Comparing the Effectiveness of LIME and SHAP in Identifying the Most Relevant Words for BERT's Sentiment Analysis Predictions

Laura Gozzo (2035565)

Abstract

This papers compares LIME and SHAP in their ability to identify the most important features for BERT's sentiment analysis of movie reviews. The result indicate that both LIME and SHAP are able to identify the core sentiment-driving words; however SHAP is better at identifying the most crucial features.

1 Introduction

Understanding why machine learning models make particular decisions is a critical challenge in the field of artificial intelligence. As machine learning models keep improving, they become more complex, and the need for tools to explain their decision grows alongside. In particular, deep learning models such as BERT (which achieves state-of-the-art performance in many natural language processing tasks (Koroteev, 2021)), are often compared to black-boxes, meaning that it is very difficult to interpret why they make specific predictions. Deep neural networks outperform humans on many tasks; however, their applicability in real life settings, in which they could be extremely valuable, is often limited by their lack of interpretability. When decisions lead to important consequences, as for example in the medical field, one cannot trust predictions made from machine learning models with blind faith. Models are evaluated based on their accuracy, however, inspecting which instances lead to each prediction is also necessary to increase trust in the model's predictions. The goal of this research is to compare the effectiveness of LIME and SHAP (two explainable AI tools), in identifying the most relevant features for sentiment classification by BERT.

Such research could also provide further insight into how models work, and help refine them. Models are trained and evaluated on train/test/validation datasets which all come from the same source; however, real life data can often differ significantly, in ways which are hard to predict and control. This can mean that the accuracy achieved on the validation set is not necessarily transferable to real life settings; inspecting predictions and their explanations could give insight into whether the model is making sensible decisions (Ribeiro et al., 2016).

A BERT model was trained to predict the sentiment of a set movie reviews, then, SHAP and LIME were be used to identity which words of the sentence were most useful to BERT's prediction. The results of this project indicate that SHAP is better at identifying the most important words in a sentence that explain BERT's predictions. Overall, although SHAP performs better, words identified by both LIME and SHAP retain the core sentiment-driving information in the text.

2 Related Work

Explaining a prediction in this context means providing qualitative measures of the relationship between the input features (in this case words of a sentence) and the model's prediction (Szczepański et al., 2021). Being able to explain a model's prediction is a crucial step to get humans to trust and implement machine learning effectively, if the explanations are faithful and easy to interpret (Szczepański et al., 2021). Practitioners tend to overestimate the accuracy of their models, therefore we should not rely on accuracy alone to trust a model (Alonso et al., 2021). There are multiple reasons why this tends to happen, data leakage refers to when there is unintentional leakage of signal into the training and validation datasets, that would not appear with real life scenarios (Bakshy et al., 2015).This type of issue would be very hard to identify by looking only at the model's performance, however, identifying which features led to the model prediction allows us to verify whether the model is using inappropriate data leaked in the dataset to make predictions. An example of this is mentioned by (Bakshy et al., 2015), where the patient's ID was heavily correlated with the target labels. Another issue hard to detect, known as dataset shift, refers to when training and test data are different (Bakshy et al., 2015).

LIME and SHAP are model agnostic tools that can be used to interpret the behavior of any machine learning model. LIME works in the following way: it generates perturbed samples of the input data by slightly modifying the original version. It passes the perturbed samples to the model (in this case BERT), to get predictions. Then, it uses a simple interpretable model to approximate the behavior of the original model; and finally, it generates values to identify the most important features for prediction (Szczepański et al., 2021). LIME is locally faithful to the classifier it is being used with, this means that its interprets single predictions of the model by locally approximating the model around the prediction (Lundberg, 2017).

SHAP (short for SHapley Additive exPlanations) is based on the game-theory concept of Shapely values (Shapley, 1953), and considered as an essential contribution to the field of explainable artificial intelligence. Shapely values were originally deployed as a method to distribute a reward among a set of players depending on their contribution to a specific outcome (Shapley, 1953); in this case the players are represented by the features of the input and the outcome is the prediction of the model. SHAP works in the following way: it computes the model's prediction for a given input, then, it considers all possible subsets of features of the input and calculates how they change the model's prediction. Finally it calculates how each feature contributes to the model's prediction and assigns each feature a value; known as the SHAP value. Compared to LIME, SHAP is more computationally expensive and works better with more complex models.

A BERT architecture for Sentiment classification (TFBertForSequenceClassification) was chosen for its ability to effectively solve sentiment analysis tasks, given its ability to capture semantic and contextual relationships within text. Furthermore BERT can work with masked features, which is important for this task (Koroteev, 2021). Another reason for choosing BERT is the fact that pre-trained BERT models are easily accessible online.

3 Methods

To address the research question of comparing the effectiveness of SHAP and LIME in interpreting BERT's predictions in a sentiment analysis task, the following approach was adopted.

A pre-trained BERT (Bidirectional Representation for Transformers) model was fine-tuned on a dataset consisting of movie reviews, the task was to classify the review's sentiment as either positive or negative.

The dataset used for this task consists of movie reviews taken from the IMDB website, it contains 5000 movie reviews and it is balanced, meaning half of the reviews are positive and the other half negative. A dataset of movie reviews was chosen for this task because reviews can be quite nuanced and contain both positive and negative elements in the same text. This makes it harder to classify them as either positive or negative since the model has to weigh the importance of the positive and negative attributes. To pre-process the dataset, the reviews were cleaned to remove noise and irrelevant elements such as: html tags, anything enclosed in square brackets, and special characters. Then, the dataset was split into training, test and validation sets, and tokenized. Tokenization is needed to convert text into numerical input and perform some additional pre-processing so that the data can be fed to the BERT model.

To interpret the model's decisions, a subset of the test dataset containing only the first 200 instances was selected to be interpreted. This is because using the entire dataset would have been too computationally intensive and slow to process. Because the code was still very slow to run, the number of perturbed samples to analyze with LIME, and number of subsets of features to analyse with SHAP, were both set to 100. Furthermore the batch size was set to 32. LIME and SHAP identified the most relevant words in each review, and these words were ranked based on their respective importance scores. Then, the reviews were modified to retain only the top 50% most important words while all the remaining words were replaced with a [MASK] token, this ensures they have minimal influence on BERT's prediction but keep intact the structure of the sentence. This created two filtered versions of each review, one filtered by LIME and the other by SHAP. To compare LIME and SHAP, two metrics were used: fidelity and prediction accuracy. Fidelity refers to the degree to which the filtered sen-

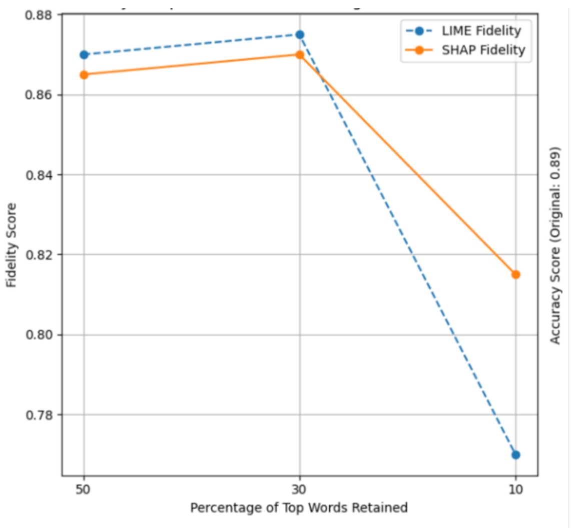

Figure 1: Fidelity Comparison Across Percentages of Retained Words

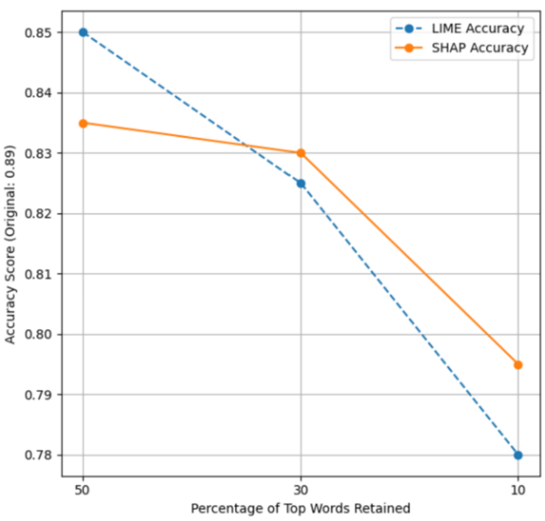

Figure 2: Accuracy Comparison Across Percentages of Retained Words

tences preserve the original predictions computed by BERT from the original sentences. Higher fidelity indicates that the words identified capture the important feature used for the model's predictions. Prediction accuracy refers to the proportion of filtered sentences that are correctly classified when compared to their target sentiment. To summarize, fidelity refers to the ability to identify words that are important to match BERT's output (regardless of whether the prediction is correct); and prediction accuracy refers to the ability to identify words important to produce correct predictions. To get more complete results, the same process was repeated with the top 30% and 10% most important words, this gives more insight into how LIME and SHAP performances change when having to identify only the most crucial features.

TensorFlow was used to implement BERT using Python; the libraries 'shap' and 'lime' were used to implement the explanatory models.

4 Results

The results in Fig[1] show that when retaining only the top 50% and 30% most important words, LIME produces a slightly higher fidelity compared to SHAP (0.87 vs 0.865 at 50% and 0.875 vs 0.87 at 30%), although the difference is very small, only 0.005. When retaining only the top 10% of words, both SHAP and LIME's fidelity drops significantly; however, SHAP produces a much higher fidelity compared to LIME (0.77 vs 0.815). This indicates that SHAP identifies more impactful words when fewer are retained.

The baseline accuracy achieved by BERT on the 200 sentences used to evaluate LIME and BERT is 0.89. Fig[2] shows how When only 50% of words are being retained, the masked sentences produced by LIME are correctly classified by BERT more often than the ones produced by SHAP (0.85 vs 0.835). However, when retaining fewer words, the opposite is true and SHAP's accuracy becomes higher than LIME's. As the percentage of words being retained decreases; this gap widens (0.83 vs 0.825 at 30% and 0.795 vs 0.78 at 10%).

These results suggest that while LIME is slightly better at retaining fidelity when identifying a larger subset of important words, SHAP performs better when having to identify the most critical subset of words. Furthermore, the accuracy of BERT's predictions remains relatively high on the sentences filtered by SHAP, and only drops by 0.1 (from 0.89 to 0.795) when only 10% of the input's features are preserved.

4.1 Discussion

The objective of this project was to compare LIME and SHAP in their ability to identify important words for BERT's sentiment analysis classification. The results indicate that SHAP is better at identifying the most crucial words to explain BERT's predictions. LIME achieves higher fidelity at higher word retention levels, suggesting it may be slightly better at approximating the importance of features in BERT's predictions when the task is less restrictive. However, the difference in performance is very small (0.005), and almost negligible. Furthermore, SHAP significantly outperforms LIME at

10% word retention, indicating its better at identifying a smaller, more impactful subset of features

The accuracy achieved by BERT on filtered sentences, even at reduced word retention levels, remained relatively high. This indicates that both LIME and SHAP are able to identify the words that retain the core sentiment of the text. However, the words identified by SHAP are better at this than the ones identified by LIME.

Given the fact that BERT uses contextual embeddings to analyze text, a better approach would have been to consider n-grams instead of singular words as features, and the contextual relationships between words. This would preserve the semantic structure of the original sentences better in the filtered sentences. Furthermore, more accurate results would have been achieved by not putting a limit to the number of perturbed sentences and sampled features analyzed by LIME and SHAP respectively, although this would be very computationally expensive given SHAP's complexity.

This study investigates the trade-off between computational efficiency and interpretability when applying LIME and SHAP to complex models like BERT. Further studies could explore the impact of incorporating better contextual preservation techniques.

References

Miguel A Alonso, David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. 2021. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144.

Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games*, 2.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.