Evaluating Feature Attribution Methods: an Usecase on a Neural Fact-checking Model

Annabel Simons[1]
Supervisors: Avishek Anand[1], Lijun Lyu[1], Lorenzo Corti[1]
[1]EEMCS, Delft University of Technology, The Netherlands

TU Delft

1. INTRODUCTION AND BACKGROUND

In today's society, claims are everywhere, in the online and offline world. Fact-checking models can check these claims and predict if a claim is true or false, but how can these models be checked?

Explainable Artificial Intelligence (XAI) can offer a solution here. Two XAI approaches were used in our research:

1. The **post-hoc feature attribution methods**. They give scores indicating the influence of the individual tokens on the model's decision-making; see Figure 1.
2. Another XAI approach is to make a model **interpretable-by-design**, like ExPred [1]. This kind of model gives an explanation for every prediction.

The research question for our research is:
How do feature attribution methods for XAI compare with each other in the context of fact-checking models using ExPred [1]?

Claim: Harris Jayaraj is Indian.	Prediction: Supports
harris jayaraj - lrb - born 8 january 1975 - rrb - is an india n film composer from chennai , tamil nadu . he composes sc ...	

Figure 1: The figure shows how the explanations were presented in the survey (the user study). The claim and prediction are offered at the top. At the bottom, a heatmap on the context is presented.

2. METHODOLOGY

LIME [2]
- Perturbation based
- Model-agnostic
- Locally faithful

Kernel SHAP [3]
- Perturbation based
- Model-agnostic
- Locally faithful

Integrated Gradients [4]
- Gradient based
- Model-specific
- Locally faithful

Implemented with the Captum library

Pseudo Ground Truth Evaluation
- The **Jaccard Distance** between:
  - ExPreds explanations
  - Three methods explanations
- **Kendall's Tau** between the three methods to see if there's a consensus.

Human-grounded Evaluation
- **Rate** the explanations on:
  - **Understandability**
  - **Insightfulness**
  - **Rely on the model**
- **Rank** the explanations on:
  - **Insightfulness**

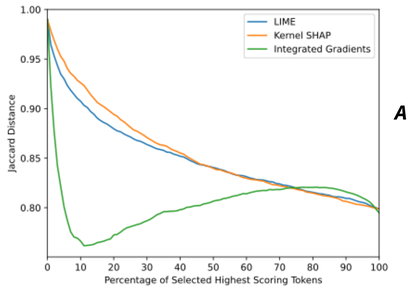
3. RESULTS

3.1 Pseudo Ground Truth Evaluation Results

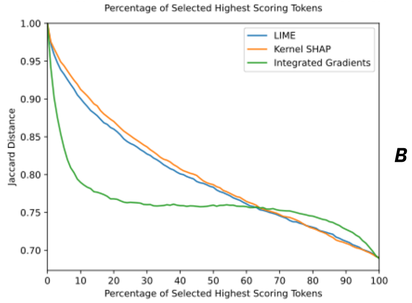
- **Integrated Gradients is more similar** to the pseudo ground truth (ExPreds explanations); see Figure 2a and 2b.
- The Kendall's Tau results showed **an absence of association between the three methods' explanations**; see Tables 1a and 1b.

3.2 Human-grounded Evaluation Results

- For the rating section, **Integrated Gradients had more variation in how it was rated** than LIME and Kernel SHAP; see Figure 3.
- In the ranking section, **Integrated Gradients won three out of five times. LIME and Kernel SHAP** had similar rankings; see Figure 4.

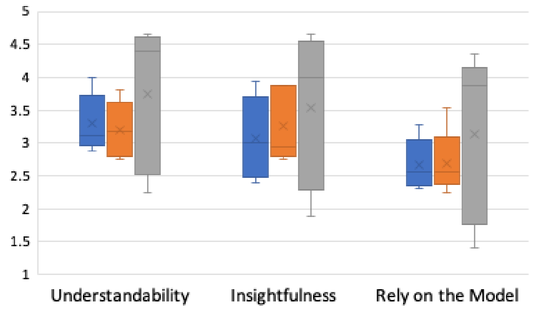


A



B

*↑ Figures 2a and 2b: Jaccard Distance between the feature attribution methods percentage of selected highest scoring tokens and the explanation from ExPred with 100 instances from **A: the train set and B: the test set**. Close to 0 means similar, and close to 1 means dissimilar.*



*← Figure 3: The results of the **rating** of the explanations of the feature attribution methods on understandability, insightfulness, and relying on the model. The y-axis is a 5-point Likert scale, ranging from "1 - strongly disagree" to "5 - strongly agree".*

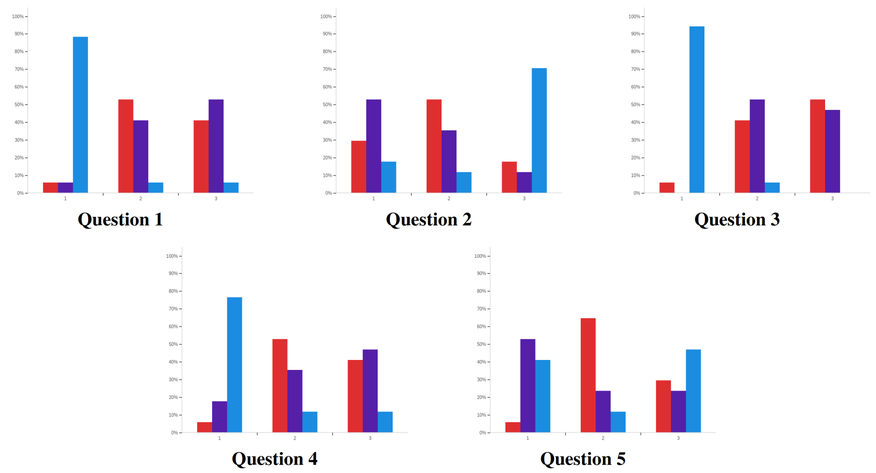


*Figure 4: The results for the questions related to **ranking** the explanations of the feature attribution methods on how much insight they gave. The x-axis is the ranking 1 up to 3, and the y-axis is the normalized count of the choice in percentages. Red = LIME; Purple = Kernel SHAP; Blue = Integrated Gradients.*

A

Compared methods	Kendall's Tau	p-value
LIME/Kernel SHAP	0.0204	0.4556
LIME/IG	0.0281	0.4103
Kernel SHAP/IG	0.0101	0.5030

B

Compared methods	Kendall's Tau	p-value
LIME/Kernel SHAP	0.0545	0.3851
LIME/IG	0.0444	0.3983
Kernel SHAP/IG	0.0207	0.3661

*Tables 1a and 1b: Kendall's Tau for two feature attribution methods looks at 100 instances from **A: the train set and B: the test set**. For Kendall's Tau: close to 1 means the two methods are similar, and -1 means the two methods are dissimilar.*

4. DISCUSSION

- Integrated Gradients seems to outperform the rest.
- On the other hand, **all feature attribution methods are never very similar to the pseudo ground truth.**
- Additionally, **Integrated Gradients** is overall better rated, but also **quite poorly for the second question.**
- The results indicate that **the iterations** should have been **higher** for the perturbation-based methods.

5. CONCLUSION AND FUTURE WORK

The main findings are:

- There is **no consensus** among the explanations from the feature attribution methods.
- Integrated Gradients **seems to outperform** LIME and Kernel SHAP in the pseudo ground truth and human-grounded evaluation, but maybe this is only the case due to **the number of iterations**.

For future research, it would be beneficial to run more instances and more iterations per instance.
Additionally, it would be beneficial to see if similar results are achieved for other tasks and models.

REFERENCES

[1] Z. Zhang, K. Rudra, and A. Anand, "Explain and predict, and then predict again," in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, ser. WSDM '21, Virtual Event, Israel, 2021, pp. 418–426. DOI: 10.1145/3437963.3441758.
[2] 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. DOI: 10.1145/2939672.2939778.
[3] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," CoRR, vol. abs/1705.07874, 2017. arXiv: 1705.07874.
[4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," CoRR, vol. abs/1703.01365, 2017. arXiv: 1703.01365.