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

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 30, 2023

Name of the student: Annabel Simons
Final project course: CSE3000 Research Project
Thesis committee: Avishek Anand, Lijun Lyu, Lorenzo Corti, Marco Loog

Abstract

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? Post-hoc XAI feature attribution methods can be used for this. These methods give scores indicating the influence of the individual tokens on the model's decision-making. In our research, we evaluate three popular feature attribution methods in the context of fact-checking: LIME, Kernel SHAP, and Integrated Gradients. We used the NLP architecture ExPred as a fact-checking model in our research. The feature attribution methods were evaluated using a human-grounded and pseudo ground truth evaluation. The results from these evaluations indicate that Integrated Gradients enables humans to form an opinion better and performs better in our pseudo ground truth evaluation. A potential explanation is that the iterations should be increased for LIME and Kernel SHAP. Our findings suggest that Integrated Gradients performs better in our study. Still, more research for other tasks and models would be beneficial to ensure that these results apply to other cases.

1 Introduction

Deep learning models have been successfully used for various Natural Language Processing (NLP) tasks in recent years, but these models are often "black boxes," making it hard to demystify them [1]. One of these models is the language representation model: BERT [2]. BERT can be finetuned and used for diverse purposes, such as for the NLP task of sentiment analysis or question answering [2]. A disadvantage of BERT is that it can be seen as opaque due to a large number of parameters, making it a so-called "black box" [1]. The rise of "black box" models, like BERT, leads to interpretability problems [1]. This clashes with the GDPR Recital 71, which says there is a right: "to obtain an explanation of the decision reached" [3, Recital 71]. Therefore, transparency in the decision-making of these models is of great importance.

So the advent of these NLP models comes with a new problem, why does the model make the decisions it does? Explainable Artificial Intelligence (XAI) methods endeavor to tackle this problem by explaining what happens in these "black boxes" [1]. One of these XAI methods is the post-hoc feature attribution method, seen in tools like LIME [4], Kernel SHAP [5], and Integrated Gradients [6]. Feature attribution methods can be applied to models that work with, for example, images, tabular data, and text. For text, they give scores to the separate words, tokens, in the input [7]. Feature attribution methods can provide insight into the importance of each token for the classification, which can show if a vital word is missing in the decision-making [7]. Feature attribution methods used for NLP tasks provide scores for individual tokens, indicating their influence on the classification of an instance [7]. Next to post-hoc XAI methods, there is also the *interpretable-by-design* method, like the NLP architecture: ExPred [8]. ExPred can be used for several tasks after being trained on a dataset [9]. *Interpretable-by-design* methods do not only give the classification, but additionally, an explanation [9].

A field in NLP where an explanation is needed is fact-checking. In the last few years, fact and fiction became harder to distinguish, for instance, during the Brexit campaign [10, pp. 1-11]. One may argue that we have entered an era of post-truth [10, pp. 1-11]. ExPred can help solve this problem; it can be used for fact-checking after training it with a dataset such as the *Fact Extraction and VERification* dataset (FEVER) [9], [11]. Fact-checking tools can help decide if a claim is accurate, but their decisions must be explainable.

As stated earlier, there are several feature attribution methods, and currently, there is a lack of understanding of how these tools compare to each other in the context of fact-checking models. Therefore this paper will investigate the research question:

RQ How do feature attribution methods for Explainable Artificial Intelligence (XAI) compare with each other in the context of fact-checking models using ExPred [9]?

This research question (RQ) can be split up into two subquestions, these are:

SQ1 Which of the three feature attribution methods (LIME, Kernel SHAP, and Integrated Gradients) outperforms the rest looking at a pseudo ground truth evaluation?

SQ2 Which of the three feature attribution methods (LIME, Kernel SHAP, and Integrated Gradients) outperforms the rest looking at a human-grounded evaluation?

The contributions of our work are, in short:

- The results from our pseudo ground truth evaluation (SQ1) show that the feature attribution method Integrated Gradients gives proportionally higher scores to the tokens compared to the pseudo ground truth than LIME and Kernel SHAP, and there seems to be no consensus in explanations among the three feature attribution methods: LIME, Kernel SHAP, and Integrated Gradients.
- The results from our user study (SQ2) indicate that the method Integrated Gradients enables users to form an opinion better than LIME and Kernel SHAP.

For reproducibility, the code with the implementation for the feature attribution methods and how to evaluate them are available at https://github.com/Herbje/Evaluating-Feature-Attribution-Methods.

2 Problem Description

Feature attribution methods can be used to explain a decision made by a model [1]. They use a post-hoc approach [1]. For post-hoc approaches, the model is already trained, and post-hoc methods try to make this model interpretable [9], [12]. According to Zhang, Rudra, and Anand [9], the developer can avoid making the trade-off between performance and interpretability when using a post-hoc method.

However, the question arises if these feature attribution methods give the correct explanation when demystifying a model and how these feature attribution methods compare with each other (see RQ). The problem of not having a clear ground truth makes evaluating feature attribution methods hard, and complicated [7], [9]. Humans are often used to assess these methods, which leads to a human-grounded evaluation [1]. Here *interpretable-by-design* models, like ExPred, can offer a solution [9]. When an *interpretable-by-design* model gives its prediction, it also provides an explanation for this decision together with it [9]. Rudin [12] says that these explanations are "faithful to what the model actually computes," which is better, in her opinion, than post-hoc approaches.

Therefore in our research, we use the explanations from ExPred, in our case, a trained *interpretable-by-design* fact-checking model, as a pseudo ground truth. This makes a pseudo ground truth evaluation possible (see SQ1). The input for this pseudo ground truth evaluation is the explanations from the three feature attribution methods (LIME, Kernel SHAP, and Integrated Gradients) and the explanations from ExPred. The output for this pseudo ground truth evaluation is the similarity between the methods' explanations themselves and with explanations from ExPred. Next to the pseudo ground truth evaluation, we conducted a user study for a human-grounded evaluation (see SQ2). The explanations from the three feature attribution methods (LIME, Kernel SHAP, and Integrated Gradients) were presented to the participants of the study (the input). The outcome of this user study was a set of explanations scored on the understandability of the explanation, insightfulness of the explanation, and if the user would rely on the model.

3 Related Work

The current literature on the topic investigated in this paper can be divided into feature attribution methods, machine learning for fact-checking, XAI for fact-checking, and the evaluation of feature attribution methods. Firstly, subsection 3.1 will describe feature attribution methods in more detail. Then, subsection 3.2 will discuss research on XAI, specifically in the context of fact-checking. Subsection 3.3 describes how machine learning and fact-checking relate to each other. In the last subsection 3.4, we describe in further detail research on how feature attribution methods can be evaluated and the results of previous research.

3.1 Feature Attribution Methods

Feature attribution methods are used in the field of XAI to demystify deep learning models. They do this by giving scores to the attributes based on the individual attributes' contribution to the prediction [7]. There are different types of feature attribution methods. Firstly, a feature attribution method can be model-agnostic or model-specific. Model-agnostic feature attribution methods treat the model as a "black box" and only look at the in- and output [1]. Model-specific feature attribution methods do the opposite; they look at the internal parameters of a model for their explanation [1]. Lastly, feature attribution methods can be globally or locally faithful. Globally faithful means the feature attribution method can explain the behavior of a model on all instances in the dataset [1]. On the contrary, locally faithful say something about a specific instance [4]. The feature attribution methods we chose for our research are described in more detail in subsection 4.1.

3.2 Fact-checking and Machine Learning

With the rise of misinformation, there is also a rise in fact-checking tools [13]. According to Zeng, Abumansour, and Zubiaga [13], these tools help to keep up with the fast-moving internet, because doing this by hand would take too long. There are several datasets to train these fact-checking models [13], for instance, FEVER [14], SCIFACT [15], and UKP Snopes [16]. FEVER [14], and SCIFACT [15] retrieve evidence, and with this evidence, they validate the claim [13]. UKP Snopes [16] does this differently. It first retrieves documents related to the claim, extracts sentences of importance, stances these sentences, and then validates the claim with this evidence [16].

SCIFACT [15], and UKP Snopes [16] are both relatively small. SCIFACT is an expert-annotated dataset and has 1,409 claims [15]. It was made during the COVID-19 pandemic to check claims on COVID-19 [15]. UKP Snopes has 6,422 claims, and to build the corpus, they used information from places on the web where many false claims are present like social media [16]. A more detailed description of FEVER can be found in subsection 5.2.

3.3 Fact-checking and XAI

Literature on XAI in fact-checking is currently scarce and focuses mainly on presentation methods rather than on what a model does. Lim and Perrault [17] and Linder, Mohseni, Yang, *et al.* [18] did an XAI study in the context of fact-checking. Lim and Perrault [17] conducted a user study with 22 participants. They showed these participants the explanations from an XAI method in five different ways. Their results showed that explanations with more text were preferred [17]. Linder, Mohseni, Yang, *et al.* [18] found similar results in their research. On the other hand, they noticed a clear trade-off when participants were given more information; it took a lot more time to assess explanations [18].

To conclude, research on XAI in the context of fact-checking is mainly focused on presentation methods. Lim and Perrault [17] and Linder, Mohseni, Yang, *et al.* [18] looked at what the best way was to present an explanation to users; what the model does is not essential. This current research shows that more text helps participants with assessing but also increases the time needed for the assessment [17], [18]. The gap in the literature about looking at a way to evaluate XAI methods using a ground truth shows where our research can contribute to this research field.

3.4 Evaluating Feature Attribution Methods

The evaluation of feature attribution methods can be split in two: with or without humans [4], [7], [19]. When not using humans, the main struggle when evaluating feature attribution methods is the lack of a ground truth; therefore, this is often 'created' in research [7].

Zhou, Booth, Ribeiro, *et al.* [7] solved the problem of a lacking ground truth by modifying the dataset they used.

They advocate that their results show that using a rationale model, which is a model with a selector and a predictor part, does not guarantee that the results from the selector help understand the model [7]. Something similar was done by Yang and Kim [20]. They developed a framework called BAM, including a unique BAM dataset. Next to that, they included models and metrics in their framework [20]. Both Zhou, Booth, Ribeiro, *et al.* [7] and Yang and Kim [20] focused on image datasets.

Another way of evaluating feature attribution methods is with a human-grounded truth. An example of research with humans is the research from Dieber and Kirrane [21]. They did interviews with six people, and they asked questions about the usability of LIME. The results showed that the results of LIME were hard to understand, but a machine learning background helped [21]. The authors of LIME held a human study, too, and participants had a lot less trust in the 'poor' models after seeing the explanations from LIME [4]. According to Zhou, Booth, Ribeiro, *et al.* [7], using rationale annotation is a bad idea because a human and a model can select the same information but still handle this information differently. An example of research that did this is Bastings, Aziz, and Titov [22], in which they used human rationale annotation to evaluate their rationale model.

Our research includes a pseudo ground truth and a human-grounded evaluation to ensure that both types are investigated. Both offer different capabilities, so doing both will improve the comparison of the feature attribution methods.

4 Methodology

In this section, we will describe the methodology in three subsections. First, in subsection 4.1, the selected feature attribution methods will be mentioned and described briefly. Then in subsection 4.2, the chosen model will be discussed. The last subsection 4.3 entails a description of the evaluation.

4.1 Selected Feature Attribution Methods

The following three feature attribution methods were chosen for the comparison:

- **LIME** [4] is a feature attribution method used for the comparison. It is locally faithful and model-agnostic, reflected in its name: "Local Interpretable Model-agnostic Explanations" [4]. It is perturbation-based, meaning in the case of textual input, it will change the input by removing or covering tokens and looking at the potential change in output [4]. It will try to minimize the weighted square loss, $\mathcal{L}(f, g, \pi_x)$, to make sure it meets its goal of interpretability and local fidelity [4]; see the equations 1 and 2 below:

$$\xi(x) = \underset{g \in G}{\arg\max}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z^{'} \in Z} (f(z) - g(z^{'}))^2 \tag{2}$$

  as seen in [4].

- **Kernel SHAP** [5] is a locally faithful and model-agnostic method based on Shapley value, a game theory. Shapley values are used to determine the contribution of individual members to a game's win, and Kernel SHAP is a combination of LIME and Shapley values [5]. Computing Shapley values takes a lot of time; therefore, combining it with LIME makes it possible to approximate Shapley values and reduce the time [5]. Equation 2 is adapted to equation 3:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z^{'} \in Z} [f(h_x^{-1}(z^{'})) - g(z^{'})]^2\ \pi_{x^{'}}(z^{'}) \tag{3}$$

  as seen in [5].

- **Integrated Gradients** [6] is a gradient-based method based on two axioms: *Sensitivity* and *Implementation Invariance*. The method approximates "the integral of integrated gradients" [6]. Sundararajan, Taly, and Yan [6] do this by summing the gradients from several points close to the baseline (benchmark). This benchmark needs to be selected before calculating the gradients; often, this baseline consists of all-zeros [6]. Integrated Gradients is model-specific because it needs the model's internal parameters to compute the gradients [6]. It calls the gradient operation several times [6]. For approximating "the integral of integrated gradients," it uses equation 4:

$$\begin{aligned} &IG_i^{approx}(x) ::= \\ &(x_i - x_i^{'}) * \textstyle\sum_{k=1}^{m} \frac{\partial F(x^{'} + \frac{k}{m} * (x - x^{'}))}{\partial x_i} * \frac{1}{m} \end{aligned} \tag{4}$$

  as seen in [6].

4.2 Model

The model selected for our research is ExPred. ExPred is an NLP architecture that can be used for fact-checking. In our case, it was trained with the FEVER dataset from ERASER benchmark [8], [14], [23]. ExPred is an *interpretable-by-design* model; it gives an explanation with it is classification [9]. In this way, it differs from post-hoc methods, which get their explanation after the model was trained [1]. To make it an *interpretable-by-design* model, the authors made ExPred entail two parts: a selector/explanation and a predictor part; see Figure 1. The input in the model first gets mapped to an "explanation mask" by the explanation generation part [9]. Then, this mask gets mapped to a prediction by the prediction network [9].

The explanation generation network uses a shared encoder architecture, where it uses BERT as an encoder [9], see Figure 1. The choice for BERT leads to a maximum length of 512 tokens [9]. There are two decoders in the explanation generation network, the *Auxiliary Task Prediction* and the *Extractive Explanation Generation* [9]. The accuracy of the *Auxiliary Task Prediction* and the *Extractive Explanation Generation* are both parts of the loss function, which is used for training the explanation generation network [9]. Only the *Extractive Explanation Generation* is part of the output of the entire model [9], which Figure 1 visualizes. The *Extractive Explanation Generation* output consists of boolean values, indicating if the word was essential or not according to ExPred [9].

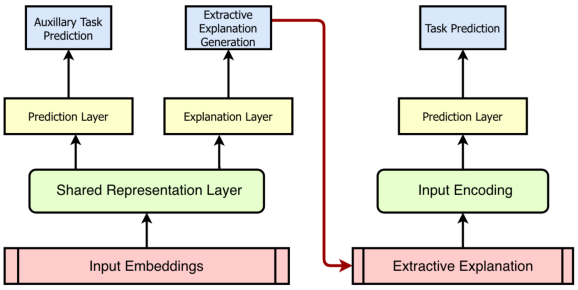

Figure 1: A visualization of ExPred, with the Explanation Generation on the left and the Task Prediction on the right [9]. The green box on the left is the shared encoder [9].

This output is used in our research to compare with the results of the post-hoc feature attribution methods.

The prediction network gets a masked version of the input, where the wildcard '.' is used to mask tokens [9]. For the network, BERT was also used in the second part of the model [9]. The result of the *Task Prediction* is part of the final result of the model combined with the *Extractive Explanation Generation* [9], see Figure 1.

4.3 Evaluation

To answer the research question (RQ), we chose to compare three feature attribution methods that are popular: LIME [4], Kernel SHAP [5], and Integrated Gradients [6]. These were evaluated in two ways: automatic (subsection 4.3.1) and manual (subsection 4.3.2).

4.3.1 Automatic Comparison

The use of ExPred enabled us to compare its explanation with the results from the feature attribution methods [7]. The explanations from ExPred were used as a pseudo ground truth. The feature attribution methods' top-n highest-scoring tokens were compared with the ExPreds explanation. Next to this comparison, the feature attribution methods were compared with each other by translating scores to ranks. The rank correlations [24] between the explanations of the methods were computed. The results of this automated pseudo ground truth evaluation, with ExPred and each other, gave us **similarity scores** (for SQ1).

4.3.2 Manual Comparison

On the other hand, more than the automatic evaluation would be needed to compare the feature attribution methods thoroughly. Hence we conducted a small user study in the form of a survey (for SQ2). This user study entailed two parts: (1) questions about individual explanations from the feature attribution methods and (2) ranking the three feature attribution methods. In the survey, these explanations were combinations of the claim, the colorized context, and the prediction (see Figure 2). The survey questions' focus was not on how the explanation was presented but on the information in this explanation. The first part's questions were about the **understandability** of the explanation, the **insightfulness** of why the prediction was made, and if the participant would **rely on the model** in the future. These were 5-point Likert scale questions; see Figure 7 in Appendix A. In the second part, three explanations, from every method one, needed to be ranked on **insightfulness**; see Figure 8 in Appendix A.

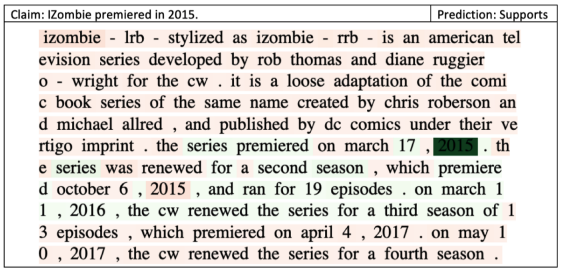

Figure 2: The figure shows how the explanations were presented in the survey (the user study). The claim and prediction are offered at the top. At the bottom, a heatmap on the context is presented.

5 Experimental Setup

The experimental setup consists of three main parts. First, we will discuss how the feature attribution methods from the Captum library[1] were implemented (subsection 5.1), followed by the dataset used (subsection 5.2). Lastly, the setup and the decisions made for the evaluation are described (subsection 5.3).

5.1 Implementation

Before the feature attribution methods could be evaluated, they had to be implemented. For this implementation, the Captum library[1] was used with Python version 3.8.

For the implementation of the perturbation-based feature attribution methods, LIME [4] and Kernel SHAP [5], we implemented a wrapper to ensure the in- and output from Captum[1] and ExPred were compatible [9]. Additionally, we used the wildcard '.' from ExPred for the padding or masking of the words by the feature attribution methods instead of the pad token that Captum[1] uses [9]. The claim remained untouched in the masking by feature attribution method. Only the context went into the feature attribution method. Therefore one input of ExPred changed during the perturbations.

To implement the gradient-based feature attribution method, Integrated Gradients [6], only the predictor part of ExPred was used by the method [9]. We used the output of the selector part of ExPred as an input for the Integrated Gradients. In addition, the maximum length of 512 that ExPred had for the input of the predictor part was taken into account [9]. Two versions were implemented for Integrated Gradients: (1) the claim and context as input, and (2) only the context as input. The latter was implemented for comparison with the perturbation-based methods.

After implementing the feature attribution methods using Captum[1], we implemented a script to run the feature attribution methods on multiple instances automatically and in parallel. LIME and Kernel SHAP had 300 iterations for each instance, and Integrated Gradients had 50.

[1] https://github.com/pytorch/captum

5.2 The Dataset

The authors of ExPred trained their model with the FEVER dataset from ERASER benchmark [8], [14], [23]; therefore, we used this dataset as well. FEVER consists of claims and contexts to train a model and is a 'big' dataset compared to others; see [13]. It uses Wikipedia pages for its context [14]. For every claim in FEVER, there are three options: (1) supported, (2) refuted, and (3) not enough info. In total, there are 185,445 claims, according to Thorne, Vlachos, Christodoulopoulos, *et al.* [14]. A problem with FEVER is that it is pretty unbalanced according to Zeng, Abumansour, and Zubiaga [13]. In the training set, there are around 80,000 claims labeled true, 30,000 claims labeled false, and 35,000 claims labeled not enough information [14]. For the ERASER benchmark, only the supported and refuted claims are kept in the dataset [23].

In our research, 100 claims of the test set of the ERASER benchmark FEVER dataset [23] were run, and 100 claims of the train set. For the test set, the division was 35 supported claims and 65 refuted claims. For the train set, the division was 67 supported claims and 33 refuted claims.

5.3 Pseudo Ground Truth and Human-grounded Evaluation

The evaluation consisted of two parts pseudo ground truth (SQ1) and human-grounded (SQ2). For the pseudo ground truth evaluation (SQ1), we compared the results from LIME [4], Kernel SHAP [5], and Integrated Gradients [6] with each other and with the explanation from ExPred [9]. We used Kendall's Tau [24] for the comparison between the feature attribution methods and Jaccard distance (dissimilarity) [25] for the comparison with ExPreds explanation. Both were implemented using the SciPy library[2]. For the comparison between the feature attribution methods, the lengths of the lists with the tokens were limited to the size of the Integrated Gradients result when being compared to Integrated Gradients. For the comparison with ExPred, all methods were limited to the length of Integrated Gradients.

For the human-grounded evaluation (SQ2), we conducted a small user study with 19 participants in the Netherlands. The participants were found through convenience sampling; our personal networks were used. 19 participants reached the end of the survey, where 17 out of 19 answered all questions. The participants were students from the Technical University of Delft or employees from ICT companies. The participants were selected on having a technical background and understanding of the English language. They had to fill in an online survey, which consisted of twenty questions: fifteen were about rating the explanations from feature attribution methods, and five were about ranking the explanations. Twenty different instances from the test dataset were used for these twenty questions.

6 Results

In this section, we will present the results found from the pseudo ground truth and human-grounded evaluation. In subsection 6.1, the pseudo ground truth evaluation results will be discussed, and the discussion of these results is separated into the test and train set. In the following subsection 6.2, the results from the human-grounded evaluation are described in the two parts of the user study: the rating and the ranking of the explanations of the feature attribution methods.

[2]https://scipy.org

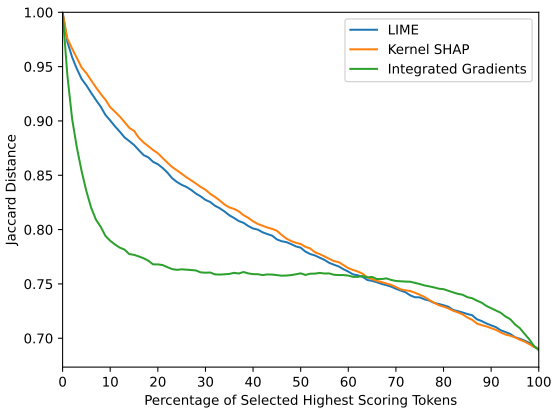

Figure 3: Jaccard Distance between the feature attribution methods percentage of selected highest scoring tokens and the explanation from ExPred with 100 instances from the **test set**. Close to 0 means similar, and close to 1 means dissimilar.

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

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

6.1 Pseudo Ground Truth Results

6.1.1 Test Set

The pseudo ground truth evaluation (part of SQ1) for the test set, data is divided into two parts the comparison between the methods and the comparison with ExPred, see Table 1 and Figure 3.

From the results of Kendall's Tau, it seems that all methods are a little bit similar to each other. Close to -1 would mean the methods are dissimilar, and close to 1 that the methods are similar. Kendall's Tau for LIME and Kernel SHAP is 0.0545, for LIME and Integrated Gradients 0.0444, and Kernel SHAP and Integrated Gradients 0.0207 (see Table 1). Meanwhile, these Kendall's Tau's results are not statistically significant ($p = 0.385$, $p = 0.398$, $p = 0.366$, respectively) because all p-values are above 0.05. Therefore, we cannot reject the null hypothesis of an absence of association.

The results from Jaccard Distance, when comparing the individual methods with ExPreds explanations, show that none of the methods is very similar to ExPreds explanations. In Figure 3, the Jaccard Distance over the percentage of selected tokens is plotted. When this percentage increases, the Jaccard

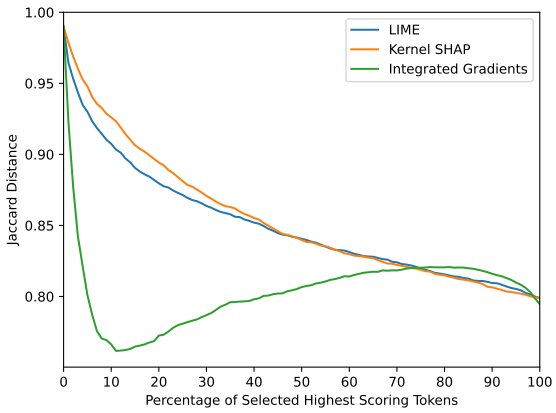

Figure 4: Jaccard Distance between the feature attribution methods percentage of selected highest scoring tokens and the explanation from ExPred with 100 instances from the **train set**. Close to 0 means similar, and close to 1 means dissimilar.

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

Table 2: Kendall's Tau for two feature attribution methods looks at 100 instances from the **train set**. For Kendall's Tau: close to 1 means the two methods are similar, and -1 means the two methods are dissimilar.

Distance decreases, and more tokens are selected, so more overlap with ExPreds explanations, but it is never close to 0 (very similar). It steadily decreases for LIME and Kernel SHAP; this is not the case for Integrated Gradients. Integrated Gradients decreases faster at first (become more similar to ExPreds explanation) and then stagnate.

In short for the **test set**, the results from Kendall's Tau for comparing the methods without ExPred for the test set show that they are a little bit similar but this result is not statistically significant. Therefore, we cannot reject the null hypothesis of an absence of association. The comparison with ExPred for the test set shows that the method Integrated Gradients gives proportionally higher scores to tokens in ExPreds explanation than LIME and Kernel SHAP.

6.1.2 Train Set

The pseudo ground truth evaluation (part of SQ1) for the train set, data consists of two parts: the comparison between the methods and the comparison with ExPred, like with the test set, see Table 2 and Figure 4.

The results from Kendall's Tau seem to indicate that all methods are a little bit similar. The Kendall's Tau for LIME and Kernel SHAP is 0.0204, for LIME and Integrated Gradients 0.0281, and Kernel SHAP and Integrated Gradients 0.0101 (see Table 2). However, the results from Kendall's Tau are not statistically significant ($p = 0.456$, $p = 0.410$, $p = 0.503$, respectively) because all p-values are above 0.05.

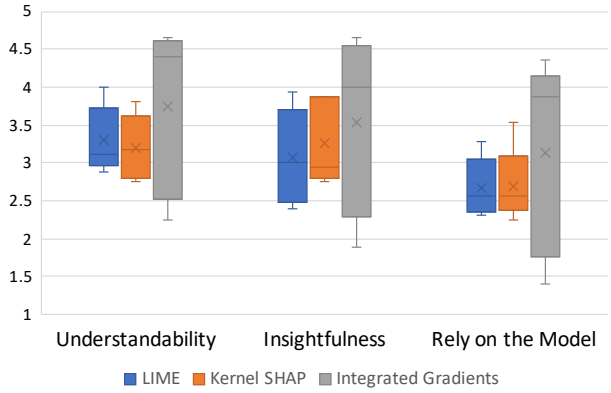

Figure 5: The results of the **rating of the explanations of the feature attribution methods** on understandability, insightfulness, and relying on the model. The y-axis is a 5-point Likert scale, ranging from "1 - strongly disagree" to "5 - strongly agree". In the boxes, the crosses are the means, and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

Therefore, we cannot reject the null hypothesis of an absence of association.

The results from Jaccard Distance, for the comparison of the individual methods with the explanation of ExPred, show that none of the methods' explanations is very similar to ExPreds explanations. In Figure 4, the Jaccard Distance over the percentage of selected tokens is plotted. The Jaccard Distance decreases when the percentage of chosen tokens increases because there is more overlap with ExPreds explanation, but it is never near 0 (very similar). It steadily decreases for LIME and Kernel SHAP. For Integrated Gradients, the dissimilarity decreases fast at the start and then stagnates, even increasing slightly.

In summary for the **train set**, the result of the comparison without ExPred that the feature attribution methods are a little bit similar to each other is not statistically significant, according to Kendall's Tau. Therefore the null hypothesis of an absence of association cannot be rejected. The comparison with ExPred shows that the method Integrated Gradients gives proportionally higher scores than LIME and Kernel SHAP to tokens in ExPreds explanation.

6.2 Human-grounded Results

6.2.1 Rating Part

In the rating section of the user study (part of SQ2), participants were asked to rate five explanations from all three feature attribution methods on a 5-point Likert scale (1 - strongly disagree to 5 - strongly agree); see Figure 2 for how the explanation was presented or the more detailed Figure 7 in Appendix A. They rated the explanation on if it was understandable, how much insight it gave, and if they would rely on the model.

The results in Figure 5 show LIME and Kernel SHAP lie close to each other with their results. The medians for both are very similar for all three categories and close to "3 - neither agree or disagree"; see Figure 5. For Integrated Gradients, the result is different; there is more variation among

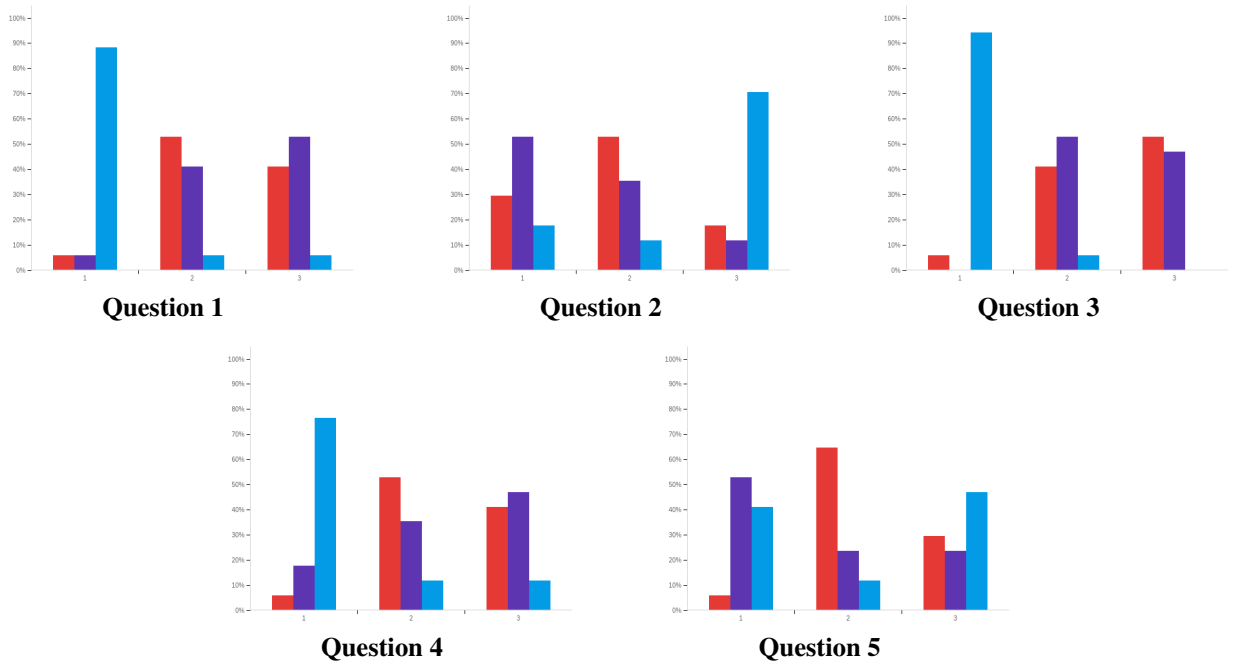

Figure 6: The results for the questions related to **ranking the explanations of the feature attribution methods** on how much insight they gave. The x-axis is the ranking 1 up to 3, and the y-axis is the normalized count of the choice in percentages. Red = LIME; Purple = Kernel SHAP; Blue = Integrated Gradients.

the ratings, see Figure 5. Additionally, the median in all categories is a lot higher. A potential explanation is the number of iterations for LIME and Kernel SHAP. The lengths of the contexts differed for both, and the participants seemed to have stronger opinions about more concise contexts; the colors were more intense here.

To conclude, the results from Integrated Gradients have more variation than LIME and Kernel SHAP. The median and mean from Integrated Gradients are higher than LIME and Kernel SHAP, see Figure 5.

6.2.2 Ranking Part

In the second part of the user study (part of SQ2), we wanted to see if the participants preferred one of the feature attribution methods when asked to rank them on how much insight their explanations gave. They were asked five times to rank the methods without knowing which; see Figure 8 in Appendix A.

The results show Integrated Gradients three times as the clear winner; it was ranked first for the first, third, and fourth questions, see Figure 6. It had the most votes for the last place for questions two and five. The second and fifth questions had a shorter context, indicating that the number of iterations (300) was not enough for contexts with more text, which could explain the difference. The results of LIME and Kernel SHAP for the ranking lie close to each other in most of the questions, as shown in Figure 6.

Lastly, Integrated Gradients won three out of five times. Besides this, the rankings results for LIME and Kernel SHAP lay close to each other, and when the context was shorter, they were ranked better.

7 Responsible Research

This section will discuss and reflect on the ethical aspects of this research. In subsection 7.1, the ethical aspects of human participants are discussed. In the following subsection 7.2, the accessibility of feature attribution methods' results are discussed.

Due to the importance of reproducibility in research, the code from the feature attribution methods and the evaluation is made available on GitHub, including the explanation on how to run it: https://github.com/Herbje/Evaluating-Feature-Attribution-Methods.

7.1 Human Participants

Our study involved human participants, which makes it essential for our research to be responsible. Our research falls under an approval from the Human Resource Ethics Committee of the Technical University of Delft. In the Code of Conduct, *responsibility* is described as: "acknowledging the fact that a researcher does not operate in isolation and hence taking into consideration – within reasonable limits – the legitimate interests of human and animal test subjects" [26]. Therefore we took several measures to mitigate potential risks.

First, all survey responses in Qualtrics[3] were anonymized entirely; we stored no identifiable information. Additionally, we did not ask for or store personal data from the participants. We were not interested in looking at the differences between humans, which made it unnecessary to ask questions about age or gender, for example.

[3]https://www.qualtrics.com

Secondly, the participants were informed via a consent form at the start of the survey. This consent form stated that they could stop at any point during the survey by closing the tab or window. Additionally, they were informed about the risk of a data breach and what was done to minimize this risk (anonymizing the data). They also needed to confirm they were above 18 years old. Besides this, all respondents lived in the Netherlands. The invitations were sent to both people studying and working in the technical field to make the group of participants a bit diverse and to prevent a bias (not only students or people working in the technical field).

7.2 Color Blindness Accessibility

The Web Content Accessibility Guidelines [27, Success Criterion 1.4.1] stresses the importance of the "use of color" for accessibility. During the user study, we also got the feedback from some participants that the colors were hard to distinguish; see Figure 2 for an example. Color blindness is quite common in the human population, especially among men [28]. The most common color blindness is red-green color blindness, which 8% of the Northern European men have [28]. Red-green color blindness makes it hard to distinguish the difference between red and green due to the absence of retinal photoreceptors for these colors [28].

The heatmaps in our experiment used the colors green and red; see Figure 2. These colors seemed intuitive for positive and negative influence scores from the feature attribution methods, but for the accessibility, these colors are a problem [28]. Therefore to mitigate the issue of color blindness with presenting feature attribution methods' explanations, different colors than red and green could be chosen in future research. Wong [28] says that changing red to magenta helps and changing green to turquoise. For future research, magenta and turquoise could be used, and it can be tested if these colors help people with color blindness.

8 Discussion

The goal of this research was to compare three post-hoc feature attribution methods in the context of fact-checking with each other (RQ): LIME [4], Kernel SHAP [5], and Integrated Gradients [6]. This comparison was made in two ways a pseudo ground truth evaluation and a human-grounded evaluation. We used ExPred [9] as a fact-checking model for applying the feature attribution methods.

Our results for the **pseudo ground truth evaluation with ExPred (SQ1)** show that Integrated Gradients performs a bit better at the start when comparing to the explanations from ExPred for the test and train set; see Figures 3 and 4. Before 10%, the dissimilarity to the explanation with ExPred drops rapidly. LIME and Kernel SHAP decline as well but a lot slower. On the other hand, at no point did the feature attribution methods have a low dissimilarity score with the explanation from ExPred, as seen in Figures 3 and 4. Thus, Integrated Gradients gives tokens that overlap with ExPreds explanation proportionally higher scores than LIME and Kernel SHAP.

The results for the **pseudo ground truth evaluation without ExPred (SQ1)** show that the explanations of the methods are a tiny bit similar to one another, but this result is not significant; see Tables 1 and 2. Therefore it is not possible to draw definite conclusions from these numbers. A bigger dataset could potentially improve these results, but due to time constraints, this was not possible.

The **human-grounded evaluation (SQ2)** results show that overall the feature attribution methods helped to give some insight into why the model made its decision, see Figure 5. For Integrated Gradients, the ratings and rankings fluctuated much more than for LIME or Kernel SHAP, and participants agreed or disagreed more strongly. When looking at the explanations presented to the participants, we noticed that the explanations with the shorter contexts made it easier for participants to form an opinion about the explanation. LIME and Kernel SHAP had intenser colors in these cases. LIME and Kernel SHAP also became more of a competition for Integrated Gradients in these cases. The reason for this could be the number of iterations (300) chosen for LIME and Kernel SHAP. Potentially, it was not high enough for all claims and contexts. This number of iterations was selected due to constraints such as time and computational power; with 300 iterations, it took around half an hour to run a single instance.

For this research, we relied heavily on ExPred [9]. ExPred does not obtain a 100% classification accuracy. Potential misclassifications or other problems in Expred could lead to a mistake in our results. Additionally, in the comparison with ExPreds explanation, we looked at the *Extractive Explanation Generation* result [9]. The feature attribution methods looked at the *Task Prediction* [9], see Figure 1. This means that if ExPred handles its explanation wrongly in the prediction network, the explanation would be less helpful [9]. Zhou, Booth, Ribeiro, *et al.* [7] give the example of focusing a lot on "grammatical idiosyncrasy" instead of meaning in their paper. The results of our research can not be generalized immediately to all other models and tasks; the steps of this research need to be conducted on other tasks and models.

9 Conclusions and Future Work

This research aimed to evaluate post-hoc feature attribution methods in the context of fact-checking. Feature attribution methods are part of XAI and give scores to tokens indicating their importance to the decision of a model [7]. For this evaluation, three popular feature attribution methods were selected: LIME [4], Kernel SHAP [5], and Integrated Gradients [6]. Evaluating feature attribution methods is often a struggle due to the lack of a ground truth. The NLP architecture ExPred [9] was used as the model after being trained with the FEVER dataset from ERASER benchmark [8], [14], [23]. Due to that ExPred is an *interpretable-by-design*, its explanations were used as a pseudo ground truth in our evaluation. Additionally, we did a small user study for a human-grounded evaluation.

The results of the pseudo ground truth evaluation (SQ1) show that Integrated Gradients gives proportionally higher scores to tokens that are in the explanation of ExPred and that the three feature attribution methods show no consensus in their explanations, see section 6. In the human-grounded evaluation (SQ2), the participants seemed more opinionated regarding the explanations from Integrated Gradients; the

scores the participants gave differed more per question. The results from both evaluations (RQ) indicate that Integrated Gradients performs better and enable a human to form an opinion better. On the other hand, these results could also mean that the number of iterations of LIME and Kernel SHAP was too low for some instances.

For future research, it would be beneficial to look at other models and tasks to see if similar results could be found there. Additionally, more iterations and instances could be run, but because of the time constraint, this was not possible in this research.

References

[1] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," *CoRR*, vol. abs/2101.09429, 2021. arXiv: 2101.09429.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805.

[3] Intersoft Consulting, *General Data Protection Regulation (GDPR): Recital 71 - Profiling*. [Online]. Available: https://gdpr-info.eu/recitals/no-71/ (visited on 11/20/2022).

[4] 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.

[5] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017. arXiv: 1705.07874.

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *CoRR*, vol. abs/1703.01365, 2017. arXiv: 1703.01365.

[7] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" *CoRR*, vol. abs/2104.14403, 2021. arXiv: 2104.14403.

[8] Z. Zhang and M. Reimer, *ExPred*, original-date: 2020-12-27T13:01:03Z, Jul. 2022. [Online]. Available: https://github.com/JoshuaGhost/expred (visited on 11/19/2022).

[9] 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.

[10] L. C. McIntyre, *Post-truth* (The MIT Press Essential Knowledge Series). Cambridge, Massachusetts: The MIT Press, 2018.

[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, *Fact Extraction and VERification*, original-date: 2018-03-09T20:12:11Z, Oct. 2022. [Online]. Available: https://github.com/awslabs/fever (visited on 11/19/2022).

[12] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," 2018. DOI: 10.48550/ARXIV.1811.10154.

[13] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, e12438, 2021. DOI: 10.1111/lnc3.12438.

[14] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for Fact Extraction and VERification," *CoRR*, vol. abs/1803.05355, 2018. arXiv: 1803.05355.

[15] D. Wadden, K. Lo, L. L. Wang, *et al.*, "Fact or fiction: Verifying scientific claims," *CoRR*, vol. abs/2004.14974, 2020. arXiv: 2004.14974.

[16] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, "A richly annotated corpus for different tasks in automated fact-checking," *CoRR*, vol. abs/1911.01214, 2019. arXiv: 1911.01214.

[17] G. Lim and S. T. Perrault, "Explanation preferences in XAI Fact-Checkers," in *European Society for Socially Embedded Technologies*, ser. 2, vol. 6, Coimbra, Portugal: European Society for Socially Embedded Technologies (EUSSET), 2022. DOI: 10.48340/ECSCW2022_P02.

[18] R. Linder, S. Mohseni, F. Yang, S. K. Pentyala, E. D. Ragan, and X. B. Hu, "How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking," *Applied AI Letters*, vol. 2, no. 4, e49, 2021. DOI: 10.1002/ail2.49.

[19] K. Abhishek and D. Kamath, *Attribution-based XAI methods in Computer Vision: A review*, 2022. DOI: 10.48550/ARXIV.2211.14736.

[20] M. Yang and B. Kim, "Benchmarking attribution methods with relative feature importance," *CoRR*, vol. abs/1907.09701, 2019. arXiv: 1907.09701.

[21] J. Dieber and S. Kirrane, "Why model why? Assessing the strengths and limitations of LIME," *CoRR*, vol. abs/2012.00093, 2020. arXiv: 2012.00093.

[22] J. Bastings, W. Aziz, and I. Titov, "Interpretable neural predictions with differentiable binary variables," *CoRR*, vol. abs/1905.08160, 2019. arXiv: 1905.08160.

[23] J. DeYoung, S. Jain, N. F. Rajani, *et al.*, "ERASER: A benchmark to evaluate rationalized NLP models," *CoRR*, vol. abs/1911.03429, 2019. arXiv: 1911.03429.

[24] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938, Publisher: [Oxford University Press, Biometrika Trust]. DOI: 10.2307/2332226. (visited on 01/08/2023).

[25] P. Jaccard, "The Distribution of the Flora in the Alpine Zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912, ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

[26] KNAW, NFU, NWO, TO2-Federatie, V. Hogescholen, and VSNU, *Nederlandse gedragscode wetenschappelijke integriteit*. Data Archiving and Networked Services (DANS), 2018. DOI: 10.17026/DANS-2CJ-NVWU.

[27] A. Kirkpatrick, J. O Conner, A. Campbell, and M. Cooper, Eds., *Web Content Accessibility Guidelines (WCAG) 2.1*, 2018. [Online]. Available: https://www.w3.org/TR/2018/REC-WCAG21-20180605/#use-of-color (visited on 01/24/2023).

[28] B. Wong, "Points of view: Color blindness," *Nature Methods*, vol. 8, no. 6, pp. 441–441, Jun. 2011. DOI: 10.1038/nmeth.1618.

A Figures

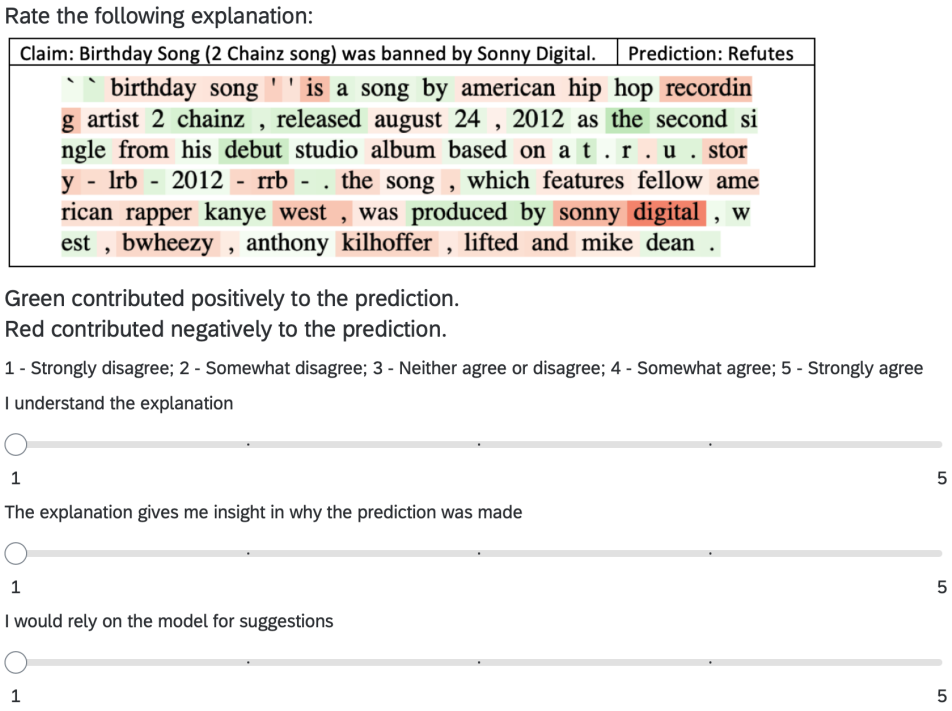

Figure 7: An example of a question in the user study. In this part of the user study, the participant was asked to **rate** an explanation from a feature attribution method.

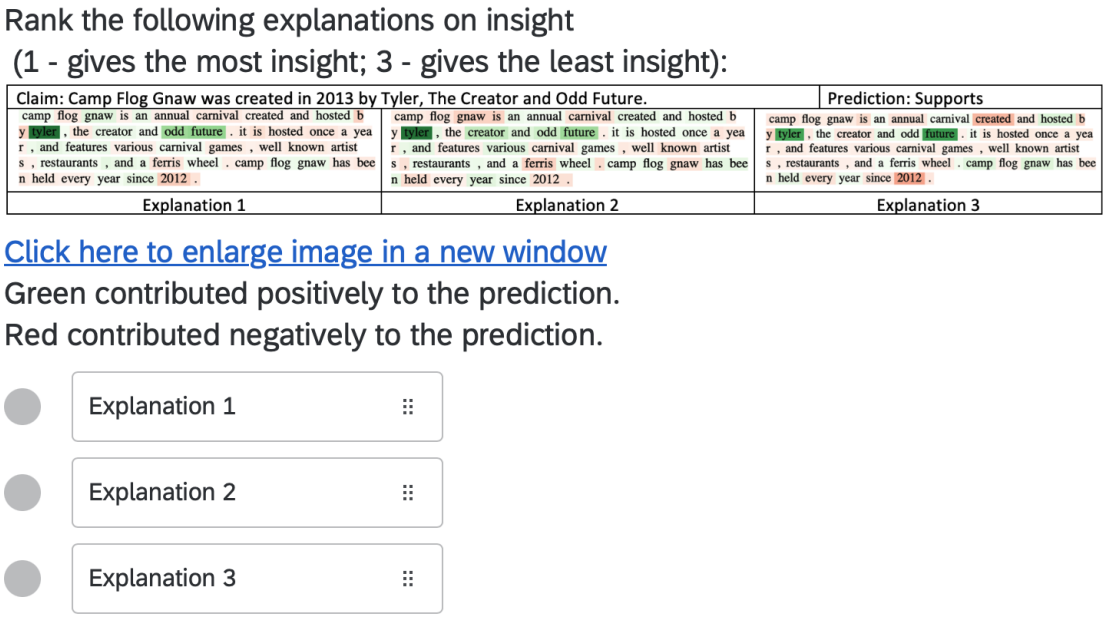

Figure 8: An example of a question in the user study. In this part of the user study, the participant was asked to **rank** the explanations from the feature attribution methods.