Phrase-Based LIME

Anonymous ACL submission

Abstract

This paper introduces Phrase-Based Local Interpretable Model-Agnostic Explanations (Phrase-Based LIME), a novel posthoc interpretability method in natural language processing. Building upon the foundational LIME framework, this research extends its capabilities by focusing on phrases as the basic unit of interpretability, rather than individual words. This approach aligns more closely with contextualized word embeddings, offering enhanced explainability for complex NLP models. We conduct a human-grounded experiment using sentiment models trained on the IMDb Movie Review corpus, demonstrating that non-expert users can more accurately and confidently discern between machine learning models using Phrase-Based LIME compared to traditional word-based methods.

1 Introduction

The field of machine learning and natural language processing increasingly employs models whose inner workings are opaque, making interpretability an important concern in many applications. Popular interpretability methods, like Local Interpretable Model-Agnostic Explanations (LIME), have made significant strides in elucidating model decisions in NLP applications but often fall short in providing contextual understanding due to their word-based nature.

This paper proposes an innovative approach, Phrase-Based LIME, which enhances the interpretability of models by utilizing phrases as the fundamental unit for explanations. We also develop a new human-grounded interpretability evaluation method, asking non-expert participants to choose the better machine learning model given two local explanations on the same input. Phrase-Based LIME is designed to provide more contextually rich and human-comprehensible explanations, bridging the gap between complex model outputs and their interpretability.

2 Background and Motivation

2.1 Definition of Terms

Interpretability of neural network models is a growing research area in natural language processing and machine learning in general. As such, most terms in interpretability are not well-established, so we will begin by defining our terms. At a high level, we will adopt Doshi-Velez and Kim's (2017) definition of interpretability as the "ability to explain or to present in understandable terms to a human." In this paper, we use explainability and interpretability interchangeably, although some researchers distinguish between these (Rudin, 2019; Gilpin et al., 2019).

We believe explainability is increasingly important in natural language processing as machine learning models become more complex. Explainability allows non-experts to interpret and predict model behavior, detect shortcomings or limitations of a model, and build trust in models. Since NLP models impact our lives in many ways, it is important that we all have opportunities for insight into these models and their behavior.

2.2 Types of Explainability

Some types of machine learning models lend themselves to intrinsic interpretability, meaning that the system itself can explain the decisions or steps that lead to a given outcome, like simple decision trees. In our case, however, we are interested in explainability for models without intrinsic interpretability.

Specifically, we are already using successful machine learning RoBERTa-style models, a very popular NLP approach. RoBERTa, like other contextualized word embeddings, are trained to represent word meaning based on the context surrounding the word. Our goal is to develop explainability that leads to increased transparency about model work-

ings not only for in-house technical improvements but to better answer user questions about model score. It is crucial for users to be able to interpret machine learning models, especially when users are relying on the output of these models' predictions.

To accomplish this, we needed to develop a posthoc model to interpret the existing models we had built for our product. Perhaps the most popular posthoc approaches are LIME and SHAP (Ribeiro et al., 2016; Lundberg and Lee, 2017). Our contribution in this paper is an extension of LIME.

LIME (Local Interpretable Model-Agnostic Explanations) is very versatile and designed to work with any type of machine learning model. LIME is a type of local explanation, providing an explanation for a particular input to the model. LIME takes the local input and creates vectors of feature perturbations from the input, usually words in the NLP context. Then a simple model is trained to mimic the behavior of the original model, constrained to be locally faithful to the original model. LIME calculates a numerical explanation value for each feature, based on the overall impact each feature has on the model's prediction–to what degree and in what direction did this feature impact the score. Other variations of LIME exist, notably DLIME (Deterministic LIME), which reduces randomness by deterministically selecting the neighborhood samples, and SP-LIME (Submodular Pick LIME), which aggregates multiple local LIME explanations to produce a global explanation (Zafar and Khan, 2019; Ribeiro et al., 2016).

Some of the difficulty with these explanation approaches as they are used today is that they are word-based, so the lack of other textual context, especially in models that use contextual word embeddings or other kinds of contextual information, can bring about misleading results.

2.3 Evaluation of Interpretability Methods

There is a lot of variation in how interpretability methods are evaluated. Doshi-Velez and Kim (2017) separate the evaluation of interpretability into three categories, functionally-grounded, human-grounded, and application-grounded. Functionally-grounded interpretability checks how true the explanation is to the original model, application-groundedness checks how much the explanation improves the downstream application results, and human-groundedness checks whether the explanation is useful to humans. In our case, the purpose of our explanation is to allow non-expert humans, particularly job applicants and hiring teams, to interpret the model, so we will evaluate our methods with a human-grounded simulation annotation approach, similar to Hase and Bansal (2020).

3 Methods

3.1 Our Goal

In this work, we present a posthoc interpretability approach which uses an adjusted phrase-based LIME technique to label the impact of phrases on a model's score of a local document. Our goal is to determine whether humans are able to distinguish between a machine learning model and a slightly worse machine learning model using this phrase-based LIME technique more effectively than with classic word-based LIME. We follow the concept of cognitive chunks as outlined by Doshi-Velez and Kim (2017). We present explanations using derived features, specifically phrases, which we believe align more effectively with contextualized word embedding features than word-based explanations. Our hypothesis is that phrase-based explanations provide additional semantic context, therefore enhancing human understanding of model behavior.

3.2 The Model to Explain

We began by training two IMDb sentiment models for our experiment. The IMDB Movie Review corpus contains 50,000 movie reviews (half for training, half for testing) with a score greater than or equal to 7 out of 10 or less than or equal to 4 out of 10 (Maas et al., 2011). We chose this task because it is a popular task with a public research dataset, making this work transparent and easily reproducible.

Both sentiment models' features include bag-of-words and RoBERTa embeddings. We train a binary sentiment analysis regression model over the IMDb Movie Review dataset. Because we want to have human participants compare a good machine learning model and an inferior model based on explanations alone, we trained two regression models on the IMDb dataset. Model A has 0.92 area under the curve (AUC), and we randomly shuffled 15% of the labels of the data for Model B, resulting in 0.87 AUC.

3.3 Phrase-Based LIME Methods

Then, we train our posthoc interpretability model. We select all 25,000 training examples from the IMDb Movie Review dataset to analyze for results. First, we run these review documents through a phrase parser built off of SpaCy's dependency parser and an in-house grammar. We contextualize these phrases by reattaching syntactic heads and then remove any entirely overlapping phrases. This means that the phrases are not always continuous text in the original sentence.

Once we have a set of phrases, we train our phrase-based LIME model. Our input perturbations are of phrases instead of words, in this case, and we use 5,000 perturbations of these phrases per local example. The phrase-based LIME model outputs each phrase with the source sentence and a LIME score.

Now each phrase from the training data is associated with a LIME score, which we can use to highlight the original text (in this case, an IMDb review) with shades of green or red to show the impact of the phrase on the overall score. Darker shades of the color show phrases with larger positive or negative LIME scores and lighter colors show less impactful LIME scores.

This explainability approach is currently only built and tested in English, but the approach is designed to be extensible to other languages. This extension of LIME only applies to models that use text as input. As with any explainability methods, there is always a risk that the explanation could be incomplete or otherwise misleading, and explainability methods should be used as a tool to understand machine learning models rather than understood as a perfect representation of the model and its behavior.

4 Experiments

4.1 Hypothesis

The goal of our experiment is to determine whether phrase-based or word-based LIME is more interpretable to humans, using a human-groundedness evaluation of phrase-based LIME by comparison to the already-proven word-based LIME. This was inspired by similar interpretability method approaches by Hase and Bansal (2020) called "forward simulation," where participants guess the prediction of the model with and without a learning phase that includes input examples with explanations provided. The main difference in our simulation is that participants are not guessing the prediction of the model but are comparing two models and choosing the more accurate model. We make this change for two reasons: (1) we believe this is a practical way non-experts can use explanations and (2) the prediction of the model often becomes very evident in the explanation itself–in the comparison task we are looking at nuances between the behavior of two models, not considering the positive-vs-negative model prediction. Also, in our evaluation we always show explanations because we are comparing two explainability methods, not attempting to evaluate whether explanations add value over no explanations.

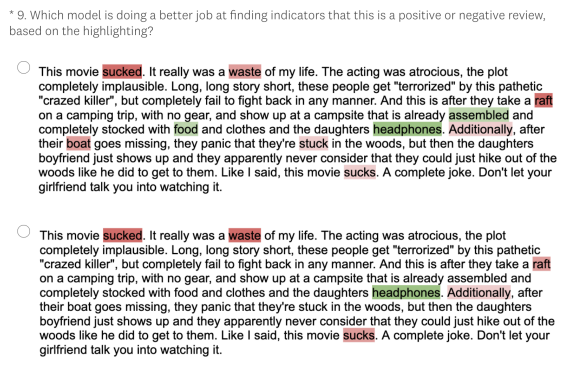

Figure 1: Example annotation page with Word-Based LIME

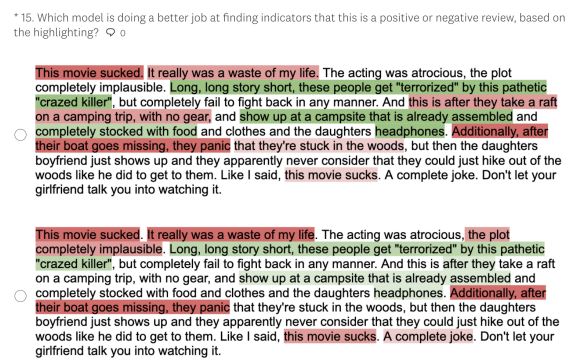

Figure 2: Example annotation page with Phrase-Based LIME

4.2 Human Evaluation Experiment

We set up a human evaluation simulation annotation experiment, where participants are shown the same IMDb review twice, scored by two different models. Model A has a 10.9% higher R-squared value and a 5.3% higher area under the curve (AUC) score (0.92 vs 0.87 AUC) in labeling IMDb movie review sentiment over Model B

(achieved by adding noise to the training data for Model B). The participant is shown the LIME explanation for an IMDb review for each model and asked to choose the better model. In some cases, the explanation is a phrase-based LIME explanation and in other cases, it is a word-based LIME explanation. Explanations are shown via green and red highlighting, where darker highlighting means more impactful LIME scores and where green is positive sentiment score and red is negative sentiment score. Because participants are not machine learning experts, the instructions state that the highlighting represents how important the model thinks these phrases are to the overall sentiment of the movie review, and the participants are asked to choose which model did a better job finding indicators this was a positive or negative review. No distinction is made between phrase and word based explanations in the instructions, but every participant is shown both kinds of explanation, though not side by side. After choosing the better model based on the explanation highlighting, the participant was asked to use a sliding scale (0-100) to rate their confidence in their answer. The full instructions for the participants can be found in Appendix A.

A general sample of 210 people was gathered by Prolific and randomly split into two groups, which we will call Group 1 and Group 2. Every participant was a native English speaker. The average participant age was 35.7, 51% female, . The racial demographics were 67% white, 10% black, 10% mixed, 7% Asian, and 6% other. Each participant who completed the survey was paid $7, with an average survey completion time of 15 minutes means an average hourly wage of over $27, which certainly exceeds the minimum wage requirements across the country of residence for all participants (United States).

We used Prolific to handle sourcing and splitting our participants randomly into two groups. Group 1 reviewed 26 LIME explanations, 50% of which were phrase-based and 50% of which were word-based. Group 2 saw the complement of Group 1's experiment, meaning that if Group 1 saw the phrase-based explanation of a particular movie review, Group 2 saw the word-based explanation of the same review. Examples can be seen in Figures 1 and 2.

The questions were shown and completed through a SurveyMonkey survey and questions were randomly shuffled for each participant to account for participants improving or fatiguing in the task as they went along. We removed from the results any participants who did not complete all questions. Overall, we had 208 completed surveys that we analyzed, 109 in Group 1 and 99 in Group 2.

5 Results

5.1 Human Results on Phrase-based vs. Word-based LIME

To analyze our results, we were mostly interested in how successful participants were in choosing the more accurate machine learning model when shown our phrase-based LIME vs the classic word-based LIME. Overall, humans were 14% more accurate ($p < 0.005$) at choosing the better machine learning model when shown the phrase-based LIME explanation and 9% more confident in their choice ($p < 0.001$). Comparison results by individual IMDb movie reviews can be seen in Figure 3 (accuracy) and Figure 4 (confidence), the average participant metrics over the word-based and phrase-based explanations for the same review are shown in a pair of blue and red bars, respectively.

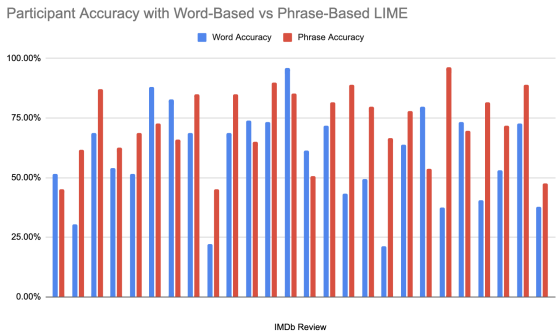

Figure 3: Participant Accuracy for each IMDb review with Word-Based vs Phrase-Based LIME

Interestingly, participants' reported confidence in their own choice was highly correlated with their accuracy in picking the better model (Pearson's R = 0.656).

5.2 Comparing the Groups

We compared metrics across each group to ensure that results were consistent and one group was not somehow much better at the task than the other group. Groups were comparable with Group 1 having 16.58% higher accuracy on phrase-based LIME and Group 2 having 9.48% higher accuracy on phrase-based LIME explanations than word-based LIME explanations. Both Group 1 and 2 were more

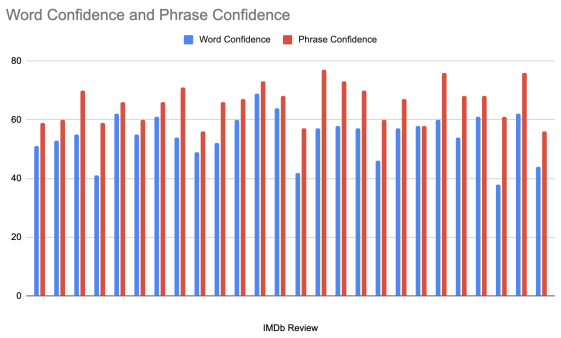

Figure 4: Participant Confidence for each IMDb review with Word-Based vs Phrase-Based LIME

confident in their choices on phrase-based LIME (17% and 5% respectively). Given the consistency in findings across groups, we consider the reported results to be accurate.

One interesting group distinction is between machine learning experts and non-machine learning experts. We ran a small pilot of this experiment internally and found that the 6 data scientists who participated had the same accuracy between word and phrase-based LIME, and that this accuracy was much higher than the pilot accuracy of the general public, at 89% and 88% accuracy in choosing the better model for word-based and phrase-based LIME, respectively. Of course, this is a very small experiment with machine learning experts, but it seems to indicate that while phrase-based LIME makes a model more interpretable to people in general, specialists can understand both word and phrase-based LIME equally well due to their familiarity with how these models work.

6 Conclusion

Our study concludes that Phrase-Based LIME significantly improves the interpretability of machine learning models for non-expert users. By shifting the focus from individual words to phrases, this method provides more contextually rich explanations that align closely with the way NLP models process language. The results from our human-grounded experiment using the IMDb Movie Review corpus reveal that participants were more accurate and confident in distinguishing between models using Phrase-Based LIME compared to traditional word-based explanations. This finding underscores the potential of Phrase-Based LIME in enhancing transparency and trust in machine learning applications. Future research directions include extending this methodology to non-English languages and developing global explanations using an ensemble of local explanations, which could further democratize understanding of machine learning models across diverse user groups.

References

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. Explaining explanations: An overview of interpretability of machine learning.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

Muhammad Rehman Zafar and Naimul Mefraz Khan. 2019. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems.

A Instructions for Annotators

Choose the Better Machine Learning Model

The job of machine learning models is to make predictions. In this project, we trained two machine learning models to predict whether a movie review is a positive or negative review. If you give the model a movie review, it will predict "positive" or "negative."

One of these models does its job better than the other. We are hoping you can spot the better model just by looking at the parts of the movie review the model thought was important.

In this survey, you will look at two images at a time. Both will be images of the same movie review with highlighting. The highlights in green are parts of the review that the model thought were positive, the highlights in red are parts the model thought were negative. Darker colors mean the model thought they were more positive or negative, lighter colors mean the model thought they were less. For each pair of explanations, look at the highlighting and select which model is doing a better job at what it considers signs of a positive or negative review.

You have to choose one of the models. We know it's tricky and often unclear, just do your best! In these multiple choice questions, "Model A" and "Model B" refer to one of the two models randomly, not the same model throughout the examples.

1. Which model is doing a better job at finding indicators that this is a positive or negative review, based on the highlighting?

2. How confident are you that you got this answer right?