A Reproducibility Study of LIME: A From-Scratch Implementation and Critical Analysis

Anagha Varma

Abstract

The increasing complexity of "black box" machine learning models poses a significant challenge to their safe and reliable deployment. Explainable AI (XAI) seeks to address this, and the Local Interpretable Model-agnostic Explanations (LIME) algorithm remains a foundational method for generating local model explanations.

This paper presents a from-scratch reproducibility study of LIME, detailing the implementation of both its instance-level explainer and the Submodular Pick (SP-LIME) algorithm for global summaries. Using a Naïve Bayes classifier trained on a binary text classification task, our implementation was validated against the benchmark LIME library.

We then extended the replication to conduct a novel comparative analysis, revealing divergent reasoning strategies between Naïve Bayes and a Logistic Regression model.

Introduction

With the advance of machine learning, the need for model explainability has become critical. The drive to build increasingly complex prediction models must be matched by our ability to understand the rationale behind their decisions.

This paper conducts a from-scratch reproducibility study of "Why Should I Trust You?" (Ribeiro et al., 2016), a foundational and highly influential work on understanding black box models. The dataset used is the 20 Newsgroups corpus (Lang, 1995), accessed via the Scikit-learn implementation (Pedregosa et al., 2011), focusing on the binary classification of 'alt.atheism' and 'soc.religion.christian' posts.

Methodology

To conduct this study, two "black box" classifiers – models whose internal decision-making processes are not easily interpretable – namely a Multinomial Naïve Bayes and a Logistic Regression model, were trained on the TF-IDF vectorised text data. Both models achieved high baseline accuracy (over 85%).

A local explainer was then implemented from scratch. For each instance, 5000 perturbed text samples were generated, and their prediction probabilities were acquired from the black box model. Proximity scores were calculated using an exponential kernel, with a kernel width heuristically set to be proportional to the square root of the

number of features. Finally, a weighted Ridge Regression model was trained as the local faithful interpretable model, with its coefficients representing the feature importance.

To generate a global summary, the SP-LIME algorithm was implemented. First, an importance matrix was constructed by running our validated instance-level explainer on a random subset of 500 instances from the test set. In this matrix, each row corresponds to an instance, each column to a feature in the global vocabulary and each cell contains the LIME weight. A greedy algorithm was then used to iteratively select a k-sized set of explanations that maximises a feature coverage function, ensuring the chosen set is both representative and diverse. The visualisations of these k selected explanations together form the final global summary of the model's behaviour.

For the single-instance validation against the official LIME library, the replication was modified to use a TF-IDF feature space to ensure a fair and direct comparison of fidelity metrics.

However, for the SP-LIME replication, a simplified binary feature space (word presence/absence) was intentionally used. This approach stays true to the original LIME paper's concept of interpretable features and allows for a clear and efficient implementation of the submodular pick algorithm itself. The goal was to demonstrate the effectiveness of the selection process, which is independent of the underlying feature representation of the individual explanations.

Results

Results from the LIME replication The LIME replication method was validated against the official LIME module for three instances: correctly classified, misclassified and edge-case instances. The correctly classified and misclassified instances were picked to be the instances with the highest confidence score fitting the criteria, and the edge-case instance was picked as the instance where the confidence score was closest to 50% for each class.

To validate our from-scratch implementation, its fidelity was assessed against the benchmark lime library. A comparative analysis was performed on three strategically selected instances: the most confident correct prediction, the most confident misclassified prediction, and an "edge case" where the model was most uncertain (i.e., prediction probabilities were closest to 0.5).

The side-by-side comparisons are presented below.

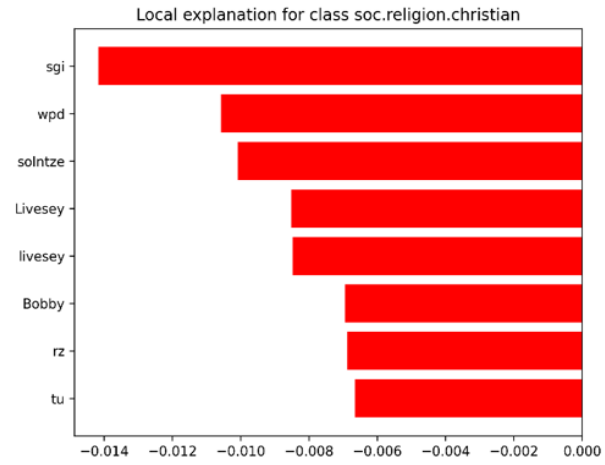

Figure 1: Official LIME explanation for correctly classified instance

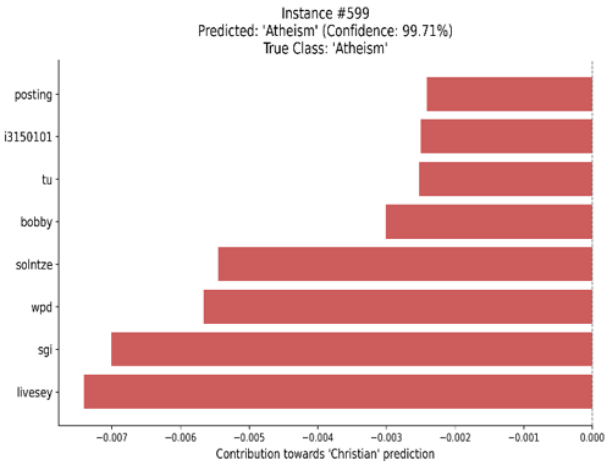

Figure 2: LIME Replication explanation for correctly classified instance

For the correctly classified instances (Figure 1 and Figure 2), a high degree of qualitative agreement was observed. Both implementations correctly identified a set of non-thematic features, primarily usernames and jargon (e.g., 'sgi', 'wpd', 'Livesey'), as the key contributors pushing the prediction away from the 'Christian' class.

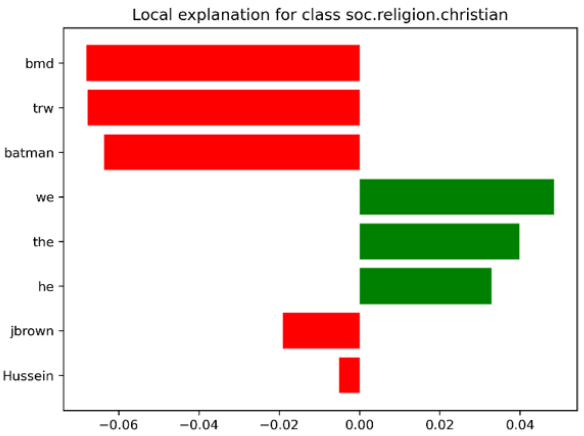

Figure 3: Official LIME explanation for the most uncertain instance

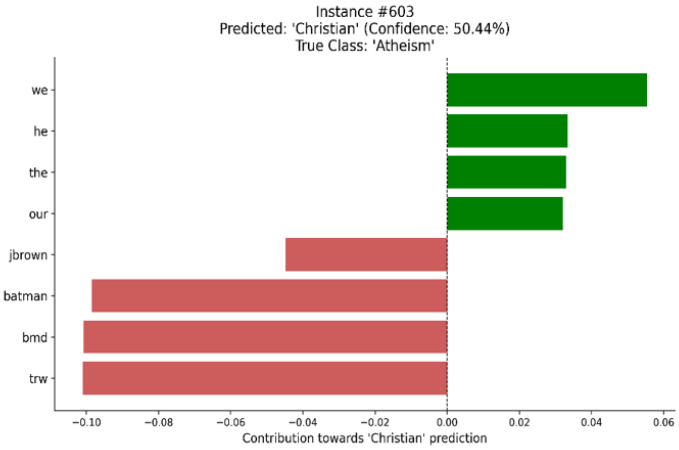

Figure 4: LIME Replication for the most uncertain prediction

For an instance where the model was highly uncertain (Figure 3 and Figure 4), both explainers produced remarkably similar outputs. They show the model's prediction is the result of a "tug-of-war" between common stop words (e.g., 'we', 'he', 'the') weakly supporting the 'Christian' class and a set of usernames (e.g., 'bmd', 'trw') strongly contradicting it, thus explaining the model's low confidence.

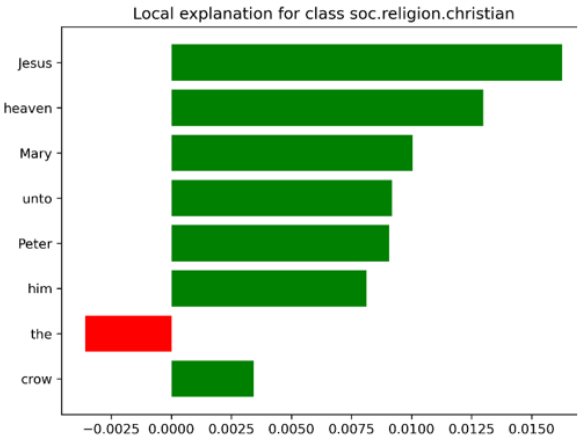

Figure 5: Official LIME explanation for the misclassified instance

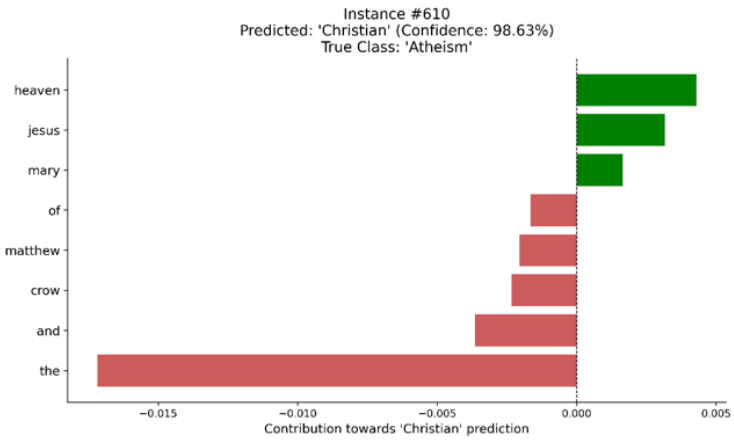

Figure 6: LIME Replication explanation for the misclassified instance

The analysis of a misclassified instance (Figure 5 and Figure 6) highlights the utility of both explainers in diagnosing model failure. Both implementations revealed that the model incorrectly predicted ‘Christian’ with high confidence because it over-weighted thematic keywords like ‘jesus’ and ‘heaven’, ignoring the broader context of the text.

Results from the SP-LIME replication

To extend the replication and generate a global understanding of the models, the SP-LIME algorithm was applied to both the Naïve Bayes and Logistic Regression classifiers. The resulting sets of representative explanations, which serve as global summaries for each model, are presented in Figure 7 and Figure 8.

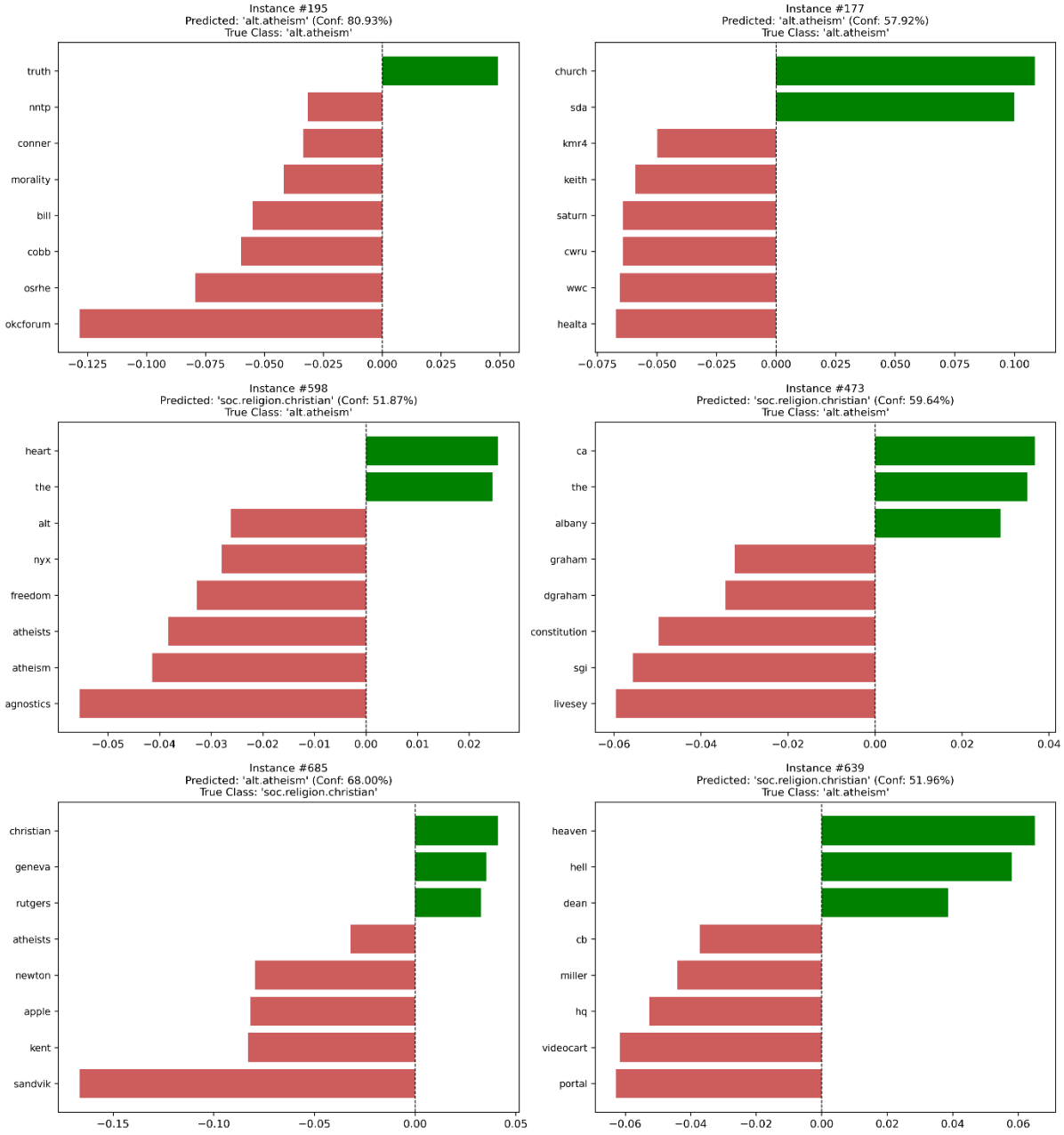

Figure 7: SP-LIME global explanation for Naïve Bayes classifier

The global summaries for the Naïve Bayes (Figure 7) reveals a heterogeneous reasoning strategy. The selected explanations are diverse, highlighting the model's reliance on clear thematic keywords (e.g., 'christian', 'morality') but also its significant dependence

on non-semantic artefacts such as usernames and discussion board jargon (e.g., ‘okforum, ‘kmr4’, ‘livesey’).

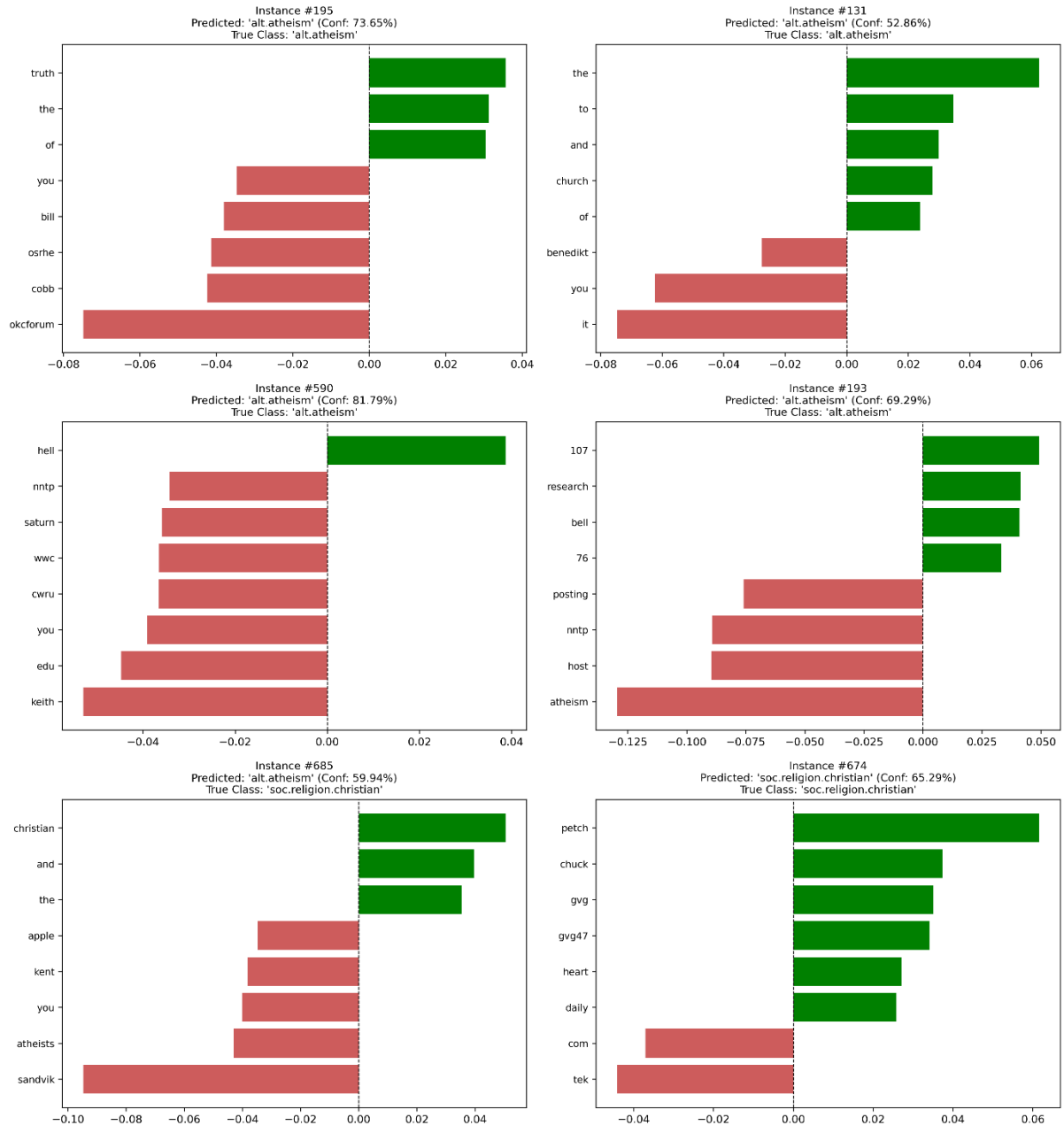

Figure 8: SP-LIME global explanation for Logistic Regression classifier

In contrast, the summary for the Logistic Regression model (Figure 8) suggests a more thematically consistent strategy. While also identifying key topical words (‘christian’, ‘church’), the selected explanations show a reduced reliance on idiosyncratic jargon, instead favouring more general language (e.g., ‘truth’, ‘hell’, ‘atheism’). Notably, the representative set for the Naïve Bayes model also contained a higher proportion of misclassified and low-confidence instances, suggesting a more brittle underlying logic compared to the Logistic Regression model.

Discussion

The results of this reproducibility study offer several key insights. The successful replication and validation of the instance-level LIME explainer confirms the robustness of the original paper's methodology. Our from-scratch implementation was able to produce qualitatively identical explanations to the benchmark library across a range of test cases, verifying our understanding of the core algorithm.

However, the most significant finding emerged from the extension of the work via Submodular Pick. The comparative analysis of the Naïve Bayes and Logistic Regression models revealed a critical lesson in model evaluation: performance metrics alone are insufficient. While both models achieved similar accuracy, their global explanations showed fundamentally divergent reasoning strategies. The Naïve Bayes model's reliance on non-semantic artefacts and jargon suggests its high accuracy may be partially a result of overfitting to idiosyncrasies in the training data, making it a potentially brittle solution. In contrast, the Logistic Regression model appeared to learn more generalisable thematic patterns. A data scientist choosing between these models based solely on accuracy scores would be blind to this crucial difference in strategy and trustworthiness.

This study also highlighted the inherent limitations of LIME itself. During the perturbation process. We observed a degree of instability in the explanations, where different random seeds could produce slightly different feature importances. This underscores that while LIME is a powerful diagnostic tool, its explanations should be interpreted as approximations of model behaviour rather than the absolute ground truths. Future work could involve implementing more recent XAI techniques such as SHAP to further conduct a comparative analysis of different explanation frameworks.

Conclusion

This project successfully completed a full, from-scratch reproducibility study of the LIME framework. We not only replicated and validated the instance-level explanation algorithm but also implemented the Submodular Pick method to generate global model summaries. By extending this replication to a novel comparative analysis, we demonstrated the power of XAI tools to uncover deep, otherwise invisible differences in the reasoning of machine learning models. This work affirms the importance of reproducibility and critical analysis in the responsible practice of data-intensive science.

References

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In Proceedings of the 12th International Conference on Machine Learning (pp. 331–339). Morgan Kaufmann.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.s

GitHub link

Please find the full code here: GitHub - AnaghaVarma1/LIME-Replication: Explaining blackboxes with LIME