PROJECT REPORT 1

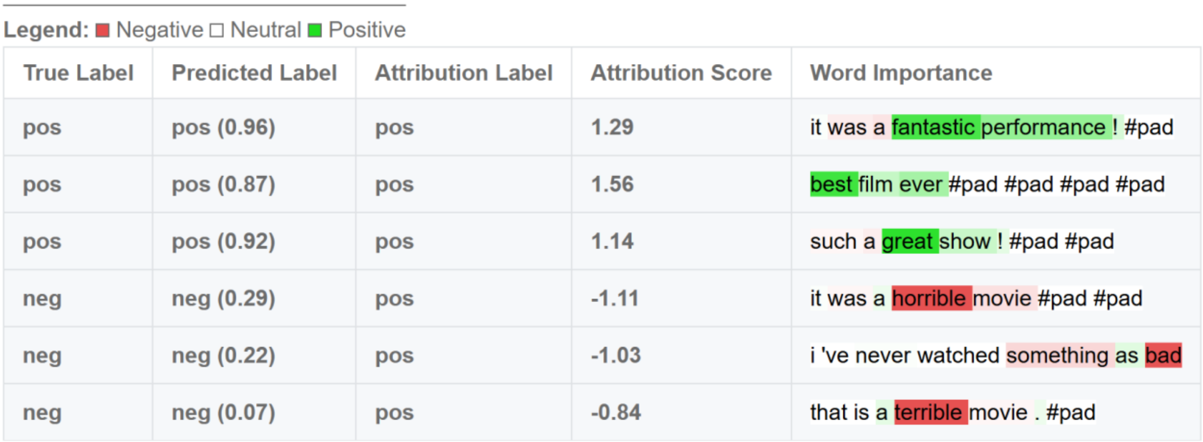
Szymon Pająk, Tomasz Ogiołda

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.

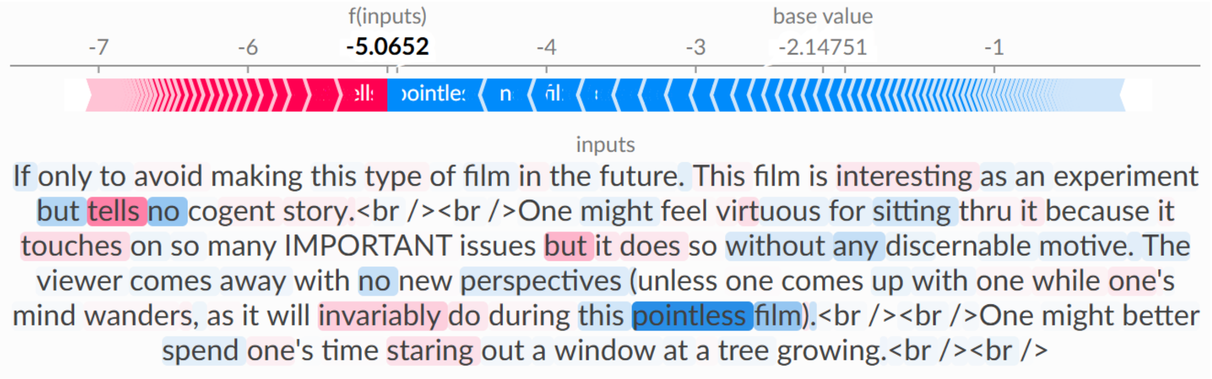
CAPTUM



Legend: 🟥 Negative □ Neutral 🟩 Positive

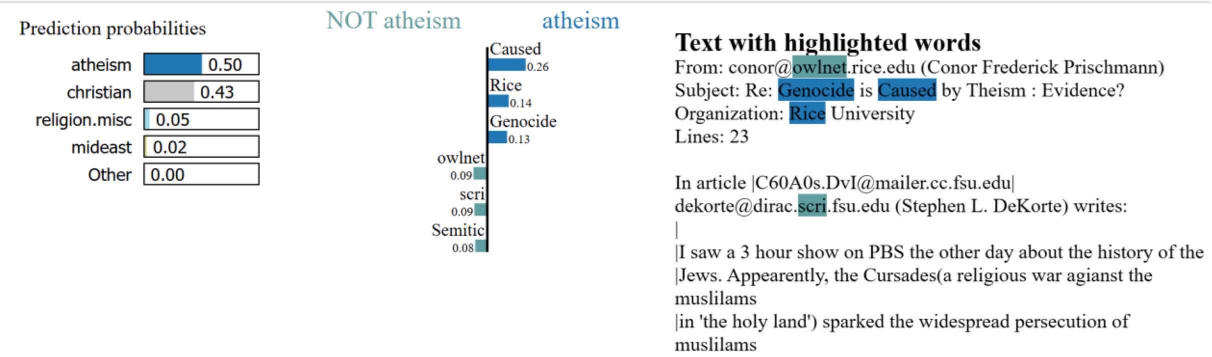
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad

SHAP



If only to avoid making this type of film in the future. This film is interesting as an experiment but tells no cogent story.<br /><br />One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so without any discernable motive. The viewer comes away with no new perspectives (unless one comes up with one while one's mind wanders, as it will invariably do during this pointless film).<br /><br />One might better spend one's time staring out a window at a tree growing.<br /><br />

LIME



Text with highlighted words

From: conor@owlnet.rice.edu (Conor Frederick Prischmann)
Subject: Re: Genocide is Caused by Theism : Evidence?
Organization: Rice University
Lines: 23

In article |C60A0s.DvI@mailer.cc.fsu.edu|
dekorte@dirac.scri.fsu.edu (Stephen L. DeKorte) writes:
|
|I saw a 3 hour show on PBS the other day about the history of the
|Jews. Appearently, the Cursades(a religious war agianst the
muslilams
|in 'the holy land') sparked the widespread persecution of
muslilams