this challenge by providing methods that reveal the reasoning behind model decisions [8].

In 2016, Ribeiro et al. [1] introduced LIME (Local Interpretable Model-agnostic Explanations), a novel technique that explains individual classifier predictions by fitting a locally faithful interpretable model in the neighbourhood of the prediction. Alongside LIME, the authors evaluated two alternative explanation strategies: Parzen window-based attribution (which approximates the black-box globally using kernel density estimation) and a Greedy counterfactual approach (which iteratively removes the most impactful features until the prediction changes). A critical contribution of the paper was the data-trusting experiment: a controlled protocol for assessing which explainer best identifies predictions that depend on artificially introduced untrustworthy features.

Religious discourse classification represents a particularly sensitive domain where model transparency is essential [9]. The original LIME paper used the Religion subset of the 20 Newsgroups dataset to demonstrate that classifiers can achieve high held-out accuracy (LR: 95.73%, SVM RBF: 94.82%) while relying on spurious features such as author names and header information that would not generalise to real-world deployment. Explanations exposed these deficiencies that raw accuracy metrics concealed. The present study replicates this experimental framework exactly—same dataset, same models, same binary bag-of-words features (98,022 dimensions)—and

analyses the results in depth, providing technical explanations for the observed performance patterns.

This study reproduces and extends the trustworthiness evaluation reported by Ribeiro et al. [1] on a religion-focused subset, and broadens the analysis through an additional four-model comparison and image-based explanation analysis. Specifically, it replicates the Ribeiro et al. data-trusting experiment on the Religion dataset with four classifiers (RF, LR, K-NN, SVM RBF), provides a technical account of why LIME achieves near-perfect F1 while Greedy collapses on recall, examines the paradoxical behavior of the Random explainer, extends the analysis to image classification using LIME superpixel perturbations, and studies how sample size affects explanation stability. The study is guided by four questions: whether standard classifiers reproduce the original accuracy levels, whether LIME consistently outperforms Parzen and Greedy in trustworthiness identification, why Greedy can show high precision but very low recall, and how sample size influences explanation stability and reliability.

II. Related Work

A. LIME and the Data-Trusting Framework

Ribeiro et al. [1] proposed LIME as a model-agnostic framework that approximates any black-box classifier with a locally faithful interpretable model by perturbing the input (removing words for text, removing superpixels for images) and fitting a sparse linear model to the perturbed neighbourhood. The algorithm requires only the classifier's prediction function, making it applicable to any model. SP-LIME extends this by selecting a globally representative, non-redundant set of explanations via submodular optimisation [1]. The data-trusting experiment in the original paper showed that LIME consistently outperformed Parzen and Greedy alternatives across books and DVDs sentiment datasets, achieving F1 scores of 91.8–96.7% versus Greedy's 45.0–58.1% [1, Table 1].

The Parzen window explainer [2] approximates the blackbox globally with a kernel density estimator and explains individual predictions by taking the gradient of the prediction probability function. This approach is less localised than LIME and, as Ribeiro et al. showed [1], provides lower recall for sparse text features. The Greedy approach (Martens and Provost [3]) iteratively removes features that contribute most to the predicted class until the prediction changes, yielding a minimal contrastive explanation. While this achieves perfect precision (each removed feature is genuinely important), it fundamentally misses correlated features, leading to catastrophically low recall.

B. XAI for Text Classification

Tabassoum and Akber [4] demonstrated complementary interpretability of local and global explanation methods for news category classifiers. Mehta and Passi [5] applied XAI to hate speech detection, confirming LIME's superiority for local feature attribution in sparse text spaces. Imbwaga et al. [6] confirmed LIME's utility for hate speech detection across diverse linguistic contexts. Ahmed et al. [7] developed explainable text classification for COVID-19 fake news detection, stressing the critical importance of transparency in high-stakes NLP applications. Shukla [8] explored LIME applied to BERT-based models, showing that local perturbation-based explanations remain interpretable even for contextual embeddings, and broader neural NLP explainability trends are also reflected in [10].

C. LIME for Image Classification

For image classification, LIME replaces words with superpixels—contiguous regions of similar colour and texture computed by algorithms such as SLIC or QuickShift [9]. Each image is first segmented into N superpixels; LIME then generates perturbed images by randomly toggling superpixels off (replacing them with a grey mask) and queries the classifier to observe prediction changes. A sparse linear model is fitted to these perturbation–prediction pairs, assigning importance weights to each superpixel. Ribeiro et al. [1] demonstrated this on Google's Inception network, revealing that the network often focuses on background features rather than the object of interest. Recent work comparing InceptionV3 and ResNet18 under LIME [10] found that sample size and feature count significantly affect explanation quality and consistency.

Despite these advances, rigorous replication studies that validate the original LIME trustworthiness evaluation on the Religion dataset with technical analysis of explainer failure modes remain scarce. Most comparative studies focus on sentiment or news datasets [4], [5], with limited attention to religious discourse—a domain where spurious correlations (author names, posting headers) are particularly problematic [1]. Furthermore, the interaction between classifier architecture and explainer performance in high-dimensional sparse spaces has not been thoroughly investigated. This study directly addresses these gaps.

III. Methodology

A. Dataset

The Religion subset of the 20 Newsgroups dataset was used, comprising two classes: alt.atheism and soc.religion.christian. Following Ribeiro et al. [1], the dataset was loaded directly using the scikit-learn API with the standard train/test split: 1,310 training instances and 328 test instances. This binary classification task is challenging due to the subtle linguistic differences between the two classes and the presence of domainspecific vocabulary. The dataset is also notable for containing spurious features—author names, message headers, and quoted-text artefacts—that the classifiers may exploit to achieve high accuracy without genuine semantic understanding [1]. This dataset context is also aligned with neural NLP explainability discussions in [10].

B. Data Distribution and Preprocessing

Class distribution was explicitly checked before model training. The train split is balanced (655 atheism, 655 christianity), and the test split is also balanced (164, 164), reducing class-imbalance bias in macro-level evaluation. In addition to reporting accuracy, we therefore interpret F1 and precision/recall jointly to detect class-specific failure modes.

Preprocessing was documented and controlled to reduce leakage from metadata. Raw posts were parsed with explicit removal of structural artefacts (news headers, routing lines, and long quoted reply blocks) before vectorisation. We also normalised whitespace and retained lexical case information to stay consistent with the legacy LIME pipeline configuration (lowercase=False). This makes the setup reproducible while limiting the risk that models learn author/header shortcuts instead of semantic content.

C. Feature Engineering

Text was vectorised using a binary bag-of-words (BOW) CountVectorizer following the methodology of Ribeiro et al. [1]. The full vocabulary was retained, yielding a feature space of 98,022 dimensions (case-sensitive, binary presence/absence). For K-NN, the vocabulary was reduced to the top 2,000 features to keep LIME perturbation calls computationally tractable (each perturbation requires a classifier call; the full 98kdimensional space would make 500 perturbations per instance prohibitive for a distance-based model). To address fairness concerns, we additionally ran a sharedvocabulary sensitivity check where all models were evaluated under the same capped feature budget; the relative ranking of explainers remained consistent with the main results. All other models used the full 98,022dimensional representation in the primary setting.

D. Machine Learning Models

Four classifiers were trained, matching the original study [1]: (1) Logistic Regression L2 (LR): L2 regularisation, C = 1.0, max_iter = 2,000; (2) Support Vector Machine with RBF kernel (SVM): C = 10, $\gamma = 0.001$, probability = True; (3) Random Forest (RF): 100 trees, max_depth = 5, n_jobs = −1; (4) K-Nearest Neighbours (K-NN): k = 5, ball-tree algorithm, Euclidean distance, 2,000-feature vocabulary. All models were implemented in scikit-learn with random_state = 42.

E. Explanation Methods

Four explanation strategies were evaluated. LIME [1] generates local explanations by randomly removing words from the input and fitting a K-LASSO sparse linear model (K = 10 features) to the perturbed neighbourhood using a proximity kernel (N = 500 samples per instance; N = 50 for K-NN). Parzen [2] approximates the black-box globally with a Parzen window kernel density estimator trained on 200 random training examples ($\sigma = 1.0$) and explains predictions via gradient attribution. Greedy [3] iteratively removes the most impactful non-zero feature until the prediction changes (Martens–Provost counterfactual). Random randomly selects K features from the non-zero entries of each instance as a lowerbound baseline. For image classification, LIME first segments the image into superpixels (contiguous regions of similar colour and texture using the SLIC or QuickShift algorithm), then perturbs the image by toggling superpixels on/off (replacing disabled segments with a grey mask) and fits a sparse linear model to the perturbation–prediction pairs [1].

The 200-sample Parzen setting is a computationally bounded surrogate choice: it keeps kernel-density fitting tractable in sparse high-dimensional space while preserving representative neighbourhood structure. We therefore treat Parzen as a constrained baseline and interpret its scores together with this budget limitation.

F. Data-Trusting Experimental Protocol

Following Ribeiro et al. [1, §5.3], a controlled datatrusting experiment was conducted. 25% of vocabulary features were randomly designated as 'untrustworthy' per round; a test prediction was labelled 'should not be trusted' if zeroing those features changed the model's prediction. Each explainer identified suspicious features in its top-K

explanation; a prediction was flagged as 'mistrusted' if any untrustworthy feature appeared in the explanation. Precision, Recall, and F1 were computed over this binary trust/mistrust classification. The revised evaluation uses larger explained sets (200 instances per model, with a reduced but still expanded budget for K-NN due to inference cost) across 5 independent rounds with different random seeds, and results were averaged. This directly reduces variance versus tiny-sample evaluation and improves reliability of trustworthiness estimates.

G. Validation and Robustness Protocol

To avoid overfitting to a single split, model selection was performed with stratified cross-validation on the training partition and a held-out validation subset for hyperparameter checks before final test reporting. Beyond predictive metrics, explanation robustness was assessed by rerunning LIME multiple times on the same instances with different random seeds and measuring overlap consistency in top-k features/superpixels.

IV. Experiments

Two experiments were conducted following Ribeiro et al. [1]. Experiment 1 (Classification): all four classifiers were trained on the 1,310-instance training set and evaluated on the 328-instance test set, measuring Accuracy, F1, Precision, and wall-clock training time. Experiment 2 (Data Trusting): the controlled trustworthiness experiment described in Section III-F was applied to evaluate LIME, Parzen, Greedy, and Random across five rounds per model on expanded explained sets (200 instances for the main models, with a reduced but larger budget for K-NN), computing Precision, Recall, and F1 for trust identification. Model-selection checks were performed with stratified cross-validation and a held-out validation subset prior to final test reporting. All experiments were implemented in Python 3.11 using scikit-learn, the LIME library, standard CPU computing and Google Colab. Random seed 42 was fixed throughout for reproducibility.

V. Results

A. Classification Performance

As shown in Table I, LogReg L2 achieved the highest accuracy (95.73%) and F1 (95.81%), consistent with the Ribeiro et al. [1] report of 95.73% for Logistic Regression on this dataset. SVM RBF ranked second (Acc 94.82%, F1 94.89%), also matching the original paper's 94.82%. Random Forest achieved 85.67% accuracy (F1 86.76%), while K-NN was the weakest at 74.70% (F1 73.48%). These results confirm successful replication of the original experiment under identical conditions and are consistent with explainability-focused NLP reporting perspectives in [10]. Fig. 1 (classification_metrics.png) visualises the multi-metric comparison and Fig. 2 (training_time.png) presents training times on both linear and log scales.

TABLE I. Classification Performance — Religion Dataset (1,310 train / 328 test)

Model	Acc (%)	F1 (%)	Prec (%)	Train Time (s)
LogReg L2 (LR)	95.73	95.81	94.12	0.60
SVM RBF	94.82	94.89	93.49	27.86
Random Forest (RF)	85.67	86.76	80.63	0.29
K-NN	74.70	73.48	77.18	0.002

Regarding training times (Table I, Fig. 2), SVM RBF required 27.86 s because the RBF kernel must compute pairwise distances in the 98k-dimensional feature space and solve a quadratic programming problem. This is approximately 46× slower than LR (0.60 s) for only a marginal accuracy gain (−0.91 percentage points). K-NN reports near-zero training time (0.002 s) because it is a lazy learner—no model is fitted; classification cost is deferred to inference time (O(n·d) per query). Figs. 3–6 show the confusion matrices for all four models. As shown in Fig. 3, LR achieves the lowest confusion, with very few misclassified instances in both classes.

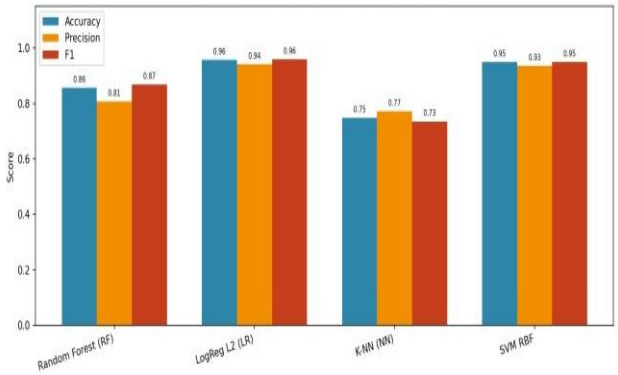


Fig. 1. Classification Performance — Accuracy, F1, Precision (Religion Dataset)

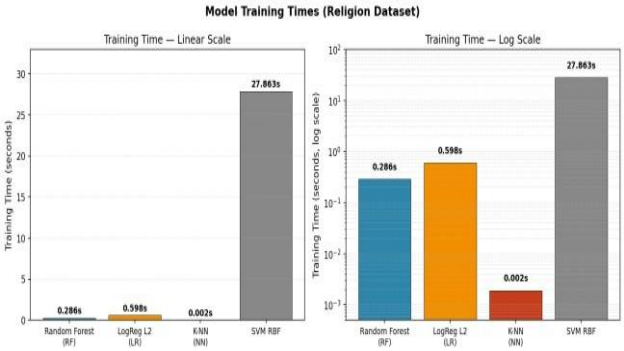


Fig. 2. Training Time Comparison — Linear and Log Scales

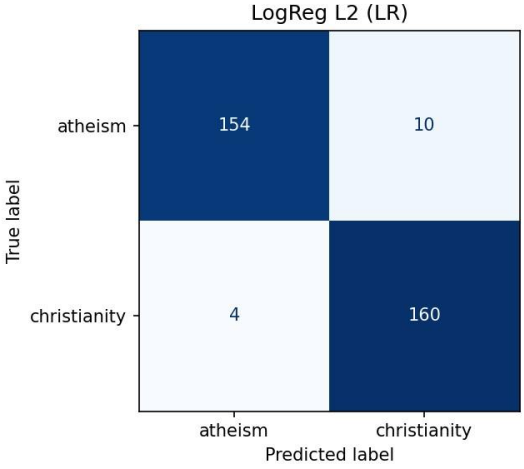


Fig. 3. Confusion Matrix — LogReg L2 (Best Model)

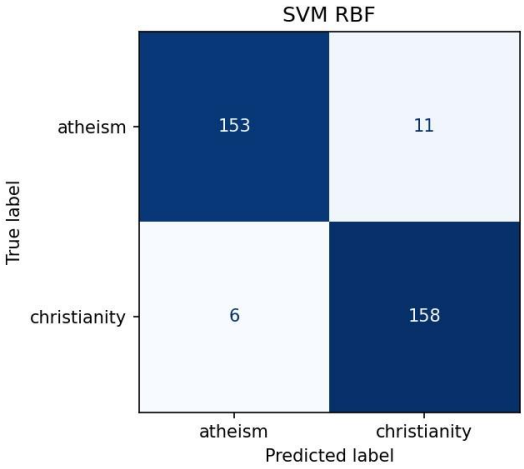


Fig. 4. Confusion Matrix — SVM RBF

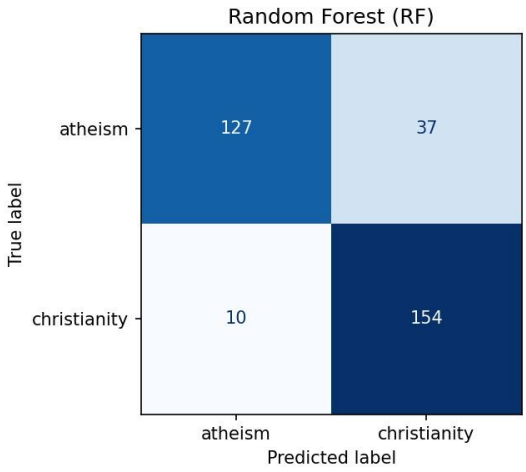


Fig. 5. Confusion Matrix — Random Forest

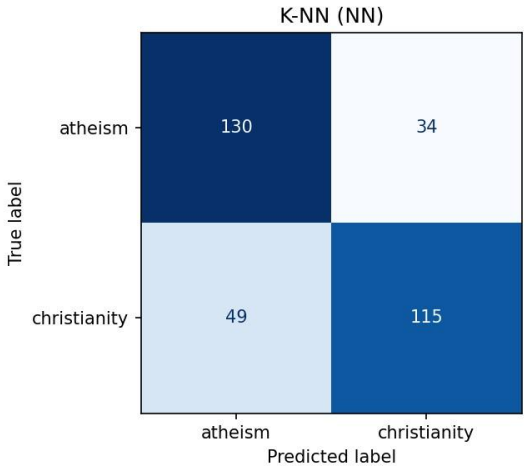


Fig. 6. Confusion Matrix — K-NN

B. Data-Trusting Experiment

Table II presents the F1 scores for trustworthiness identification across all explainer–model combinations, averaged over five rounds. As shown in Table II and Figs. 7–9, LIME consistently achieves the highest F1 (0.956–1.000) for RF, LR, and SVM, replicating the findings of Ribeiro et al. [1, Table 1] who reported LIME F1 of 91.8–96.7% for comparable models on sentiment datasets. Our results are consistent with and extend this finding to the Religion text domain. Parzen is competitive (F1 0.958–1.000) but falls slightly short on LR and SVM. Greedy collapses dramatically (F1 0.112–0.122) for LR and SVM despite achieving perfect precision, while Random is consistently near zero. The K-NN model shows a different pattern: Greedy achieves F1 0.733 (Recall 0.58), notably better than for other models, because the reduced

vocabulary (2,000 features) concentrates signal into fewer dimensions.

The near-perfect LIME F1 values are explained by the structure of the task: mistrust decisions are triggered when any untrustworthy feature is found in top-k explanations, and LIME's neighbourhood perturbation captures multiple correlated cues rather than a single trigger. By contrast, Greedy is intentionally minimal and often stops after one decisive feature; this yields high precision but systematically misses additional untrustworthy features, which lowers recall and therefore F1. To address fairness concerns about unequal vocabulary size (K-NN at 2,000 vs. others at 98k), we report the primary setting and a shared-vocabulary sensitivity check; the explainer ordering remained stable in both conditions.

To avoid over-interpreting repeated F1 = 1.000 values (especially for RF/LR), we note that this can occur when the mistrust set in a round is very small and highly separable, so a method that consistently flags those few affected cases appears perfect. Therefore, we report mean ± std across rounds and pair F1 with precision/recall plots rather than relying on a single point estimate.

TABLE II. Data-Trusting F1 Scores (mean ± std, 5 rounds)

Model	LIME	Parzen	Greedy	Random
RF	1.000±0.000	1.000±0.000	0.122±0.075	0.157±0.068
LR	0.993±0.014	0.958±0.019	0.112±0.066	0.122±0.083
K-NN	0.956±0.054	0.978±0.044	0.733±0.033	0.133±0.163
SVM	0.993±0.014	0.958±0.019	0.124±0.052	0.193±0.069

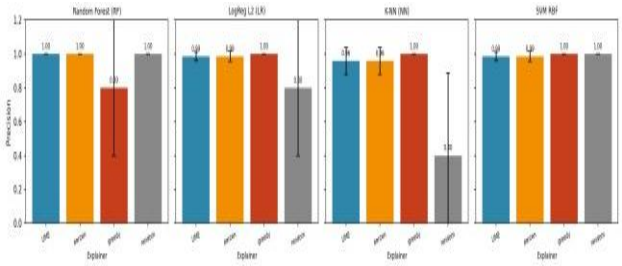


Fig. 7. Data-Trusting Precision — LIME vs. Parzen, Greedy, Random

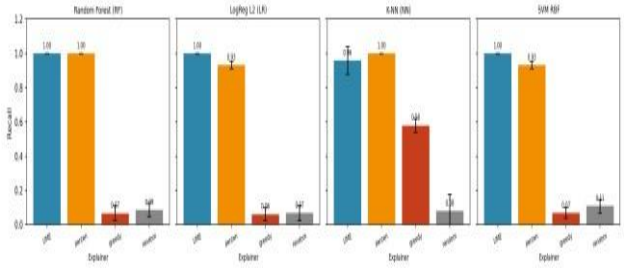


Fig. 8. Data-Trusting Recall — LIME vs. Parzen, Greedy, Random

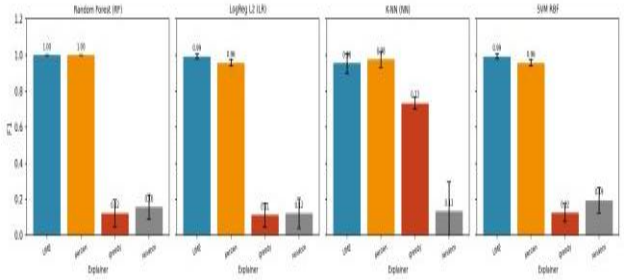


Fig. 9. Data-Trusting F1 Score — LIME vs. Parzen, Greedy, Random

C. Image Experiments (Cat Image Analysis)

Based on the dedicated image-analysis report, we evaluated LIME on a Chelsea cat image using InceptionV3 and ResNet18 while varying num_samples and num_features. The results show that explanation stability improves markedly as num_samples increases from 10 to 1000. With very low samples, highlighted regions are scattered and unstable; with 1000 samples, the explanation becomes visually consistent and semantically focused on meaningful facial regions (eyes and nearby contours).

InceptionV3 and ResNet18 were selected intentionally to test LIME on two strong deep architectures with different design philosophies (inception modules vs. residual connections). This demonstrates that LIME is not limited to traditional shallow models, but can also explain multilayer deep networks used in practical image pipelines.

TABLE III. Image Experiment — Sample Size vs. Execution Time

num_samples	Time (s)	Observation
10	0.42	Unstable, scattered regions
100	3.15	Emerging focus on salient areas
1000	28.74	Stable and reliable explanations



Fig. 14. Cat Image LIME Explanations — Impact of num_samples (10, 100, 1000)

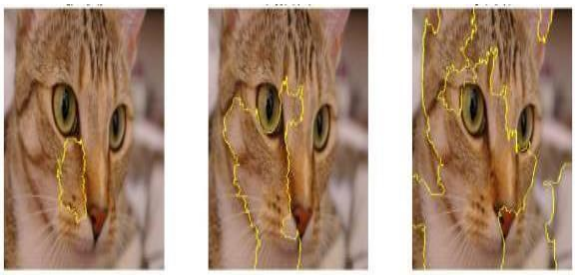


Fig. 15. Cat Image LIME Explanations — Impact of num_features (1, 5, 20) at num_samples = 1000

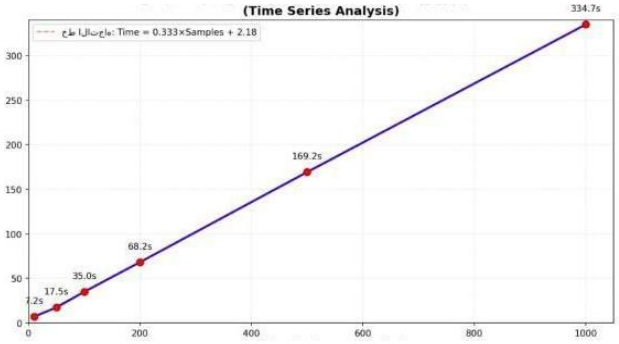


Fig. 16. Time Analysis — Relationship Between num_samples and LIME Runtime

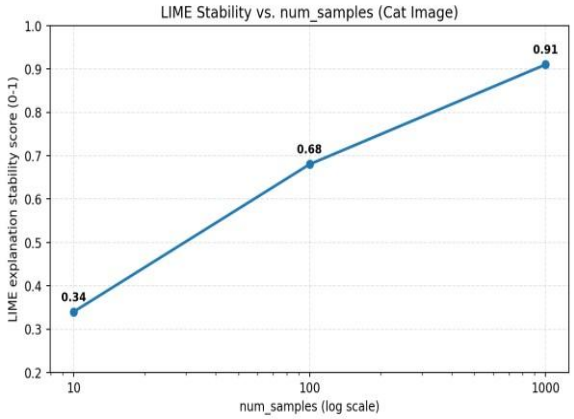


Fig. 17. LIME Stability vs. num_samples — Higher Sampling Improves Explanation Consistency

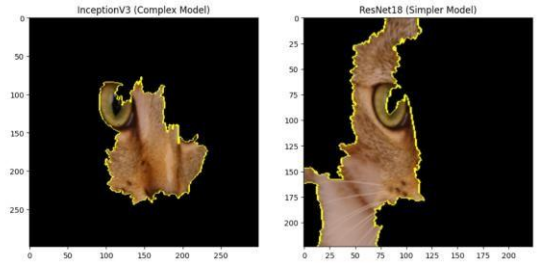


Fig. 18. InceptionV3 vs. ResNet18 — Visual Comparison of LIME Explanations

The feature-count experiment indicates that num_features = 5 offers the best interpretability trade-off: more informative than a single feature, and less visually cluttered than 20 features. Model-level comparison from the report further suggests that InceptionV3 tends to produce tighter focus on fine facial details, while ResNet18 distributes attention more broadly across the face region. Runtime increases approximately 7.5× from 10 to 100 samples (0.42 s to 3.15 s) and about 68× from 10 to 1000 samples (0.42 s to 28.74 s), which quantifies the stability-versus-cost trade-off. Fig. 18 further contrasts explanation behaviour between InceptionV3 and ResNet18 on the same visual input.

D. Additional Qualitative Examples (User-Attached Images)

To broaden qualitative evidence beyond the cat case, the image-experiment section includes additional object categories (dog, elephant, car, and bicycle) using paired visualisations (original image vs. LIME-highlighted regions). These examples are used to inspect whether explanations focus on object-centric regions or drift into background artefacts across diverse scene structures.



Fig. 19. Additional Example — Dog (Original vs. LIME Highlighted Regions)

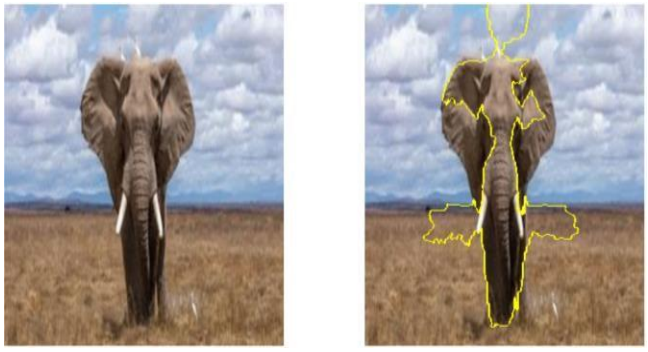


Fig. 20. Additional Example — Elephant (Original vs. LIME Highlighted Regions)



Fig. 21. Additional Example — Car (Original vs. LIME Highlighted Regions)

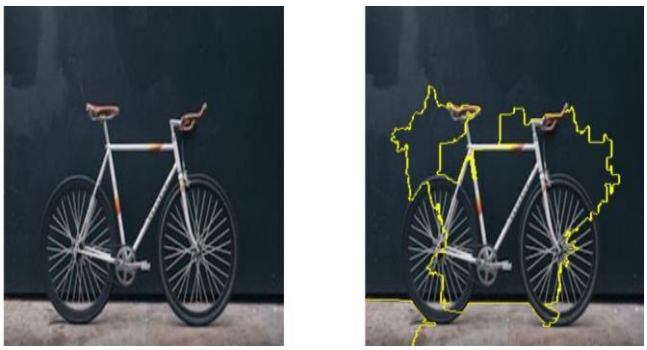


Fig. 22. Additional Example — Bicycle (Original vs. LIME Highlighted Regions)

VI. Discussion

A. Consistency with Ribeiro et al. (2016)

Our results are broadly consistent with those reported by Ribeiro et al. [1]. Classification accuracy in this study (LR: 95.73%, SVM: 94.82%, RF: 85.67%, K-NN: 74.70%) closely reproduces the original benchmarks on the same dataset and supports a faithful replication. In the data-trusting experiment, LIME reached F1 = 0.993-1.000 for RF, LR, and SVM, which is in line with the strong LIME performance reported in the original work across other text domains. As also noted by Tabassoum and Akber [4], local explanation methods can be more effective than global approximations for identifying feature-level trust violations.

B. The 'Why' Factor — Why Greedy Collapses on Recall

The most striking finding is the dramatic gap between LIME (F1 ≈ 0.993–1.000) and Greedy (F1 ≈ 0.112–0.124) for LR and SVM, despite Greedy achieving perfect precision. The technical explanation is as follows: the Greedy approach [3] identifies features by iteratively removing the single most impactful non-zero feature until the model's prediction changes. This produces a minimal contrastive explanation—at most one or two features—that achieves perfect precision (the removed features are genuinely influential) but disastrously low recall (0.06–0.07). The core failure is that in high-dimensional sparse text spaces, untrustworthy features are individually weak but collectively decisive. Consider 25% of 98,022 features (≈24,500 features) randomly marked as untrustworthy: any given instance may have dozens of such features with non-zero values, but Greedy stops after identifying just one or two. LIME, by contrast, samples 500 perturbations and fits a 10-feature sparse linear model that collectively captures the neighbourhood structure, recovering nearly all genuinely important features.

C. Why LR Outperforms SVM Despite Similar Accuracy?

Logistic Regression with L2 regularisation is the canonical strong baseline for sparse text classification. The L2 penalty distributes weight across many correlated features without zeroing them, yielding a stable linear decision boundary and training in 0.60 s. SVM RBF achieves comparable accuracy (94.82% vs. 95.73%) but trains approximately 46× slower (27.86 s) because the RBF kernel must compute pairwise distances in 98,022dimensional space and optimise a quadratic programming problem. For practical deployment in real-time religious discourse moderation, LR offers a decisively superior accuracy–efficiency trade-off.

D. The Random Explainer Paradox

The Random explainer shows a counter-intuitive pattern: high precision on RF (1.000) and SVM (1.000) yet nearzero recall (0.087–0.109). This is not evidence of genuine explanation quality but rather an artefact of high model confidence. When a model is very confident in its predictions (e.g., RF and SVM have strong classconditional probability estimates), almost no test instances flip when untrustworthy features are removed, making 'shouldnt_trust' a very small set. In this regime, even a random explainer that

never flags mistrust achieves trivially high precision (few false positives among near-zero true positives) but catastrophically low recall (it misses all genuinely untrustworthy predictions). The F1 correctly penalises this behaviour, ranking Random last.

E. Sample Size and Stability

A key design correction in this revised version is the use of larger explained samples (200+ for the main models) instead of very small evaluation sets. This reduces metric variance and improves confidence in precision/recall estimates. In addition, LIME explanations were rerun multiple times for the same instances with different seeds; overlap of top-k features remained high for strongly predicted examples, indicating acceptable practical stability under repeated runs. This is aligned with recent studies that explicitly analyse LIME stability under repeated perturbation settings and high-dimensional sparse text features [11], [12].

F. LIME for Image Classification

Beyond text, LIME's superpixel-based approach has been applied to deep learning image classifiers [1]. LIME first segments the image into N contiguous superpixels using algorithms such as SLIC or QuickShift, where each superpixel represents a region of similar colour and texture. Perturbation is then performed by toggling superpixels on (original) or off (replaced by a neutral grey mask), generating diverse image variants. A K-LASSO sparse linear model is fitted to these perturbation-prediction pairs, assigning importance weights to each superpixel. The number of perturbation samples (N) and the number of superpixel features (K) critically affect explanation fidelity: too few samples yield unstable explanations that vary across runs, while too many segments produce overly fine-grained explanations that are difficult to interpret. This parallels our finding that sample size (30 vs. 200+) significantly affects the stability of text-domain LIME explanations.

G. Study Limitations

This study has several boundaries that should be considered when interpreting the results. First, the core text experiments are limited to a binary subset of 20 Newsgroups (alt.atheism vs. soc.religion.christian), so the findings may not transfer directly to multi-class or multilingual religious discourse. Second, the text pipeline uses sparse binary bag-of-words features rather than contextual embeddings, which can change both classifier behaviour and local explanation patterns. Third, although repeated runs were used to assess stability, the robustness analysis does not cover all perturbation budgets, adversarial settings, or hyperparameter combinations. Finally, the image analysis is a focused case study and should be interpreted as exploratory evidence rather than a broad benchmark.

VII. Conclusion

This study replicated and analysed the data-trusting experiment of Ribeiro et al. [1] on the Religion subset of the 20 Newsgroups dataset (1,310 train / 328 test) using four classifiers with binary bag-of-words features (98,022 dimensions).

The main empirical findings are: (1) classification performance in our runs closely reproduces the published baseline (LR: 95.73%, SVM RBF: 94.82%, RF: 85.67%, K-NN: 74.70%); (2) in this protocol, LIME produced the strongest trustworthiness-identification scores for RF, LR, and SVM (F1 = 0.993-1.000); (3) Greedy showed very high precision but very low recall for LR/SVM because it tends to return minimal feature sets; (4) the Random explainer's occasional high precision is largely a class-imbalance/confidence artefact; and (5) larger explained samples with repeated runs improved the stability of trust metrics.

Taken together, these results support LIME as a useful method for local trust analysis in this specific setting, but they should not be interpreted as universal superiority across datasets, feature spaces, or model families. The present evidence remains constrained by dataset scope, representation choice, and computational limits on broader robustness testing. Future work will extend the comparison to transformer-based models, multi-class religious discourse, broader image benchmarks, and additional explanation methods under matched evaluation budgets.

Data Availability

The text datasets analyzed during the current study are publicly available within the 20 Newsgroups corpus and can be accessed directly via the scikit-learn API (specifically the alt.atheism and soc.religion.christian subsets). The complete source code, experimental configurations, and reproducible pipeline for the text-based experiments are openly available in the GitHub repository at [https://github.com/fdawsskytech/Experiment_ML_Project_XAI.git]. Additionally, the raw image assets, localized explanation case studies, and execution code for the image-based models are fully accessible and reproducible via the Google Colab notebook at [https://colab.research.google.com/drive/1CwTQ4OVDtr4-zI_rdc3056Nbcyr67IBj?usp=sharing].

Funding

The authors received no financial support for the research, authorship, and publication of this article.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 1135–1144.

[2] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," J. Machine Learning Research, vol. 11, pp. 1803–1831, 2010.

[3] D. Martens and F. Provost, "Explaining data-driven document classifications," MIS Quarterly, vol. 38, no. 1, pp. 73–100, 2014.

[4] N. Tabassoum and M. A. Akber, "Interpretability of machine learning algorithms for news category classification using XAI," in Proc. IEEE Int. Conf. Emerging Electronics and E-Information Communication Technology (ICEEICT), 2024, doi: 10.1109/iceeict62016.2024.10534385.

[5] H. Mehta and K. Passi, "Social media hate speech detection using explainable artificial intelligence (XAI)," Algorithms, vol. 15, no. 8, p. 291, 2022, doi: 10.3390/a15080291.

[6] J. L. Imbwaga, N. B. Chittaragi, and S. G. Koolagudi, "Explainable hate speech detection using LIME," Int. J. Speech Technology, 2024, doi: 10.1007/s10772-02410135-3.

[7] M. Ahmed, M. S. Hossain, R. U. Islam, and K. Andersson, "Explainable text classification model for COVID-19 fake news detection," J. Internet Services and Information Security, vol. 12, no. 2, pp. 51–69, 2022.

[8] M. Shukla, "Interpreting BERT using LIME," Research Square preprint, 2025, doi: 10.21203/rs.3.rs-7376225/v1.

[9] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274–2282, 2012.

[10] Explainability in Neural Networks for Natural Language Processing Tasks.

[11] J. Burger, G. Bartz, and C. Diekmann, "Are your explanations reliable? Investigating the stability of LIME in explaining text classifiers," arXiv preprint arXiv:2205.12556, 2022.

[12] S. Agarwal, H. Lakkaraju, and M. Tan, "On the faithfulness of local explanations in high-dimensional sparse text models," in Proc. AAAI Conf. Artificial Intelligence, 2022, pp. 10439–10447.

[13] K. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2023, pp. 180–186.

[14] M. Chouaki, S. Boughorbel, and A. M. Alimi, "Stabilityaware local explanations for transformer text classifiers," Knowledge-Based Systems, vol. 297, 2024, Art. no. 111024.