# **PROJECT REPORT 1**

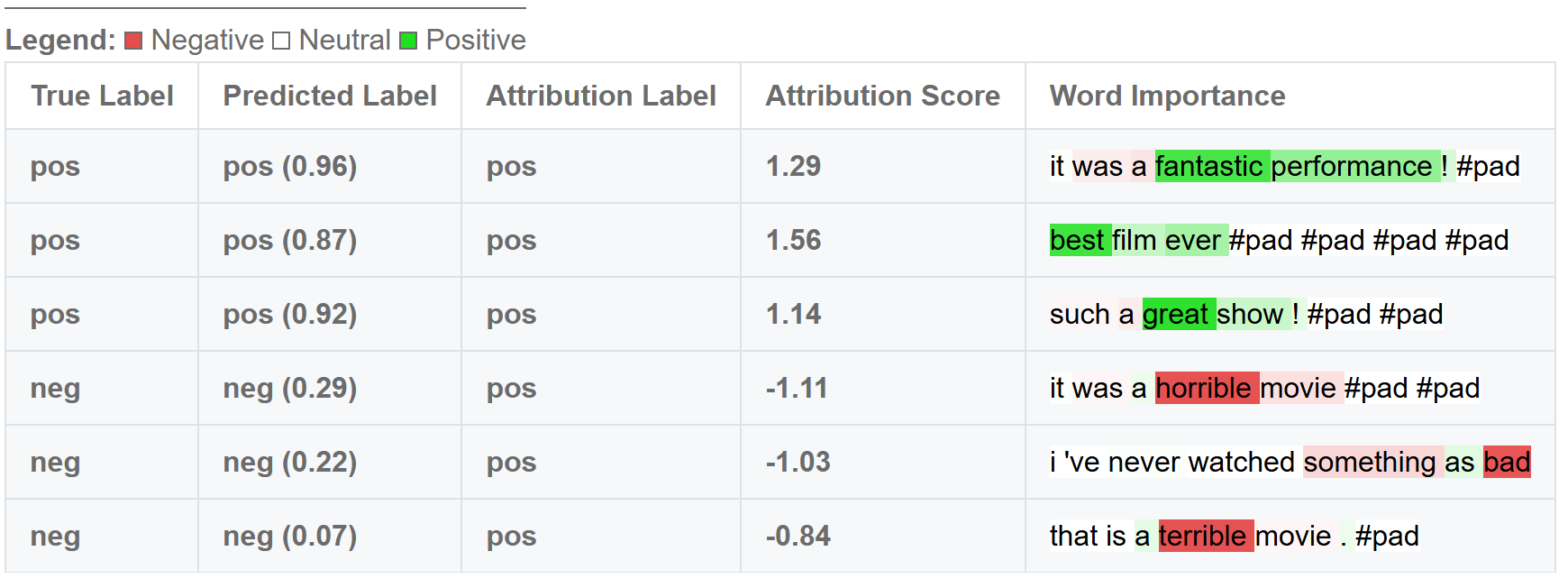
Szymon Pająk, Tomasz Ogiołda

|  |  |  |  |
| --- | --- | --- | --- |
| Aspect | Captum | SHAP | LIME |
| What it is | Captum is a PyTorch-specific library for interpretability that provides gradient-based methods to attribute input relevance in neural networks. | SHAP is a model-agnostic and model-specific explainability framework based on Shapley values from game theory. | LIME is a model-agnostic tool that approximates black-box model predictions locally using simple surrogate models like linear regression. |
| Model compatibility | Integrated with PyTorch and only supports models built in this framework. | Supports a wide range of models: deep learning, tree-based, and ensemble methods, with both model-specific and agnostic variants. | Works with virtually any model by treating it as a black box and does not require access to internal structure or gradients. |
| Explanation method | Relies on internal model gradients to compute attributions, offering precise insights into the model’s behavior layer by layer. | It attributes prediction output to features using theoretically grounded Shapley values, ensuring consistency in explanations. | It perturbs inputs and observes model responses to fit a simple interpretable model around the prediction, focusing on local behavior. |
| Local vs Global | Local and Global explanations | Local and Global explanations | Only Local |
| Computational efficiency | Efficient, and suitable for deep models where gradient access is feasible. | Can be computationally intensive, particularly with large datasets or deep models. Often requires approximations to reduce cost. | It's lightweight and relatively fast. But it relies on random sampling what can make results less stable. |
| use case | Best while working with PyTorch deep learning models. | Best for users needing model-agnostic, reliable explanations across different model types. Especially where global interpretability is needed. | Best for quick, local explanations in exploratory analysis or when working with any black-box model in early development stages. |

**Visualizations comparison**

Visualizations come from sentiment analysis tutorials provided by libraries.

**CAPTUM**

****

**SHAP**

**Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.**

**LIME**

**Obraz zawierający tekst, Czcionka, numer, zrzut ekranu

Zawartość wygenerowana przez sztuczną inteligencję może być niepoprawna.**