Enhancing Model Interpretability with Local Interpretable Model-Agnostic Explanations (LIME) for Religion Dataset Analysis

Szeyi Chan
Northeastern University
chan.szey@northeastern.edu

***Abstract*—Machine learning models are increasingly used in critical domains such as medical diagnostics, where accurate predictions are essential. However, these models often function as "black-box" algorithms, lacking transparency in their decision-making processes, which raises concerns about their reliability and trustworthiness among users. This paper addresses the need for explainable AI (XAI) techniques to enhance the interpretability of such models. Specifically, we reimplement the Local Interpretable Model-agnostic Explanations (LIME) method on a religion dataset to investigate its effectiveness compared to a baseline random K-features explanation similar to the original study. Our findings align with original study, showing that while LIME offers greater interpretability, it also highlights certain limitations when applied to complex datasets.**

I. Introduction

Machine learning is widely used today in various fields for prediction, such as in the medical domain. However, due to the lack of transparency in the predictions or recommendations made by these algorithms, these algorithms are also often called "black-box" algorithms. It affects the reliability and trustworthiness of these predictions for users. In areas like image classification for medical diagnostics, the lack of transparency in the predictions concerns medical practitioners who are trying to understand the reasoning behind a prediction (1). Therefore, improving the explainability of "black-box" models is essential. For instance, to assist domain experts in diagnosing system errors and understanding potential biases in these models (2).

There are various explainable AI (XAI) approaches, such as LIME (Local Interpretable Model-agnostic Explanations) (3) and SHAP (SHapley Additive exPlanations) (4), both of which are model-agnostic methods. These approaches can be applied to any machine learning model and provide insights into how predictions are made. They aim to increase the transparency and trustworthiness of AI systems, ultimately enhancing their usability and acceptance in critical applications like medical diagnostics.

In this project, we aims to reimplement LIME and the baseline(random K-features explanation) on a religion dataset[1] provided by the original paper by Ribeiro et al. (3) to explore these performance issues. By comparing the performance of the two explainers, we can analyze and explore how LIME enhances human understanding of the reasoning behind predictions, thereby providing transparency into the model's decision-making process.

The results of the experiment indicate that while LIME can provide explanations for the predictions, some features identified using our data are not strongly related to the predicted outcome. However, compared to the Random K-Features explanation, LIME shows a higher degree of interpretability. These findings suggest that while LIME has its limitations, it remains a more reliable tool for understanding model predictions in this context.

The structure of this paper is presented as follows: First, background information and an overview of LIME will be provided in Section II. Then, the methodology used in this study, including data processing and implementation details, will be presented in Section III. Following this, the results will be discussed in Section IV. Finally, the paper concludes with a summary of findings in Section V.

II. LIME Explainer

LIME uses a surrogate model as its explanation methodology, where the original model is first trained and used to make predictions. LIME use the local data to train the surrogate model using these predictions instead of the target values. It employs an interpretable feature space and an interpretable-by-design surrogate model–K-LASSO (K-nearest neighbor Least Absolute Shrinkage and Selection Operator), which incorporates sparsity to select only the k-most significant features and focuses on the local neighborhood. This approach provides a interpretable and concise explanation by highlighting key features from potentially thousands. Additionally, LIME allows training one or more local neighborhood models to approximate the global model's behavior around the instance being explained. The target audience for LIME is domain experts, such as medical practitioners, who need to understand the underlying reasons behind model predictions. These experts require transparency to trust the predictions made by complex machine learning models.

In summary, LIME as an explainer has the following characteristics:

1) **Interpretability of Data:** LIME highlights the features that contributed to the predictions, making it easier for users to understand the influence of each feature.

[1] http://bit.ly/4cIGwbX

2) **Local Perspective:** LIME focuses on understanding the predictions from a local point of view rather than a global perspective. This local approximation helps to discover the specific reasons behind individual (local) predictions.
3) **Model-Agnostic Method:** LIME can be applied to any machine learning model, regardless of its complexity, due to its model-agnostic nature.

III. Methodology

This section details the method used to reimplement exapliner, specifically LIME and Random K-Features explanation, for this study. In the implementation, the feature generation steps should enable the classifier to categorize each text as either atheism or Christianity. Each explainer should be able to clarify which features led to the prediction.

A. Data Preprocessing

The dataset utilized in this study comprises 819 texts related to atheism and 1,000 texts related to Christianity, totaling 1,819 instances. Prior to model training, the data was cleaned and pre-processed. This process included encoding categorical variables and ensuring proper data importation. We then split the data set into a training and a test set using an 80/20 ratio, resulting in 1,455 instances for training and 364 instances for testing. These quantities were verified through assertions in the code to confirm that there is no absence of the data.

B. Model Implementation and Explainer Integration

Following the preprocessing and representation of the text data, we implemented a decision tree classifier, denoted as f. This method was previously applied in the original study, to perform the classification task using features generated through a bag-of-words approach (5; 6). Decision trees are a tree-like graph structure commonly used for classification due to their easily interpretable predictions. For instance, Azar and El-Metwally (7) used decision trees to classify a breast cancer dataset, while Yoo et al. (8) applied decision trees for COVID-19 diagnosis based on chest X-ray imaging. This step produced the predictions and feature lists required for the later steps. Given that our dataset consists of news articles with a large volume of text, the bag-of-words method was employed to extract all words from each training file, transforming unstructured text data into tokens suitable for model training.

To gain insights into the model's predictions, we integrated two different explanation methods: the LIME algorithm and the Random K-Features explanation, providing to us the interpretable insights into the decision tree's predictions. These explanations allow us to understanding the model's decision-making process from a human perspective and for validating its predictions.

We implemented the LIME method using the algorithm provided by the original study, showing in Algorithm 1. We first generate local surrogate models and interpret the subset of features from the training step that significantly influence predictions within the local context of specific data points. LIME approximates the black-box model with a surrogate linear model around the prediction of interest, making it easier to understand the influence of individual features. This method has been widely used in various domains, including medical scenarios (9; 10; 11). Additionally, we applied the Random K-Features explanation with K set to ten features in our study, allowing us to compare its performance with LIME. This approach helped us understand the impact of different subsets of features on the model's predictions.

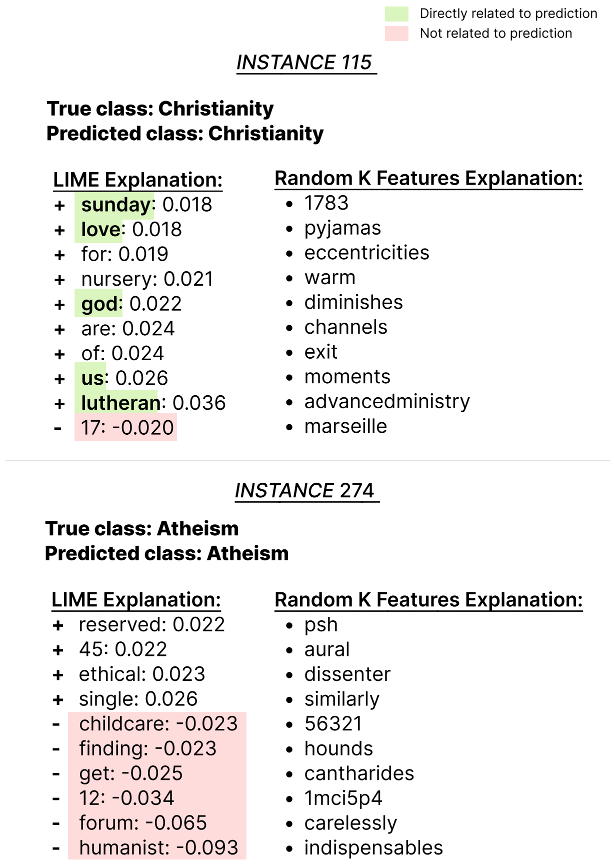
Algorithm 1 Sparse Linear Explanations using LIME

```
**Require:** Classifier $f$, Number of samples $N$
**Require:** Instance $x$, and its interpretable version $x'$
**Require:** Similarity kernel $\pi_x$, Length of explanation $K$
$\mathcal{Z} \leftarrow \{\}$
**for** $i \in \{1, 2, 3, \dots, N\}$ **do**
    $z_i \leftarrow$ sample_around$(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{(z_i', f(z_i), \pi_x(z_i))\}$
**end for**
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ //with $z_i'$ as features, $f(z)$ as target
**return** $w$
```

Before applying LIME and the Random K-Features explanation methods, we first performed perturbation on the selected instances. The perturbed instance denoted as(z_i). This step involved modifying the text by randomly removing words to create new variations of the original text. These perturbations slightly altered the feature values, allowing the exploration of the local decision boundary around each instance.

For the LIME implementation, once the perturbed samples were generated, we calculated the weights using the cosine distance between the original and perturbed texts. The prediction for that perturbed sample denoted as $\pi_x(z_i)$. This weighting method follows approaches used in the original study to emphasize perturbations closer to the original text. The next step involved calculating the probabilities of these perturbations using the trained classifier $(f(z_i))$. To do this, the perturbed texts were converted into numerical vectors using the bag-of-words method. LIME then used these vectors and their corresponding probabilities to fit a linear model that approximated the behavior of the original classifier around the instance of interest. Finally, LIME explained by highlighting the top ten most influential features by the coefficient that contributed to the classifier's prediction for each instance, denoted as w.

In contrast, the Random K-Features explanation method randomly selected ten features after transforming the perturbations into vectors and generated the explanation based on these selected features. Both methods aimed to help humans understand the model's decision-making process by revealing which features the model considered were most impactful in the predictions.



🟩 Directly related to prediction
🟥 Not related to prediction

INSTANCE 115

True class: Christianity
Predicted class: Christianity

LIME Explanation:	Random K Features Explanation:
+ **sunday**: 0.018	• 1783
+ **love**: 0.018	• pyjamas
+ for: 0.019	• eccentricities
+ nursery: 0.021	• warm
+ **god**: 0.022	• diminishes
+ are: 0.024	• channels
+ of: 0.024	• exit
+ **us**: 0.026	• moments
+ **lutheran**: 0.036	• advancedministry
- 17: -0.020	• marseille

INSTANCE 274

True class: Atheism
Predicted class: Atheism

LIME Explanation:	Random K Features Explanation:
+ reserved: 0.022	• psh
+ 45: 0.022	• aural
+ ethical: 0.023	• dissenter
+ single: 0.026	• similarly
- childcare: -0.023	• 56321
- finding: -0.023	• hounds
- get: -0.025	• cantharides
- 12: -0.034	• 1mci5p4
- forum: -0.065	• carelessly
- humanist: -0.093	• indispensables

Fig. 1. Example of comparison result of LIME and Random K Features explanations with ten features. The top explanation corresponds to the "Christianity" class, and the bottom to the "atheism" class. Features highlighted in red are unrelated to the prediction, while those in green are directly related. The numbers in the LIME explanation represent the feature coefficients.

IV. Result

A. Model Performance

The decision tree classifier was trained on the training set, with its performance evaluated on the test set. The model achieved an accuracy of 87.36% and a recall value of 87.46%, presenting good performance in identifying the target classes within the dataset.

B. Explaner Performance

Overall, LIME achieved a recall value of 100%, closely matching the recall observed in the results of the original paper. In contrast, the Random K-Features method yielded a significantly lower recall value of 10%, which is below the recall reported in the original result. Additionally, we calculated the overlap between the results of the two explainers, which revealed only a 0.03% overlap, showing the minimal agreement between the features identified by LIME and those selected by Random K-Features.

1) LIME: The top ten features influencing the classifier's predictions were identified for each instance using LIME. Figure 1 (left) illustrates examples of LIME explanations for correctly classified instances in both classes, highlighting features with positive (annotated as "+") and negative (annotated as "-") influences on the prediction. A positive influence indicates that certain words contribute to the predicted class, while a negative influence suggests that these features pull the prediction away from the correct classification.

From these examples, we observe that LIME explanations can help humans understand which words influenced the predicted class. For instance, in Instance 115, which was classified as "Christianity," the explanation includes words like "god," "sunday," and "lutheran" (highlighted in green in Figure 1), which are directly related to the theme of "Christianity," making the prediction intuitive and understandable to a human reader. However, not all words identified as contributing to the prediction are relevant to the predicted class. For example, words like "for," "are," and "for" in Instance 115 do not clearly relate to the class they are associated with. Additionally, the words with negative influence (highlighted in red in Figure 1) do not necessarily contradict the predicted class, showing potential misalignment from human understanding and machine interpretation. These misalignments could caused by different factors, such as the limitations of the bag-of-words representation, which may not fully capture the context or meaning of the text. In the result on Instance 274 illustrated in Figure 1, some explanations for both positive and negative influences appear irrelevant to the prediction.

this result is align with the results from the original study that the authors find out indicating that the dataset itself posssibly have issues that can not be tell by if its from the dataset or the classigication, which could explain these inconsistencies in the explanations. Overall, we still think that LIME reached our goal of having explanation can human to learn the reasoning behind the prediction, and break the "blackbox."

Our finding aligns with the results from the original study, where the authors noted that the dataset itself might have issues. These issues could caused by the data or the classification process, which may contribute to the inconsistencies observed in the explanations. Despite these challenges, we believe that LIME successfully achieved our goal of studying explanations from LIME can help humans understand the reasoning behind the predictions, providing transparency into the model's decision-making process.

2) Random K-Features Explanations: Figure1 also shows the result of the explanation from the Random K-Features Explaner. Since this model is using the strategy of random selecting, the explanation does not show any coeffient numbers of the numbers. from the human understanding perspective, since the explanation does not nesscary related to the predicted result, the human cannot understand the resoing behind this explanation, leaving human staying the in "balck-box" even though there are explanation provided.

Figure 1 (right) also presents the results from the Random K-Features explainer. This method, which selects features randomly, provides explanations without associated coefficient values. Due to its random nature, the features identified by this explainer may not have any meaningful connection to the predicted result. This approach presents challenges to the human interpretability of the prediction. The lack of a clear relationship between the selected features and the prediction makes it difficult for a human to understand the reasoning behind the model's decision. As a result, despite the presence of an "explanation" the user remains in a "black box," unable to gain clear insight into how the model arrived at its prediction, as well as requiring additional time and effort to understand the reasoning.

In conclusion, this method defeats the purpose of providing explanations. Unlike LIME, which aims to bridge the gap between model predictions and human understanding, the Random K-Features explainer does not facilitate transparency, leaving the decision-making process unclear to human.

C. Challenges with the Greedy Algorithm

Although we successfully implemented the LIME and Random K-Features methods, we were also interested in exploring another explainer used in the original study, which provided coefficient values for the top ten features. We further implemented the Greedy algorithm to test with our test set. However, we faced challenges with the computational complexity of the Greedy Algorithm. Given the size of our dataset, our computational resources were insufficient to fully execute the Greedy Algorithm on the test set. Given the size of our dataset, our computational resources were insufficient to fully execute the Greedy Algorithm on the test set. The computational complexity of the Greedy Algorithm arises because, in each iteration, the algorithm evaluates each remaining feature to determine which one best explains the residuals of the current model. The feature with the highest score is selected, added to the model, and then the residuals are updated by subtracting the contribution of the selected feature. This process is repeated for the specified number of features, leading to a high computational due to multiple iterations of model fitting and score evaluation. Each iteration requires fitting a linear regression model, which, while generally efficient, becomes resource-intensive when done repeatedly, especially in high-dimensional spaces with many features to evaluate. The algorithm continues to select and add features until the specified number of features is reached, potentially leading to significant computational demands if many features are needed to maintain the prediction accuracy. Consequently, we were unable to obtain results using this method within our current experimental setup. re unable to obtain results using this method within our current experimental setup.

V. Conclusion

This project reimplemented LIME and the Random K-Features explanation on a religion dataset using decision tree classification. By comparing the performance of these two explainers, we evaluated how effectively LIME can enhance human understanding of the model's decision-making process, thereby addressing the issue of non-transparency often encountered during model training.

Our findings suggest that LIME offers a significantly higher degree of interpretability than the Random K-Features explanation, making it a more reliable tool for understanding the reasoning behind predictions. However, some limitations were observed with LIME when applied to our specific dataset, which suggests areas for further improvement and optimization.

References

[1] W. N. Price, "Big data and black-box medical algorithms," *Science translational medicine*, vol. 10, no. 471, p. eaao5333, 2018.

[2] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[5] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[6] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with ga-based feature selection," in *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, 2005, pp. 136–141.

[7] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Computing and Applications*, vol. 23, pp. 2387–2403, 2013.

[8] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, M. S. Choi, I. H. Choi, C. Cung Van, N. V. Nhung *et al.*, "Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging," *Frontiers in medicine*, vol. 7, p. 427, 2020.

[9] Y. Wu, L. Zhang, U. A. Bhatti, and M. Huang, "Interpretable machine learning for personalized medical recommendations: A lime-based approach," *Diagnostics*, vol. 13, no. 16, p. 2681, 2023.

[10] I. Muralikrishna and P. Jayanthi, "Lime approach in diagnosing diseases–a study on explainable ai," in *Explainable Artificial Intelligence for Biomedical Applications*. River Publishers, 2023, pp. 17–31.

[11] J. Dieber and S. Kirrane, "Why model why? assessing the strengths and limitations of lime," *arXiv preprint arXiv:2012.00093*, 2020.