How do different explanation presentation strategies of feature and data attribution techniques affect non-expert understanding?

Explaining Deep Learning models for Fact-Checking

Shivani Singh

Supervisor(s): Avishek Anand, Lorenzo Corti, Lijun Lyu

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 29, 2023

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

Abstract

The goal of this paper is to examine how different presentation strategies of Explanainable Artificial Intelligence (XAI) explanation methods for textual data affect non-expert understanding in the context of fact-checking. The importance of understanding the decision of an Artificial Intelligence (AI) in human-AI interaction and the need for effective explanation methods to improve trust in AI models are highlighted. The study focuses on three explanation methods: interpretable-by-design model ExPred, feature attribution technique LIME, and instance attribution method k-NN. Two presentation strategies were compared for each method, and participants were presented with a set of claims and asked to indicate their understanding and level of agreement with the AI's classification. The main hypothesis is that participants will appreciate all available context and details, as long as it is presented in a structured way, and will find visual representations of data easier to understand than textual ones. Results from the study indicate that participants prefer explanations that are simple and structured, and that visual presentations are not as effective, especially when it is the first time a user interacts with this type of data. Additionally, it was found that better formatting leads to a better-calibrated understanding of the explanation. The results of this study will provide valuable insight into the best way to present XAI explanations to non-experts to enhance their understanding and reduce the deployment risk associated with Natural Language Processing (NLP) models for automated fact-checking. The study's code, data, and Figma templates are publicly available for reproducibility.

Introduction

The increase in use of Artificial Intelligence (AI) in daily life has brought attention to the issue of data bias and the risks that come with it. Data bias can reduce the accuracy of AI models, which is particularly important in the context of fact-checking [1]. Here the goal is to detect misinformation and provide a credibility check for online information using AI and Natural Language Processing (NLP) models. State-of-the-art fact-checking algorithms are trained to give one of three answers: "Support", "Refute" and "Not enough info". However, as the target user group for these tools now includes non-experts, i.e. someone is familiar with AI but not able to develop such fact-checking AI, it is crucial to consider the potential for data bias in these fact checking tools. Now there is a need for explanations that are not only accurate but equally important - understandable for the user. This is the focus of the field Explainable AI (XAI), which aims to demystify the "black box" of AI models and increase trust in them [2]. It does not only serve to protect individuals from misinformation but also holds potential relevance for a wide range of industries, from the more obvious fields of software engineering and content creation to less apparent areas such as medical testing and diagnosis or the agricultural sector [3].

In this study, we focus on three main types of XAI explanation methods: *interpretable-by-design models, feature attribution techniques*, and *instance attribution methods* - and do not include *counterfactual explanations* or *sequence mining techniques* in our analysis. A comprehensive overview of the outputs of the most common types of XAI methods can be seen in the accompanying Figure 1. The output of these methods must be tailored to human comprehension in order to be effective. Even if an explanation is very detailed and accurate, whether the human understands it, greatly influences the model performance [4]. While there is previous research that has shown that explanations that utilize text and provide a high degree of detail are often preferred by users [5], our study is unique in that it focuses specifically on making conclusions about optimal presentations within the context of different explanation methods and fact-checking.

In this study, we aim to investigate:

RQ1 How do different explanation presentation strategies of feature and data attribution techniques affect non-expert understanding?

To accomplish this, we explore a range of sub-questions, including:

RQ2 What is the preferred level of context/details in explanations?

RQ3 How significant is the data presented in terms of comprehension?

RQ4 Does the use of visual presentations enhance or hinder understanding compared to textual explanations?

RQ5 In what ways would the participant suggest improving the presentation?

Through answering these questions, we aim to gain a deeper understanding of how to most effectively present XAI explanations to non-experts.

The study found that participants generally preferred explanations that were simple and structured, and that visual presentations were not as effective, particularly for first-time users of this type of data. It was also discovered that proper formatting improved participants' understanding and comprehension of the explanation. For the sake of reproducibility the code[1] and data[2] used for the experiments, as well as the Figma templates[3] for the prototypes and future recommendations are publicly available.

1 Problem Description

The main problem addressed in this study is understanding the best way to present XAI explanations to non-experts. This way we aim to enhance their understanding and reduce the deployment risk associated with NLP models for automated fact-checking. We propose to compare different presentation strategies for specific explanation methods, including the binary selector output of the interpretable-by-design

[1] www.github.com/shivanisgithub/FactCheckingPresentations

[2] https://fever.ai/dataset/fever.html

[3] www.figma.com/community/file/1194688306111467424

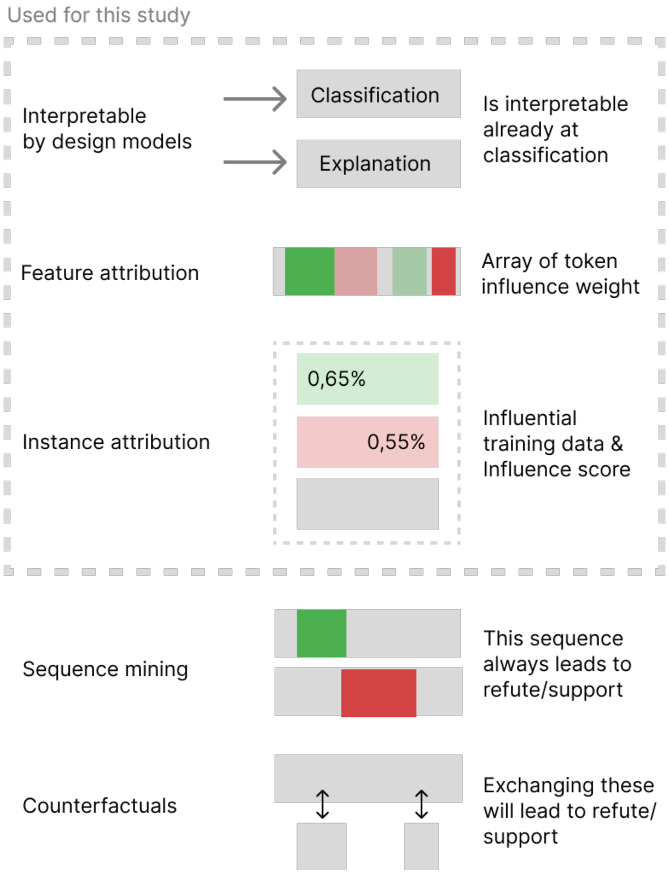

Used for this study
Interpretable by design models
Classification
Explanation
Is interpretable already at classification
Feature attribution
Array of token influence weight
Instance attribution
0,65%
0,55%
Influential training data & Influence score
Sequence mining
This sequence always leads to refute/support
Counterfactuals
Exchanging these will lead to refute/ support

Figure 1: Outputs of different XAI explanation methods

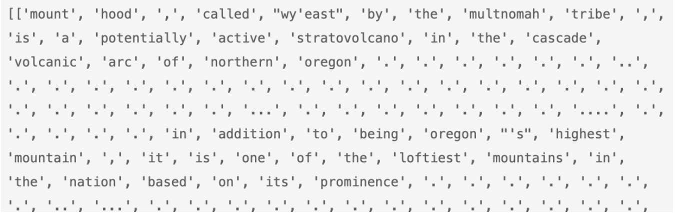
model Explain-then-Predict (ExPred) [6], the feature attribution technique Local Interpretable Model-Agnostic Explanations (LIME) [7], and the instance attribution method k-Nearest-Neighbour (k-NN) [8]. By combining existing research on XAI explanation methods with research on human-AI interaction, we aim to provide valuable insight into the optimal way to present XAI explanations to non-experts and offer recommendations that can be implemented immediately. The main hypothesis is that participants will appreciate all available context and details, as long as it is presented in a structured way, and will find visual representations of data to be easier to understand than textual ones.

2 Related Work

As necessary background knowledge for our research, we present the explanation methods that serve as the foundation for our study. Some related works include articles pertaining to user studies of explanation methods, while others focus on utilising human evaluation to improve AI models. We then showcase a broader array of works that address theories on human understanding of the text and data-driven approaches.

2.1 Explanation Methods Used in the Study

In the context of this study, we employ interpretable-by-design models, which are designed with interpretability in mind, meaning their internal workings and decision-making processes are easily understandable by humans. This allows for greater transparency of the model's predictions and decisions. The specific model we use is ExPred, as presented in the work of Zhang et al. [6], which utilizes multi-task learning in the generation of explanations for explain-then-predict models. Specifically, a binary array of all tokens in a given context is generated, with the subset of readable tokens contributing to the classification label. In Figure 2 a decoded example of this is shown. The readable tokens influenced the classification, and the dots represent non-influential data.



```
[['mount', 'hood', ',', 'called', "wy'east", 'by', 'the', 'multnomah', 'tribe', ',',
'is', 'a', 'potentially', 'active', 'stratovolcano', 'in', 'the', 'cascade',
'volcanic', 'arc', 'of', 'northern', 'oregon', '.', '.', '.', '.', '.', '.', '..',
'.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.',
'.', '.', '.', '.', '.', '.', '...', '.', '.', '.', '.', '.', '.', '.', '....', '.',
'.', '.', '.', '.', 'in', 'addition', 'to', 'being', 'oregon', "'s", 'highest',
'mountain', ',', 'it', 'is', 'one', 'of', 'the', 'loftiest', 'mountains', 'in',
'the', 'nation', 'based', 'on', 'its', 'prominence', '.', '.', '.', '.', '.', '.',
'.', '..', '...', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.', '.',
```

Figure 2: A truncated frame of the output generated by ExPred for the claim "Mount Hood is in the Andes".

In addition, we also employ feature attribution techniques, where each token in the context is assigned an influence score or weight towards a classification label. This study uses the LIME algorithm [7], which generates explanations for predictions made by classifiers or regressors. It does this by approximating the model locally with an interpretable one that can provide faithful explanations. Other popular feature attribution techniques like SHAP (SHapley Additive exPlanations) [9], which employs a game-theoretic approach to explain the output of any machine learning model, are also worth considering.

Lastly, we also utilize instance attribution methods, where the goal is to identify the training data point that most influenced the final classification label. A simple approach that we employ in this study is the k-Nearest Neighbours (kNN) algorithm [8]. A more advanced alternative is the Fast IF method, which combines k-NN with inverse Hessian-vector product estimation to identify the most influential data points [10].

2.2 User Studies on Human/AI Interaction

One study by Lim and Perrault [5] is particularly relevant to our research, as it also examines the comparison of different explanation methods in the fact-checking context. The study employs a user study with an online form, and the participants are asked to evaluate five different fact-checking processes. The study found that participants generally preferred explanations that were heavy on details and followed the organic fact-checking process of cross-referencing with other news articles.

Linder et al. [4]'s study investigates the relationship between the amount of explanatory information provided by XAI interfaces and the user's understanding of the AI model. The study concludes that there is a trade-off between the time and attention required for the user to understand the explanation and the amount of understanding required in a given situation.

In another study, Schuff et al. [11] also focuses on human perception. Their research is about randomly generated feature attribution explanations presented as heat maps or bar charts. The study found that how the explanations were presented had a significant impact on the user's comprehension.

In Nguyen et al. [12]'s paper, the design and evaluation of a mixed-initiative approach to fact-checking are presented. The user study examines the combination of human knowledge and experience with the efficiency and scalability of automated information retrieval and machine learning. Their results suggest that transparency of the models leads to the most effective human/AI interaction, particularly in cases where the model is fallible.

Overall, these studies highlight the importance of understanding human-AI interaction in the fact-checking context and the need for effective explanation methods that can facilitate this interaction.

2.3 Explanation-Based Human Debugging in AI

In the realm of AI research, various studies have been conducted to explore the concept of explanation-based human debugging. Lertvittayakumjorn and Toni [13] conducted a survey of existing work on this topic, providing an overview of the current state of the field. Additionally, the Fax-PlainAC framework presented in Zhang et al. [14] utilizes human annotation as a means of correcting the incorrect results produced by AI models. Similarly, there is the FIND framework [15] which confirms the effectiveness of word clouds as a means of debugging deep text classifiers, and understanding the decision-making process of AI models. Overall, these studies provide valuable insights into the ongoing efforts to improve the transparency and interpretability of AI models through explanation-based human debugging.

2.4 Theoretical Approaches

Papers like Madsen et al. [16] have a more theoretical focus, but their conclusions can be applied to research in XAI and fact-checking. They discuss the trade-off between presenting explanations in a more human-grounded (higher abstraction level) and functionally-grounded presentation (that reflects the model's behaviour better). On another note Chang et al. [17] examined whether the association between a document and a topic makes sense. They call this task topic intrusion, as the subject must identify a topic that was not associated with the document shown. It is a method that we think can be applied to fact-checking, as the user needs to identify whether the explanation relates to the claim to decide whether to trust the AI or not. These related studies provide a robust foundation for our research, highlighting the diversity of research in this field and offering potential avenues for further research that incorporates the concepts and conclusions of these studies.

3 Methodology

Our experimental procedure allows for both qualitative and quantitative evaluation. In this section, we describe the method used for semi-structured interviews to observe the impact of presentation strategies on trust in the AI's decision. Additionally, we detail the dataset used, as well as the explanation methods and their chosen presentation strategies.

3.1 Semi-structured interviews

To observe which presentation strategy most effectively contributes to trust in the AI's decision for each explanation method, a qualitative user study in the form of semi-structured interviews was employed. This method was chosen in light of Johs et al. [18]'s findings, which indicate that qualitative user studies are the most common in XAI research, and offer the benefits of both unstructured and structured interviews. The evaluation of presentation strategies often is influenced by the user's understanding of the data. Thus, the semi-structured interview format enabled us to steer the conversation towards the presentation of the data rather than the data itself. Moreover, it allowed for follow-up questions and clarification of misunderstandings. This would have been more challenging in a questionnaire. To enable us to compare the answers of the interviewees, we asked them to provide quantitative feedback on a Likert-type scale. The scale was chosen based on Lim and Perrault [5], a similar research question.

3.2 FEVER Dataset

For this study, we utilise data from the FEVER dataset, a corpus of claims and their corresponding evidence for verification against textual sources, extracted from Wikipedia [19]. Each data point comprises a claim, its associated context and evidence, as well as a classification label.

3.3 Explanation methods and their presentations

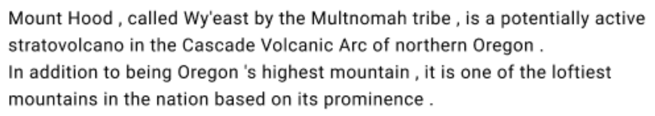
We describe the explanation methods and the corresponding presentation strategies that were chosen for our study. The approach taken is similar to that of Linder et al. [4], where the level of explanation detail is incrementally increased to determine the optimal amount for a user to understand the decision. Some of the chosen presentations are standard in the XAI community, while others were derived by applying common presentation strategies to the outputs of the explanation methods.

ExPred - Interpretable by design

For the ExPred explanation method, we present participants with both the influencing tokens only (Figure 3), and the full context with the influencing tokens highlighted (Figure 4). We decided for highlighting, as the use case of "Advise" as identified in Jacovi and Goldberg [20] fit our goal well. The goal is to assess whether participants prefer more or less context and whether the highlight aligns with their perception of what it should include. We opted for not showing the dots of the raw data (see Figure 2), as their presence can be overwhelming and thus hinder us from observing this goal.

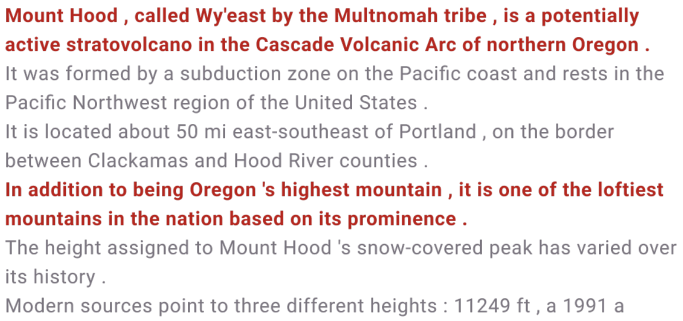
LIME - Feature attribution

The output of the LIME method is an array of values representing the weights of each token, with high absolute values indicating a high influence. We present this information to participants using both heat maps and word clouds. Heat maps are a popular medium for visually explaining feature attribution like done in Schuff et al. [11], but most commonly



Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon .
In addition to being Oregon 's highest mountain , it is one of the loftiest mountains in the nation based on its prominence .

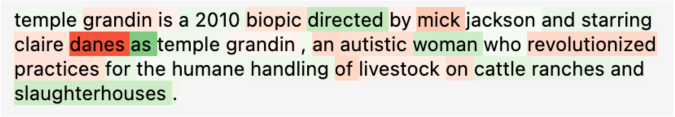
Figure 3: ExPred Freetext presentation for claim "Mount Hood is in the Andes". All tokens that influenced the classification are shown



Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon .
It was formed by a subduction zone on the Pacific coast and rests in the Pacific Northwest region of the United States .
It is located about 50 mi east-southeast of Portland , on the border between Clackamas and Hood River counties .
In addition to being Oregon 's highest mountain , it is one of the loftiest mountains in the nation based on its prominence .
The height assigned to Mount Hood 's snow-covered peak has varied over its history .
Modern sources point to three different heights : 11249 ft , a 1991 a

Figure 4: ExPred Highlighted presentation for claim "Mount Hood is in the Andes".

used for image classification [21]. An example of such a heatmap is shown in Figure 5. The entire context is shown, highlighted in different intensities of green or red (support or refute). The darker the marking, the more influence the token had on the classification. Meanwhile, word clouds are a familiar format that is commonly used for data qualitative assessment [22] and NPL content analysis [23]. For example in Lertvittayakumjorn et al. [15]'s paper, word clouds are utilised for debugging deep learning models. Figure 6 shows an example of our word clouds. Here there is one word cloud per label, the larger a token is, the more influence it had to the corresponding classification.



temple grandin is a 2010 biopic directed by mick jackson and starring claire danes as temple grandin , an autistic woman who revolutionized practices for the humane handling of livestock on cattle ranches and slaughterhouses .

Figure 5: LIME heat map over all tokens of a context for the claim "Temple Grandin stars Claire Danes as a stormtrooper".

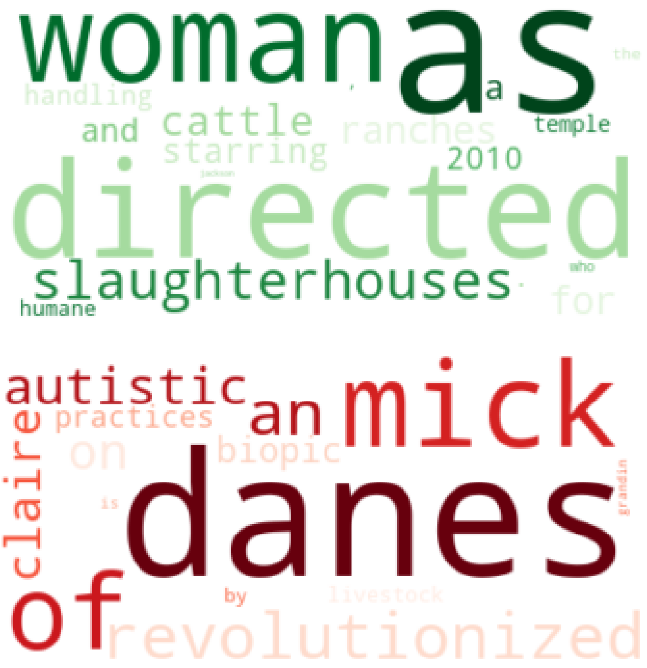
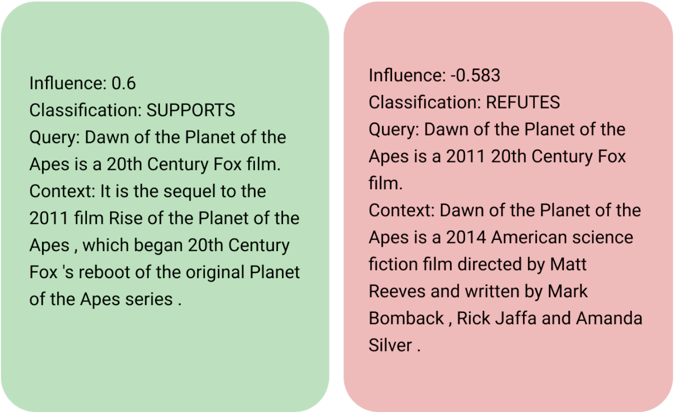


Figure 6: Word clouds for the claim "Temple Grandin stars Claire Danes as a stormtrooper" using the LIME algorithm.

kNN - Instance attribution

The output of the instance attribution method is a sorted array of influence scores and corresponding data points, where the first element is the deciding factor in the classification of the claim. We present this information to participants in both its raw form and as a plot and table. The goal is to assess whether the visual representation of the data points makes it easier to compare the influence scores. An example of two claims can be found in Figure 7. By using labels, a box for each data point and colouring the box in red or green for a supporting or refuting influence we made the data more visually appealing. The full prototype can be found in Appendix A, Figure 15. In Figure 8 a generic example of the graph is shown. Each circle represents one data point. The larger and more green/red it is, the higher the influence. We accompanied this graph with a table which showed a part of the context of each of the data points. For the full example refer to Appendix A, Figure 16.



Influence: 0.6
Classification: SUPPORTS
Query: Dawn of the Planet of the Apes is a 20th Century Fox film.
Context: It is the sequel to the 2011 film Rise of the Planet of the Apes , which began 20th Century Fox 's reboot of the original Planet of the Apes series .

Influence: -0.583
Classification: REFUTES
Query: Dawn of the Planet of the Apes is a 2011 20th Century Fox film.
Context: Dawn of the Planet of the Apes is a 2014 American science fiction film directed by Matt Reeves and written by Mark Bomback , Rick Jaffa and Amanda Silver .

Figure 7: First two claims of instance attribution boxes presentation for the claim "TakePart is a division of 20th Century Fox".

4 Experimental Setup

The details of the user study and its setup are described here. We provide a detailed account of the participants involved in the experimental setup, and then present the methodology for the creation of prototypes utilized in the study. Finally, we describe the measures used to evaluate the prototypes.

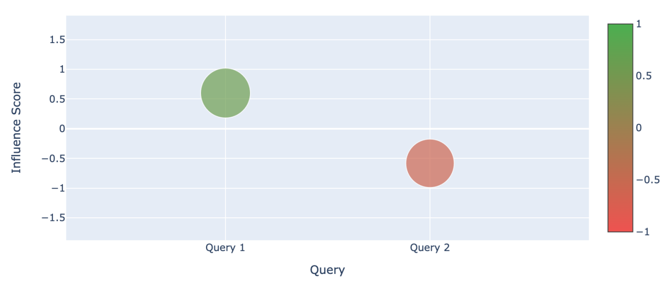


Figure 8: Generic example of instance attribution plot, where every circle represents a data point, with influence scores between (-1,1).

4.1 Participant selection

20 individuals were recruited through the use of convenience sampling, with the inclusion criteria being that participants were students at TU Delft and proficient in the English language. The sample was composed of 7 undergraduate and 13 postgraduate students, with a gender breakdown of 9 females and 11 males. The exclusion criteria for the study included inability to read or colour blindness, as these factors could potentially impact the validity of the results. The sample size of 20 participants was chosen as it provides an optimal balance between the amount of information obtained and the time required to conduct the study, as per the recommendations of Faulkner [24]. This sample size is also sufficient to obtain a minimum of 95% of the expected results.

4.2 Prototype creation

Each participant was presented with six different prototypes, resulting in a total of 120 prototypes being generated for the research. These prototypes were created using Figma [4], with the data points for the prototypes being drawn from 60 entries of the FEVER dataset [19]. The initial evaluations of ExPred, LIME, and k-NN were run on 20 data points respectively.

The process of prototype creation involved the utilization of Python code, which was used to create visualizations and presentations of the data, including heat maps, word clouds, and graphs. These visualizations were then integrated into an interactive frame within Figma. To ensure consistency and ease of understanding, the prototypes adhered to the principles outlined in Google's Material Design Guidelines [5]. An example of a finished prototype that was shown to the participant is here: Figma prototype.

To control for potential sources of bias and improve the comparability of the prototypes, we included at least one supporting and one refuting example per participant and used three different claims per participant to rule out misunderstanding due to unclear data. Additionally, we made sure that the claims used for each explanation method were kept consistent, to accurately compare the presentations. Furthermore, we took special care that the prototypes were created in a manner that would minimize any sampling bias or selection bias and ensure that the effects of earlier prototypes would not affect the later ones.[25] The ”Not Enough Info” claims were not included in the study, to ensure consistency with the test and training data of ExPred [6], which uses the version of FEVER from Eraserbenchmark[6]. Further these claims have limited relevance in real-world scenarios.

Parameter	Description
Visually Appealing	I like the design/colour/layout of the fact-checker.
Easy to Understand	I can understand the details of the fact-checker.
Useful	The details given by the fact-checker are meaningful to me.
Informative	The amount of details given by the fact-checker is acceptable to me.
Convincing	The details given by the fact-checker persuade me to believe in its veracity prediction result.

Table 1: Assessment parameters of an XAI explanation. [5]

4.3 Measures

The evaluation parameters are *Visually Appealing, Easy to Understand, Useful, Informative*, and *Convincing*. To ensure the validity of the results, we followed the definitions of the measures as defined in [5], shown in Table 1. These measures were deemed appropriate for the research as they allow for a comprehensive analysis of the fact-checking prototypes. Participants were asked to rate each measure on a Likert scale of 1-5, with 1 indicating strong disagreement and 5 indicating strong agreement. The study collected data using Likert scale single-item scores, which were analyzed using averages and Cronbach's alpha to measure reliability. This analysis provides an overall consensus and an indication of agreement among participants, which is further supported by qualitative data. This qualitative data was collected through follow-up questions to gain a deeper understanding of the participants' perceptions of the prototypes and to answer our research questions. These questions included queries about the above measures such as “What do you think of the design/layout of the fact checker?”(RQ4), “What would you improve about the fact checker?”(RQ5), and “Is the amount of details acceptable to you, or what would you add/remove?”(RQ2). Finally, we asked the interviewees to indicate which one of the presentations they prefer. The results obtained from this evaluation will aid in determining the efficiency of the different presentation strategies and provide insights for future research in the field of XAI.

5 Results

In this section we present the results of the quantitative evaluation as well as qualitative evaluation of the prototypes, which utilized a thematic analysis approach. Finally we answer our research questions and discuss our recommendations towards an optimal explanation presentation strategy based on these results.

5.1 Quantitative Evaluation

The data gathered from the interviews was analyzed using single-item scores, averages, and Cronbach's alpha as measures of reliability. The scores per measure of the presentation strategies for **ExPred** are in Figures 9 and 10. Overall,

[4] www.figma.com

[5] m2.material.io

[6] https://www.eraserbenchmark.com

there were **high** scores, except for *Visually appealing*, where the scores were **neutral** for the highlights and **low** for the free text. In Figures 11 and 12, the **LIME** scores were depicted, with word clouds scoring **highly** in *Visually appealing* and heat maps scoring **low**. The remaining measures have **low** scores for both, with word clouds scoring **lower** overall, especially for *Informative* and *Convincing*. The **kNN** scores, as depicted in Figures 13 and 14, scored similarly, with boxes receiving **higher** scores for *Informative*. Additionally, participants were asked to indicate their preference between the two presentations for each explanation method. The results showed a majority **preference for ExPred highlighted** with 70% **and instance attribution boxes** with 65%. For **feature attribution, there was a tie**. The average scores in Table 2 reflect this, with the presentation that scored the highest overall also being preferred by the majority of users.

The use of Cronbach's alpha was employed to assess the reliability or internal consistency of the set of scale or test items. It is most commonly used for Likert-type scales [26], and unlike other Inter-Annotator Agreement measures does not limit the amount of annotators, nor requires the data to be randomly sampled. The computed α of 0.928 with a 95% confidence interval [0.884, 0.961] thus indicates **high reliability and internal consistency of the five-point Likert scale.**

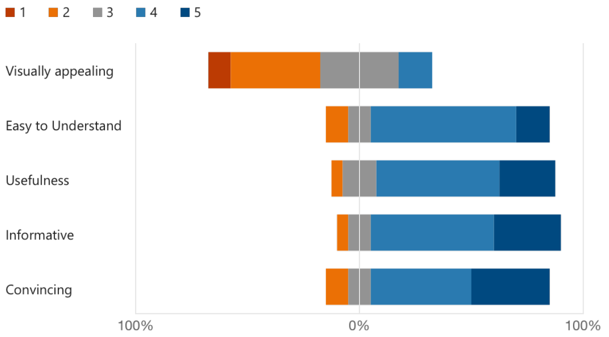


Figure 9: ExPred scores for FreeText. (1 = Strongly disagree, 5 = Strongly agree)

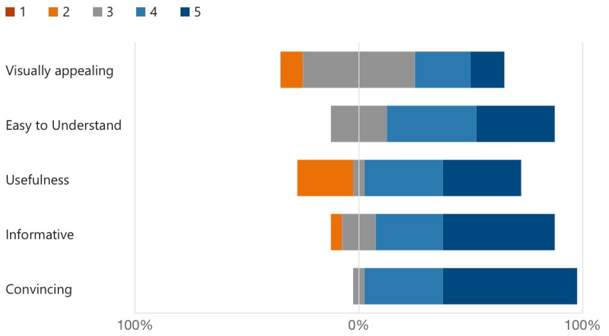


Figure 10: ExPred scores for Highlights. (1 = Strongly disagree, 5 = Strongly agree)

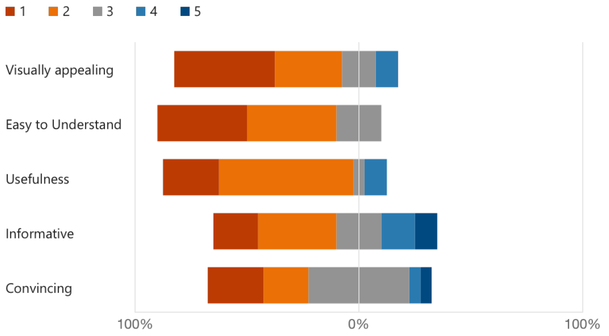


Figure 11: Feature attribution scores for Heatmaps. (1 = Strongly disagree, 5 = Strongly agree)

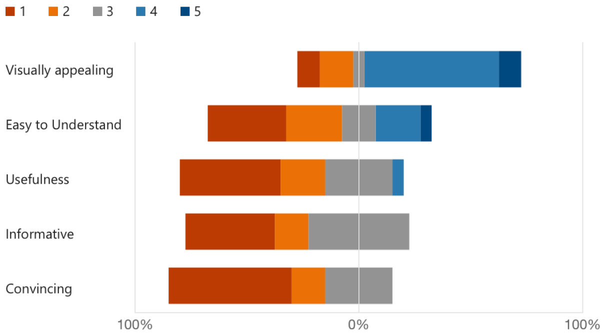


Figure 12: Feature attribution scores for Word Clouds. (1 = Strongly disagree, 5 = Strongly agree)

5.2 Thematic Analysis of Qualitative Data

The method of thematic analysis employed was an inductive, open-coding approach, where the theme development was directed by the content of the data. The aim was to organize the data into different categories and draw meaningful conclusions that relate to both the quantitative data and the research questions. We followed Clarke and Braun [27]'s approach, where we first familiarized ourselves with the data through the listening to and re-reading of the interview transcripts, leading to the identification of common themes within the data. Initial analytic observations and insights were made, with relevant data highlighted and coded to generate labels that captured the important features of the data. These labels, their definitions and examples are presented in Table 3. Throughout this process the research question was kept in mind. Common themes were identified with the labels for each of the presentation strategies and parameters, leading to the identification of dominant patterns for each of the parameters. The analysis led to several **key observations**, including (1) the importance formatting by using correct spacing, "fluff" such as icons, fonts and labels, intuitive colors (e.g. red for refute and green for support) and (2) the presence of context and relevant information being necessary for the users' understanding and focus. As well as (3) that the data presented plays a crucial role in the user's understanding of the fact checker and (4) an explanation was considered most useful when provided with an explanation directly related to the claim. Lastly (5) the number of details presented to the

	Visually appealing	Understandable	Useful	Informative	Convincing	Overall	Preferred
ExPred free text	2,55	3,85	4	4,1	4,05	3,71	
ExPred highlights	3,45	4,1	3,8	4,25	4,55	**4,03**	x
LIME heat maps	1,9	1,8	2	2,6	2,45	2,15	tied
LIME word clouds	3,45	2,35	1,95	2,05	1,75	**2,31**	tied
Instance attr. Boxes	3,45	2,6	2,3	3,55	2,25	**2,83**	x
Instance attr. Graph	3,25	2,45	2,4	2,65	2,4	2,63	

Table 2: Average scores for each parameter, overall average scores per presentation and preference for presentation per explanation method based on feedback from 20 participants.

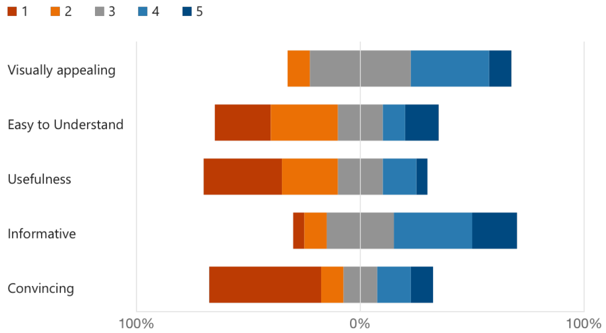


Figure 13: Instance attribution scores for Boxes. (1 = Strongly disagree, 5 = Strongly agree)

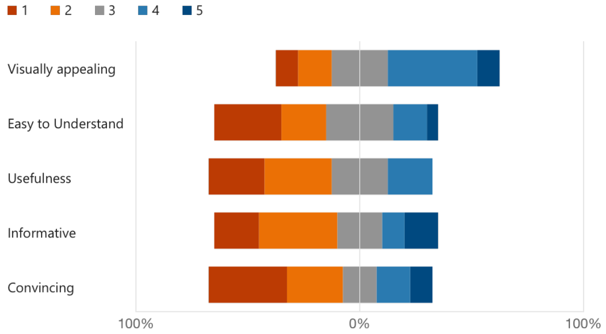


Figure 14: Instance attribution scores for Boxes. (1 = Strongly disagree, 5 = Strongly agree)

users was also a important factor, with a preference for not being overwhelmed with too much information but not not so little that individual words cannot be understood on their own. Overall, the **key takeaway from this analysis is the importance of providing relevant and well formatted context and removing irrelevant information in order to enhance the user's understanding and engagement with the fact checker.**

5.3 Outcomes and Recommendations

The results of both the quantitative and qualitative evaluations were analyzed to derive recommendations for an optimal non-expert-understandable presentation strategy. These recommendations aim at answering RQ5, as we lay out the ways the participant would improve the presentations.

The **ExPred highlighted** presentation performed the strongest in the quantitative analysis, with high scores in all parameters except for *Visually appealing*. The highlights had high scores for both *Convincing* and *Informative*, which could indicate that the preferred level of context/details in explanation is high (RQ2). However, the qualitative analysis revealed other areas of improvement for this method. It was suggested by 35% that we **show only the influential tokens** instead of the whole context and by 50% of interviewees that **the source in form of a hyperlink** should be provided as an option for verification. Further, the formatting of the text should be corrected for spacing and punctuation. A possible implementation of these suggestions is depicted in Appendix B, Figure 17.

The **feature attribution** options received mixed feedback from users. The heat maps were criticized for containing too many details to focus on. Interviewees stated that ”It feels like when you are using a highlighter to mark the important stuff but you end up marking the whole book” and ”It feels like a Christmas tree.” However, the word clouds received criticism for lacking context, making it difficult for users to make sense of the information. This strengthens the argument for more context-heavy explanations (RQ2) and indicates that the use of visual presentations does not necessarily enhance the users understanding (RQ4). Users gave the word clouds a high score on *Visually appealing*, but also mentioned that they ”cannot separate visual from explainable.” For both options, users said that they were not able to see at first glance whether the explanation was supported or refuted. Therefore, **we suggest combining the two solutions for LIME as a compromise between more context and less overwhelming data**. We increased the contrast and did not colour any tokens with a value between the $maximumInfluence/4$ and $minumumInfluence/4$ in the heat map, as well as show only the word cloud of the classification, as seen in Figure 20 in Appendix B.

Regarding the **instance attribution** method, the majority of interviewees did not find it convincing. While 50% of interviewees said they understood the concept and why the AI selected the data, they also felt that the **data and claim had no relation**. This clearly shows the significance of the data in terms of comprehension (RQ3). The type of data provided was deemed useful, but some suggested improvements such as adding icons, highlighting, or better alignment. The participants recognized the potential of the plots, but found them difficult to understand and interact with, supporting the conclusion that visual presentations may not always be effec-

Label	Definition	Examples
Formatting-	Negatively connotated comments about formatting	Spacing, Capitalisation, Colours, Fonts, Icons
Formating+	Positively connotated comments about formatting	I like the plot, Word clouds are pretty
Source	Comments about wanting a source	Need a source, It is not credible like this, I want a link
Overwhelmed	Comments about confusion/overwhelming data/presentations	Too much text/highlight, Distracted by noise, What should I focus on?, Data does not make sense
Clarity	Comments about clear information/presentation	Important parts/highlights are clear, I have all info I need, The formatting is clear, It is easy to compare
Context-	Negatively connotated comments about the context	I miss .., This does not seem relevant, I don't have the details I want
Context+	Positively connotated comments about the context	I like seeing the context, This one gives me more info, It is clearer with the context

Table 3: Open coding labels, their definition and examples from the interview data

tive (RQ4).To improve user understanding, we suggest ranking the queries from top to bottom and replacing the boxes with small graphs. Additionally, the context should be hidden and only shown when the user requests more information, as shown in Appendix B in Figure 16.

6 Responsible Research

The following chapter critically reflects on the paper in terms of the responsibility of the provided research. It discusses the replicability and integrity of the user study, how the research adheres to the principles of FAIR research, and the ethical implications in the field of XAI.

6.1 Threats to Replicability

This section discusses the potential limitations and threats to the replicability of our study results, including sample size, participant demographics, and researcher bias. While the sample size of 20 participants is sufficient for the scope of this research, it is important to note that it may not be large enough to fully capture the diversity of opinions and experiences. Additionally, factors such as room brightness, personal differences in perception, and participant demographics may have influenced the study results, and it is crucial to consider these limitations when interpreting the findings.

The study participants were fully informed of the implications of participating in the research, including the voluntary nature of participation, the use of written notes and audio recordings, and the anonymity of their personal information. They also agreed to be quoted in research outputs, as outlined in the consent form found in Appendix C.

It is also important to acknowledge the potential impact of researcher bias on the study outcomes. Assumptions were made regarding the replicability of the study results, and the risk of researcher bias was increased due to the recruiting of participants through convenience sampling and the semi-structured nature of the interviews. To minimize this impact, measures were taken such as maintaining a neutral tone in the questioning process and emphasizing that there were no incorrect answers to the questions. Additionally, the background knowledge of the participants, all students at a technical university, may also affect the reproducibility, as individuals without technical knowledge may have different observations.

Furthermore, it should be noted that the study did not take into account whether the participants knew the answer to the claim that was fact-checked or not, which could have an impact on the level of trust in the AI-generated explanations. Lastly, the study did not shuffle the order of the explanation methods shown, which could potentially create a false sense of understanding when presented with the second prototype. Therefore, it is important to consider these limitations when interpreting the findings and when replicating the study.

6.2 FAIR research

This research adheres to the FAIR guiding principles for scientific data management as proposed by Wilkinson et al. [28]. These principles, which include findability, accessibility, interoperability, and reusability, have been widely applied to various forms of research, including software development [29].

In terms of findability, the software utilized in this study can be easily located on GitHub, while the prototypes created can be accessed via a unique identifier on the Figma community page. Furthermore, the study's data, in the form of the FEVER dataset, is publicly available.

In terms of accessibility, the software and data are retrievable through standard protocols, as both Figma and GitHub are standard and open-source platforms.

Regarding interoperability, the software is designed to exchange data and allows for ease of integration with other software, as the Jupyter notebooks used in the study specify the necessary inputs.

Finally, in terms of reusability, the study's software is both usable (executable) and reusable (understandable, modifiable, and adaptable for incorporation into other software). Detailed descriptions of the prototype creation process, as well as the tools and codebase used, are publicly available, enabling other researchers to recreate and modify the prototypes for their own use.

6.3 Ethical implications

As AI technology continues to advance and become more widely available, it is important to consider the ethical implications of its use. Tools such as OpenAI's ChatGPT[7], which

[7] https://chat.openai.com/

can generate human-like responses, can make it difficult for users to distinguish between accurate and misleading information. While the benefits of AI are numerous, it is crucial to ensure that users are not misled by inaccurate information. One potential solution is to present explanations in a clear and understandable manner, allowing users to make informed decisions about the data they receive. Other alternatives include tools such as Sourcer[8], which provide users with the ability to verify the credibility of the information they receive and protect against fake news.

7 Discussion

There are several threats to the validity of the research. The present study makes conclusions about optimal instance attribution presentation strategies, but these conclusions lack supporting evidence from other relevant research, which could compromise the validity of the conclusions made about instance attribution presentations.

The study is also limited to a small subset of the FEVER dataset and specific XAI methods, so the conclusions may be different if a larger sample or a different dataset were used. While we attempted to account for data being the reason for misunderstanding by having different prototypes with different data points for each user, given the resources, the risk of invalidity could be reduced by showing more data points and different XAI methods. Specifically, there was data-related confusion for the instance attribution examples. Our conclusions about these presentations would have been more generalizable if this confusion could have been mitigated. However, it is suggested that the disagreement may be a positive outcome as it highlights the importance of providing users with sufficient insight to understand the decision-making process of the AI and to evaluate the credibility of the generated explanations. Conversely, in cases where the data does not provide a clear and convincing justification for the AI's conclusion, it may be more appropriate for the AI to indicate a lack of sufficient information rather than provide a conclusion that is not supported by the data.

Another potential limitation of this study is the possibility of imposing the researchers own biases or perspectives during the data analysis process, rather than fully understanding and interpreting the perspectives and meanings of the participants. While this research used data-driven thematic analysis to mitigate this threat, it is worth mentioning as a threat to validity.

Lastly, the study did not conduct an accessibility analysis on the prototypes before showing them to the participants, which could have an impact on the results. Despite these limitations, the study highlights the importance of presenting explanations in a digestible manner to protect users from being misinformed.

8 Conclusions and Future Work

The present study aims to investigate how different explanation presentation strategies of feature and data attribution techniques affect non-expert understanding in the fact-checking context. Through semi-structured interviews with 20 participants, with *Visually appealing, Easy to Understand, Useful, Informative*, and *Convincing* as our measures, we evaluate different presentation strategies for ExPred, LIME and kNN. The results, and thus the answer to our research question indicate that participants prefer a simple, structured and primarily textual presentation of all available context and details, rather than visual presentations, particularly for first-time interactions with this type of data. Additionally, the study finds that users find fact-checking more convincing when they are able to make the same conclusion as the AI with minimal reading effort, and thus understand how the presented data relates to the claim.

These findings provide recommendations for the design of future XAI explanation method presentations and call for further research in this field. This includes studies involving different data sets, participants from diverse backgrounds, and those with accessibility issues such as colourblindness. Additionally, future research could also apply these findings to other NLP tasks and explore other XAI explanation methods, such as counterfactual explanations or rule/sequence mining, as well as the application of common presentation strategies across different methods.

[8] https://getsourcer.com

References

[1] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL https://aclanthology.org/N19-1060.

[2] Giulia Vilone and Luca Longo. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661, August 2021. doi: 10.3390/make3030032. URL https://doi.org/10.3390/make3030032.

[3] Expert Panel®. Council post: 16 industries and functions that will benefit from ai in 2022 and beyond, Jan 2022.

[4] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K Pentyala, Eric D Ragan, and Xia Ben Hu. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*, 2(4), December 2021.

[5] Gionnieve Lim and Simon T Perrault. Explanation preferences in xai fact-checkers. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET), 2022.

[6] Zijian Zhang, Koustav Rudra, and Avishek Anand. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 418–426, 2021.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[8] Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. Explaining and improving model behavior with k nearest neighbor representations. *CoRR*, abs/2010.09030, 2020. URL https://arxiv.org/abs/2010.09030.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[10] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

[11] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human interpretation of saliency-based explanation over text. *arXiv preprint arXiv:2201.11569*, 2022.

[12] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 189–199, 2018.

[13] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021.

[14] Zijian Zhang, Koustav Rudra, and Avishek Anand. Faxplainac: A fact-checking tool based on explainable models with human correction in the loop. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4823–4827, 2021.

[15] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. Find: Human-in-the-loop debugging deep text classifiers. *arXiv preprint arXiv:2010.04987*, 2020.

[16] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*, 2021.

[17] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

[18] Adam J Johs, Denise E Agosto, and Rosina O Weber. Qualitative investigation in explainable artificial intelligence: A bit more insight from social science. *arXiv preprint arXiv:2011.07130*, 2020.

[19] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[20] Alon Jacovi and Yoav Goldberg. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9: 294–310, 03 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00367. URL https://doi.org/10.1162/tacl_a_00367.

[21] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[22] Concetta A DePaolo and Kelly Wilkinson. Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. *TechTrends*, 58(3):38–44, 2014.

[23] Johannes Fagerlind, Julia Grentzelius, Helena Gustafsson, Marcus Malmberg, Gustaf Norberg, and Carlos Palomino Casseres. Natural language processing for content analysis.

[24] Laura Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383, 2003.

[25] Pritha Bhandari. Within-Subjects Design — Explanation, Approaches, Examples, 12 2022. URL https://www.scribbr.com/methodology/within-subjects-design/.

[26] Joseph A Gliem and Rosemary R Gliem. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community . . . , 2003.

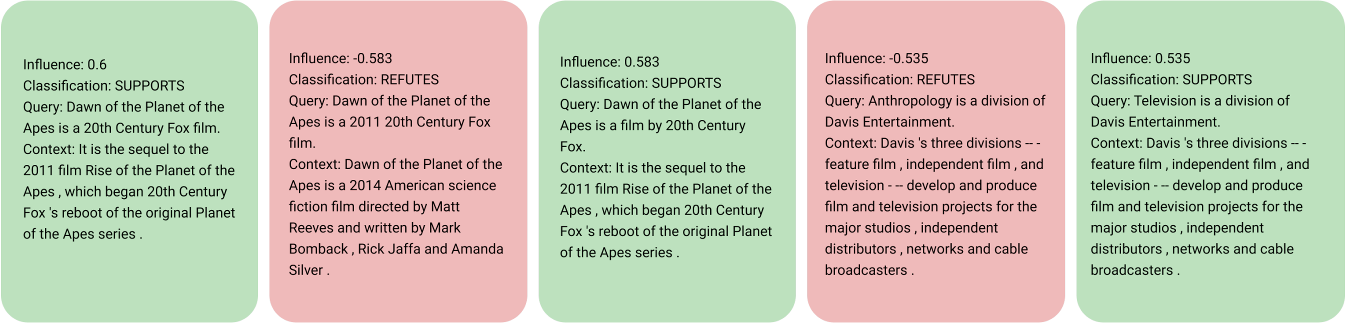
[27] Victoria Clarke and Virginia Braun. Thematic analysis: a practical guide. *Thematic Analysis*, pages 1–100, 2021.

[28] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[29] Michelle Barker, Neil P Chue Hong, Daniel S Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, et al. Introducing the fair principles for research software. *Scientific Data*, 9(1):1–6, 2022.

A Instance attribution full designs

Here we present the complete designs of instance attribution prototype snippets shown in Chapter 3



Influence: 0.6
Classification: SUPPORTS
Query: Dawn of the Planet of the Apes is a 20th Century Fox film.
Context: It is the sequel to the 2011 film Rise of the Planet of the Apes , which began 20th Century Fox 's reboot of the original Planet of the Apes series .

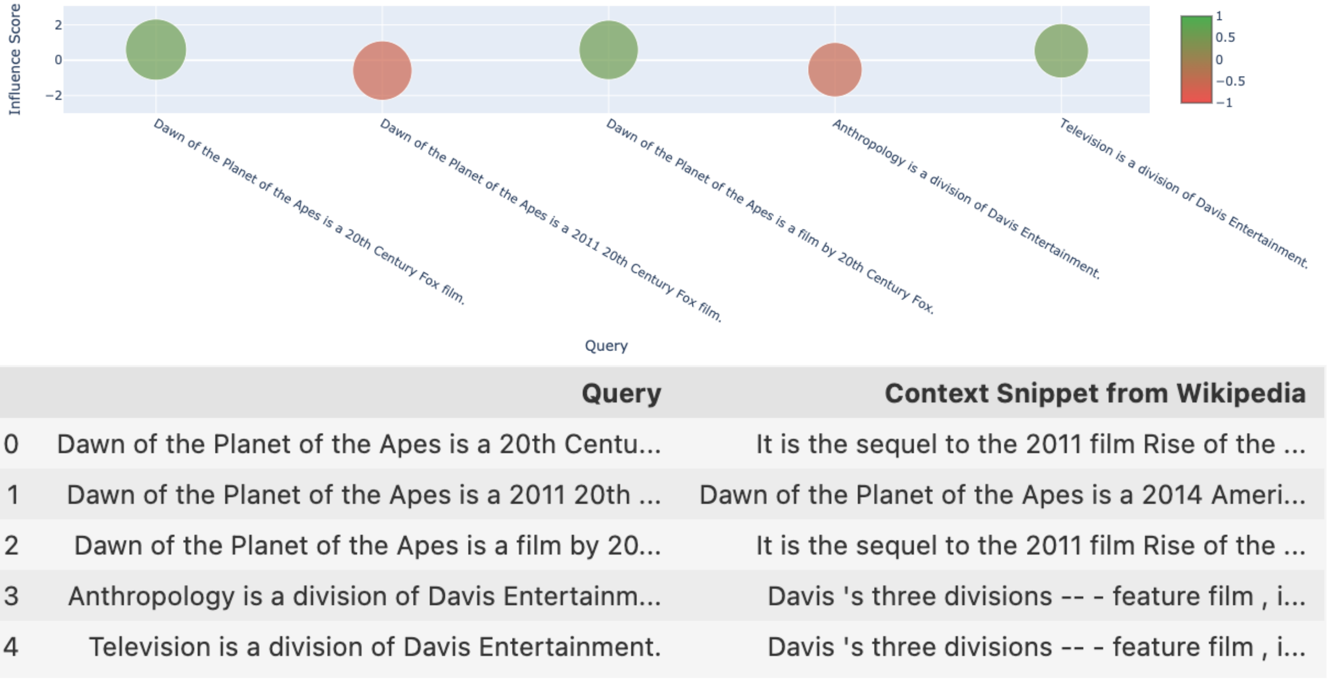
Influence: -0.583
Classification: REFUTES
Query: Dawn of the Planet of the Apes is a 2011 20th Century Fox film.
Context: Dawn of the Planet of the Apes is a 2014 American science fiction film directed by Matt Reeves and written by Mark Bomback , Rick Jaffa and Amanda Silver .

Influence: 0.583
Classification: SUPPORTS
Query: Dawn of the Planet of the Apes is a film by 20th Century Fox.
Context: It is the sequel to the 2011 film Rise of the Planet of the Apes , which began 20th Century Fox 's reboot of the original Planet of the Apes series .

Influence: -0.535
Classification: REFUTES
Query: Anthropology is a division of Davis Entertainment.
Context: Davis 's three divisions -- - feature film , independent film , and television - -- develop and produce film and television projects for the major studios , independent distributors , networks and cable broadcasters .

Influence: 0.535
Classification: SUPPORTS
Query: Television is a division of Davis Entertainment.
Context: Davis 's three divisions -- - feature film , independent film , and television - -- develop and produce film and television projects for the major studios , independent distributors , networks and cable broadcasters .

Figure 15: Instance attribution boxes presentation for the claim “TakePart is a division of 20th Century Fox.” The full version of Figure 7 with all 5 data points and their influence. Again sorted, with the first one having the highest score, and thus being the deciding one.

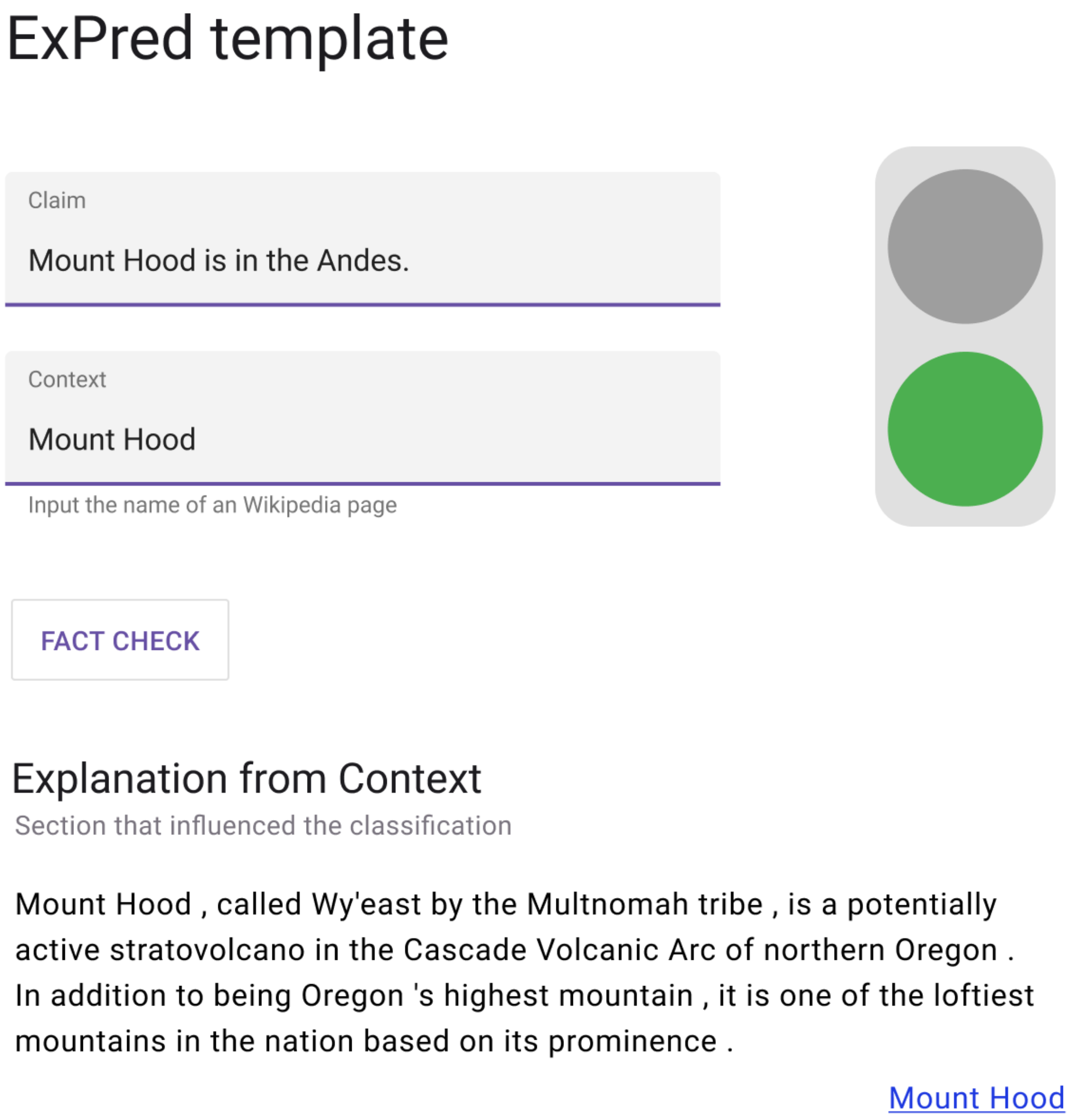


	Query	Context Snippet from Wikipedia
0	Dawn of the Planet of the Apes is a 20th Centu...	It is the sequel to the 2011 film Rise of the ...
1	Dawn of the Planet of the Apes is a 2011 20th ...	Dawn of the Planet of the Apes is a 2014 Ameri...
2	Dawn of the Planet of the Apes is a film by 20...	It is the sequel to the 2011 film Rise of the ...
3	Anthropology is a division of Davis Entertainm...	Davis 's three divisions -- - feature film , i...
4	Television is a division of Davis Entertainment.	Davis 's three divisions -- - feature film , i...

Figure 16: Instance attribution plot and table for claim “TakePart is a division of 20th Century Fox”, with accompanying context and scores for influential training data points. The data points are sorted in descending order, though in this case, all have a similar influence score. The first claim “Dawn of the Plant of the Apes is a 20th Century Fox film.” has the highest, and thus decides on the classification.

B Recommended final presentations

Figma prototypes of possible final presentations for each explanation method, as presented in Chapter 5, as part of the outcomes.



ExPred template

Claim

Mount Hood is in the Andes.

Context

Mount Hood

Input the name of an Wikipedia page

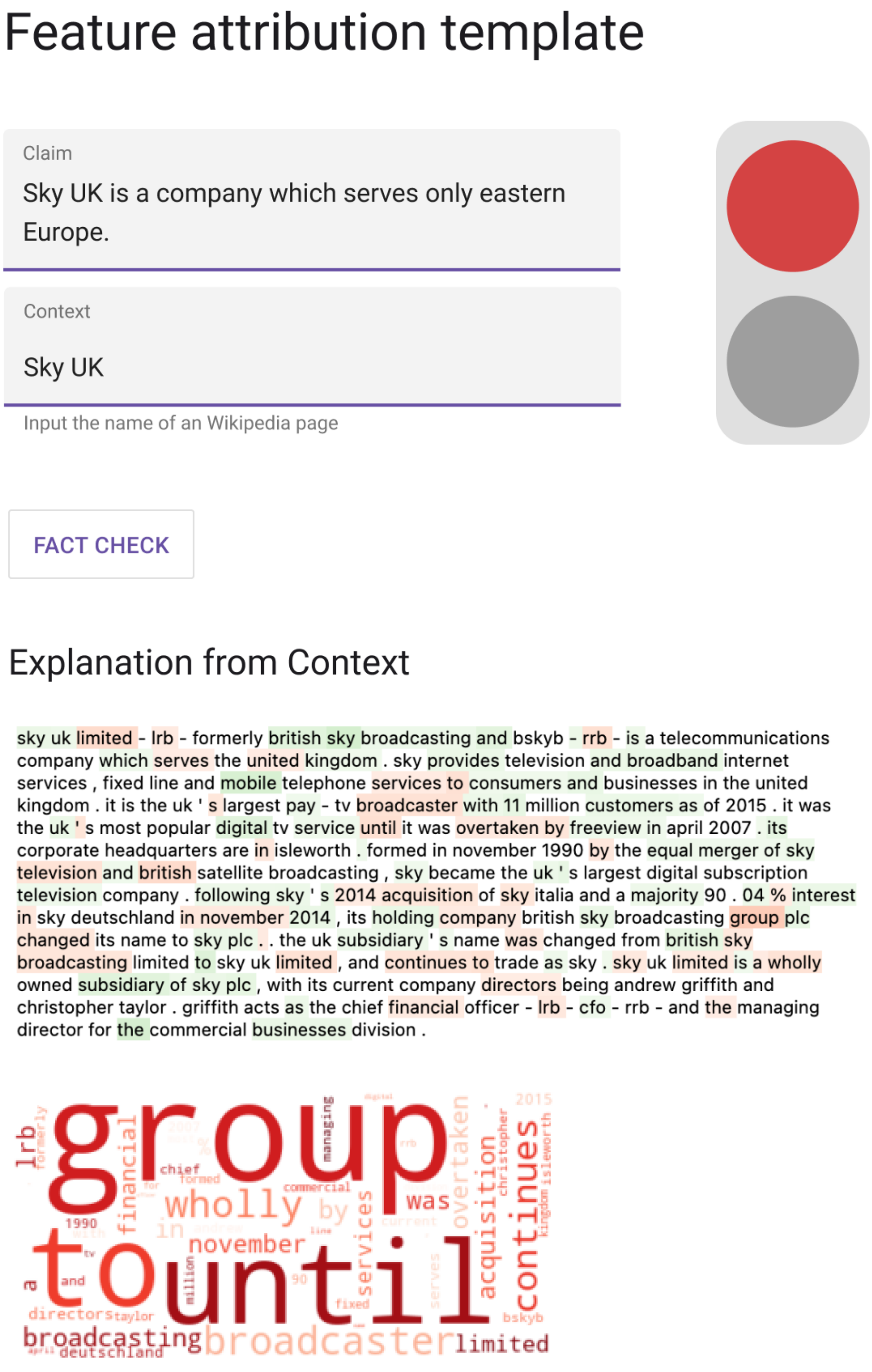
Explanation from Context

Section that influenced the classification

Mount Hood , called Wy'east by the Multnomah tribe , is a potentially active stratovolcano in the Cascade Volcanic Arc of northern Oregon . In addition to being Oregon 's highest mountain , it is one of the loftiest mountains in the nation based on its prominence .

Mount Hood

Figure 17: ExPred recommended presentation as explained in Section 5.3. Showing only the influential tokens, and a link to the source/full context.



Feature attribution template

Claim

Sky UK is a company which serves only eastern Europe.

Context

Sky UK

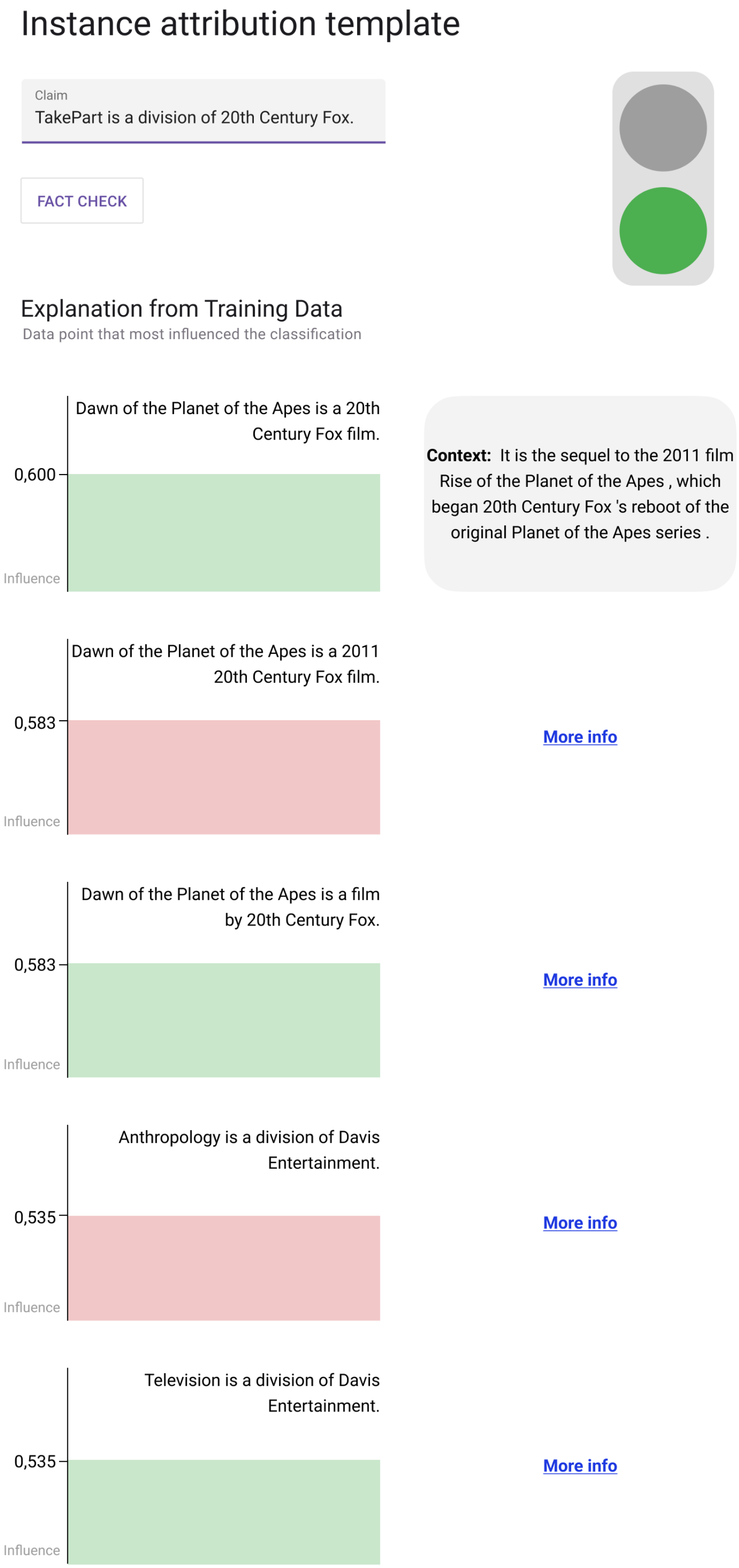
Input the name of an Wikipedia page

FACT CHECK

Explanation from Context

sky uk limited - lrb - formerly british sky broadcasting and bskyb - rrb - is a telecommunications company which serves the united kingdom . sky provides television and broadband internet services , fixed line and mobile telephone services to consumers and businesses in the united kingdom . it is the uk ' s largest pay - tv broadcaster with 11 million customers as of 2015 . it was the uk ' s most popular digital tv service until it was overtaken by freeview in april 2007 . its corporate headquarters are in isleworth . formed in november 1990 by the equal merger of sky television and british satellite broadcasting , sky became the uk ' s largest digital subscription television company . following sky ' s 2014 acquisition of sky italia and a majority 90 . 04 % interest in sky deutschland in november 2014 , its holding company british sky broadcasting group plc changed its name to sky plc . . the uk subsidiary ' s name was changed from british sky broadcasting limited to sky uk limited , and continues to trade as sky . sky uk limited is a wholly owned subsidiary of sky plc , with its current company directors being andrew griffith and christopher taylor . griffith acts as the chief financial officer - lrb - cfo - rrb - and the managing director for the commercial businesses division .

Figure 18: Feature attribution recommended presentation as explained in Section 5.3. Combination of both presentations, with everything in the range $minumumInfluence/4$ to $maximumInfluence/4$ coloured neutral, and only the word cloud matching the final classification shown.



Instance attribution template

Claim: TakePart is a division of 20th Century Fox.

Explanation from Training Data

Data point that most influenced the classification

Dawn of the Planet of the Apes is a 20th Century Fox film.

Context: It is the sequel to the 2011 film Rise of the Planet of the Apes , which began 20th Century Fox 's reboot of the original Planet of the Apes series .

Dawn of the Planet of the Apes is a 2011 20th Century Fox film.

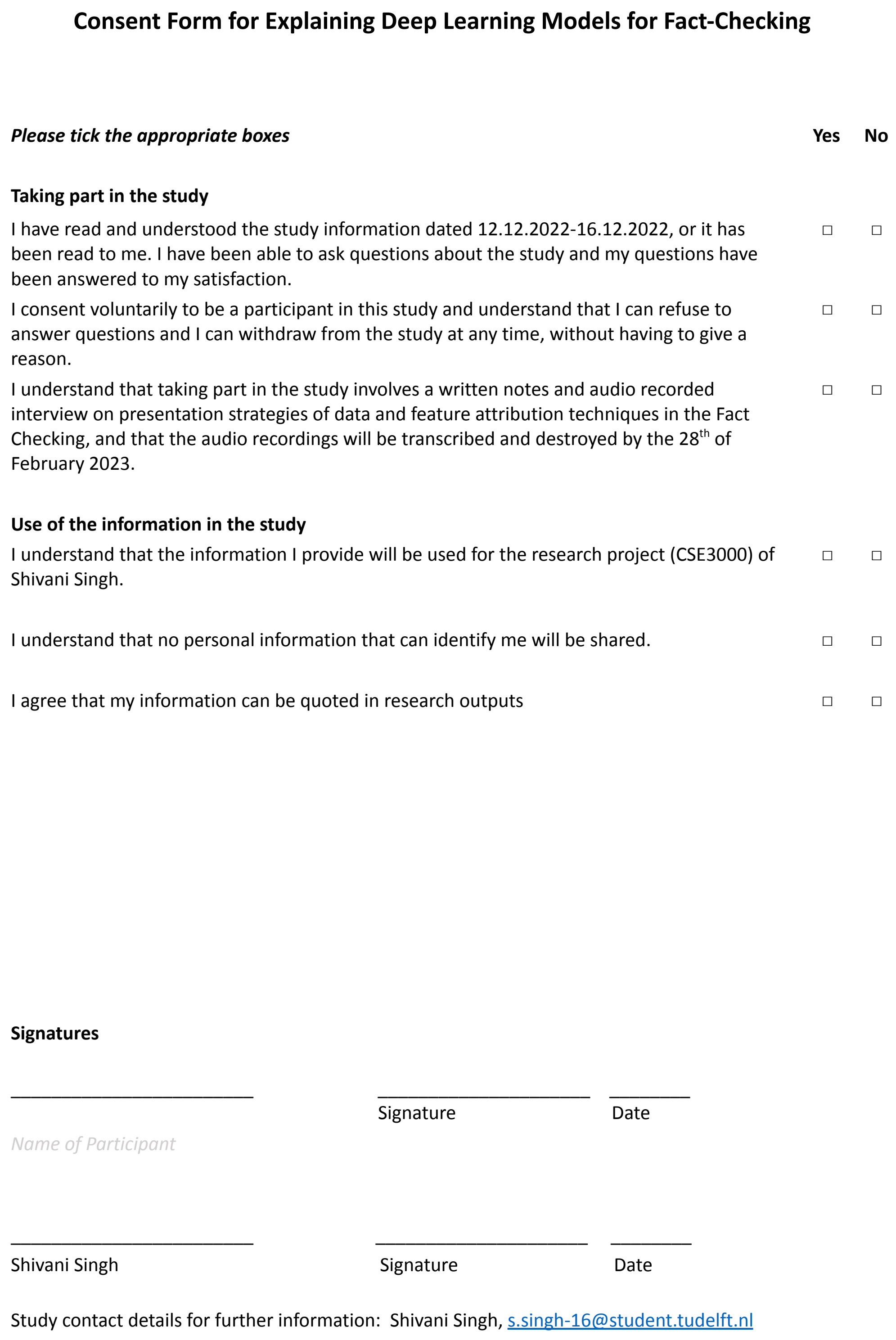
Dawn of the Planet of the Apes is a film by 20th Century Fox.

Anthropology is a division of Davis Entertainment.

Television is a division of Davis Entertainment.

Figure 19: Instance attribution recommended presentation as explained in Section 5.3. A list of bar graphs, each corresponding to one data point, sorted from top to bottom. The colour of the graph indicates refuting or supporting influence. Upon pressing the "More info" button a grey text bubble with the context is shown.

C Consent Form



Consent Form for Explaining Deep Learning Models for Fact-Checking

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated 12.12.2022-16.12.2022, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
I understand that taking part in the study involves a written notes and audio recorded interview on presentation strategies of data and feature attribution techniques in the Fact Checking, and that the audio recordings will be transcribed and destroyed by the 28th of February 2023.	□	□
Use of the information in the study		
I understand that the information I provide will be used for the research project (CSE3000) of Shivani Singh.	□	□
I understand that no personal information that can identify me will be shared.	□	□
I agree that my information can be quoted in research outputs	□	□

Signatures

____________________________ ________________________ _________

Signature Date

Name of Participant

____________________________ ________________________ _________

Shivani Singh Signature Date

Study contact details for further information: Shivani Singh, s.singh-16@student.tudelft.nl

Figure 20: Consent form for user study