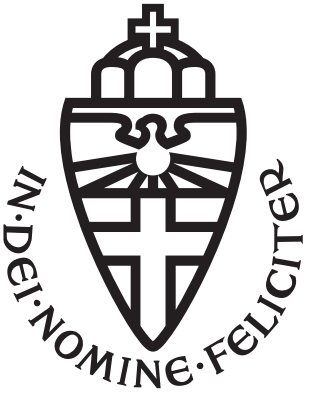
Bachelor's Thesis Computing Science



Radboud University Nijmegen

Evaluating as a human

A functionally grounded evaluation approach to measuring human understanding of SHAP feature relevance explanations of natural language classifications

Author:
B.C. Kapteijns
s1052111

First supervisor/assessor:
prof. M.A. Larson

Second assessor:
dr. I.H.E. Hendrickx

Second supervisor:
J.A. dela Cruz

June 28, 2023

Abstract

Machine learning models that are too complex to interpret can be made more interpretable by creating explanations for them. The type of explanation that this thesis focuses on is one where all input features get a relevance value. These values represent the magnitude of their contributions to getting the output of the model.

Explanations of machine learning models can be evaluated by two types of methods: functionally grounded (automatic, not involving humans) and human-grounded (involving humans) evaluation methods. There is, however, a discrepancy between those methods. Functionally-grounded evaluation methods are biased toward the accuracy of explanations, which is how well the explanation matches the actual reasons for the model's predictions. Human-grounded evaluation methods, on the other hand, are biased toward understandability. Besides, human-grounded evaluations are costly as they involve people.

This thesis tries to strike a balance between these two different methods. This was achieved by investigating the way people understand explanations and by developing a new evaluation method that minimizes human involvement but is still optimized for human understanding. The evaluation method is applicable to post-hoc feature relevance explanations of natural language classifications.

This new evaluation method was developed in three steps. First, the existing literature on evaluation methods and on the social science of human understanding was investigated to obtain an initial evaluation method. Then structured interviews were held to improve the method. Last, the evaluation method was tested using questionnaires.

Contents

1 Introduction **4**

2 Literature review **7**
2.1 What is an explanation (from a social-scientific perspective)? 7
2.1.1 Cognitive process 7
2.1.2 Product 9
2.1.3 Social process 10
2.2 How do humans understand explanations? 11
2.2.1 Coherence, simplicity, and generality 11
2.2.2 Leake's principles 12
2.2.3 Miller's principles 13
2.2.4 Honegger's principles 13
2.3 What is an explanation as the result of an explanation method? 14
2.3.1 Carvalho et al. 14
2.3.2 Gilpin et al. 15
2.3.3 Vollert et al. 16
2.3.4 SHAP 17
2.3.5 Other explanation methods 18
2.4 What are current evaluation methods? 19
2.4.1 Stability and fidelity 19
2.4.2 Identity, separability, and stability 20
2.4.3 Senstivity and implementation invariance 21
2.4.4 Completeness, correctness, and compactness 21

3 Axioms from the literature **23**
3.1 Existing functionally grounded axioms 23
3.1.1 Stability and fidelity 23
3.1.2 Identity, separability, and stability 25
3.1.3 Sensitivity and implementation invariance 26
3.1.4 Completeness, correctness, and compactness 27
3.2 The new axioms 29
3.3 Scope 30
3.3.1 Input data 30

3.3.2 Explanation method 30
3.3.3 Measuring the axioms 31

4 Axioms from Human Surveys **33**
4.1 Designing the interview 33
4.1.1 Why interviews 33
4.1.2 The interview 33
4.1.3 Process of designing 34
4.1.4 Sampling method 37
4.1.5 Explanations used in the interview 37
4.2 Findings 38
4.2.1 Participants 38
4.2.2 Evaluating axioms from the literature 38
4.2.3 Axioms from participants 39
4.2.4 Analysis of human explanations 41
4.2.5 Analysis of machine explanations evaluations 41
4.2.6 Effect of experience 43

5 Validating the evaluation method **45**
5.1 The new evaluation method 45
5.1.1 Compared to people's axioms 46
5.1.2 Compared to the literature principles 47
5.2 Validating the new evaluation method 48
5.2.1 Selection of explanations 49
5.2.2 Selection of participants 49
5.2.3 Results 49
5.3 Limitations 50
5.3.1 Models used 50
5.3.2 Purpose of explanation 50
5.3.3 Measurement of axioms 53
5.3.4 When are explanations necessary 53
5.3.5 Dependence on model and input data 54

6 Conclusions and outlook **55**
6.1 Research questions 55
6.1.1 What is an explanation from a social scientific perspective? 55
6.1.2 How do humans understand explanations? 56
6.1.3 What are current functionally grounded evaluation methods? 56
6.1.4 What functionally grounded evaluation axioms can be extracted from the social sciences? 56
6.1.5 How can the functionally grounded evaluation methods be improved, according to people? 56

6.1.6 How does the new functionally grounded evaluation method perform compared to other functionally grounded evaluation methods? . . . 57
6.2 Future work . . . 57

A Appendix **61**
A.1 The interview . . . 61
A.2 Explanations shown . . . 65
A.3 Questionnaire . . . 77
A.4 Explanations shown in questionnaire . . . 79
A.5 Evaluation method from literature . . . 95
A.5.1 Contrast . . . 95
A.5.2 Distinctiveness . . . 96
A.5.3 Fidelity . . . 97
A.5.4 Realism . . . 98
A.5.5 Compactness . . . 99
A.6 Evaluation method from interviews . . . 100
A.6.1 Realism . . . 101
A.6.2 Compactness . . . 102
A.6.3 Correctness . . . 102
A.6.4 Nuance . . . 102
A.7 Validating the evaluation method . . . 103
A.7.1 Evaluating the new method . . . 103
A.7.2 Evaluating fidelity and stability . . . 104

Chapter 1

Introduction

Machine learning algorithms are more accurate than humans in some instances, such as shown in a paper by Tschandl et al., (2019). Consequently, it makes sense to implement machine learning in practice. In situations where there is little risk or the problem is sufficiently well-defined and studied, machine learning models can already be used to make predictions and decisions. However, in critical situations that are not completely formalized (such as in health care or finance), these models cannot always be used yet (Doshi-Velez & Kim, 2017).

Machine learning models cannot be trusted in those situations and humans must make the final decisions. Using a model's decisions requires people to trust the decisions. Understanding why a decision was made contributes to trusting this decision (Schmidt & Biessmann, 2019), for example in news recommendation (Shin, 2021). However, the nature of many machine learning models is that they are very hard to understand. This is where interpretability comes in.

Interpretability is the ability to explain to a human in understandable terms. Interpretability has more benefits than just giving trust. Adadi et al. distinguish four goals of explaining: to justify, to control, to improve, and to discover (Adadi & Berrada, 2018). Explanations to justify the model give us more trust in the model. Because of that, we focus on explaining to justify in this thesis. Explanations to control and improve can help us identify biases or mistakes in those predictions. These biases and mistakes can then be fixed to improve the machine learning implementation (and get more trust in its predictions). Explanations to discover allow us to learn from machine learning. Unidentified correlations may be hidden in the data that the machine learning implementation has detected (Doshi-Velez & Kim, 2017).

Some machine learning models do not require a separate explanation to be interpretable, such as decision trees. These models are called inherently interpretable machine learning models. Using inherently interpretable

machine learning models is the easiest way to achieve interpretability, but using those models has drawbacks. Most notably, the complexity of the data patterns that these models can describe is usually limited (Carvalho et al., 2019). To increase predictive power, more complex models need to be used. In that case, we can resort to explaining the model using an explanation method. An explanation method creates an explanation of the more complex model.

We can look at their three goals to judge explanation methods: accuracy, understandability, and efficiency (Rüping, 2006). Accuracy is about how much the explanation corresponds to the reasons for the predictions of the model; understandability is about how well humans can understand the explanation; efficiency is about how fast humans can understand the explanation. In this thesis, the term understandability is used to refer to both understandability and efficiency for humans in general.

Whether an explanation method meets those goals can be determined by an evaluation method. Some evaluation methods involve humans: the application- or human-grounded evaluation methods (which we will summarize as human-grounded evaluation methods). Others do not involve humans: the functionally grounded evaluation methods (Doshi-Velez & Kim, 2017). Because humans prefer simpler and more understandable explanations (Gilpin et al., 2018), human-grounded evaluation methods are generally biased towards understandability. Besides, human-grounded evaluation methods require humans, which makes them expensive. On the other hand, functionally grounded evaluation methods are generally biased towards accuracy and insufficiently measure understandability. To the best of our knowledge, current functionally grounded evaluation methods do not use any human judgments.

This thesis aims to make a compromise between these two types of evaluation methods. Its contribution is twofold. First, it describes how humans evaluate explanations in general. And second, it develops an evaluation method for explanations of natural language classification for its relevance in areas such as disaster risk management and sentiment analysis (Behl et al., 2021). The evaluation method evaluates the result of the explanation method (the explanation itself), not the explanation method. The explanations that are evaluated were created using SHAP (Lundberg & Lee, 2017) because SHAP is a popular explanation method that is applicable to natural language classification. It is an explanation method that generates feature relevance explanations.

As said in the third paragraph of this introduction, we focus on explaining with the goal to justify. So for feature relevance explanations with the goal of justification, this thesis develops an evaluation method. The created method reduces the human's task in evaluating explanations to identifying keywords.

In natural language classification, we define keywords as words that are

highly correlated with one of the classes. For example, if we want to classify whether a text is about a flood, the word 'flood' itself would be a keyword. These keywords can be identified without seeing the predictions or the explanations and can be used across multiple explanations, which minimizes human involvement in the evaluation process, while still measuring human understandability.

The results are based on existing functionally grounded evaluation methods, knowledge of human understanding from the social science domain, and interviews. To describe how humans evaluate explanations and to develop the evaluation method, the following questions are answered:

1. What is an explanation from a social scientific perspective?
2. How do humans understand explanations?
3. What are current functionally grounded evaluation methods?
4. What functionally grounded evaluation axioms can be adopted from the social sciences?
5. How can the functionally grounded evaluation methods be improved, according to people?
6. How does the new functionally grounded evaluation method perform compared to other functionally grounded evaluation methods?

The first two questions describe how humans evaluate explanations in general. Combined with the third question, they form the basis for the bottom three and they are answered in the literature review (chapter 2). The literature review is then used as a basis to answer the fourth research question to create a draft of the evaluation method (chapter 3). After, the fifth question is answered to improve the evaluation method using an interview (chapter 4). In addition, the interview answers the second question again as a complement to the literature review. The following chapter proposes the new evaluation method and validates it using a questionnaire (chapter 5). And finally, we will conclude and discuss the results (chapter 6).

Chapter 2

Literature review

In this chapter, the first three research questions from the introduction are answered. These are: "What is an explanation from a social scientific perspective?" (section 2.1), "How do humans understand explanations?" (section 2.2), and "What are current functionally grounded evaluation methods?" (section 2.4). To answer the third question effectively, section 2.3 gives a topology of explanations of machine learning.

2.1 What is an explanation (from a social-scientific perspective)?

Before we look at the way explanation methods explain machine learning models, we first look at how humans explain. When we, humans, want to explain, we first identify the causes of this event (the cognitive process). This cognitive process results in an explanation (the product). This product then needs to be transferred from one person to the next: from the explainer to the explainee (the social process) (Miller, 2019). Let us now discuss these two processes and the product (the explanation).

2.1.1 Cognitive process

Identifying the causes of an event consists of two steps in humans (Miller, 2019):

- Causal connection: identifying the causes of an event using past knowledge;
- Explanation selection: selecting a subset of all causes as the explanation of the event based on the pragmatic goals of the explanation.

Causal connection

There are two (partly overlapping) processes of attributing causes to an event: abductive reasoning and simulation. Abductive reasoning works as follows. The explainer thinks of multiple hypothetical causes for the event. These hypotheses are then tested using past evidence. It is analogous to the scientific method in the sense that you observe a phenomenon, create hypotheses for that phenomenon and get more confidence in that hypothesis as you observe more phenomena that confirm the hypothesis (Miller, 2019).

Simulation is considering different counterfactuals to get a good explanation. Counterfactuals are situations that would have resulted from events that did not happen. For example, imagine you were late to work because your car did not start. In this case, you got to the explanation (because your car did not start) by simulating a world in which your car did start and you were on time. Because there are too many counterfactual cases to simulate all of them, we use heuristics to mutate some events. These mutated events are the counterfactuals that are considered in the (mental) simulation. The heuristics include (Miller, 2019):

- Abnormality: when something is different from usual, it is often considered more mutable;
- Temporality: people undo or change more recent than distant events;
- Controllability and intent: actions that are more controllable by deliberate actors are considered more mutable;
- Social norms: socially inappropriate actions are considered more mutable.

When the causes are clear to the explainer, the most important causes are selected to be used in the explanation.

Explanation selection

When people ask for an explanation, they generally ask the question "Why does P hold?". However, what they mean by this question is "Why does P hold, instead of Q?". Miller, (2019) calls P the fact and Q the foil. Besides expecting an explanation for why the fact holds, the explainee implicitly expects an explanation for why the foil does not hold as well. An explanation that considers both the fact and the foil is a contrastive explanation. It is important that we make a correct assumption about what the foil is. Giving a complete explanation, where all possible foils are explained, will be too overwhelming for a human. So, we need to infer the foil(s) from the context, from human intuition, and from the tone of the explainee's question. People are very good at assuming what the foil is, but it can be hard for computers.

When we know the foils, the best explanation highlights the highest number of differences between the fact and the foil (Miller, 2019).

Besides differentiating between the fact and the foil there are several other (soft) criteria the explainer has for selecting causes for their explanation (Miller, 2019):

- Abnormality: a common fact is hardly an explanation, but an abnormal situation is;
- Intentionality and functionality: intentional actions are better than unintentional actions, but unintentional actions are better than natural causes when it comes to explanation quality;
- Necessity, sufficiency, and robustness: necessary causes are preferred over sufficient causes. If a necessary cause was not the case, the event that is being explained cannot have happened, but if a sufficient case was not the case, the event that is being explained still could have happened. Robust causes, causes that explain multiple events, are desired;
- Responsibility: a cause is more responsible when it has a higher effect on the event. Responsible causes are desired;
- Preconditions, failure, and intentions: an action fails when its preconditions are not met. In this case, mechanistic explanations (explanations that explain the direct mechanism by which something happens) are better than intentional explanations, because failing is generally not intentional.

The selected causes are then used in the product (the explanation).

2.1.2 Product

Now that we understand the process of how an explainer makes an explanation, we get to the next question: what is this explanation that they make? Generally speaking, an explanation is defined as an answer to a 'why'-question that has an implicit inner 'whether'-question. For example, consider the question "Why does P hold?". This question has an implicit inter 'whether'-question "Does P hold?" and the answer to this question is assumed to be "yes". As said in the end of the last question, the answer to that question (which is the explanation) is assumed to be contrastive: the explainee expects the explainer to explain why P holds, rather than something else (Miller, 2019).

Typologies of explanations

Different typologies of explanations exist. The typology of Aristotle considers four modes of explanation (Miller, 2019). Which type of explanation is best depends on what is explained.

- Material: an explanation considering the material of an object;
- Formal: an explanation considering the form or shape of an object;
- Efficient: an explanation considering who or what caused a change to the object;
- Final: an explanation considering the end goal of an object.

Miller, (2019) mentions that other topologies have been proposed, for instance: Dennett separated physical, design, and intention explanations; Marr (building on Poggio) separated computational, representational, and hardware explanations for explaining computational problems; and Kass and Leake proposed a classification of intentional, material, and social explanations. It is important to identify the type of question a human asks to give the right type of answer (Miller, 2019). This means that an explanation can have different types.

2.1.3 Social process

Explanations are different from mere causal attributions in the sense that explanations are a social process. This means it involves a (potentially one-sided) conversation between an explainer and an explainee. This explanation should convey the causes for an event, so a causal attribution is part of an explanation. The explanation should also be relevant to the question and be according to the rules of cooperative conversation (Miller, 2019).

Grice has developed a model that describes how people engage in cooperative conversation. His model consists of four maxims (Miller, 2019):

- Quality: a contribution to the conversation should be of high quality: it should be true. So do not say things that are false or have too little evidence;
- Quantity: a contribution should be as informative as required, but not more informative than required;
- Relation: a contribution should be relevant to the conversation. This is also related to quantity;
- Manner: a contribution should not be obscure or ambiguous, but brief and orderly.

These are not hard criteria in the sense that a conversation cannot be cooperative if some of the principles are violated.

2.2 How do humans understand explanations?

In an explanation, causes are more effective than statistical relationships (Miller, 2019). When it comes to selecting these causes, the most likely cause is not always the best explanation. For example, when we want to explain why a house burned down, the most likely cause is that there was oxygen in the air. However, this is not a satisfactory explanation. A better explanation is that there was a gas leak. The probability that there was a gas leak is lower than the probability that there was oxygen in the air, but the practical implications of a gas leak are a lot greater. So, people evaluate explanations based on practical implications, rather than probability that explanations are correct (Miller, 2019).

The rest of this subsection contains sets of social scientific principles for understanding explanations.

2.2.1 Coherence, simplicity, and generality

In general, explanations are evaluated based on coherence, simplicity, and generality (Miller, 2019). Regarding coherence, Thagard says explanatory coherence is the 'holding together' of an explanation because of explanatory relations. Thagard distinguished four possibilities for coherence. Propositions P and Q cohere if there is an explanatory relation (Thagard, 1989):

- P is part of the explanation of Q; or
- Q is part of the explanation of P; or
- P and Q are both part of the explanation of proposition R; or
- P and Q are analogous in the explanations they give of propositions R and S, respectively.

Two propositions incohere if they contradict each other or are incompatible according to background knowledge. Thagard explains that if an explanation is coherent with someone's personal beliefs, that person confidently believes in the explanation. If an explanation is incoherent with someone's personal beliefs, that person does not believe the explanation (Thagard, 1989).

Explanatory coherence can be seen as a relation between two propositions, as the property of several related propositions, or as a property of one proposition. But in the last case, we do not speak of coherence, but rather of acceptability. Acceptability is the degree to which a proposition is coherent with other propositions.

Simplicity and generality (or breadth, as they call it) are discussed in the paper of Read and Marcus-Newhall. Generality states that (all things being equal) an explanation that explains more facts is more coherent and

thus better than an explanation that explains fewer facts. An explanation that explains many facts is called broad and an explanation that few facts is called narrow. Simplicity is about the number of assumptions an explanation requires. The explanation that requires the fewest assumptions is the simplest (Read & Marcus-Newhall, 1993).

These principles will be used to create the initial version of the evaluation method in chapter 3.

2.2.2 Leake's principles

Besides coherence, simplicity, and generality, Leake proposed nine metrics in four categories for evaluating explanations (Leake, 1991):

- Evaluating for predictions
  - Predictive power: a cause does not always have to imply the event. If we take the burning house again, having oxygen in the air does not always imply that houses burn down. "If all the rules connecting the causes to the outcome are predictive, the causes are considered predictive" (Leake, 1991, p. 22);
  - Timeliness: when a cause is an early warning, rather than an indication that something is happening, it is timely. This is much more useful because then we can recognize and influence the event before it happens the next time;
  - Distinctiveness: indicates whether a surprising event had a surprising cause (which can then be used as a prediction for that event);
  - Knowability: indicates how easy it is to know when the event occurred. There are three levels: observable, testable, and undetectable.
- Evaluating for repair
  - Independence: indicates whether a cause is dependent on other causes;
  - Causal force: indicates whether a reason is a cause or just predictive of the event. For example, when explaining why a ball is red, saying all balls are red is not a cause, but it is still predictive of the color of the ball;
  - Repairability: do we know how to fix the cause of a problem?
- Evaluation for control
  - Blockability: do we know how to avoid the cause of a problem?

- Evaluation of actor's contributions
  - Desirability: did the actor want the outcome? Praise or blame can also be ascribed.

As mentioned in the introduction, we focus on explaining to justify. Evaluating for predictions is assumed to align the best with this goal of explaining, so we focus on predictive power, timeliness, and distinctiveness. We do not focus on knowability, because machine learning predictions are always observable (the predictions are always shown). These principles will be used in chapter 3 to create the initial version of the evaluation method.

2.2.3 Miller's principles

In the paper of Carvalho et al., Miller's criteria for an understandable explanation are summarized as follows (Carvalho et al., 2019):

- Contrastive: why doesn't the model predict something else;
- Selective: select only the main causes;
- Social: take the target audience into account;
- Focus on abnormal: mention rare reasons (when there are low odds for some input, it is abnormal);
- Truthful: the explanation makes sense in the real world;
- Consistent: it is in conformation with prior beliefs;
- General and probable: it explains many cases.

All of these principles are relevant to machine learning explanations. They will be used in chapter 3 to create the initial version of the evaluation method.

2.2.4 Honegger's principles

According to Honegger, there are three conditions that the evaluation methods need to meet to be appropriate: they need to be findable, applicable, and common practice (Honegger, 2018). These are not social-scientific principles but are still principles that the ideal evaluation method should adhere to, so they have been included here.

- Axioms are findable when we can find the axioms to compare different explanations. In this thesis, axioms are considered findable when they have a basis in existing research;
- Axioms are applicable when they are feasible to be used in practice;

- And axioms are common practice when "it is non-extraordinary to use axioms in the field of interpretable machine learning" (Honegger, 2018, p. 15).

These conditions will also be used to judge the new evaluation method.

2.3 What is an explanation as the result of an explanation method?

Before we discuss existing functionally grounded evaluation metrics of explanations, it makes sense to first understand what types of explanations are generated by explanation methods. The goal of an explanation method is to explain a machine-learning model or a prediction of that model. Explanation methods do this in a variety of ways. In this section, we will discuss a few taxonomies of explanation methods. We give multiple taxonomies, of which the taxonomy by Vollert et al., (2021) will be used in the rest of this paper because it focuses on the result of the explanation method more than other taxonomies.

2.3.1 Carvalho et al.

The first taxonomy is by Carvalho et al. The authors define 4 different dimensions along which to classify interpretability. Rather than just classifying explanation methods, they take the broader scope of interpretability in general (Carvalho et al., 2019).

First, they separate pre-model, in-model, and post-model interpretability. Pre-model interpretability is achieved by interpreting the data before making the model (for example using visualization); in-model interpretability is achieved by using an inherently interpretable machine learning model, such that the explanation of the model is the model itself (for example decision trees are inherently interpretable: you can understand a decision tree's prediction by looking at it); post-model interpretability is achieved by using an explanation method to explain the model.

Second, they separate intrinsic vs post-hoc interpretability. Intrinsic interpretability is baked into the machine learning model; post-hoc interpretablity is achieved by applying explanation methods to the model. Intrinsic interpretability is associated with in-model interpretability and post-hoc interpretability is associated with post-model interpretability.

Third, model-specific vs model-agnostic interpretability. Model-specific interpretability is specific to one machine learning method; model-agnostic interpretability is not limited to one model but can explain any model. In-model interpretability is always model-specific; post-model interpretability can be both model-specific and model-agnostic, depending on the explanation method that is used.

And last they differentiate between the different results an explanation can have. There are more results than only the four options given, but the four options cover most explanation methods. The four different types of results are:

- Feature summary: a statistical fact about each feature of the input. This can be a single number for every feature, explaining that feature's importance, but it can also be a visualization of dependencies between features.
- Model internals: the output for intrinsically interpretable machine learning models.
- Data point: example-based explanation methods explain predictions by showing other (potentially already existing) data points.
- Surrogate intrinsically interpretable model: an inherently interpretable (local or global) model is trained to approximate the model to be explained.

SHAP is an explanation method that returns feature relevances from a prediction of an existing model. It is thus a post-model, post-hoc explanation method that generates feature summaries. In addition, SHAP does not look at the internals of the model, so it is a model-agnostic explanation method.

2.3.2 Gilpin et al.

Another taxonomy is given by Gilpin et al. The authors distinguish three types of models: processing models, representation models, and explanation-producing models. Processing models answer the question "Why does this particular input lead to that particular output?"; representation models answer the question "What information does the network contain?"; explanation-producing models aim to simplify some aspect of their behavior (this can be processing, representation, or another aspect). Gilpin et al. give several options for processing, representation, and explanation-producing models (Gilpin et al., 2018):

- Processing models:
  - Linear proxy methods: a linear (potentially local) model is trained such that it approximates the model to be explained. This model is then used as an explanation;
  - Decision trees: a decision tree is used as a proxy model and used as an explanation;
  - Automatic-rule extraction: a list of rules that explain the model;

  - Salience mapping: show which parts of the input have an effect on the output.

- Representation models:
  - Role of layers: all information going through a layer is considered together;
  - Role of neurons: single neurons are considered individually;
  - Role of representation vectors: groups of single neurons are considered.

- Explanation-producing models:
  - Attention networks: functions that provide a weighting over the inputs or internals of a model;
  - Disentangled representations: describe independent meaningful factors of variation;
  - Generated explanations: the model learns to make "because"-sentences to explain predictions.

SHAP creates a salience mapping to get to its explanation because it looks at the direct correlation between input and output. It represents the model as one big representation vector because it does not take the model internals into account. The explanation is presented as an attention network.

2.3.3 Vollert et al.

The last taxonomy that is discussed is made by Vollert et al. and is the taxonomy that will be used in this thesis, because it fits this thesis the best: the typology focuses on the result of the explanation method, rather than the method itself. Different post-hoc explanation types are described (Vollert et al., 2021):

- Visual explanations: use graphical plots to explain the model, for example, decision boundary plots;
- Example-based explanations: use another data point to explain a prediction. This other data point can be of the same or of a different class, depending on the explanation method. It can also be a data point that did not exist in data yet;
- Feature-relevance explanations: quantify the contribution of features. This can be combined with a visual representation;
- Knowledge-extraction explanations: use an inherently interpretable model to approximate the black-box model.

Special cases

Because not all explanation methods fit directly in one of the types, in this section some edge cases are described.

Combinations of different types can occur. In those cases, we look at the intention of the explanation. For example:

- Imagine a graph of the feature relevances. The graph's function is likely just to show the feature relevances, so we classify the graph of feature relevances as a feature-relevance explanation.
- Now imagine a list of examples, based on feature relevances. In this case, we used the feature relevances to find relevant data points. So we classify it as an example-based explanation.
- If we have a simple linear model that assumes feature independence, we can view this as both a knowledge-extraction explanation and a feature-relevance explanation (the weights of the features in the linear model can be seen as relevances). If the linear model is used to show which features have a big impact on the outcome of the black-box model, we count it as a feature-relevance explanation. If it is intended as a simplification of the black-box model, it is seen as a knowledge-extraction explanation.
- Using a plot to show a knowledge-extraction explanation (for example plotting a decision tree or linear model) does not make it a visual explanation, because it is just a way of clearly giving a knowledge-extraction explanation.

The explanations generated by SHAP that were used in this thesis are a visual representation of feature-relevance explanations. Because the explanations intend to bring forward the feature relevances, we classify them as feature relevance explanations.

2.3.4 SHAP

As mentioned in the introduction, the explanation method that this thesis focuses on is SHAP. SHAP is a popular post-hoc, model-agnostic, feature-relevance explanation method. This means it assigns importance scores to features, based on how much they affect the output of the model. It does this by approaching so-called 'Shapley values' (Lundberg & Lee, 2017).

Shapley values come from game theory. Given a set of input values that produces an output value, Shapley values can tell us how much each input value influenced the output value. To compute the Shapley value for input value x, we need to know the output values of all possible combinations of input values without input value x and compare them to the output values of all possible combinations of input features with input value x.

Applied to natural language classification, the input value x would be a feature (for example a word) of the input text and the output would be the class that the model predicted for that input text. In practice, background data is used for x, because most models do not allow removing features. The mean of the difference between the output of all combinations with and without x is the Shapley value of x (Lundberg & Lee, 2017).

The difficulty in this is interactions between different input values. The number of possibilities grows exponentially over the number of input features. Consequently, there are different ways of approaching these Shapley values. Three approximations are described here (Lundberg, 2018):

- Permutation explainer: "approximates the Shapley values by iterating through permutations of the inputs" (Lundberg, 2018). It guarantees local accuracy by iterating through permutations of features. It guarantees exact SHAP values for higher-order interactions between features (depending on the number of iterations).
- Partition explainer: "computes Shapley values recursively through a hierarchy of features" (Lundberg, 2018). Rather than the Shapley values, this method results in the Owen values. Owen values are similar to Shapley values, except computing feature relevance of groups of features, rather than for each feature separately (López & Saboya, 2009).
- Sampling explainer: "computes Shapley values under the assumption of feature independence" (Lundberg, 2018).

These approximations have been used in this thesis.

2.3.5 Other explanation methods

To provide more context into the different explanation methods, this section introduces some of them. Besides SHAP, LIME is also a popular feature relevance explanation method. It works by creating local linear models around a prediction. These local linear models are trained using randomly generated data and the predictions of the models, given that data. The gradients of these linear models can be used to estimate the relevances of the input features (Ribeiro et al., 2016).

Partial dependence plots (Friedman, 2001) aim to give the user insight into the prediction of the machine learning model. They show the dependence between variables, for example by showing their feature relevances. Depending on the type of plot, it can be discussed whether it should be classified as a feature relevance explanation or a visual explanation.

Example-based explanations include counterfactual explanations (Wachter et al., 2017). Counterfactual explanations consist of example data points that are close to the data point to be explained but that have a different

prediction. This can show the explainee by example how different inputs create different outputs.

And finally, an example of a knowledge-based explanation would be a decision tree (Guidotti et al., 2018). Decision trees are interpretable models that are used in machine learning. Because of their inherent white-box nature, they can also be used as surrogate models for more complex black-box models, such as (deep) neural networks).

2.4 What are current evaluation methods?

Functionally grounded metrics are generally sets of axioms along which explanations are evaluated. These axioms are formulas that measure some value that should be indicative of the quality of the explanation. In this section, we will explain a few existing functionally grounded evaluation methods.

2.4.1 Stability and fidelity

Two commonly used axioms that measure the quality of an explanation are stability and fidelity (Velmurugan et al., 2020).

- **Stability** is the degree to which similar data and predictions generate similar explanations. For feature relevance explanations, stability can be measured in three ways, depending on the result of the form of the explanation (Kalousis et al., 2007):
  - Stability by weight: when the explanations are weightings of the features (each feature is associated with a weight or attribution to the prediction), the stability between two data points can be computed using Pearson's correlation coefficient between those two points (see equation 2.1);

$$S_W(w, w') = \frac{\sum_i (w_i - \mu_w)(w'_i - \mu_{w'})}{\sqrt{\sum_i (w_i - \mu_w)^2 \sum_i (w'_i - \mu_{w'})^2}} \tag{2.1}$$

  - Stability by rank: when the explanations are rankings of the importance of the features, the stability is Spearman's correlation coefficient between two vectors r and r', where r_i is the rank of feature i in one data point and r'_i is the rank of feature i in the other data point (see equation 2.2);

$$S_R(r, r') = 1 - 6 \sum_i \frac{(r_i - r'_i)^2}{m(m^2 - 1)} \tag{2.2}$$

  - Stability by subset: when the explanations are subsets of the important features, the stability can be calculated using an adaptation of the Tanimoto distance between the two sets. The Tanimoto distance depends is a combination of the cardinalities of the two sets and the intersection between the two sets (see equation 2.3).

$$S_S(s, s') = 1 - \frac{\|s\| + \|s'\| - 2\|s \cap s'\|}{\|s\| + \|s'\| - \|s \cap s'\|} \tag{2.3}$$

- **Fidelity** measures how 'faithful' an explanation is to the model, which is how well an explanation approximates the model. Two ways of measuring fidelity are defined, for some types of explanations:
  - External fidelity: measures the similarity of the decisions of the black-box model and the surrogate model;
  - Internal fidelity: measures whether the decision-making process is the same for the black box and surrogate model. This can be done by removing features and investigating what happens; by investigating the decision-making process of the surrogate (white box) model; or by using another explanation method on both the black box and surrogate model.

In the case of SHAP, stability by weight and internal fidelity would be appropriate implementations of these axioms. External fidelity cannot be used because not every sentence consists of the same features.

2.4.2 Identity, separability, and stability

Another set of functionally grounded metrics includes identity, separability, and stability. The axioms work for all feature relevance explanations and are grounded on human intuition according to the author (Honegger, 2018):

- **Identity** states that identical objects should have identical explanations (so there should not be a random component in the explanation method);
- **Separability** means that non-identical objects should have non-identical explanations;
- **Stability** states that similar objects should have similar explanations. It can be computed in the same way as in the other set of principles.

The author claims that these measures are "necessary but not sufficient to achieve interpretability," although they should be used as soft criteria. The metrics have all been formalized in the paper of Honegger, (2018). However, studies such as the one by Lertvittayakumjorn et al., (2019) show that LIME

still produces the best explanations in some tasks, even though the method does not satisfy identity because of the inherent randomness (Honegger, 2018).

2.4.3 Senstivity and implementation invariance

Other functionally grounded metrics include sensitivity and implementation invariance (Sundararajan et al., 2017):

- **Sensitivity** has two types.
  - When two samples differ in only one feature but have differing predictions, the relevance of that feature should not be zero.
  - If a feature is not important for the prediction, its attribution should be zero.
- **Implementation invariance** means if two models are trained on the same data and have the same predictions, the attributions should be the same for both models.

The authors have developed an explanation method that satisfies the axioms of sensitivity and implementation invariance: integral gradients. But Abeyagunasekera et al., (2022) found that in some cases "[integral gradients] aren't intuitive as LIME." Efficiency (how fast humans can understand an explanation) is not considered by these metrics according to Carvalho et al., (2019).

2.4.4 Completeness, correctness, and compactness

That last set of functionally grounded metrics that are discussed in this thesis contains completeness, correctness, and compactness, which Silva et al., (2018) summarize as the three Cs of interpretability:

- **Completeness** indicates the degree to which the explanation can be used for other cases (so how much of the training set is covered by the explanation);
- **Correctness** is about trust. It measures how many instances covered by the same explanation have the same label/prediction. This is analogous to the simple accuracy metric when evaluating a machine learning model ($\frac{correct}{total}$);
- **Compactness** states that the explanation should be succinct. This can be the compressed size of an explanation.

These three axioms are not directly applicable to feature relevance explanations, because (Silva et al., 2018) assumed the explanation to be a

simple model (knowledge-based explanation). However, similar metrics can be defined for feature-relevance explanations, as will be shown in section 3.2.

In chapter 3, the functionally grounded metrics that have been presented in this section (2.4), will be measured against the social-scientific principles for human understanding of explanations that have been presented in section 2.2. This comparison will be used as inspiration for the hypothesis for the functionally grounded evaluation method that is presented in section 3.2.

Chapter 3

Axioms from the literature

This chapter answers the fourth question from the introduction: "What functionally grounded evaluation axioms can be obtained from social sciences?" To answer this question, we look at the social-scientific principles given in section 2.2 and the existing evaluation methods given in section 2.4.

3.1 Existing functionally grounded axioms

In table 3.1, on the next page, we can see which of the different principles from the social sciences are used in existing functionally grounded axioms. While functionally grounded explanation methods generally do not directly measure understandability, the accuracy of the explanation also impacts its understandability. For example, this is shown by Leake's principle of truthfulness (if the model makes sense in the real world).

The table works as follows. Let us take the table's top-left (non-bold) cell that says "Coherence". In this case, it means that the axioms of stability and fidelity can be used as a measure of coherence (but not for simplicity and generality). The results of the table are mainly based on chapter 2 (especially sections 2.2 and 2.4), combined with the author's intuition. An explanation of the table is given in the following subsections (3.1.1 to 3.1.4).

3.1.1 Stability and fidelity

The axioms of stability and fidelity have been explained in section 2.4.1.

Coherence, simplicity, and generality

The social-scientific principles of coherence, simplicity, and generality have been explained in section 2.2.1. The axioms of stability and fidelity make sure that the explanations of predictions of the model are as close as possible to what seems to be happening inside the model. This means that the different explanations of predictions of the model are all close to the

Axioms	**Coherence, simplicity, and generality (2.2.1)**	**Leake's principles (2.2.2)**	**Miller's principles (2.2.3)**	**Honegger's principles (2.2.4)**
Stability and fidelity (2.4.1)	Coherence	Timeliness, predictive power	Consistent	Findable, applicable, common practice
Identity, separability, and stability (2.4.2)	Coherence	Timeliness, predictive power	Consistent	Findable, applicable, common practice
Sensitivity and implementation invariance (2.4.3)	Coherence		Consistent	Applicable, common practice
Completeness, correctness, and compactness (2.4.4)	Coherence, simplicity, generality	Timeliness, predictive power	Consistent, selective, general and probable	Applicable
Unused principles		Distinctiveness	Contrastive, social, focus on abnormal, truthful	

Table 3.1: What principles from the social sciences are captured in the axioms of current functionally grounded evaluation methods?

model itself, which means that those explanations are coherent (Thagard, 1989). The axioms of stability and fidelity have no measure of the number of assumptions of explanations. Stability could be seen as a measure of generality, in the sense that similar explanations are used for similar data. However, because the definition from Read et al., (1993) says that generality is about one explanation explaining multiple scenarios, these axioms do not fit generality. So coherence is measured by stability and fidelity, but simplicity and generality are not.

Leake's principles

Leake's social-scientific principles have been explained in section 2.2.2. The principles from Leake that are captured in the axioms of stability and fidelity are timeliness and predictive power. When the explanations are faithful to the model (fidelity), we can use the explanation to predict future predictions. For example, when a faithful explanation indicates that a high feature A indicates prediction X, we have reason to think a new prediction is X as well when we see a high A. So, that explanation is indicative of the prediction of the model (which implies predictive power). Distinctiveness is not measured, because there are no measurements for normality or abnormality. The sta-

bility metric is not needed to satisfy Leake's principles, but it can be argued that it measures predictive power because when we see an explanation that is similar to one we have seen before, we can predict a similar outcome of the model. So timeliness and predictive power are measured by stability and fidelity, but distinctiveness is not.

Miller's principles

Miller's social-scientific principles have been explained in section 2.2.3. Stability and fidelity are a proxy for consistency in the sense that explanations for similar instances need to be similar as well. The explanations then also fit with our 'prior belief' of a previous similar instance. Stability and fidelity cannot confirm whether the explanations are contrastive, because neither of the axioms considers the foil. Selectivity is also not measured because the size of the explanation does not impact the measure of fidelity or stability. The axioms are the same for all audiences, so they are not social. There is no knowledge of normality or abnormality encoded in the axioms, so a focus on abnormality cannot be measured. The axioms do not consider common knowledge, so truthfulness cannot be measured. We could argue that if the model is in accordance with the real world, fidelity would measure truthfulness because it measures the fidelity of the explanation to the model (and by extension to the real world). However, because of the assumption that the model is in accordance with the real world, it is not satisfactory. They are also not general and probable, with similar reasoning of generality. In conclusion, only consistency from Miller's principles is measured to a sufficient degree by stability and fidelity.

Honegger's principles

Honegger's principles for evaluation methods have been explained in section 2.2.4. Stability and fidelity are commonly used metrics as Velmurugan et al., (2020) say. This means they are applicable (because they can be used) and common practice (because they are used commonly). Because they are used in papers before (Velmurugan et al., 2020), they are also findable. So the axioms of stability and fidelity are findable, applicable, and common practice.

3.1.2 Identity, separability, and stability

The axioms of identity, separability, and stability have been explained in section 2.4.2.

Coherence, simplicity, and generality

The social-scientific principles of coherence, simplicity, and generality have been explained in section 2.2.1. Identity, separability, and stability are a proxy for coherence, because of the stability and identity axioms. When similar predictions are given similar explanations, the explanations will play an analogous role in explaining those predictions. This is a form of coherence, according to (Thagard, 1989). Using the same reasoning as for stability and fidelity, we can conclude that identity, separability, and stability also do not measure simplicity and generality. So coherence is measured by identity, separability, and stability, but simplicity and generality are not.

Leake's principles

Leake's social-scientific principles have been explained in section 2.2.2. If we have new data that is similar to existing data, explanations for existing data could be good explanations for the new data because of stability. This means the axioms are proxies for both timeliness (because we can guess the prediction before we know it) and predictive power. The same reasoning from stability and fidelity, explaining why distinctiveness is not captured by the axioms, holds for identity, separability, and stability as well. So timeliness and predictive power are measured by identity, separability, and stability, but distinctiveness is not.

Miller's principles

Miller's social-scientific principles have been explained in section 2.2.3. For the same reasons that stability and fidelity measure consistency, identity, separability, and fidelity do as well. And similarly, the other principles also do not hold. So only consistency from Miller's principles is measured by the axioms.

Honegger's principles

Honegger's principles for evaluation methods have been explained in section 2.2.4. Honegger, (2018) himself claims that the axioms are findable, applicable, and common practice. So all of Honegger's principles are met by identity, separability, and stability.

3.1.3 Sensitivity and implementation invariance

The axioms of sensitivity and implementation invariance have been described in section 2.4.3.

Coherence, simplicity, and generality

The social-scientific principles of coherence, simplicity, and generality have been explained in section 2.2.1. The sensitivity metric makes sure that different predictions do not get the same explanation. If the explanations were the same, they would have been incoherent. In that sense sensitivity is a proxy for coherence. There is no measure of simplicity. It could be argued that implementation invariance is a measure of generality: the same explanation can be used across different models, so it explains different predictions. However, in this thesis, we see generality as explaining different predictions of one model so the axioms are not general.

Leake's principles

Leake's social-scientific principles have been explained in section 2.2.2. Because sensitivity and implementation invariance do not measure how accurate an explanation is to the model or previous explanations, the axioms are no proxies for timeliness and predictive power. Distinctiveness is not measured, for the same reasons as for the previous axioms. So none of Leake's principles are captured in the axioms of sensitivity and implementation invariance.

Miller's principles

Miller's social-scientific principles have been explained in section 2.2.3. Because neither of the axioms measures anything related to the user's prior beliefs, they cannot measure consistency. The rest of Miller's principles do not hold for the same reasons as for the previous axioms. So none of Miller's principles are measured by sensitivity and implementation invariance.

Honegger's principles

Honegger's principles for evaluation methods have been explained in section 2.2.4. Although the principles do make sense, the authors have not directly provided scientific research validating the sensitivity and implementation invariance axioms. This means they are not findable. Sensitivity and implementation invariance can be tested by checking all predictions, so the axioms are applicable. The paper presenting the axioms by (Sundararajan et al., 2017) has over 3300 citations. This could be a good indication that the axioms are common practice. This means the axioms are applicable and common practice.

3.1.4 Completeness, correctness, and compactness

The axioms of completeness, correctness, and compactness have been described in section 2.4.4.

Coherence, simplicity, and generality

The social-scientific principles of coherence, simplicity, and generality have been explained in section 2.2.1. The combination of completeness and correctness measures coherence. If the explanation for one prediction explains more predictions and these explanations are correct, the explanation better explains the model. If all explanations are as complete and correct as possible, then they all explain a part of the model correctly and they are coherent. Compactness measures simplicity, because the more compact an explanation is, the fewer reasons it can use. The combination of completeness and correctness also explains generality, because if an explanation is complete and correct, then it can be used for many predictions. This means it is general. So coherence, simplicity, and generality are all measured by the axioms completeness, correctness, and compactness.

Leake's principles

Leake's social-scientific principles have been explained in section 2.2.2. Completeness and correctness show how close an explanation is to the other predictions. This increases the accuracy of that explanation and shows it is closer to the model. If an explanation is correct, we can predict the predictions of the model for data points it has not seen yet, so the axioms measure timeliness and predictive power. There is no measure for distinctiveness in completeness, correctness, or compactness. We could even argue that completeness hinders the measurement of distinctiveness because distinctive explanations (that explain only a few cases, but accurately) are considered worse than general explanations. So timeliness and predictive power are measured by completeness, correctness, and compactness, but distinctiveness is not.

Miller's principles

Miller's social-scientific principles have been explained in section 2.2.3. Explanations are selective because of the combination of compactness and correctness. The more compact and explanation (the fewer reasons/simpler), the better. But to ensure we pick only the most accurate reasons, we have correctness. Explanations are also general and probable, using the same reasoning as for the generality principle. Because completeness and correctness measure how many predictions are in accordance with an explanation, consistency is also measured: it is consistent with prior predictions/beliefs. For the same reasons as the other axioms, explanations are not measured to be contrastive, social, focusing on abnormalities, or truthful. So using completeness, correctness, and compactness, we can measure whether explanations are consistent, selective, and general and probable, but we cannot

measure whether they are contrastive, social, truthful, or focus on the abnormal.

Honegger's principles

Honegger's principles for evaluation methods have been explained in section 2.2.4. The authors have not directly provided scientific research backing the axioms, so they are not findable. The axioms are applicable, as the authors have shown, although they are not very useful for all explanations. Some explanations are given in the form of feature attributions, but they are specific for each case (so completeness and correctness cannot be measured). The axioms are not common practice, because the paper only has 13 citations and the axioms are quite unique compared to the other measures presented.

3.2 The new axioms

This section will present a hypothesis for the new evaluation method, based on section 3.1 (which was in turn based on 2.2 and 2.4). It consists of six general axioms that measure the explanation's quality. In the next section, an implementation of the axioms will be given that can be used for feature relevance explanations.

A few of the axioms can be copied from the table and its explanation presented in section 3.1. They can already be used as proxies for the social-scientific principles for a good explanation. For the principles that were not used in existing functionally grounded methods, new axioms were created. The following list is a hypothesis of the relevant axioms, according to the literature, answering the fourth question from the introduction "What functionally grounded evaluation axioms can be obtained from social sciences?"

- Fidelity: how faithful is an explanation to the model? How well does the explanation approximate actual 'reasons' for the prediction of the model? This axiom has been taken from existing functionally grounded evaluation methods and measures the principles of coherence, timeliness, predictive power;
- Completeness: how many cases/data points are explained by the explanation? This axiom has been taken from existing functionally grounded evaluation methods and measures the principles of generality, and generality and probability;
- Compactness: how large is the explanation? This axiom has been taken from existing functionally grounded evaluation methods and measures the principles of simplicity, and selectivity;
- Distinctiveness: how unique are the causes? This axiom is new and measures the principles of distinctiveness, and focus on abnormality;

Figure 3.1: Example of a feature-relevance explanation generated by SHAP.

- Contrast: how contrastive is the explanation? Does the explanation mention why the model did not predict something else? This axiom is new and measures the principle of contrastivity;
- Realism: does the explanation make sense in the real world? Is the explanation plausible? A (reusable) human judgment is necessary in order to measure this axiom. What this human judgment looks like depends on the explanation and model. In section 3.3.3, an example implementation of this axiom can be found. This axiom is new and measures the principles of coherence, truthfulness, and consistency.

The evaluation method is then the combination of these axioms.

3.3 Scope

The implementations of the new axioms depend on the explanation method used. In this section, we detail the data, model, and explanation and how the axioms are measured.

3.3.1 Input data

In this thesis, we focus on text classification. More specifically, we focus on explanations for the predictions of neural networks that take in small pieces of text (64 words or fewer) and make a binary classification.

3.3.2 Explanation method

Ideally, we would create and test our axioms for evaluation using all explanation methods. The introduction and literature research describe explanations in general. However, due to time constraints, we need to focus on only one explanation method. As said in the introduction, the thesis focuses on one explanation method: SHAP, a feature-relevance explanation method. We made this decision because of SHAP's popularity, existing implementations, and applicability to natural language classification. Figure 3.1 is an example of an explanation that has been generated using SHAP.

The goal of the explanations was to justify the underlying machine learning model, as mentioned in the introduction.

3.3.3 Measuring the axioms

For feature-relevance explanations that are generated using SHAP (like the one in figure 3.1), an implementation of the axioms was made. This implementation is largely based on the author's understanding of the different axioms. The exact algorithms that were used for measuring the axioms can be found in the appendix. Here is a general description of them.

Fidelity Velmurugan et al. describe multiple ways of measuring fidelity (Velmurugan et al., 2020). Because in this case, we do not have a surrogate model as an explanation, the only suitable measurement is the following: change features in the explanation one at a time and for each change run the model again to see how it impacts the prediction. The more a prediction changes with changing features, the higher the feature relevance should be. The drawback of this method is that it assumes feature independence. However, because testing all feature combinations is computationally too expensive, this is still the best option.

Completeness Feature-relevance explanations are only applicable to one single prediction in natural language processing. This means the completeness axiom is not applicable in this case.

Compactness For feature-relevance explanations, compactness is defined as the inverse of the number of outliers in the explanation. The number of words used cannot be used as a measure, because the number of words used depends on the length of the sentence. Using this method means that if we have few outliers, the explanation is more compact. This makes sense because the explainee's attention is directed to the outliers and having fewer outliers means having fewer things to pay attention to.

Distinctiveness Explanations are distinctive if they use unique features. A feature is unique if it has a value it has in only a few cases. In the case of natural language classification, a unique feature would be a word that is not used often. According to the distinctiveness axiom, a more unique feature should have a higher attribution. The way this is measured is: the words in the sentence are ranked on uniqueness. The distinctiveness of the explanation can then be measured by multiplying the relevance of each feature by the inverse of the rank. So a word that comes early in the ranking (for example 1st place) should have a higher relevance than a word that comes late in the ranking to maximize distinctiveness.

Contrast A feature-relevance explanation is contrastive if the features (words in the case of natural language classification) with high relevance

make a difference to the prediction. In this thesis, it is measured by removing the word and running the model again. If the prediction has changed, the feature should have high relevance. This way of computing contrast assumes feature independence and might not work well if the prediction of the model depends on multiple words. However, for simplicity as well as for computational efficiency, combinations of features are not considered.

Realism For all classes of a prediction, a set of keywords related to that class is selected. For example, when one of the classes is electronic products, keywords would include names of electronic brands or specifications specific to electronic products (e.g. 4G, Wi-Fi, etc.). Identification of keywords is done by a domain expert, potentially with the help of texts that are related to each of the classes of the prediction. For all words used in the explanation, the similarity between that word and the prediction class's words is computed using WordNet. WordNet is a large database of English words. We can use WordNet to compute similarities between words. For words that are not in WordNet, we can test whether they are exactly equal to one of the keywords. The more similar a word is to one of the keywords, the higher the relevance should be for a realistic explanation.

Chapter 4

Axioms from Human Surveys

The next step is answering the question "How can the functionally grounded evaluation methods be improved, according to people?"

4.1 Designing the interview

4.1.1 Why interviews

In this thesis, the aim is to create a functionally grounded evaluation method that sufficiently measures understandability. Because human-grounded evaluation is biased towards understandability, interviews will help us achieve this goal because they allow us to learn how people evaluate these explanations. Interviews allow us to go into detail on the thought processes of the participants, while they are evaluating the explanations. This allows us to mimic (part of) their evaluation process in the functionally grounded evaluation method. A benefit of questionnaires would be that they can test a bigger number of subjects. However, the reasoning behind an evaluation is important in order to create an evaluation method, so interviews are better suited for this purpose.

4.1.2 The interview

The interview consists of two parts. The first part contains general questions about the quality of explanations. The second part contains specific questions about explanations of machine learning models from varying domains. The second part includes asking the participant to select the best explanation and asking the participant to create explanations. The interview lasts around 45 minutes like in (Felt et al., 2012) to accommodate for the participant's attention span.

Before the first part, an introduction describes XAI and its purpose. Then two questions are asked. The first question asks the participant what the characteristics of explanations should be, given the purpose of XAI. The

second question asks the participant to evaluate the characteristics from section 3.2. This first section was intended to roughly indicate what the evaluation method should look like in general. Besides, it functioned as an independent opinion on the validity of the literature review.

Next, the participants were given an explanation of four different domains in which AI models and explanations were created. These were disaster risk management, product classification, sentiment analysis, and spam filtering. These domains were selected based on their applicability to natural language classification and the quality of existing datasets. Then the second part started and two questions were asked. The first question asked the participant to take the role of the AI models and make classifications and explanations. The purpose of this question was to learn how humans make explanations (which were assumed to be understandable). The second question asked the participants to evaluate the explanations from the AI models by ranking them. For the same sentence and classification, multiple explanations were made. The participants were asked to explain why some explanations were preferred over others. This part gave us insight into the participants' thinking process in evaluating and creating explanations. This helped us create the evaluation method.

Last, the participants were asked whether their answers in the first part were also true for the feature relevance explanations that were shown to them. Also, they were asked whether there are characteristics specific to feature relevance explanations.

To reiterate, we asked the participants:

- To think about the characteristics of a good explanation; (Part 1)
- To rank the axioms mentioned in section 3.2 (fidelity, completeness, compactness, distinctiveness, contrast, and realism); (Part 1)
- To take the role of the machine learning model and create classifications and explanations for sentences; (Part 2)
- To rank the sets of explanations and explain their ranking; (Part 2)
- Whether their answers in part 1 also apply to the explanation type in part 2 (feature relevance explanations).

The full interview can be found in the appendix, sections A.1 and A.2. An analysis of the results can be found in section 4.2.

4.1.3 Process of designing

In the beginning stages of designing the interview, a few questions were created based on intuition and the literature section, with the goal of understanding how people evaluate explanations in mind. Here follows a prioritized list of the questions, from most to least valuable question:

1. **Which explanation is better?** - The participant will get a sentence with its classification combined with some explanations for that same sentence that fit different axioms. The way these explanations are made is explained in section 4.1.5. The participant then is asked to rank these explanations. We specifically asked for a ranking, rather than a rating of each explanation separately because a ranking will give more consistent results between users and over time (Freund et al., 2003).

2. **Replicate this explanation** - An explanation is shown to the participant that they can study. They are then given a different task as a distraction. After that task, they are asked to replicate the explanation. When an explanation is more understandable, it is expected to be easier to remember. Craik and Lockhart(Craik & Lockhart, 1972) state that "later stages [of perception] are concerned with pattern recognition and the extraction of meaning" (page 675). The depth of processing is defined to be the number of layers of perception that are involved. They state that a greater depth of processing "implies a greater degree of semantic or cognitive analysis" (page 675) and a greater depth of processing implies a stronger memory trace (Craik & Lockhart, 1972). This question/task is preferred because it has the participant's memory as a proxy of understanding, rather than their interpretation of their own understanding. This is expected to be a more reliable measure.

3. **How would you describe a good explanation?** - The participant is asked to come up with axioms. When the axioms proposed in this thesis are mentioned often by the participants, the axioms gain credibility. This question can be answered quickly.

4. **Would you say axiom X helps you understand explanations?** - The participant is directly asked whether they think an axiom is good. The importance of axioms from both existing as well as new evaluation methods can be asked. There will be a lot of subjectivity and differences between participants because the term 'important' is open to interpretation. But because the question can be answered in very little time, the limited benefit is worth the cost.

5. **Create an explanation** - The participant is asked to create an explanation to see what axioms fit their explanation. Humans are expected to give understandable explanations. If the axioms are correct, the explanations from participants should also fit them.

6. **How would you improve this explanation?** - The participant is given one explanation (fitting some of the axioms) and the question of

how to improve this explanation. If they (indirectly) mention one of the axioms in their improvements, the axiom gets credibility.

7. **What process did you use to understand the explanation?** - The participant is given an explanation with an associated task and is asked to explain their thinking process. The participant is expected to show which parts of the explanation are important and how they use them in order to reason. This may give us indirect information on what axioms are important for their thinking process.

8. **What do you like about the explanation?** - The participant is given an explanation and asked what they like about it. They can mention the axioms in their judgment.

After having made the questions, the first pilots were done with a convenience sample of three people. We concluded that participants needed more context in order to answer the questions properly. So an introduction was added.

Then the interview was improved over a number of iterations. By discussing with the team, it first became clear that the questions in the interview should be applied to a domain. This would make the results more consistent and more valid because it helps the participants understand the purpose of the explanation. We initially chose to base the interview on the domain of disaster risk management because of the supervisors' experience in the domain. After a few iterations of improving the background story and questions, we came to the conclusion that the results would be too biased because of the domain. Consequently, we decided to include more domains in the interview, besides disaster risk management: these were e-commerce, sentiment analysis, and spam classification. The natural language used did not require the participants to be domain experts.

With the new domains added, the questions were divided into two parts. The first part contains general questions. These questions are about all types of explanations. The second part contains questions that are specific to explanations generated using SHAP. This means that the questions involve a participant evaluating an explanation. We decided to remove question 2 ("Replicate this explanation"), as it will take too much time to fit within an interview of 45 minutes. Asking only a few questions of this type could be a solution. However, this increases bias to the particular explanations presented in the question, so we decided to skip the question. Questions 6 ("How would you improve this explanation?") and 7 ("What process did you use to understand the explanation?") were also removed because they provide too little value compared to the time required. Question 8 ("What do you like about the explanation?") was adapted to fit within question 1 ("Which explanation is better?"), asking the participant to justify their prioritizing of the questions. The final interview contained question 3

("How would you describe a good explanation?"), question 4 ("Would you say axiom X helps you understand explanations?"), question 5 ("Create an explanation"), and question 1 ("Which explanation is better?").

Further improvements were made in providing the participants with context over the course of meetings with the team. Also, an informed consent form was written.

4.1.4 Sampling method

We interviewed 12 people in total (Guest et al., 2006) (Moitra et al., 2022). Half of them had experience in machine learning. The other half did not. Experience with machine learning was expected to affect the results because it gives the participant an expectation of how machine learning explanations work or should work. We used a combination of convenience sampling and purposive sampling (Oates, 2006).

4.1.5 Explanations used in the interview

One of the questions in the interview ("Which explanation is better?"), requires the use of explanations. In this paragraph, we will explain how these explanations were picked. First, for every sentence 5 explanations were generated for 50 sentences for each domain. Three of those explanations were made by the permutation explainer (because of its ability to explain higher-order interactions), one was made by the partition explainer, and one was made by the sampling explainer. These explainers are all approximations of SHAP, as described in section 2.3.4. The axioms that were developed in chapter 3 result in a number, as explained in sections 3.2, 3.3.3, and A.5 (from least to most specific). When the axiom for one explanation exceeded the 3rd quartile of that axiom for all explanations, that one explanation is considered to fit the axiom. For example, if we have eight explanations with fidelities of 0, 0.3, 0.3, 0.5, 0.5, 0.7, 0.8, and 0.8, then the last two explanations would fit the axiom fidelity because they exceed the 3rd quartile (which is 0.7). Similarly, when the axiom associated with an explanation is lower than the 1st quartile of that axiom for all explanations, it is considered not to fit that axiom.

In the case of realism and compactness, there are clear distinctions between explanations that do not adhere and explanations that do adhere. For realism, the thresholds are 0.4 or lower to not adhere to the axiom and 0.6 or higher to adhere to the axiom. If an explanation has 4 or more 'outliers' (words with a relatively high relevance), the explanation is not compact. If it has 2 or fewer 'outliers', it is compact. In this case, an outlier was defined as being higher than 60 percent from the lowest to the highest value. For example: if you have values spanning from 0 to 1, an outlier is defined as any value above 0.6. With this value, about half of the explanations were

compact and the other half were not.

4.2 Findings

In this section, the findings from the interviews will be described. Analogous to the two parts of the interview - the general and specific part -, this section will discuss people's general opinions on the axioms of good explanations and then discuss the qualities of specific explanations. First a list of the participants. The 'e' in the label of the participant indicates experience with machine learning.

4.2.1 Participants

P1e Has a Ph.D. based on machine learning.

P2e Has a master's and Ph.D. in explainability and transparency of machine learning systems.

P3e Has 10 years of experience in working with machine learning.

P4e Master's student in the machine learning domain, including explainability.

P5e Master's student in the machine learning domain.

P6e Has 4 years of experience in working with machine learning, including explainability research.

P7 Knows about machine learning from the media.

P8 Knows about machine learning from social media.

P9 Has some conceptual knowledge of machine learning, but never used it in practice.

P10 Has taken a course on the ethics of machine learning and uses AI chatbots.

P11 Uses AI chatbots.

P12 Knows nothing about AI.

4.2.2 Evaluating axioms from the literature

First, let us describe the participants' opinions on the axioms for explanation quality that were found in the literature, answering the question "Would you say axiom X helps you understand explanations?" Everyone strongly agreed with the axiom of 'fidelity'. Many put it as a first priority when evaluating

explanations. All people agreed with the axiom 'contrast'. P6e refined the axiom 'contrast' by saying that an explanation should "show what distinguished this case [...] from a group [of cases, but] not from individuals." 4 of the people with experience, and all people without experience in machine learning agreed strongly with 'realism', P2e even claimed that "explanations need to make sense in the real world [because] otherwise they're useless". All people with experience in machine learning agreed with 'distinctiveness,' however only one of the people without experience in machine learning agreed with it. Only 5 of the 12 people deemed 'compactness' important, and 'completeness' was considered important by only 4 of the 12 people. P9 argued that the ranking of these axioms depends on the goal of the explanation: "the order is dependent on what kind of explanation you want. So, sometimes you need compactness, for example, but if you want to have a thorough understanding of something, then completeness might be more important." After showing the types of explanations in part 2 of the interview, one participant removed the axiom completeness, while another participant added it. This could be due to a difference in interpretation of the axiom. In aggregate, the participant's opinions on the axioms stayed the same.

4.2.3 Axioms from participants

Participants' responses to the question "How would you describe a good explanation?" could be categorized into three main themes: factual, user dependent, and contextual.

Factual

The first category was that the explanation should be based on true facts. People without experience in machine learning defined facts as true when they hold in the real world. People with experience in machine learning defined facts as true when they hold according to the model. These two definitions of truth are analogous to the axioms realism and fidelity, respectively.

P7, P8, P9, and P10 said that an explanation needs to be based on the truth. P7 and P9 even claimed that the definition of an explanation is a set of true facts that are used as proof of the classification. P11 and P12 explained this differently. They said that explanations should be based on scientific facts or theory. In addition, P11 said that the explanation should use a lot of statistics. P11 said about statistics that "you can see exactly what [the model] values more and I would [also] say that is part of transparency."

P6e defined an explanation as being 'true' as follows: an explanation "needs to identify what was actually the basis of the decision." Or as phrased

more informally by P2e: it should "actually explain what it's trying to explain." P4e even indicated that accuracy trumps understandability: "It should explain the actual decisions the model does, even if that means giving an explanation that the human doesn't find intuitive".

User dependent

The second category is that explanations should be user dependent: the quality of an explanation depends on the user. As explained in sections 2.1.3 and 2.2.3, the social-scientific literature also states that explanations should take the target audience in mind.

P2e said: "if the explanations don't match what [users] think, they will call it a bad explanation." This means that understandability and people's expectations of explanations are also important. People's backgrounds have an effect on their expectations: P1e said that different people have different expectations of explanations. For example, "as a computer scientist [...] we expect some numerical values when we explain something [...] but when we go to doctors, they expect some kind of natural way of explaining decisions." P4e summarized this by saying: "A good human explanation is an explanation that harmonizes with the human you're talking to. Then you can see the feedback [...] and you know the background [...] and you're trying to adapt your message to the person you're talking to."

Contextual

The last category of axioms was that it should take into account the context or human expectations of the context. For example, P1e indicated that in the natural language domain, the "context of this whole text is more important than individual words." P5e said something similar when he said that "it's really important that the model not only considers the words, but it should also figure out the meaning." So that explanations should go beyond just the words, and also explain their meaning and context. P7 and P9 suggested solving this by having the explanation give a certain context in which to understand the classifications, saying "first you need an introduction" and "it would set the scene", respectively.

When it comes to the context of the use case, P3e said that "in critical use cases [...] you really need to understand why the model came to this decision", but that "for other things, it might be enough to show shallow explanations like heatmaps". Three of the participants related the dependence on the use case to the axioms of completeness and compactness. A more complex use case would then require completeness, while a more simple use case would then require compactness.

4.2.4 Analysis of human explanations

In this part of the interview, participants made classifications and explanations for sentences in the given domains (disaster risk management, product classification, sentiment analysis, and spam filtering). It answers the question called "Create an explanation." Explanations made by the participants could be distinguished into three categories. The explanations consisted of either a set of keywords that were the reason for the classification; a description of the context or meaning of a sentence that related to the classification; or an absence of keywords or context.

In the disaster risk management, e-commerce, and sentiment analysis domain, the explanations consisted largely of sets of keywords. These were assumed to be the correct keywords because people have an understanding of natural language. When an explanation consists of the right set of keywords, it can be considered factual. The explanations in the spam domain mainly involved the context of the text, which makes the explanations contextual. We hypothesized that the difference in explanation type was mainly due to the fact that our examples in the spam domain did not have clear keywords. For example: in the sentiment analysis domain, the word 'awesome' clearly indicates a positive sentiment and the word 'uncomfortable' clearly indicates a negative sentiment. The words in our examples in the spam domain were not as clear. For example, the word 'cheap' in and of itself does not necessarily make a text spam text: it depends on the context.

Besides the difference between spam and the other domains, there also appear to be individual differences in proclivity to use keywords or context. This could be due to explanations being user dependent. There could be a difference in familiarity with the domain or show that different people prefer different explanations. However, more research is needed in order to draw reliable conclusions about why people use different ways of explaining.

4.2.5 Analysis of machine explanations evaluations

In this section, participants' opinions on the quality of machine learning explanations are analyzed. It answers the question "Which explanation is better?" Section 3.2 hypothesized that explanations could be judged along six axioms: fidelity, completeness, compactness, distinctiveness, contrast, and realism. According to most of the interview participants, of those six axioms, fidelity, contrast, distinctiveness and realism were expected to be able to measure the quality of explanations, as described in section 4.2.2.

However, from the evaluations of machine explanations it became clear that none of these axioms were good predictors of the quality of the explanation in practice: at least using the measurements described in section 3.3.3 and in the appendix, section A.5. There could be other ways of measuring the same axioms where they would have been able to predict the quality

Contrast	Compactness	Fidelity	Realism	Distinctiveness
	-	-		-
	+	+		-
	×	×		
		×		×
-	-	-		×
×				
	-			
-				
+				
		×		×
		+	-	-
	-		×	×

Table 4.1: Accuracy of the prediction for each of the axioms. Every row represents one of the classifications. '+' means there is a positive correlation between axiom and explanation quality; '-' means a negative correlation; '×' means no correlation; blank means not measured.

of an explanation more accurately, however, this needs more investigation. For compactness and distinctiveness, the inverse of the metric even seemed to be a better predictor. This is shown in table 4.1. The method explained in section 4.1.5 determines whether an explanation satisfies an axiom or not: an explanation adheres to an axiom when its metric is above the third quartile, it does not adhere when its metric is below the first quartile. For example, when an explanation is preferred and it does not fit an axiom, but another explanation of the same classification does fit one, then there is a negative correlation.

In this section, we will describe the reasons why participants liked or disliked an explanation. In section 5.1 a new evaluation method will be proposed for feature relevance explanations in the natural language processing domain that is based on these reasons.

The three categories have been described for liking or disliking an explanation in general (as in section 4.2.3). This section discusses four common reasons for liking or disliking a feature-relevance explanation of the type shown in figure 4.1. These four categories are the high relevance of keywords, high relevance in punctuation or incorrect words, specificity, and nuance. Having a high relevance of keywords is assumed to be an obvious reason for liking an explanation because keywords give a strong indication of a certain prediction. When punctuation marks had a high feature rel-

Figure 4.1: Example of a feature-relevance explanation generated by SHAP.

evance, the participants found the explanation of lower quality. They did not consider punctuation crucial to the meaning of a text, so they found it should not have had a high relevance. When a high feature relevance was attributed to incorrect words (these are words that the participants did not associate with the given prediction), the explanations were also considered to be of lower quality. The last two categories, specificity and nuance, seemed to be related to the first category, the high relevance of keywords. When a sentence has keywords, explanations that gave a high feature relevance to the keywords were praised for their specificity (while explanations, where all words had similar feature relevances, were called unclear). However, when there were no keywords, explanations that attributed similar feature relevances to all words were considered better and more nuanced (while explanations with greater variety in feature relevances were called wrong).

In addition, some people did not like explanations where all words had exactly the same relevance. However, for some other people, this did not seem to be a problem.

Participants also mentioned that explanations they deemed to be of low quality gave them less trust in the underlying machine learning model. As mentioned in the introduction and section 3.3.2, the goal of the explanation was the justify the machine learning model. Explanations that are intended to justify the model are intended to increase trust in the model, not decrease it. This again shows that it is important that we have explanations that are of high quality.

4.2.6 Effect of experience

The main difference between people with experience in machine learning compared to people without experience was in answering the question "How would you describe a good explanation?" (section 4.2.3). While the underlying categories were similar (factual, user-dependent, and containing context), the ways in which the two groups presented the categories were different. People without machine learning experience all said that a good explanation should have the following format: first, a context should be given to set the scene and help the user understand what is going to be explained, or what question is going to be answered; then should follow a set of true facts that help the user understand what is going to be explained

or answer that question. People with machine learning experience had a much more specific idea of what an explanation was, so their axioms were more tailored to machine learning explanations (not necessarily to feature relevance explanations), rather than explanations in general.

The way explanations were made and evaluated seemed similar for both groups.

Chapter 5

Validating the evaluation method

5.1 The new evaluation method

From section 4.2, and especially 4.2.5, it became clear that the quality of the type of explanations used in the interview (feature relevance explanations of small texts) could be evaluated along the following simple formula. This formula uses the concept of keywords. These are words that are highly associated with a particular class in the classification model. The formula works as follows: if there are keywords in the text, these keywords should have significantly higher relevance than the other words in the sentence (see equation 5.1); and if there are no such keywords in the text, then there should not be words with significantly higher relevance than the rest.

$$\text{quality} = \frac{\#\text{keywords selected}^2}{\#\text{keywords in input} \times \#\text{words selected}} \tag{5.1}$$

In case there are keywords in the text, as many of them as possible should be selected, without having other words selected. We define a word to be selected when its relevance is more than 60 percent of the range of word relevance values higher than the lowest relevance. For example, if we have a sentence for which the word relevance values range from 0 to 1, then a word is called 'selected' when its relevance is higher than 0.6. The way the formula ensures that as many keywords as possible should be selected can be compared to the recall measure. The number of keywords selected can be compared to the true positive rate, which is then divided by the total number of keywords in the input that can be compared to the true positive rate plus the false negative rate. The way that the formula ensures that as few other words as possible are selected can be compared to the precision measure. Again the number of keywords selected can be compared to the true positive rate and the number of words selected can be compared to the

true positive rate plus the false positive rate.

As learned in section 4.2.5, people prefer explanations without a focus on any particular word at all when there are no keywords in the sentence. We need to measure how uniformly the feature relevance values are distributed. One way of measuring this is using the standard deviation. The implementation of the complete evaluation metric can be found in the appendix, section A.6.

The evaluation method reduces the problem of evaluating these explanations to identifying keywords. These keywords can be identified once per domain by a domain expert. The benefit of this approach is that one human judgment can be reused for multiple explanations. Depending on the domain, different types of keywords exist. For example, if we take the 'electronic product' class from product classification, the list of keywords can include brand names, types of electronic products (such as 'keyboard', 'phone', or 'dishwasher') or specifications inherent to electronic products (such as 'kWh', 'Wi-Fi', or '4G'). Eventually, every class has a list of keywords.

Words in the text that are equal to one of the keywords in the list should then have high relevance for the explanation to be good. Besides being equal to a word in the list, keywords can also be recognized by using WordNet, which has been introduced in section 3.3.3, under the realism paragraph. WordNet can then be used to calculate the similarity of a word to the keywords in the list. When the similarity exceeds a threshold (set to 0.75 in our implementation), it is also classified as a keyword. The approach we take of having a predefined list of keywords to identify them would be similar to the suggestion of P8, who suggested measuring keywords by having people list them.

There are other ways of computing the quality of an explanation, besides using formula 5.1 and the standard deviation to compute uniformity. We used formula 5.1 and standard deviation because we found them the most intuitive. The implementation of the evaluation method can be found in the appendix, section A.6.

5.1.1 Compared to people's axioms

This simple formula fits three of the four common reasons for liking or disliking an explanation from section 4.2.5: high relevance of keywords, specificity, and nuance. This gives people the best understanding and most trust in the decision of the model, according to their own judgment. The last reason, avoiding high relevance in punctuation, has not been included in the evaluation method to keep it simple. Looking at section 4.2.3, we can recognize one of the three categories for explanation quality: factual. The method judges whether an explanation is factual because it looks for the high relevance of keywords. By definition, keywords are words that are

highly associated with the predicted class so the formula measures whether the explanation is factual in the real world. The formula is not considered to measure whether an explanation is contextual because it only looks at certain words in the sentence and not at the definition of the entire sentence. It is also not user dependent, because the formula is the same for every user.

5.1.2 Compared to the literature principles

This section discusses which of the literature principles are measured by the evaluation method. Similar to section 3.1, it is largely based on the author's intuition, but explanations for the intuition are given as well.

Going back to the social-scientific principles from the literature (section 2.2), we can see that the new method covers the following principles from the literature:

- Coherence: people that use the method are assumed to have an understanding of the domain and the English language. This would mean that they agree with the keywords, so because the evaluation method focuses on what the keywords are, it measures whether it is coherent with the user's prior beliefs.
- Simplicity: only keywords should have high relevance, according to the method. The other words should not. So this validates whether the explanations fit the simplicity principle.
- Consistency: users of the evaluation method are assumed to believe that a class's keywords are associated with that class. Because the evaluation method depends on keywords, it measures whether an explanation is consistent with the user's prior beliefs.
- Selectivity: only keywords should have high relevance, according to the method. The other words should not. This means that the evaluation method tests whether the explanation fits selectivity.
- Truthful: keywords are assumed to be associated with the predicted class in the real world. Because the evaluation method judges the quality of an explanation based on keywords, it measures whether the explanation is truthful.
- Applicable: the method has been applied, so it is applicable.
- Common practice: we do not consider the method to be extraordinary, so it fits the common practice principle.

However, it does not cover these principles:

- Generality: the method looks at one explanation for one classification. It does not use other explanations, so the generality of an explanation is not measured.
- Predictive power: the evaluation method does not consider the model itself. This means we will not know whether the explanation is useful in predicting other texts. The evaluation method does not measure predictive power.
- Timeliness: the evaluation method does not consider the model itself. This means we will not know whether the explanation is useful in giving indicators for the same classification. The evaluation method does not measure timeliness.
- Distinctiveness: according to the method, the focus should be on keywords. Keywords are not necessarily abnormal, so distinctiveness is not measured.
- General and probable: the method looks at one explanation for one classification. It does not use other explanations, so it does not measure whether an explanation is general and probable.
- Contrastive: we do not look at alternative predictions of the model. This means we cannot know which words would change the prediction, so the evaluation method does not measure whether the explanation is contrastive.
- Social: the method is the same for every user and does not take the target audience in mind. It does not measure whether an explanation is social.
- Focus on abnormal: according to the method, the focus should be on keywords. Keywords are not necessarily abnormal, so focus on abnormality is not measured.
- Findable: this is, to the best of our knowledge, the first evaluation method that combines the benefits of human-grounded and functionally grounded evaluation methods. This means there is no existing basis for it in the literature, so it is not findable.

More research is required to incorporate the other principles into the evaluation method.

5.2 Validating the new evaluation method

The newly created evaluation method was tested using a small questionnaire. Before the questionnaire started, participants got an introduction to the

topic and the purpose of the thesis. The questionnaire was similar to the question that was analyzed in section 4.2.5, "Which explanation is better?". Generated explanations were shown and participants could vote on which explanation they found the best (again defined as most understandable and useful). The full questionnaire can be found in the appendix, sections A.3 and A.4.

5.2.1 Selection of explanations

The same domains were used in the questionnaire as in the interview. For every domain, explanations of four sentences were picked. Two of those sentences had keywords, the other two did not (as the presence of keywords changes the evaluation method). We picked sentences whose explanations had the biggest difference in quality, according to the new evaluation method. None of the explanations were used in the interview.

5.2.2 Selection of participants

Participants were selected using convenience sampling. None of the participants participated in the interview. In total, eleven people participated in the questionnaire who all answered 16 questions.

5.2.3 Results

Results of the questionnaire are in table 5.1. The left side of the table contains the qualities of the explanations according to the evaluation method. The right side of the table contains the number of votes each explanation got. According to the code used in the appendix, section A.7, the evaluation method performed better than random at picking the best explanation (sample size: 11 votes for every one of 16 classifications, p-value: 0.009). The way this code works is as follows. For every classification, two, three, or four explanations were made. One of those explanations is the best according to the evaluation method. The total number of votes for the best explanation is summed and divided by the total number of votes. This is the evaluation method's accuracy. Then the random accuracy needs to be computed, which is the summation for each classification of the number of votes divided by the number of explanations for that classification. The code then uses a binomial test, where the sample size is the total number of votes ($11 \cdot 16 = 176$ in this case).

When using the same code to compare the evaluation method to random choice per domain, the method worked better than random choice with p-values of 0.01, 0.27, 0.36, and 0.54 for disaster risk management, product classification, sentiment analysis, and spam filtering, respectively (sample size: 11 votes for every one of 4 classifications). So the method works best for the disaster risk management domain. When comparing the evaluation

method to random choice for sentences with keywords and sentences without keywords, the method worked better than random choice with p-values of 0.20 and 0.01, respectively (sample size: 11 votes for every one of 8 classifications). So especially for sentences without keywords, the evaluation method works better than random choice.

The new evaluation method is compared to one other evaluation method. The other evaluation method is a combination of fidelity and stability, from section 2.4.1. The other evaluation methods presented in section 2.4 do not focus on particular explanations (but on the explanation method). The quality of the other evaluation method is assessed in table 5.2. The values in the table were based on the metrics for fidelity and stability, defined in section 2.4.1. The algorithms that were used have been defined in the appendix, section A.7.2. For each question, the table contains either a '-', a '+', a '×', or a blank cell. For each question, we determined the explanation with the most votes in the questionnaire. If the difference between the explanation with the second most votes was 1 or less and one of them contained the highest or lowest value for fidelity or stability, a '*times*' was put in the table. If the explanation with the most votes got the highest value for fidelity or stability, a '+' was put in the table, unless there was another explanation with a value of at most 0.1 less. Similarly, if that explanation got a lower value for fidelity or stability, a '-' was put in the table, unless there was another explanation with a value of at most 0.1 more.

Compared to the functionally grounded evaluation method involving the axioms of stability and fidelity (Kalousis et al., 2007), the new evaluation method works better. As we have seen during the interviews, (internal) fidelity is not a good predictor of explanation quality. This was also the case in the questionnaire as can be seen in table 5.2. The stability axiom also does not seem to be a good predictor of explanation quality for feature relevance explanations of short sentences.

5.3 Limitations

5.3.1 Models used

While the explanation method, SHAP, is widely applicable, the interviews have reviewed only the classification of short texts (of at most 64 words). We chose this to keep the interviews shorter and more interesting for the participants. The drawback is that the axioms could, however, work differently for longer texts or different types of NLP tasks (such as translation).

5.3.2 Purpose of explanation

As mentioned in the literature research, explanations in artificial intelligence can have several purposes: justifying, controlling, and improving the model,

Evaluation method's quality				Participants' preference			
Opt 1	Opt 2	Opt 3	Opt 4	Opt 1	Opt 2	Opt 3	Opt 4
0	0.75	0		0	7	4	
0.34	0.22			9	2		
0.31	0.35	0.21		1	7	3	
0.5	1	0.25	0.5	4	3	0	4
0	0.04	0.06	0.07	2	2	2	5
0.62	0.61	0.57	0.42	4	5	1	1
0.63	0.61	0.42	0.56	3	0	5	2
0	0.33	0	0	1	2	0	8
0.36	0.44	0.31	0.48	2	4	2	3
0.52	0.40	0.50	0.51	1	5	4	1
0.2	1	0	0.07	7	1	2	1
0	0.5	0	0.5	1	8	2	1
0.52	0.55	0.65	0.52	1	0	6	4
0.13	0.06	0.5	0.13	0	7	0	4
0.4	0.38	0.25	0.14	2	0	7	2
0.56	0.50	0.67	0.55	2	2	5	2

Table 5.1: Evaluation method's quality (left size) vs number of votes indicating participants' preferences (right side)

Fidelity	Stability
-	+
-	+
×	-
	-
-	-
+	+
+	-
	-
	×
	-
-	
-	+
-	-
+	+
-	+

Table 5.2: Accuracy of the evaluation for each of the axioms. Every row represents one of the classifications. '+' means there is a positive correlation between axiom and explanation quality; '-' means a negative correlation; '×' means no correlation; blank means not measured.

and learning from the model (Adadi & Berrada, 2018). The explanations presented in the interview were presented in the context of justifying the model. Explanations in the context of controlling, improving, or learning from the model could require different characteristics than explanations to justify.

5.3.3 Measurement of axioms

The axioms 'distinctiveness', 'contrast' and 'fidelity' could not be measured or known by the participants in question 4 of the interview. Distinctiveness depends on the other training samples, which the participants had not seen. Contrast and fidelity both depend on the model, which cannot be known by the participants, because the model is a black box model. Consequently, the distinctiveness, contrast, and fidelity axioms could not be accurately validated in the interview (we could only ask general questions about those axioms). A solution for distinctiveness could have been to show the participants multiple data samples in order to give them the context needed to judge distinctiveness. We did not choose to do that to keep the interviews shorter and more interesting for the participants.

5.3.4 When are explanations necessary

During the interviews, four of the participants questioned the effectiveness of explanations in general. P2e said that "one of the good points of AI [...] is that it is able to make connections which we generally don't see." Explanations of those models may not make sense to people seeing them. The model may not make sense to us, but we need to think about the question: "if the model works, isn't that more important than if the explanation makes sense or not?"

P3e brought up a contradicting point, saying that "if you provide explanations, they might think: ah now I can trust [the model], even if you cannot." So the mere existence of an explanation may convince people that the model is correct, even if it is not and the explanation doesn't make sense. He claimed that "you need transparency throughout the whole system, from data collection until the explanation" in order to be able to trust the system. This includes the data distributions in the train data and real world, the used preprocessing techniques, and how the model was trained and evaluated.

In addition, P6e said that explanations require time to evaluate and that "explanations are [only] useful in the setting when you have enough time to evaluate." This means explanations are only helpful when it is important that every single prediction is correct.

Last, P9 indicated that the explanation method used in this thesis "lack[s] a lot of sophistication, like splitting up things into these basic blocks without

evaluating the combination of these blocks seems like a missed opportunity." And that "it doesn't explain why it uses various words", meaning that P9 does not consider feature relevance attributions valid explanations. These criticisms are inherent to the explanation method and could be mitigated using a different explanation method. However, this thesis focuses on feature relevance explanations and comparing different explanation methods is thus out of scope.

5.3.5 Dependence on model and input data

P2e said that the quality of the model is also important to the perceived quality of an explanation. "A good explanation for a bad model is a bad explanation." This means that the explanation will not justify the model. It could be argued that this explanation is good in the sense that it shows the model is bad. However, this requires the user to see a lot of different explanations from different models. This is because if you see one bad explanation, it is impossible to know if the model is bad or if the explanation method itself is bad.

P3e experienced that the quality of the explanation also depends on the quality of the input data, while evaluating the generated explanations. "The [text] itself is already not making sense, so it's hard making sense from the explanations. It shows well that explanations are also connected to the data. Garbage in is garbage out." In summary, a good explanation method does not necessarily imply a good explanation: the quality of the model and the input data likely impact the quality of the explanation as well.

Chapter 6

Conclusions and outlook

In this thesis, we have made a functionally grounded evaluation method for feature-relevance explanations in natural language classification For explanations of classification models, the method depends on keywords. Keywords are words that are highly one of the classes. Given an explanation consisting of a feature relevance for every word in the input text, the method works as follows. If the input text has keywords that are strongly associated with the predicted class, the quality of the explanation is the number of keywords that have a high relevance squared divided by the number of all words with high relevance and by the number of keywords in the input text, as shown in equation 6.1. If the input text does not contain any of these keywords, the quality is determined by how uniformly feature relevance values are distributed along the words. One way of doing this is by computing the standard deviation over the feature relevance values.

$$\text{quality} = \frac{(\#\text{keywords highlighted})^2}{\#\text{keywords in input} \times \#\text{words highlighted}} \tag{6.1}$$

This explanation method was found by answering the 6 questions from the introduction.

6.1 Research questions

In this section, we will provide short answers to the questions from the introduction.

6.1.1 What is an explanation from a social scientific perspective?

An explanation is an answer to a 'why'-question with an implicit 'whether'-question. Different topologies for explanations exist. For example, Aristotle distinguishes material explanations (considering the material of an object),

formal explanations (considering the form or shape of an object), efficient explanations (considering who or what caused a change to an object), and final explanations (considering the end-goal). This explanation is formed in our minds by attributing causes and selecting the best of these causes. It is then transferred socially, according to the rules for cooperative conversation.

6.1.2 How do humans understand explanations?

There are many sets of principles that describe how humans understand explanations. For example, explanations should be coherent, simple, and general.

6.1.3 What are current functionally grounded evaluation methods?

Some examples of functionally grounded evaluation methods include the combination of the axioms of stability and fidelity. Stability is the degree to which similar data and predictions generate similar explanations, and fidelity measures how well an explanation approximates the actual model. Other functionally grounded evaluation methods include the sets of metrics identity, separability, and stability; sensitivity and implementation invariance; or completeness, correctness, and compactness.

6.1.4 What functionally grounded evaluation axioms can be extracted from the social sciences?

From the social sciences and existing functionally grounded evaluation methods, we extracted six axioms that should be able to determine the quality of an explanation. These are fidelity, completeness, compactness, distinctiveness, contrast, and realism. In the interviews, it was found that most people agreed with fidelity, contrast, distinctiveness, and realism. However, when applied to natural language classification, none of these axioms could determine the quality of an explanation.

6.1.5 How can the functionally grounded evaluation methods be improved, according to people?

From the interviews, it was found that people mainly focused on keywords (words that are strongly associated with one of the classes), when it comes to feature relevance explanations in natural language classification. These keywords should have higher relevance in the explanation than other words for the explanation to be considered of higher quality. The new metric was based mainly on that finding. In case there were no keywords, people generally preferred explanations where the relevance was spread more evenly across words.

6.1.6 How does the new functionally grounded evaluation method perform compared to other functionally grounded evaluation methods?

The new evaluation method mimics human evaluations of feature relevance explanations better than randomly guessing. The evaluation method that combines fidelity and stability does not perform better than random. We can thus conclude that the new evaluation method performs better than some of the existing evaluation methods.

6.2 Future work

The domain in which the evaluation method works is in the classification of short texts. An important addition to this method is broadening the scope in which it works. This means having the method work on longer texts and with other explanation methods. Another addition is extending the method such that it takes into account other factors to the quality of the explanation. For example, one of the results from the interview which has not been incorporated into the method for simplicity was the highlighting of punctuation. Involving this (and potentially other) metrics could improve the evaluation method.

Bibliography

Abeyagunasekera, S. H. P., Perera, Y., Chamara, K., Kaushalya, U., Sumathipala, P., & Senaweera, O. (2022). Lisa : Enhance the explainability of medical images unifying current xai techniques. *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 1–9. https://doi.org/10.1109/I2CT54291.2022.9824840

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. (2021). Twitter for disaster relief through sentiment analysis for covid-19 and natural hazard crises. *International Journal of Disaster Risk Reduction*, *55*, 102101. https://doi.org/https://doi.org/10.1016/j.ijdrr.2021.102101

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). https://doi.org/10.3390/electronics8080832

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Felt, A. P., Ha, E., Egelman, S., Haney, A., Chin, E., & Wagner, D. (2012). Android permissions: User attention, comprehension, and behavior. *Proceedings of the Eighth Symposium on Usable Privacy and Security*. https://doi.org/10.1145/2335356.2335360

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, *4*(Nov), 933–969.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.2307/2699986

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. A., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *CoRR*, *abs/1806.00069*. http://arxiv.org/abs/1806.00069

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? *Field Methods*, *18*, 59–82. https://doi.org/10.1177/1525822X05279903

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models.

Honegger, M. (2018). *Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions* (Master's thesis).

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems*, *12*, 95–116. https://doi.org/10.1007/s10115-006-0040-8

Leake, D. B. (1991). Goal-based explanation evaluation. *Cognitive Science*, *15*(4), 509–545. https://doi.org/https://doi.org/10.1207/s15516709cog1504_2

Lertvittayakumjorn, P., & Toni, F. (2019). Human-grounded evaluations of explanation methods for text classification.

López, S., & Saboya, M. (2009). On the relationship between shapley and owen values. *Central European Journal of Operations Research*, *17*, 415–423. https://doi.org/10.1007/s10100-009-0100-8

Lundberg, S. (2018). Shap api reference [Accessed: 2023-05-09].

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Moitra, A., Wagenaar, D., Kalirai, M., Ahmed, S. I., & Soden, R. (2022). Ai and disaster risk: A practitioner perspective. *Proc. ACM Hum.-Comput. Interact.*, *6*(CSCW2). https://doi.org/10.1145/3555163

Oates, B. J. (2006). *Researching information systems and computing* (1st ed.). SAGE Publications Ltd.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429–447. https://doi.org/10.1037/0022-3514.65.3.429

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Rüping, S. (2006). *Learning interpretable models* (Doctoral dissertation). Universität Dortmund.

Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International*

Journal of Human-Computer Studies, *146*, 102551. https://doi.org/https://doi.org/10.1016/j.ijhcs.2020.102551

Silva, W., Fernandes, K., Cardoso, M., & Cardoso, J. (2018). Towards complementary explanations using deep neural networks [1st International Workshop on Machine Learning in Clinical Neuroimaging, MLCN 2018, 1st International Workshop on Deep Learning Fails, DLF 2018, and 1st International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, iMIMIC 2018, held in conjunction with the 21st International Conference on Medical Imaging and Computer-Assisted Intervention, MICCAI 2018 ; Conference date: 16-09-2018 Through 20-09-2018]. In Z. Taylor, M. Reyes, M. Cardoso, C. Silva, D. Stoyanov, L. Maier-Hein, S. Pereira, S. Kia, I. Oguz, B. Landman, A. Martel, E. Duchesnay, T. Lofstedt, A. Marquand, & R. Meier (Eds.), *Understanding and interpreting machine learning in medical image computing applications - first international workshops mlcn 2018, dlf 2018, and imimic 2018, held in conjunction with miccai 2018, proceedings* (pp. 133–140). Springer Verlag. https://doi.org/10.1007/978-3-030-02628-8_15

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., ... Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, *20*, 938–947. https://doi.org/10.1016/S1470-2045(19)30333-X

Velmurugan, M., Ouyang, C., Moreira, C., & Sindhgatta, R. (2020). Evaluating explainable methods for predictive process analytics: A functionally-grounded approach.

Vollert, S., Atzmueller, M., & Theissler, A. (2021). Interpretable machine learning: A brief survey from the predictive maintenance perspective. *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 01–08. https://doi.org/10.1109/ETFA45728.2021.9613467

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr.

Appendix A

Appendix

A.1 The interview

Artificial intelligence (AI) has the ability to improve lives in many different ways. However, in many cases, it is important to be able to trust the technology and understand its limitations. For example, we do not want bias in our system. We can do this by letting the computer explain itself.

During this interview, we would like to learn when such explanations are good. The interview consists of two parts. During the first part, we will ask you some open-ended questions about explanations. We would like to learn how you would evaluate the quality of an explanation, in general. In the second part, we would like to learn how you evaluate explanations in more specific cases. We will show you computer-generated explanations from different domains: disaster risk management, e-commerce, spam filtering, and sentiment analysis.

Before we begin, we would like to know about your experience with machine learning.

0. What is your current experience with machine learning?

Let us begin with the first part: investigating explanations using general questions.

1. We just gave you a vague description of such an explanation: something that conveys how the computer works in order to understand and trust it; to understand why the computer made the predictions it did. What would such an explanation require? Or phrased in a different way: what are the characteristics of a good explanation, according to your intuition?

2. Before the interview started, we found a few characteristics as well. Without having seen any explanations, can you give your considerations in ranking the following six statements?

(a) Explanation of a model should say what the model bases its decision on. [fidelity]

(b) One explanation should explain as many cases as possible. [completeness]

(c) Short explanations are generally better than long ones. [compactness]

(d) Unique characteristics of an event should be used in an explanation. [distinctiveness]

(e) Explanations of a prediction should mention what separated the prediction from other predictions. [contrast]

(f) Explanations need to make sense in the real world (also if the model does not know the real world). [realistic]

For the second part of the interview, we created four artificial intelligence models in four domains: disaster risk management, e-commerce, spam filtering, and sentiment analysis. Let us first explain the use cases for each of the models.

- **Disaster risk management:** Imagine there has just been a flood in Queensland, a state in Australia. You are part of a disaster risk management team and you want to get the latest updates about the flood from social media. This can give your team crucial information about the flood. You use artificial intelligence to filter the posts, such that you get only posts that are relevant to the flood. This saves much-needed time, compared to filtering these posts by hand. We let this artificial intelligence explain itself to justify whether you are getting all social media posts that are relevant to the flood (not just a small part of them).

- **E-commerce:** Imagine you have an online store that automatically resells books and electronics from other stores. Your competitor (another online store) does not categorize its products, but you want your website to do so for a better customer experience. You want your website to automatically categorize a product based on its description. This saves you a lot of time, compared to doing it manually. You let this artificial intelligence explain itself to justify whether its prediction is based on valid reasons.

- **Spam filtering:** Imagine your children get a lot of emails from robots or scammers. You do not want your children to be exposed to the harmful content in some of those emails, so you use an app to filter them. The app uses an AI to make this decision. The app also gives an explanation of its decision. You test the app on your own inbox and

you investigate the explanation to justify whether the app's prediction is based on valid reasons.

- **Sentiment analysis:** Imagine you want to decide to buy stocks of an airline company. The diligent investor you are, you want to know what people on social media think of the company. Investigating social media content allows you to see the company from multiple angles. You use an AI that automatically determines whether a review is positive or negative. Because there are thousands of reviews, you cannot do this by hand. You let an artificial intelligence model explain itself to understand why the prediction was made and whether the prediction was based on valid reasons.

Before we go to evaluating the actual explanations, we would like to know how humans would make such an explanation. So we are putting you in the computer's place.

3. We will give you a sentence that the computer can use to make a prediction and explain why it made that prediction. Keep in mind that for some of the sentences, the prediction may be obvious to you. But we are interested in the explanation anyway, because the AI can make mistakes in explaining predictions that are obvious to people.

During the next part of the interview, one type of explanation will be investigated. We will show you some explanations of this type that the computer made. We would like to ask you to evaluate them based on how useful and understandable they are. The explanations that the computers made are the words with a background color that indicates how important the word is to the prediction.

4. We will give you a few pages of explanations generated by a computer. A little clarification is needed to make clear how these explanations work. Before making the prediction, the computer removes certain abundant words (such as 'a', 'to', and 'but') because they generally add little to the meaning of the sentence. The color behind a word indicates how strongly it affects the prediction of the model. Given this information, we ask you to rank the explanations on each page. We would like to ask you to evaluate whether the prediction of the computer was based on the right words. You are encouraged to explain your thinking process in detail (so as to explain why you prefer some explanations over others).

Now you have seen what the computer's explanations look like. We would like to ask you the same questions as in the beginning. But this time, we ask them about the type of explanation we presented in the previous question.

5. You indicated that the characteristics [CHARACTERISTICS] are important to the quality of an explanation. Is this also the case for the type of explanation that we presented, and why?

6. What are other characteristics of good explanations of the type we presented in the last question (so with the color marking)?

7. How would you determine/measure [CHARACTERISTICS]? If it helps, you can look at some explanations from the previous question.

Thank you for participating.

A.2 Explanations shown

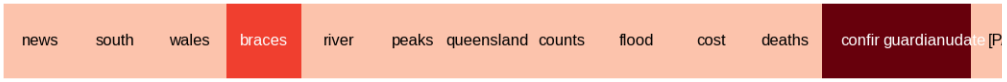
Please rank the following explanations for this prediction:
Domain: disaster;
Prediction: not relevant
Input: "@TW_Baron my picks 1,4,8,7 for this race SirBaron,will miss the race tonite working"

- 

- 

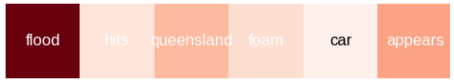
Please rank the following explanations for this prediction:
Domain: disaster;
Prediction: relevant
Input: “#news: New South Wales braces for river peaks as Queensland counts flood cost: Four deaths confir... http://t.co/LPfxlnWZ #guardianudate”

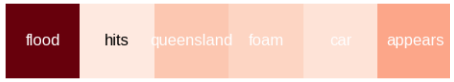
- 

- 

- 

Please rank the following explanations for this prediction:
Domain: disaster;
Prediction: relevant
Input: “Flood hits Queensland and foam car appears out of nowhere http://t.co/qYmY64v3”

- 

- 

- 

Please rank the following explanations for this prediction:
Domain: e-commerce;
Prediction: electronic
Input: “Lenovo Tab4 8 Plus Tablet (8 inch, 64GB, Wi-Fi + 4G LTE + Voice Calling), Aurora Black”

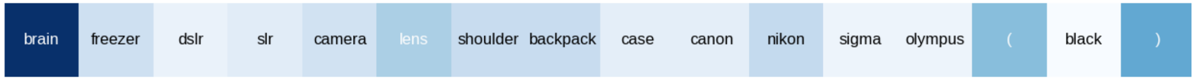
- 

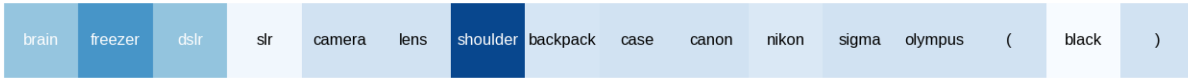
- 

- 

Please rank the following explanations for this prediction:
Domain: e-commerce;
Prediction: book
Input: “Brain Freezer J DSLR SLR Camera Lens Shoulder Backpack Case for Canon Nikon Sigma Olympus (Black)”

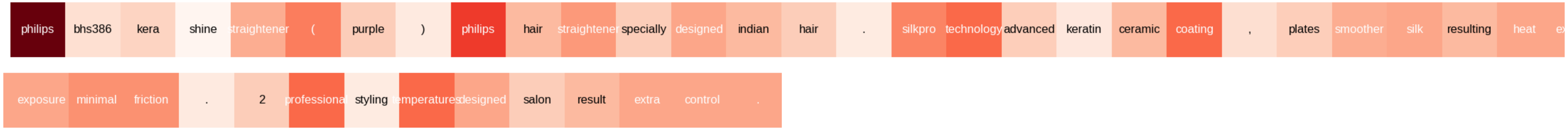
- 

- 

- 

- 

Please rank the following explanations for this prediction:
Domain: e-commerce;
Prediction: electronics
Input: “Philips BHS386 Kera Shine Straightener (Purple) This Philips hair straightener is specially designed for Indian hair. With SilkPro Care technology and advanced keratin ceramic coating, plates smoother than silk resulting in less heat exposure and minimal friction. 2 professional styling temperatures designed for salon result with extra care and control.”

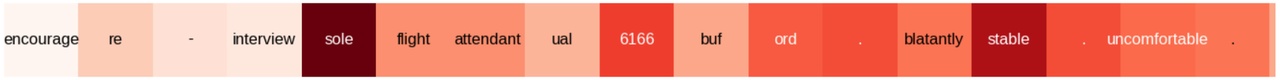
- 

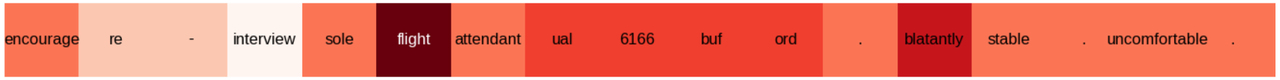
- 

- 

Please rank the following explanations for this prediction:
Domain: sentiment analysis;
Prediction: negative
Input: “@united I would encourage you to re-interview the sole flight attendant on UAL 6166 from BUF to ORD. Blatantly not stable. Uncomfortable.”

- 

- 

- 

- 

Please rank the following explanations for this prediction:
Domain: sentiment analysis;
Prediction: Positive
Input: “@JetBlue oh. And thank you for responding”

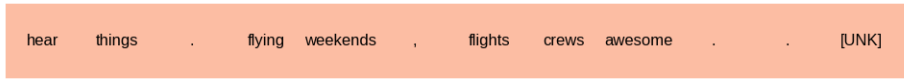
- 

- 

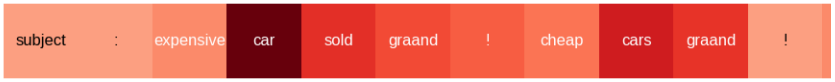
- 

Please rank the following explanations for this prediction:
Domain: sentiment analysis;
Prediction: negative
Input: “@SouthwestAir you only hear about the bad things. Flying the last to weekends, the flights and crews were awesome. Thank you. [thumbs up emoji]”

- 

- 

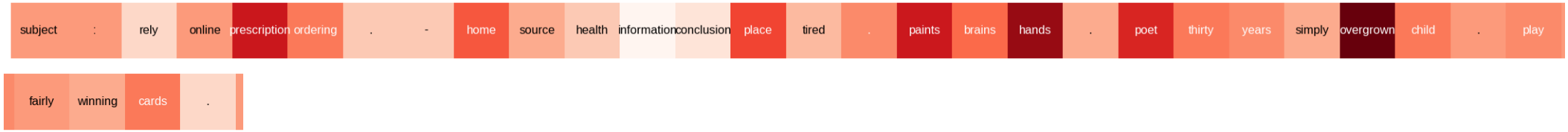
Please rank the following explanations for this prediction:
Domain: spam filtering;
Prediction: spam
Input: “Subject: the most expensive car sold in graand ! cheap cars in graand !”

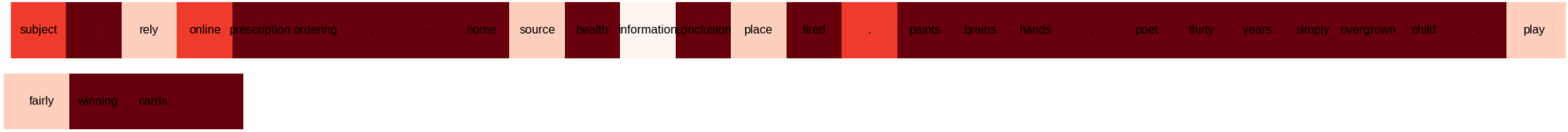
- 

- 

- 

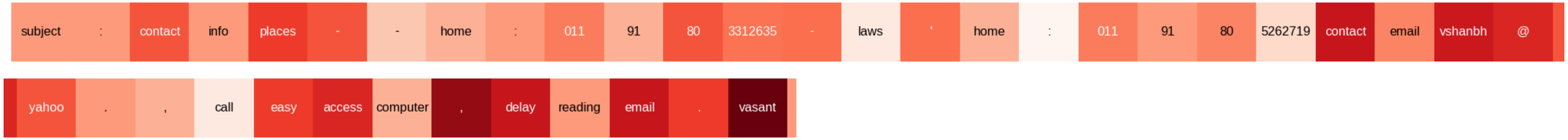
Please rank the following explanations for this prediction:
Domain: spam filtering;
Prediction: spam
Input: “Subject: rely on us for your online prescription ordering . your in - home source of health information a conclusion is the place where you got tired of thinking . a man paints with his brains and not with his hands . a poet more than thirty years old is simply an overgrown child . one shoud always play fairly when one has the winning cards .”

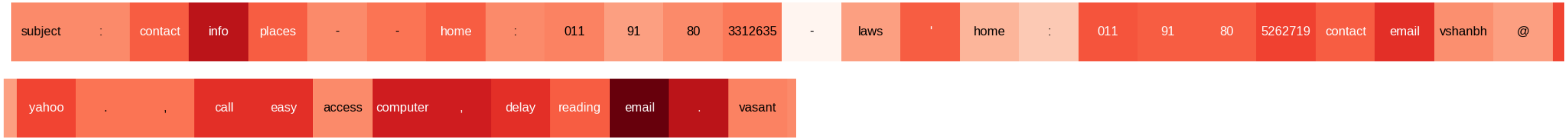
- 

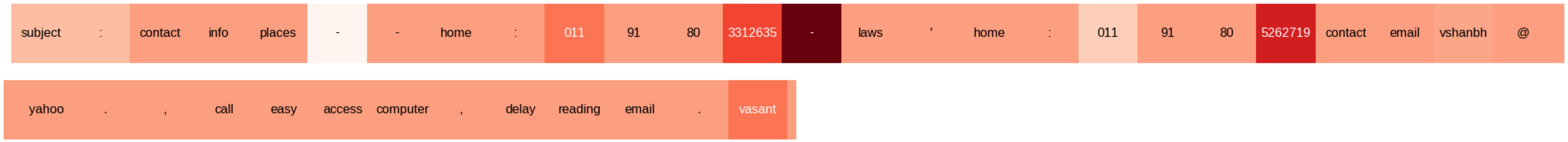
- 

- 

Please rank the following explanations for this prediction:
Domain: spam filtering;
Prediction: spam
Input: “Subject: contact info i will be in one of these two paces - - my home : 011 91 80 3312635 my in - laws ’ home : 011 91 80 5262719 you can also contact me by email at vshanbh @ yahoo . com , but it is better to call since i do not have easy access to a computer , and there may be a delay with reading email . vasant”

- 

- 

- 

A.3 Questionnaire

The questionnaire was made using Google Forms and included the following sections.

Section 1 Dear participant, here is some information about the questionnaire.

This questionnaire is part of a bachelor thesis at the Radboud University. The thesis tries to answer the question "What is a good explanation?" The answer to this question forms the basis of the product of this thesis: a metric that judges whether a (particular type of) explanation is good. This means the explanation is understandable and makes sense. Your answers will be used to validate the method that judges the quality of an explanation automatically.

Section 2 - Introduction Artificial intelligence (AI) has the ability to improve lives in many different ways. However, in many cases, it is important that we are able to trust the technology and understand its limitations. For example, we do not want bias in our system. One way of doing this is by letting the computer explain itself.

During this questionnaire, we would like to learn when such explanations are good. The questionnaire consists of sixteen questions. One type of explanation will be investigated. We will show you some explanations of this type that the computer made. We would like to ask you to evaluate them based on how useful and understandable they are.

The thesis focuses on AI models that make a prediction based on text. The explanations will show you how important different words from that text are to making the prediction.

Section 3 - Informed consent When you participate in the interview:

- You are allowed to skip any question.
- You can decide to stop your participation at any time, for any reason.
- You can download the thesis from https://www.cs.ru.nl/bachelors-theses/ once it has been completed.

Text in this interview may contain explicit content.
The questionnaire is expected to take about 10 to 15 minutes.
Do not forget to submit after you are done.

Section 4 - the four AI models The four AI models:

- Disaster risk management: Imagine there has just been a flood in Queensland, a state in Australia. You are part of a disaster risk management team and you want to get the latest updates about the flood from social media. This can give your team crucial information about the flood. You use artificial intelligence to filter the posts, such that you get only posts that are relevant to the flood. This saves much-needed time, compared to filtering these posts by hand. We let this artificial intelligence explain itself to justify whether you are getting all social media posts that are relevant to the flood (not just a small part of them).

- E-commerce: Imagine you have an online store that automatically resells books and electronics from other stores. Your competitor (another online store) does not categorize its products, but you want your website to do so for a better customer experience. You want your website to automatically categorize a product based on its description. This saves you a lot of time, compared to doing it manually. You let this artificial intelligence explain itself to justify whether its prediction is based on valid reasons.

- Sentiment analysis: Imagine you want to decide to buy stocks of an airline company. The diligent investor you are, you want to know what people on social media think of the company. Investigating social media content allows you to see the company from multiple angles. You use an AI that automatically determines whether a review is positive or negative. Because there are thousands of reviews, you cannot do this by hand. You let an artificial intelligence model explain itself to understand why the prediction was made and whether the prediction was based on valid reasons.

- Spam filtering: Imagine your children get a lot of emails from robots or scammers. You do not want your children to be exposed to the harmful content in some of those emails, so you use an app to filter them. The app uses an AI to make this decision. The app also gives an explanation of its decision. You test the app on your own inbox and you investigate the explanation to justify whether the app's prediction is based on valid reasons.

Following sections The following sections were all multiple-choice questions where the participants were asked to select the best explanation out of a set of explanations shown. The sets of explanations that were shown are in section A.4.

A.4 Explanations shown in questionnaire

Please rank the following explanations: Domain: disaster risk management
Prediction: relevant
Input: “Photos: Flood water rises in Australia”

1. 

2. 

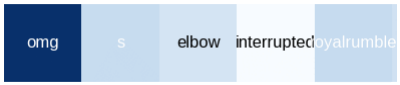
3. 

Please rank the following explanations: Domain: disaster risk management
Prediction: not relevant
Input: “I’m back :) (@ Metropolitan South Institute of TAFE) http://t.co/KfRsYzbt”

1. 

2. 

Please rank the following explanations: Domain: disaster risk management
Prediction: not relevant
Input: “WHAT OMG PEOPLE’S ELBOW INTERRUPTED WHO IS IT RoyalRumble”

1. 

2. 

3. 

Please rank the following explanations: Domain: disaster risk management
Prediction: relevant
Input: “@TIME Major Flood in brisbane,qld,australia!#qldflood http://t.co/V6CtOQ6y”

1. 

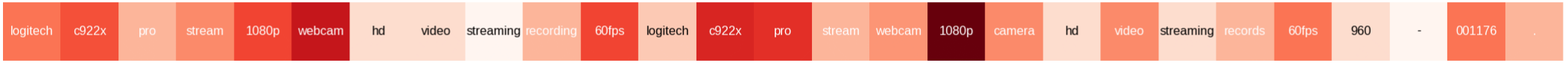
2. 

3. 

4. 

Please rank the following explanations: Domain: product classification
Prediction: electronic
Input: “Logitech C922x Pro Stream 1080p Webcam for HD Video Streaming Recording At 60Fps Logitech c922x pro stream webcam 1080p camera for HD video streaming records 60fps 960-001176.”

1. 

2. 

3. 

4. 

Please rank the following explanations: Domain: product classification
Prediction: electronic
Input: “AmazonBasics 1/2-Male to 2-Male RCA Audio interconnects - 8 feet, 2-Male to 2-Male”

1. 

2. 

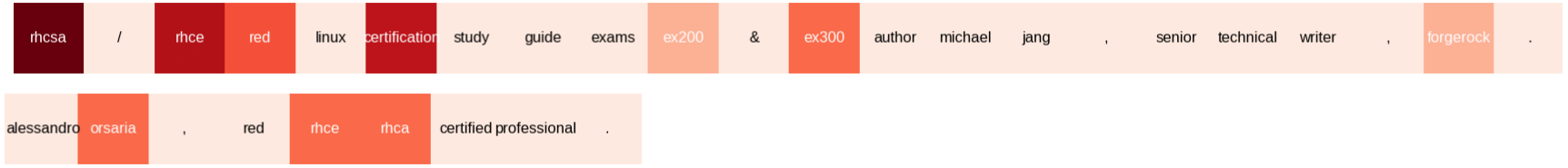
3. 

4. 

Please rank the following explanations: Domain: product classification
Prediction: electronic
Input: “RHCSA/RHCE Red Hat Linux Certification Study Guide Exams EX200 EX300 About the Author Michael Jang, Senior Technical Writer, ForgeRock. Alessandro Orsaria, Red Hat RHCE and RHCA certified IT Professional.”

1. 

2. 

3. 

4. 

Please rank the following explanations: Domain: product classification
Prediction: electronic
Input: “NISUN External Optical Drive USB 2.0 CD/DVD-RW(Read-Write) Drive for Laptop and Desktop PC -Black”

1. 

2. 

3. 

4. 

Please rank the following explanations: Domain: sentiment analysis
Prediction: negative
Input: “@united Usually an issue with Express our of SFO. Positive note: Mainline p.s. was enjoyable.”

1. 

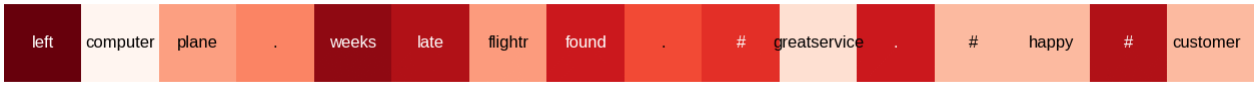
2. 

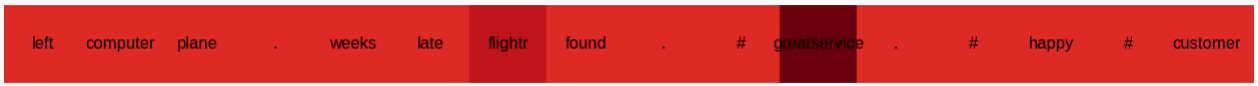
3. 

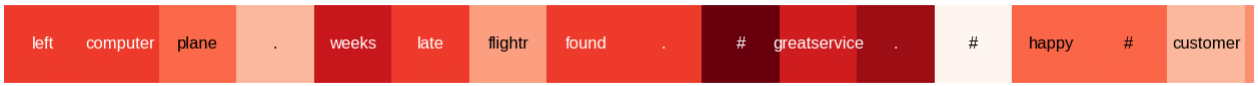
4. 

Please rank the following explanations: Domain: sentiment analysis
Prediction: negative
Input: “@SouthwestAir Left my computer on the plane. Two weeks Late Flightr they found it and sent it to me. greatservice. happy customer”

1. 

2. 

3. 

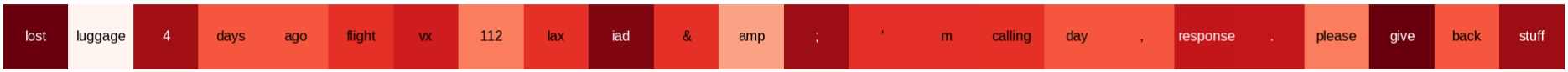
4. 

Please rank the following explanations: Domain: sentiment analysis
Prediction: negative
Input: “@VirginAmerica lost my luggage 4 days ago on flight VX 112 from LAX to IAD amp; I’m calling every day, no response.Please give me back my stuff”

1. 

2. 

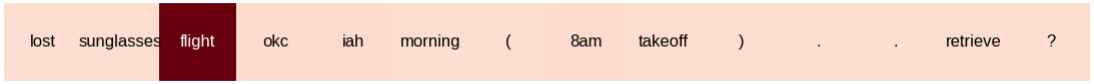
3. 

4. 

Please rank the following explanations: Domain: sentiment analysis
Prediction: negative
Input: “@united I lost my sunglasses on the flight from OKC to IAH this morning (8am takeoff) .. is there any way to retrieve them?”

1. 

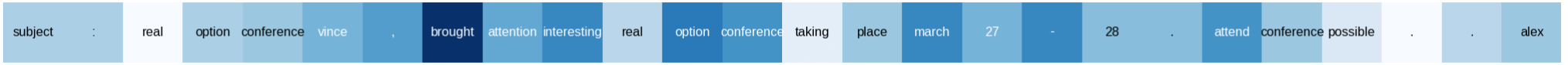
2. 

3. 

4. 

Please rank the following explanations: Domain: spam filtering
Prediction: no spam
Input: “Subject: real option conference vince , i was brought to my attention that an interesting real option conference is taking place in march 27 - 28 . i would like to attend this conference if possible . thanks . alex”

1. 

2. 

3. 

4. 

Please rank the following explanations: Domain: spam filtering
Prediction: spam
Input: “Subject: re : congratulations right back at you . . . . . great job”

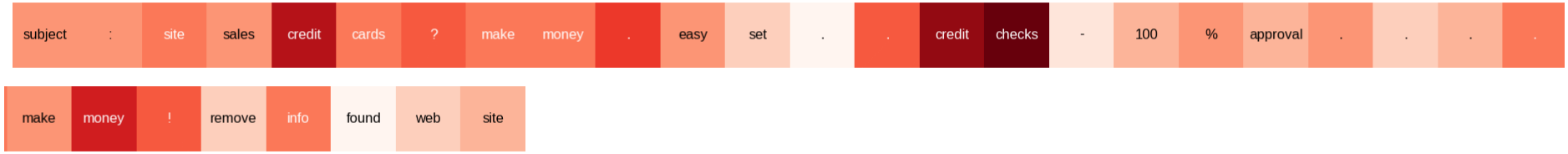
1. 

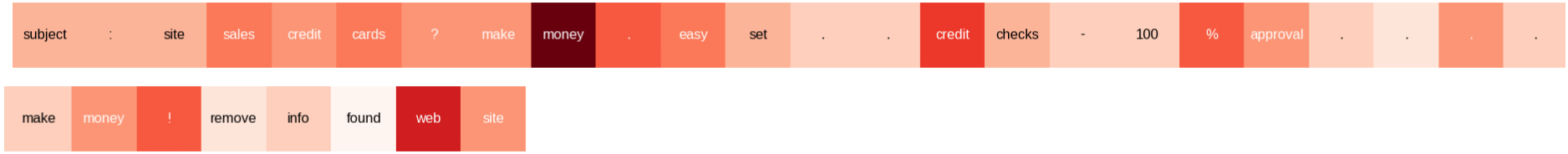
2. 

3. 

4. 

Please rank the following explanations: Domain: spam filtering
Prediction: spam
Input: “Subject: more site sales do you take credit cards ? if you do you will make more money . easy set up . . no credit checks - 100 % approval . . . . make more money now ! try now remove info is found on web site”

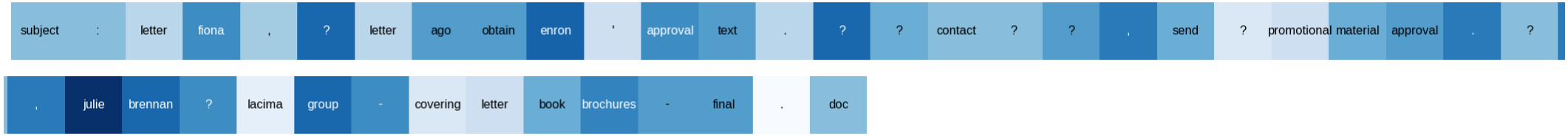
1. 

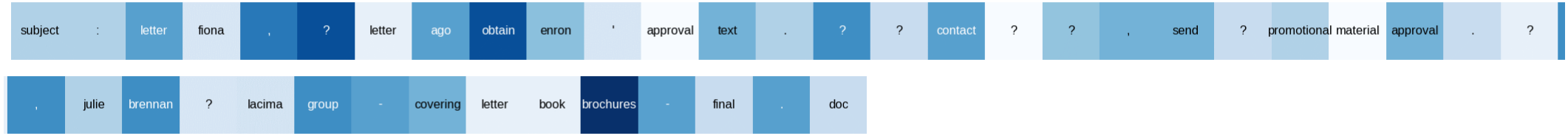
2. 

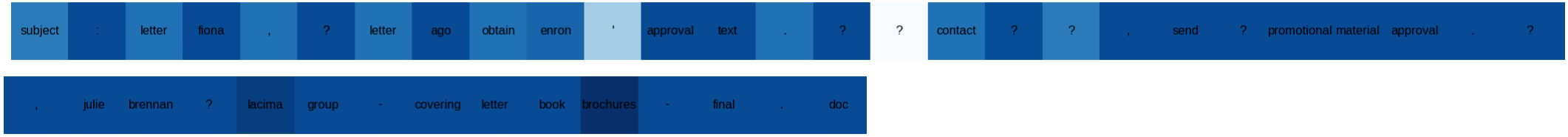
3. 

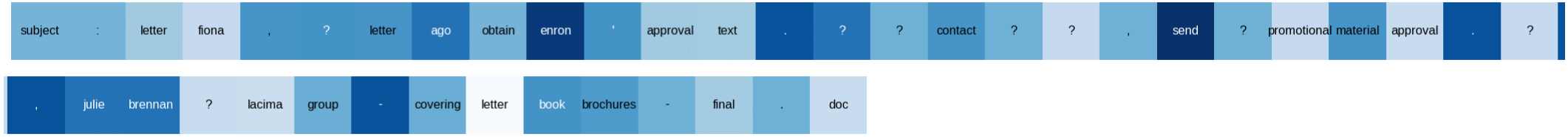
4. 

Please rank the following explanations: Domain: spam filtering
Prediction: no spam
Input: “Subject: a letter i sent fiona , ? i sent you a letter a while ago to obtain enron ’ s approval of the text . ? are you still ? my contact for this ? ? if not , please let me know who i should send the ? promotional material for approval . ? thanks , julie brennan ? lacima group - covering letter for book brochures - final . doc”

1. 

2. 

3. 

4. 

A.5 Evaluation method from literature

The evaluation method is defined as a combination of the following axioms:

A.5.1 Contrast

```
import numpy as np

def ablate(x, i):
    x[i] = 0
    for j in range(i, len(x)-1):
        x[j] = x[j+1]
    x[-1] = 0
    return x

def compute_contrast(X, model, explanation):
    """
    X: the word ids in the shape of #elements x #words
    model: the model that is being explained
    explanation: the feature relevances in the shape of #elements x #features

    returns an array of length #elements
    """
    contrast = [0]*len(X)
    for i in range(len(X)):
        old_prediction = model(np.array([X[i]]))
        for j in range(len(X[i])):
            x = ablate(copy.deepcopy(X[i]), j)
            new_prediction = model(np.array([x]))
            if (new_prediction != old_prediction):
                contrast[i] += explanation[i][j]
    return contrast
```

A.5.2 Distinctiveness

```
import numpy as np

def compute_distinctiveness(X, train_data, explanation):
    """
    X: the word ids in the shape of #elements x #words
    explanation: the feature relevances in the shape of #elements x #features
    train_data: the word ids from the train_data

    returns an array of length #elements
    """
    distinctiveness = [0] * len(X)

    frequencies = []
    for x in np.unique(train_data):
        frequencies.append((np.mean(np.array(train_data) == x), x))
    sorted_frequencies = sorted(frequencies)
    word_specialty = {0: 0}
    for i, (f, x) in enumerate(sorted_frequencies):
        word_specialty[x] = 1 - i / len(sorted_frequencies)

    for i in range(len(X)):
    for j in range(len(X[i])):
        try:
            distinctiveness[i] += word_specialty[X[i][j]] * explanation[i][j]
        except:
            distinctiveness[i] += explanation[i][j]

    return distinctiveness
```

A.5.3 Fidelity

```
import numpy as np

def compute_fidelity(X, explanation, model):
    """
    X: the word ids in the shape of #elements x #words
    explanation: the feature relevances in the shape of #elements x #features
    model: the model that is being explained

    returns an array of length #elements
    """
    fidelity = [0] * len(X)
    samples = 5
    for i in range(len(X)):
        change_ratings = [0]*len(X[i])
        x = 0
        while x < len(X[i]):
            change_rating = 0
            for change in range(samples):
                new_X = copy.deepcopy(X[i])
                new_X[x] = new_X[x] + (100 * change / samples)
                if (model(np.array([X[i]])) == model(np.array([new_X]))):
                    change_rating += 1/samples
            change_ratings[x] = change_rating
            if (X[i][x] == 0):
                x = len(X[i])
            x += 1

        for j in range(len(X[i])):
            fidelity[i] += change_ratings[j] * explanation[i][j]
    return fidelity
```

A.5.4 Realism

```
import numpy as np
from nltk.corpus import wordnet as wn

def compute_realism(sentences, explanation, positive_words, negative_words, predictions):
    """
    sentences: the words in the shape of #elements x #words
    explanation: the feature relevances in the shape of #elements x #features
    positive_words: a list of words (wordnet synsets) associated with prediction 1
    negative_words: a list of words (Wordnet synsets) associated with prediction 0
    predictions: a list in the shape of #elements containing the predictions

    returns an array of length #elements
    """
    metrics = [0] * len(sentences)
    for i in range(len(sentences)):
        if (predictions[i] == 1):
            for j in range(len(sentences[i])):
                vals = [0]
                for w in positive_words:
                    try:
                        vals.append(w.wup_similarity(wn.synsets(sentences[i][j])[0]))
                    except:
                        pass
                metrics[i] += explanation[i][j] * np.max(vals)
        else:
            for j in range(len(sentences[i])):
                vals = [0]
                for w in negative_words:
                    try:
                        vals.append(w.wup_similarity(wn.synsets(sentences[i][j])[0]))
                    except:
                        pass
                metrics[i] += explanation[i][j] * np.max(vals)
    return metrics
```

A.5.5 Compactness

```
import numpy as np

def compute_compactness(explanations):
    """
    explanations: the feature relevances in the shape of #elements x #features

    returns an array of length #elements
    """
    results = []
    for data in explanations:
      q0, q4 = np.percentile(data, [0, 100])
      iqr = q4 - q0
      upper_bound = q4 - (0.4 * iqr)
      outliers = len([x for x in data if x > upper_bound])
      if outliers == 0:
        results.append(0)
      else:
        results.append(max(0, 1.25 - (0.25 * outliers)))
    return results
```

A.6 Evaluation method from interviews

The evaluation method can be defined as:

```
def evaluate(words, ids, model, positive_words, negative_words, explanation):
    """
    words: the list of words for a prediction
    ids: the list of token ids for a prediction
    model: the prediction model
    positive words: words associated with the 'positive' class (1)
    negative words: words associated with the 'negative' class (0)
    explanation: the explanation of the prediction with size

    returns the quality of an explanation (generally a number between 0 and 1)
    """
    # remove padding
    num_words = np.sum(np.array(words) != "[PAD]")

    # Find keywords
    keywords = realism(words, positive_words, negative_words, model(np.array([ids])), num_words)
    quality = None

    # The algorithm for quality
    if len(keywords) > 0:
        if (compactness(keywords, explanation, num_words) == 0):
            quality = 0
        else:
            quality = correctness(keywords, explanation, num_words) *
                correctness(keywords, explanation, num_words) /
                compactness(keywords, explanation, num_words) /
                len(keywords)
    else:
        quality = nuance(explanation, num_words)

    return quality
```

Where the functions `realism`, `compactness`, `correctness`, and `nuance` are defined as:

A.6.1 Realism

```
def realism(sentence, positive_words, negative_words, prediction, num_words):
    """
    sentence: the list of words for a prediction
    positive words: words associated with class 1
    negative words: words associated with class 0
    prediction: the prediction (1 or 0)

    returns the indices of the realistic words, if there are any
    """
    keywords = set()
    for (i, word) in enumerate(sentence):
        if (i >= num_words):
            return keywords
        if (prediction == 0):
            for nw in negative_words:
                try:
                    if (wn.synsets(word)[0].wup_similarity(wn.synsets(nw)[0]) > 0.75):
                        keywords.add((i, word))
                except:
                    if (word == nw):
                        keywords.add((i, word))
        if (prediction == 1):
            for pw in positive_words:
                try:
                    if (wn.synsets(word)[0].wup_similarity(wn.synsets(pw)[0]) > 0.75):
                        keywords.add((i, word))
                except:
                    if (word == pw):
                        keywords.add((i, word))
    return keywords
```

A.6.2 Compactness

```
def compactness(keywords, explanation, num_words):
    """
    explanation: the feature attributions
    """
    explanation = explanation[:num_words]
    q0, q4 = np.percentile(explanation, [0, 100])
    iqr = q4 - q0
    upper_bound = q4 - (0.4 * iqr)
    outliers = len([x for x in explanation if x > upper_bound])
    return outliers
```

A.6.3 Correctness

```
def correctness(keywords, explanation, num_words):
    """
    keywords: the list of keywords + their indices
    explanation: the feature attributions
    """
    explanation = explanation[:num_words]
    q0, q4 = np.percentile(explanation, [0, 100])
    iqr = q4 - q0
    upper_bound = q4 - (0.4 * iqr)
    result = 0
    for (index, word) in keywords:
        if explanation[index] > upper_bound:
            result += 1
    return result
```

A.6.4 Nuance

```
def nuance(explanation, num_words):
    """
    explanation: the feature attributions
    """
    explanation = explanation[:num_words]
    if (np.max(explanation) == np.min(explanation)):
        return 1
    return 1 - np.std(explanation)/np.std([np.max(explanation), np.min(explanation)])
```

A.7 Validating the evaluation method

A.7.1 Evaluating the new method

```
def compare_predictor_vs_random(tuples):
    """
    tuples: a list of tuples (chance, predictor, actual), where:
        chance: random chance
        predictor: normalized quality according to the evaluation method (where the highest explanatio
        actual: the number of votes for that explanation
    """
    total = len(tuples) * n_answers
    correct_predictor = 0
    correct_random = 0

    for chance, predictor, actual in tuples:
        if predictor == 1:
            correct_predictor += actual
        correct_random += actual * chance

    predictor_accuracy = correct_predictor / total
    random_accuracy = correct_random / total

    p_value = stats.binom_test(correct_predictor, total, random_accuracy)

    if predictor_accuracy > random_accuracy:
        return f"Evaluation method performs better than random chance (p-value: {p_value})."
    elif predictor_accuracy < random_accuracy:
        return f"Evaluation method performs worse than random chance (p-value: {p_value})."
    else:
        return f"Evaluation method performs the same as random chance (p-value: {p_value})."


result = compare_predictor_vs_random(flat_chance_evaluation_actual)
print(result)
```

A.7.2 Evaluating fidelity and stability

```
def fidelity(x, model, num_words):
    """
    x: the input words
    model: the prediction model
    """
    keywords = []
    samples = 5
    for (i, w) in enumerate(x):
        if (i >= num_words):
            return keywords
        if (w != 0):
            change_rating = 0
            for change in range(samples):
                new_x = copy.deepcopy(x)
                new_x[i] = new_x[i] + (100 * change / samples)
                if (model(np.array([x])) != model(np.array([new_x]))):
                    change_rating += 1/samples

            if (change_rating > samples * 3 / 4):
                keywords.append((i, w))

    return keywords

def stability(X, explanation, explainer, n_words):
    """
    X: the input ids
    explanation: the explanation of the prediction of X
    explainer: the SHAP explainer used to make the explanation
    n_words: the number of words in the sentence
    """
    new_explanations = []
    distances = []
    for n in range(int(n_words/4)): # /4 is for performance reasons
        X[n] += 10
        new_explanations.append(explainer(np.array([X])).values[0])
        X[n] -= 10
    for n_exp in new_explanations:
        distances.append(np.linalg.norm(explanation - n_exp))
    return np.mean(distances)
```