Finding Shortcuts to a black-box model using Frequent Sequence Mining

Explaining Deep Learning models for Fact-Checking

Jean-Paul Smit[1]

Supervisor(s): Avishek Anand[1]**, Lijun Lyu**[1]**, Lorenzo Corti**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2023

Name of the student: Jean-Paul Smit
Final project course: CSE3000 Research Project
Thesis committee: Avishek Anand, Lorenzo Corti, Lijun Lyu, Marco Loog

Abstract

Deep-learning (DL) models could greatly advance the automation of fact-checking, yet have not widely been adopted by the public because of their hard-to-explain nature. Although various techniques have been proposed to use local explanations for the behaviour of DL models, little attention has been paid to global explanations. In response, we investigate whether a frequent sequence mining (FSM) tool finds sequence patterns, that act as shortcuts, to a state-of-the-art model in the context of fact-checking. By studying the connections between a model's input and output, association rules (ARs) can be used as a global explanation for the interpretation of the model. The shortcuts were evaluated using a heuristic-based *minimum support value*, the strength of each rule was determined using *confidence*, and the *support* value indicates the global coverage of rules. Shortcuts help to form an interpretation for creating counterfactual prompts, which can be used as a risk assessment tool for DL models. Other applications for rule-based global explanations are left for future work.

***Keywords* — Explainability; Explainable Artificial Intelligence; Association Rule mining; Frequent Sequence Mining; Fact-checking**.

1 Introduction

Deep-learning (DL) models have gained significant popularity recently due to their impressive performance in a variety of domains such as natural language processing, sentiment analysis, and computer vision[1–3].

Surprisingly, the adaptation of DLs has been held back in areas with impacts on decision-making, such as fact-checking [4, 5]. The capacity to explain predictions of neural networks is limited by their complexly layered character of them, creating a lack of interpretability. [6] This 'black-box' characteristic of DL models is what makes them susceptible to bias, which can have adverse effects when adopted in areas where the stakes are high: which is especially true for fact-checking [7].

Currently, a wide range of *local* (individual prediction) explanations exist for neural network predictions. Feature attributions weigh the importance of attributions to the prediction of a model, such as LIME [8], Kernel SHAP [9] and Integrated Gradients [10] methods. Other approaches are instance attributions, which utilise a subset of useful attributes that are required to exist in order to keep or eliminate a change in the prediction of a model [11] or counterfactual approaches, which change part of an input such that it flips the prediction of the model on that specific input. [12] Another approach is rule-based, where the decisions of DL models are locally explained by creating if-then rules for black-box models [13]. But because of their locality, all of the aforementioned strategies can only explain a local population of a model.

Global explanations, however, can explain entire prediction populations [14] by the use of structures that are regarded as interpretable, such as trees[15] or decision sets [16]. Another work uses summaries of local explanations as global explanations [17]. While those methods currently make up the landscape of global explanations, there is limited work on rule-based global approaches.

In response to that gap of research of global explanations, we investigate whether sequences or patterns, created by sequence mining, are able to globally explain the decisions made by a complex black-box model. These patterns, known as frequent sequences, may be indicative of underlying trends or relationships, with which if-then rules or *shortcuts* can be identified [18]. By identifying patterns in a model's input data that are connected to output data, it may be possible to gain a global understanding of how a model is making decisions and to identify potential biases in the model's predictions.

In this paper, we answer the following question: *can frequent sequence mining find shortcuts to a complex black-box model?* To answer this question, we resolve the following sub-questions:

1. Can frequent sequence mining create rules that capture global behaviour in a model for fact-checking?
2. To which categories can the rules found in question 1 be generalised?
3. What are the applications of global shortcuts for black-box models?

In summary, **our main contributions** are:

- A novel method for creating a model-agnostic explanation using association rules, which is investigated and applied to a fact-checking dataset and model. The rules are tested by three different methods to show how it creates global explanations for the model.
- An interpretation of the model which uses part-of-speech categories for global explanations.
- A risk assessment tool for DL models using association rules.

The remaining part of this paper is organised as follows. First, in section 2, a description of the problem is discussed regarding its input and desired output. Second, the contemporary research on the topics is explored in section 3, and the methodology is elaborated upon in section 4. Afterwards, in section 5 the experimental set-up and evaluation metrics will be highlighted, and in section 6 the results of the experiment are displayed. Lastly, the outcomes are discussed and reflected upon in section 7 and section 8, and the research is concluded in section 9.

The code used to perform the experiments detailed in this paper will be made accessible at https://github.com/jpsmit/Short-Cuts-for-Deep-Neural-models for reproducibility purposes.

2 Problem Description

Given a black box model $f : X \rightarrow Y$ and a dataset of instances $x \in X$, we seek to find sequence patterns in a black-box model: that is to use sequence mining to find frequent

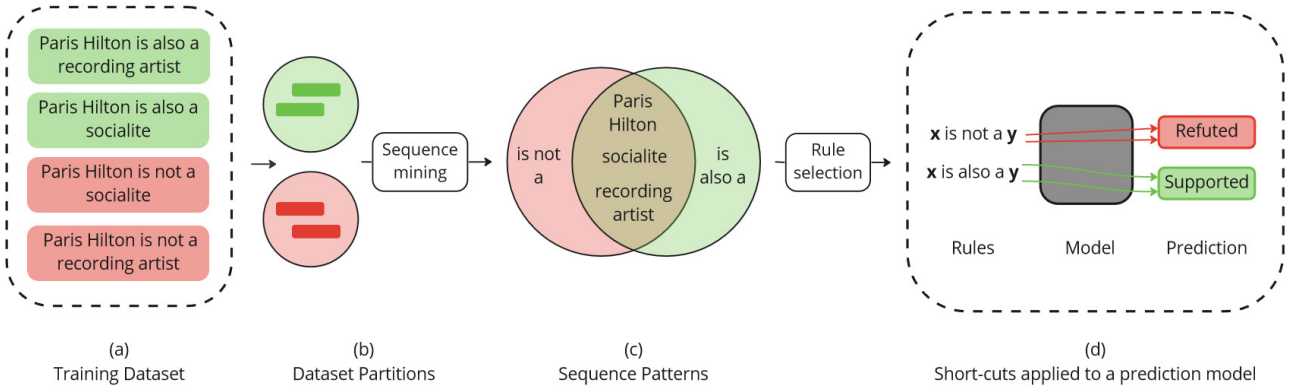

Figure 1: low-level overview of the short-cut creation process with a few examples taken from its training dataset (a) concerning the 'celebrity-socialite-artist' Paris Hilton. After partitioning the claims (b) according to training labels and performing the mining algorithm on the series of claims, the best-performing sequence patterns (c) are selected, resulting in shortcuts that 'bypass' the black-box model (d). See Figure 2a for a detailed instance of the training set.

items or rules, for which $f(x)$ maps to the same value of Y for any instance of a subset of X, where $f(x)$ is the individual prediction for the instance x.

In case a rule makes the model classify an instance on the same category, even on unseen instances that are intended to be classified as 'supported', the rule can successfully 'bypass' the model and a valid shortcut is found for the model prediction. Those rules will be able to aid developers in improving the transparency and accountability of machine learning models, making their decisions more interpretable for humans.

The main hypothesis is that we will find high-quality shortcuts using a sequence mining tool on a dataset. We expect to reveal that the black-box model that we focus on in this study, has developed a global reliance on the sequence patterns found by our research and that they expose a bias in the model's ability to predict.

Leaving the discussion of what makes a pattern into a good shortcut for later, we now demonstrate the creation process and application of shortcuts to a black-box model in the domain of fact-checking.

2.1 Shortcuts

Within the context of validating facts, the pipeline of developing shortcuts can be illustrated with the example given in Figure 1. We focus on a black-box model trained on binary classification training data, which contains training input and corresponding classification labels. Given the set the model is trained on, which consists of a claim, evidence and classification label (see Figure 2), the prediction model has two options: it will either support (*green*) or refute (*red*) on a claim. Take for example the input claim in Figure 1 *"Paris Hilton is not a recording artist"* which is a refuted claim[1]. The claim has evidence that the model uses to make a decision, as can be seen in Figure 2a.

Now from that training data, partitions can be derived using a trivial selection tool which divides the set into two partitions, which are the input of the pattern mining algorithm. The sequence patterns found by sequence mining can then be modelled into rules as illustrated in Figure 1. From the rules, we can recognise how some words or word groups in the claim of Figure 1 commonly appear in the dataset among both classification labels (*'Paris Hilton', 'recording artist'*) while other word groups only appear in refuted claims (*'is not a'*, in the red-coloured circle half of 1) creating a rule for the black-box model. Applied to the example of Figure 1 that means the rules will show a model's reliance on input data created with any of the rules ('x *is not a* y', 'x *is also a* y') for any value of x or y.

After evaluating the patterns the most significant rules can be found, and by modifying unseen input data with the found rules, adversarial input can be crafted to see whether the model always refutes the found rule, bringing cases for model debugging to light. Figure 2b illustrates how the model can be misled by the short-cut 'x *is not a* y', and therefore the model developer knows what input a model should be (re)trained on.

3 Related Work

We acknowledged that the approach of this study draws from existing work in fact-checking models and explanation methods. Therefore, the related work is divided into three parts starting with the relevance of explaining models in the context of fact-checking. Then we elaborate on concepts and categories of explainability methods that can be used to gather explanations. Lastly, rule-based approaches as a way of explaining a model in the context of fact-checking are discussed.

3.1 Fact-checking

The potential for automated fact-checking lies in its adoption to decision-making, which is why it is important that neural models can be interpreted [5]. The errors of a wrong classification are unwanted if not disastrous [4]. Even though hard to interpret because of their black-box nature, AI models can greatly advance the process of verifying claims from textual sources. Thus if the neural models used in fact-checking would be more interpretable, they could be widely adopted by decision-makers.

[1] See her single (www.youtube.com/watch?v=2YkCIrJKAY4).

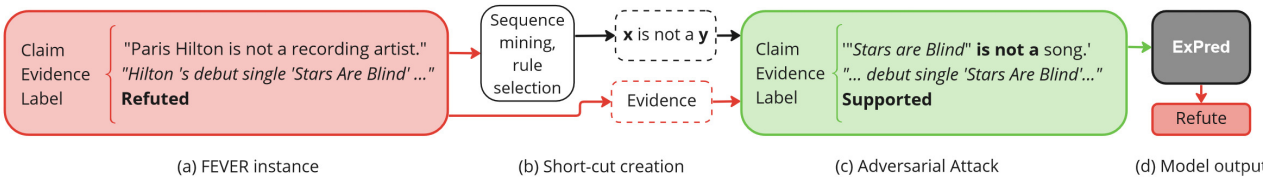

Figure 2: rule-based prompt creation process for risk assessment in the ExPred model. Example instance of the FEVER (DeYoung et al.[19]) dataset (a) and an example adverse input (c) crafted with a rule (b). The model classifies the input as refuted (d) while the evidence supports the claim. A model's training data (a) can be used to create rules (b), that may be modified into adversarial attacks (c) to show how a model such as ExPred[20] can be 'misled' using an adverse instance based on shortcuts.

Neural models learn how to classify a task from a training set with ground-truth examples, and are then optimised using a test set [21]. After being trained and tested, a model can then be deployed to perturb unseen data.

The FEVER dataset is a valuable resource for research in the field of natural language processing and fact-checking [22]. Instances in the set are classified into claims, the labels 'supported', 'refuted', or 'not enough information' and a reference to one or more reference(s) to a collection of potentially relevant Wikipedia texts, known as evidence. An example instance taken from FEVER is illustrated in Figure 2a. DeYoung et al.[19] adjusted the dataset to only contain claims with *refuted* or *supported* classifications. The adjusted dataset still contains a training set of 98.000 annotated claims.

The state-of-the-art model ExPred by Zhang et al.[20] can be applied to aid in the quest for the automation of fact-checking. It scores higher than other fact-checking models: it performs significantly better than current explain-then-predict methods such as [8] for providing explanations. The model is trained on the fact-checking dataset adjusted by DeYoung et al.[19]. Even though the model is *interpretable-by-design*, it gives explanations based on single instances and running it on big datasets would take a lot of resources. Therefore our approach investigates whether a global post-hoc explanation can be made to it using a less-consuming tool.

3.2 Explainability methods

Artificial intelligence (AI) models like neural networks are difficult to understand by the nature of their design. Therefore, AI models are considered black boxes, because it is rather unclear what the inner workings are of the models [7]. By design, models that are interpretable are linear or logistic regression, decision trees, k-nearest neighbours, Bayesian models and rule-based models. To give insights into the inner workings of black-box models, interpretable models can be used to explain a model. That is known as explainable AI (XAI)[6].

Explainable AI methods can be divided into three divisions based on the explanation phase: pre-model, in-model or post-model [6]. Pre-model explanations use data exploration and visualisation to create an explanation, whereas in-model does that by design. *Interpretable-by-design* models explain themselves, which is common in complex models like ExPred[20] or the globally faithful decision tree algorithm by Frosst and Hinton[15]. The main idea behind it is that a model can become more transparent by adding explainable layers to it.

Post-model or post-hoc explainability is the interpretation of a black-box model's outputs by looking at the inputs. As post-hoc explanations seek to provide explanations that are independent of the inner workings of the specific model being investigated, they can be applied to a wide range of models [23]. This study focuses on the latter to create a post-hoc explanation for an existing model in the context of fact-checking.

3.3 Post-hoc XAI

The main advantage of post-hoc techniques is their independence from the AI model. There are several post-hoc techniques, but no method 'rules all'. It may be suitable to involve a combination of methods. Here we outline the important aspects of post-hoc XAI methods.

- *Feature attributions* weigh the importance of attributions to make a prediction to a model instance. Examples are the Local Interpretable Model-agnostic (LIME) technique [8], SHapely Additive exPlanation (SHAP) [9] and Integrated Gradients [10]. These methods are widely used because of their simplicity, take for example the usability report on LIME: [24]. While praised for their high simplicity, feature attribution does not scale well with large sets of data. The methods make a linear perturbation from local instances, meaning that it can only be run on individual cases at a single time. This is problematic in terms of fact-checking, as large datasets need to be processed, which would take immense resources.
- *Instance attributions* focus on a subset of useful attributes required to exist in order to keep or eliminate a change in the prediction of a model. An example is the k-nearest neighbour algorithm by Rajani et al. [11].
- *Counterfactual approaches* change part of an input such that it flips the prediction of the model on that specific input. Counterfactuals are a risk assessment tool for a model by attacking a model with adversarial prompts. The technique by Zhang et al.[12] uses a combination of gradient-based and prompt-based approaches to perform adversarial attacks. Normally speaking a type of security threat, attacks involve intentionally crafting input data that is designed to trick a machine learning model into making an incorrect prediction [25]. Prompt-based attacks are more natural-sounding than gradient-based ap-

proaches, yet they have to be manually provided according to heuristics[26].

- *Association rules* observe the connection between in- and output data and generalise if-then rules that make a prediction of the model output. Most attribution-based approaches rely on user studies for measuring their explainability, while rule-based techniques are considered explainable themselves by nature[27]. Notable examples of rule-based techniques are Anchors[13] by the same authors as LIME, a decision tree model [28], or similarity neighbourhoods [29]. For more examples of association rule mining techniques, see the literature surveys by Guidotti et al. [30] and Zhao and S Bhowmick [27].

Concluding from this, it can be said that rule-based approaches are a great candidate for creating a global explanation for fact-checking problems. Rules are interpretable by nature, and in the context of fact-checking, they might be used to show bias in a model by the same methods used in counterfactuals [12].

3.4 Frequent Sequence Mining

The main advantage of association rules (AR) lies in the fact that it is traditionally considered interpretable and therefore applicable to rule-based explanations [27]. ARs can be mined using *Frequent sequence mining* (FSM), which identifies frequent sequences in the input data that are connected to corresponding outputs [18]. In FSM, rules are selected by comparing an (in some cases arbitrary [31]) threshold value or *minimum support value* [32] to a *support* value. The *support* is a measurement of the amount of input item covered by a rule [33]. The probability of a rule predicting a trend is measured by the *confidence* of a rule. See figures 4 and 5 for the formulae of support and confidence.

Building upon the strengths of the FSM algorithm from [18], an advanced FSM algorithm is Declarative Sequential Pattern Mining (DESQ), as proposed in [34]. As its name implies, it allows for the declaration of sequence and hierarchical constraints, improving the method so it can generate broadly generalised rules. The hierarchical constraints enable the use of dictionaries as input for pattern mining, from which the algorithm can generalise category patterns. We can use this advanced version of FSM and apply it to fact-checking for textual generalisations of rules in a dataset.

4 Methodology

The main idea behind our exploratory study is to explain a black-box model by performing frequent sequence mining to find rules, similar to the example in section 2.1. First, we split the FEVER[19] dataset into a 'Refuted' set and a 'Supported' set. Then both sets are run on the DESQ[34] Frequent Sequence Mining tool to generate a set of candidate rules. Then, the rules that are most likely to globally approximate predictions of the black-box model ExPred[20] are chosen by comparing their *support* to the *minimum support value* established. Lastly, the strength of shortcuts will be illustrated by comparing the *confidence* to the success rate of rule-based attack prompts on the model.

If the results of the experiment show that the study is able to find high-quality shortcuts that lead to accurate association rules for the model, it would suggest that the study is effective at approximating the predictions of the black-box model.

Alternatively, if the results of the experiment do not show that the study is able to find high-quality shortcuts, it may indicate that the model does not learn shortcuts and the study is not effective at approximating the predictions.

4.1 Mining rules

Since the ExPred model is trained on the FEVER dataset [22] adjusted by DeYoung et al.[19], that training set will be manipulated for the purpose of applying sequence mining to it. The training set can be observed in the imbalance of the prevalence of 'supported' items (73%), compared to the amount of 'refuted' items (27%) [19].

Because our dataset is imbalanced and will be split into two sets, a *minimum support threshold* θ will be calculated for each class. The minimum support value is a decision-making tool in the sense that it denotes a threshold parameter, which will be written in this experiment as θ. The threshold of the 'Refutes' set of sequences is denoted as θ_r from now on, and the set of 'Supports' sequences as θ_s.

According to the survey by Hikmawati et al.[32], the choice of a minimum support value depends on the design of a dataset. Given that 27% of items in the dataset are classified as 'refuted', compared to 73% 'supported' claims, we will have different thresholds for θ_r and θ_s respectively. Items classified as 'refuted' with a coverage that is less than 0.01% will not be tolerated as they are too specific, setting the threshold value at $\theta_r = 0.001$. The least tolerated coverage value for the other set will be $\theta_s = \frac{27}{73} * \theta_r \approx 0.0027$ according to its proportion in the dataset.

$$\mathbb{U}_{\text{FEVER[19]}} = \mathbb{I}_{refutes} \cup \mathbb{I}_{supports} \quad (1a)$$

$$\mathbb{O}_{supp.}, \mathbb{O}_{ref.} = FSM(\mathbb{I}_{supp.}), FSM(\mathbb{I}_{ref.}) \quad (1b)$$

$$\mathbb{S}_{neutral} = \mathbb{O}_{refutes} \wedge \mathbb{O}_{supports} \quad (1c)$$

$$\mathbb{S}_{refutes} = \mathbb{O}_{refutes} \setminus \mathbb{S}_{neutral} \quad (1d)$$

$$\mathbb{S}_{supports} = \mathbb{O}_{supports} \setminus \mathbb{S}_{neutral} \quad (1e)$$

Figure 3: data partition process to extract the 'neutral', 'refutes' and 'supports' set. U is the universe of all items in the FEVER[19] dataset, which will be used as the input $\mathbb{I}$ sets to the FSM tool, resulting in two $\mathbb{O}$ output sets (b). Combining the results of the two output sets, a neutral set can be deducted (c). The rule subsets S are extracted in equations (d) and (e). See Figure 8 for a coloured visualisation of the rule sets.

Two subsets were created from the training data of the FEVER dataset[19] using the approach shown in Figure 1. First, the data will be split into claims that have been proven to be false and claims that have been proven to be true, creating two input sets: a 'refutes' and 'supports' set. The sets are the input data to the FSM algorithm as illustrated in Figure 1b.

A textual hierarchy dictionary for the DESQ[34] tool will be created using natural language processing technology. It

would be exhausting to categorise the subset of mined rules by hand, by going through a dictionary and manually classifying the rules. Therefore we use NLTK[2], a technology for Part-of-Speech (POS) Tagging. This technology minimises the time spent on creating a dictionary for a declarative sequential pattern mining algorithm.

The POS algorithm will be run on the sequences found in the pattern mining experiment, and a dictionary will be generated using the output. We expect to find textual categories by running the DESQ[34] tool on the constructed dictionary, which can be interpreted by consulting part-of-speech definitions from [35–37].

The rule sets $\mathbb{S}_{refutes}, \mathbb{S}_{supports}$ are extracted by taking the set differences between the neutral set and the refutes set as illustrated in figures 1d and 1e after the neutral set $\mathbb{S}_{neutral}$ can be created by taking the duplicates (figure 1b) from the conjunction two output sets $\mathbb{O}_{supports}$ $\mathbb{O}_{refutes}$. For example, if a pattern is found to be common in both output sets, it does not indicate a rule and therefore belongs to the neutral set. Likewise, when a condition can only be found in the $\mathbb{O}_{refutes}$ set, it becomes a rule for 'refuted' classification.

Lastly, we will test the strength of the rules from rule sets $\mathbb{S}_{refutes}$ and $\mathbb{S}_{supports}$ by measuring their *support* and *confidence*. We select the ten strongest rules, by comparing the confidence of the rules, and further investigation to see how they can be used to create adverse prompts.

4.2 Rule-based prompt generation

To investigate whether the rules can be used as counterfactuals, targeted adversarial attacks are performed, by a prompt-based heuristic, with the intent of generating the highest possible success rate on the ruleset.

The heuristic for creating adversarial prompts from the mined rules is described as follows: take a set of claims $\mathbb{S}$, a claim $c \in \mathbb{S}$ and a rule or sequence s. The input claim c containing a rule s can be rewritten as $c = (a)s(d)$, where a and d are the remainders of the input claims. Now for all instances of $c \subset S$, swap out sequence s for either a suitable synonym or antonym such that $c = (a)\bar{s}(d)$ where $\bar{s}$ denotes the change of the prompt. The adversarial prompts will be denoted as $\bar{s}$ from now on.

Now which placeholder for $\bar{s}$ should be exchanged in a prompt to flip its meaning? For that, we use a corresponding antonym from the opposite ruleset. The meaning of a phrase or sentiment can be flipped by using an antonym for a specific word in a sentence [37]. The naturally expected result, in case of an antonym flipping the meaning of a query, is that the prediction label changes as well, and the naturally expected result of using a synonym to retain the meaning should not change the prediction label.

The rule-based prompt creation is illustrated in the example from Figure 2c and 2d. Take the claim $c =$ "*Paris Hilton is not a recording artist*" from the set of 'refuted' claims. After sequence mining and rule selection (figure 2b) the sequence $s =$'is not a' can be applied to other claims in S with 'supported' labels. Take $a =$'Stars are Blind' and $d =$'song' , where the original input evidence is retained, meaning compared to the dataset and human-based ground truth that the label should flip (figure 2c).

[2]https://www.nltk.org/

5 Experimental Set Up

In this section, we investigate how frequent sequence mining can find shortcuts to a complex black-box model and the required setup and parameters for reproducing the research. As described in section 4, short-cuts are verified training dataset patterns for capturing global behaviour in a trained model. In this study, a few verification techniques common to the domain of data mining are used for selecting those shortcuts from the mined patterns.

5.1 Metrics

There are plenty of metrics for assessing the quality of association rules, four methods were combined to prove the correctness of the study. First, the *confidence* and *support* are used as measures on the dataset to ensure high-quality rules. Second, the percentage covered by the rules was computed to reveal how the model predicts the algorithm, denoted as *confidence*. Third, the model will be run on unseen, adversarial rules to see how it generalises using the mined subsequences, measured as the *attack success rate*. This mixed approach to the experiment will allow us to test the performance of our study under different conditions and compare its results to the training data.

$$Supp(A \rightarrow B) = P(A \cup B) \tag{2}$$

Figure 4: Support formula as the probability of the union between the antecedent A and consequent B [33]

Support or coverage can be used to identify the occurrence of a pattern [33]. The support is counted as the occurrence of a subset of instances over the entire dataset (see Figure 4) and support can be identified as the relative number of items which contains a pattern relative to the total amount of items. Low support can indicate a limited amount of training data available or the presence of noise or other confounding factors.

$$Conf(A \rightarrow B) = P(B|A) \tag{3}$$

Figure 5: Confidence formula as the probability of a consequent B given antecedent A [33]

Confidence measures the strength of a rule[33], which is used on both the dataset and the AI model. Independent from the support value, it measures the conditional probability of B happening given the known occurrence of A. It shows the strength of a rule by revealing how high the probability is for a consequent B, given an antecedent A. In the context of sequence mining, take input items as the antecedent, and sequence mining patterns as the consequent. Then a rule will be strong when there is a high probability that it can predict the *generalisation of the dataset* given the rule. In the context of explainable AI, take input items as the antecedent,

and model predictions as the consequent. Then a rule will be strong when there is a high probability that it can predict the *outcome of the model* given the rule.

Moreover, confidence is also used to calculate the proportion of a rule that is correctly predicting the *model*. It is calculated as the number of true predictions given a rule divided by the total number of predictions made by the model. High confidence means that the model is relying heavily on a rule, whereas low confidence means a model does not rely on a rule at all in making its predictions.

$$Success(\bar{s}) = \frac{\text{successful adversarial examples}}{\text{all adversarial examples}} \quad (4)$$

Figure 6: Attack Success Rate Formula, denoted as $Success(\bar{s})$ where $\bar{s}$ is adverse input modified using short-cuts.

Attack success rate is the metric that adversarial attacks are measured with. The metric gives insight into the vulnerabilities of a model [38]. The rate of success is the probability that the model incorrectly classifies a query after performing an adversarial attack on it, by flipping the meaning, as seen in Figure 6.

5.2 Effect of θ

To achieve higher support values and better rules, the proportion of the instances on the total data was judged against the minimum support value, which is essentially our only hyper-parameter.

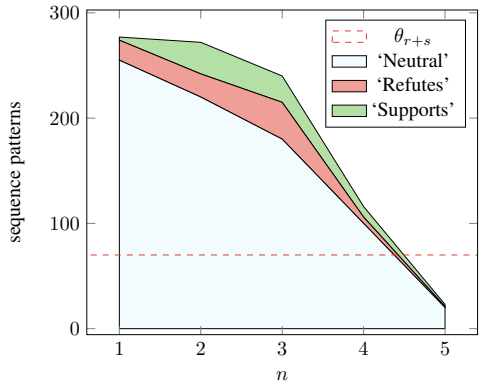

Figure 7: Number of sequence patterns for each classification set, given the number of items n per sequence. The number of subsequence patterns drops with the increasing steps of n. After $n = 4$, sequence patterns are below the minimum support value $\theta_{r+s} = 0.004$ indicated by the red line.

We combined the value of θ_{r+s} to decide how many items from FSM were relevant to the study. The values for θ_r and θ_s are set at $\theta_r = 0.001$ and $\theta_s = 0.003$. As can be seen in Figure 7, the first four stages of 4-item mining yielded relevant results within the support value of $\theta_{r+s} = 0.004$. It can be observed in Figure 7 that the number of sequences, found in all three prediction classes, decreases significantly below the threshold value after each iteration of n-item mining. An example of the relevance of using the threshold is further explained by analysis of Table 2, 3,4, 5: where can be observed how specific rules become in the 'refutes' set when they drop below 0.01%.

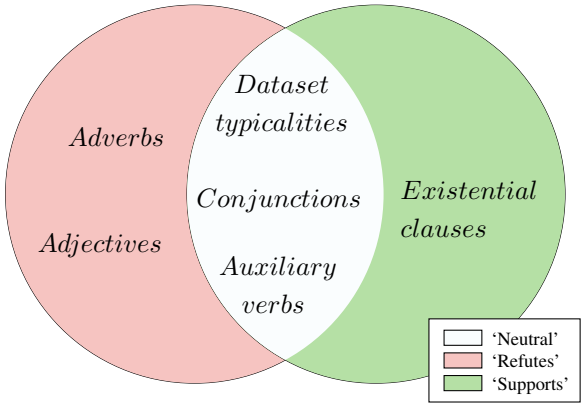

Figure 8: Venn-diagram depicting dataset partitions and the attributed part-of-speech categorisation of sequence patterns found in the FEVER dataset[19].

6 Results

We present the results of the experiment using the set-up and metrics as described in section 5. Unless stated otherwise, all experiments produced the same results.

During the exploration of the FEVER[19] dataset, four steps of the DESQ[34] tool were performed, and results were compared to the threshold value $\theta_{r+s} = 0.004$. That implies a process in which we first performed mining on sequences containing one item, then two, then three, and four items. The plot clearly illustrates that from $n = 4$ onward, results below the minimum support threshold are so few, and from Table 5 in the Appendix can be derived that the results from four-item sequences are already below the minimum threshold value.

Once the mined sequences of the two query classes were extracted, they were divided according to their classification labels, from two into three datasets, to gather three class-specific datasets. As a result, there were now three datasets containing Refutes-, Supports-, and Neutral-class-specific subsequences.

$s \in$ **'Refutes'**		$s \in$ **'Neutral'**		$s \in$ **'Supports'**	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
refused	0.36%	and	78%	acted	0.67%
yet	0.35%	the	70%	contains	0.29%
exclusively	0.31%	is	58%	birth	0.29%
unable	0.19%	a	57%	helped	0.05%

Figure 9: selection of four single-item rule patterns, per query class, with the highest coverage (support) in the FEVER[19] dataset.

6.1 Rule categories

After that, a dictionary was used as a hierarchy constraint with the data mining tool from which the results are presented

in the following section. By utilising speech-tagging technology, a categorisation was constructed to retrieve more general patterns in the landscape of found rules. A part-of-speech dictionary was used as a hierarchy constraint with the data mining algorithm, to catalogue the patterns in a more general way. Figure 8 highlights the observed categorisations in the context of the fact-checking training set.

Neutral rules. The 'neutral' set, containing the overlapping subsequences, contains dataset-specific words (mostly FEVER[19]-typicalities) or trivial language tools needed to construct phrases (think of auxiliary verbs and articles). Auxiliary verbs, also known as *helping verbs*, provide information about the tense and mood of a phrase, and conjunctions provide tools to combine words and phrases [36]. For example, the auxiliary verb "will" can be used to form the future tense, while the auxiliary verb "do" can be used to form the negative or to add emphasis to a verb phrase.

'Refuted' rules. As illustrated in Table 9, most words from the one-item mining experiment results indicating a 'refutes' prediction label are either adverbs or adjectives: words that typically modify or describe other words in a phrase. Adverbs and adjectives usually alter nouns and provide details on their properties or traits [35]. Both might provide details on nuances in a statement or claim. For example, the adverb "beautifully" modifies the verb "sings" in the sentence "She sings beautifully" to characterise the style of her singing. The adjective "beautiful" modifies the noun "voice" in the sentence "She has a beautiful voice" and identifies a feature of her voice.

'Supported' rules. In contrast to the 'Rejected' classification set, the 'Supports' dataset contains primarily claims with existential clauses. Testing ExPred to validate a claim starting with *There is a [...]* has a high chance of being accepted with a 'Supports' label.

6.2 Adversarial attacks using shortcuts

Based on the confidence and support values, shortcuts were selected from studying the pattern categorisations found in section 6.1. Table 1 lists the selection of shortcuts with confidence values of more than 85%.

After running the ExPred model on the queries containing the subsequences, rule confidence scores are reasonably high at an average of 95%. Thus, the model efficiently generalises queries containing the sequences found in the model's training data.

The confidence values of the selected shortcuts can be observed in Table 1. The adversarial attacks also show how flipping the meaning of the sequences '*was incapable of*' and '*does not have*' has a high probability of inverting the prediction label. On the other hand, it is hard to find a heuristic that is powerful enough to flip sequence pattern '*is only a(n)*' since it only has a 60% success rate, which is low compared to the other success rates. That leaves options for future investigation into an automated approach for rule-based attacks.

7 Discussion

In the following section, our results will be interpreted and compared to answer the research questions. Additionally, we pave the road for future work with some recommendations.

7.1 Interpretation of results

First of all, the method used creates rules that capture the global behaviour of the ExPred model in the FEVER dataset. The high support values indicate that the shortcuts can give global explanations to the dataset. The rules found by the method have a high support value and therefore, they cover high populations of the dataset.The model relies heavily on the sequences to make its predictions, as can be seen by the high confidence values in Table 1.

The found rules enable interesting generalisations about the nature of our dataset that we want to highlight in a linguistic context. For interpretation and analysis, the works of Huddleston[35] and Kroeger[36] were consulted to gain an interpretation of the found categories.

The sequences found in the neutral set give insight into the content of the dataset and the kinds of topics that are being covered by it. Even though the sequences found in the neutral step do not add towards an explanation for risk assessment, it explains the model nonetheless. For example, the FEVER dataset contains a lot of queries on the movie industry and has many dataset-specific words related to film terminology. These words help understand the focus of the dataset and can be used as features in natural language processing tasks such as classification or clustering. Other words found in the neutral set, such as auxiliary verbs, can be useful for tasks such as text summarising or machine translation, where it is important to accurately convey the tense and mood of the original text.

As our study points out, the ExPred[20] model is vulnerable to some adversarial attacks, where small changes to the input data can significantly alter the model's output. The shortcuts can be used to create adversarial attacks on the model, which can be used to debug the model. That essentially means that the rules found are connected to the counterfactual interpretations in [12]. Similarly, our findings can help in the risk assessment of neural models as the *attack success rate* indicates focus areas for model debugging.

7.2 Future work

The specific route taken by this study is not the only approach. Here we want to highlight that there might be other purposes for the shortcuts than those shown in our approach. Moreover, the study relies on an estimated threshold value and was only tested on one dataset and model.

First, another application for the shortcuts might be investigated. For example, the counterfactual prompts could be automatically generated or gradient-based [12, 38].

Second, a lower minimum support value threshold would have increased the number of subsequences, decreasing confidence but increasing coverage in the model. The estimation of an adequate minimum support value is difficult [32] yet impacts the outcomes of the experiment. Likewise, a higher threshold would have decreased coverage but increased confidence values.

Furthermore, we advise fellow researchers to investigate whether the method applies to other datasets and models. The experiment's results are influenced by the typicalities of the dataset, its imbalances, and the design of the model. A notable drawback of the study is that it only examined one par-

s	$\rightarrow$	$r(s)$	$Conf(s \rightarrow \text{FEVER})$	$Conf(s \rightarrow \text{ExPred})$	$\bar{s}$	$Success(\bar{s})$
is incapable of being	$\rightarrow$	Refutes	100%	94%	*was in a*	78%
has only ever been	$\rightarrow$	Refutes	100%	99%	*has **also** been*	62%
does not have	$\rightarrow$	Refutes	100%	85%	*does have*	83%
is exclusively	$\rightarrow$	Refutes	100%	99%	*is*	60%
is not a(n)	$\rightarrow$	Refutes	100%	100%	*is **also** a(n)*	74%
has yet to	$\rightarrow$	Refutes	100%	100%	*has **acted***	90%
is only a(n)	$\rightarrow$	Refutes	100%	99%	*is **not** a(n)*	77%
was unable to	$\rightarrow$	Refutes	100%	95%	*was **in the***	76%
was incapable of	$\rightarrow$	Refutes	100%	97%	*was **nominated for***	89%
There is a	$\rightarrow$	Supports	100%	90%	*There is **not** a*	89%

Table 1: selection of the 10 strongest rules to the ExPred[20] model $s \rightarrow r(s)$ using the data from FEVER[19], confidence of connections between the dataset and the rule $Conf(s \rightarrow \text{FEVER})$, model prediction confidence $Conf(s \rightarrow \text{ExPred})$, adverse inputs $\bar{s}$, and adversarial attack success-rate values $Success(\bar{s})$.

ticular dataset and model design, which may not be typical of real-world datasets or models.

8 Responsible Research

In the following section, we will critically reflect on the responsibility of the study. The method is assessed using the principles of FAIR research. Apart from that, we will take into account the ethical implications in light of our discoveries.

8.1 Reproducibility

The study was tested against the FAIR protocol for standardising data management by Chue Hong et al.[39] which states that research should be 'Findable, Accessible, Interoperable and Reusable'. We conclude that our research satisfies all criteria of the FAIR protocol.

The study's findings are *findable* and *accessible*. The code was uploaded to Github[3], which is an open-source platform. The ExPred model can be found on GitHub as well, and the dataset is open-source too.

The study results are *reusable* and *interoperable* because the experiment is described in detail and all input and output data is fully commented with motivations inside several Jupyter notebooks, one for each shortcut. Additionally, the notebooks show the database files of the subsets that were used in calculating the support and confidence values and in forging the adversarial attacks. Anyone who wishes to can recreate the process on the same model and dataset can make use of that.

8.2 Ethical implications

The findings of our experiment expose the vulnerabilities of a fact-checking model. That means the model can be manipulated and exploited. We did not have the intention to pave the way for such attacks.

To mitigate the risk of adversarial attacks, we advise developers to debug the model in a way that is more resistant to such attacks by designing and (re)training it. This could involve using techniques such as adversarial training, which involves training the model on examples that are specifically designed to be difficult to classify or predict, or adding additional constraints or regularisation terms to the model's objective function[38].

It may also be helpful regularly perform evaluations of the model's robustness and vulnerability, but that is beyond the scope of this study.

9 Conclusions

In this paper, we proposed a method for creating shortcuts for a deep-learning black-box model trained on a fact-checking dataset, by the use of frequent sequence mining.

We found that (1) sequence mining is a tool that can capture global behaviour in a training dataset, by creating shortcuts that act as global rules to a neural model. The rules were tested with a strict minimum support threshold, and the shortcuts with a coverage of at least 85% correctly predicted 95% of the model's predictions on average.

Additionally, (2) the rules could be visually categorised using their part-of-speech indication. The findings reveal categories of grammar attributes, that can globally generalise the input of the model to interpret global model decisions. Furthermore, (3) we showed how global shortcuts for black-box models were discovered as a method for creating basic counterfactuals which can be used for model debugging.

There are many possible routes for future work. First, we would want to extend the manual prompt-based adversarial attacks and see the results of automated attacks. Secondly, a lower minimum support value threshold would have increased the number of subsequences and a lower restriction on confidence would have increased the number of shortcuts. Finally, an important open question that this work prompts is whether rule-based approaches can be applied to other fact-checking datasets and models.

[3] https://github.com/jpsmit/Short-Cuts-for-Deep-Neural-models.

References

[1] Y. Bengio, *Deep Learning*, ser. Adaptive Computation and Machine Learning series. London, England: MIT Press, Nov. 2016.

[2] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170

[3] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2012. [Online]. Available: https://doi.org/10.1109/cvpr.2012.6248110

[4] C. Silverman, "Lies, damn lies and viral content," 2015. [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D8Q81RHH

[5] L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga, "Toward automated factchecking," *Digital Threats: Research and Practice*, vol. 2, no. 2, pp. 1–16, Apr. 2021. [Online]. Available: https://doi.org/10.1145/3412869

[6] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, Dec. 2021. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.05.009

[7] ——, "Classification of explainable artificial intelligence methods through their output formats," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, Aug. 2021. [Online]. Available: https://doi.org/10.3390/make3030032

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?"," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[9] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[10] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.01365

[11] N. F. Rajani, B. Krause, W. Yin, T. Niu, R. Socher, and C. Xiong, "Explaining and improving model behavior with k nearest neighbor representations," *CoRR*, vol. abs/2010.09030, 2020. [Online]. Available: https://arxiv.org/abs/2010.09030

[12] Z. Zhang, V. Setty, and A. Anand, "SparCAssist," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3477495.3531677

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://doi.org/10.1609/aaai.v32i1.11491

[14] M. Ibrahim, M. Louie, C. Modarres, and J. W. Paisley, "Global explanations of neural networks: Mapping the landscape of predictions," *CoRR*, vol. abs/1902.02384, 2019. [Online]. Available: http://arxiv.org/abs/1902.02384

[15] N. Frosst and G. E. Hinton, "Distilling a neural network into a soft decision tree," *CoRR*, vol. abs/1711.09784, 2017. [Online]. Available: http://arxiv.org/abs/1711.09784

[16] H. Lakkaraju, S. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," pp. 1675–1684, 08 2016.

[17] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 598–617.

[18] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*. ACM Press, 1993. [Online]. Available: https://doi.org/10.1145/170035.170072

[19] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," 2019. [Online]. Available: https://arxiv.org/abs/1911.03429

[20] Z. Zhang, K. Rudra, and A. Anand, "Explain and predict, and then predict again," 2021. [Online]. Available: https://arxiv.org/abs/2101.04109

[21] D. Ron, "Glossary of terms," *Machine Learning*, vol. 30, no. 1, pp. 5–6, 1998. [Online]. Available: https://doi.org/10.1023/a:1007411609915

[22] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and verification," *CoRR*, vol. abs/1803.05355, 2018. [Online]. Available: http://arxiv.org/abs/1803.05355

[23] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2101.09429

[24] J. Dieber and S. Kirrane, "A novel model usability evaluation framework (muse) for explainable artificial intelligence," *Information Fusion*, vol. 81, pp. 143–153, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002402

[25] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, Apr. 2020.

[26] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld, "Polyjuice: Automated, general-purpose counterfactual generation," *CoRR*, vol. abs/2101.00288, 2021. [Online]. Available: https://arxiv.org/abs/2101.00288

[27] Q. Zhao and S. S Bhowmick, "Association rule mining: A survey," p. 135, 01 2003.

[28] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complex.*, vol. 2021, pp. 6 634 811:1–6 634 811:11, 2021.

[29] D. S. Rajapaksha, C. Bergmeir, and W. L. Buntine, "Lormika: Local rule-based model interpretability with k-optimal associations," *Inf. Sci.*, vol. 540, pp. 221–241, 2019.

[30] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," 2018. [Online]. Available: https://arxiv.org/abs/1802.01933

[31] I. Sanchez, T. Rocktaschel, S. Riedel, and S. Singh, "Towards extracting faithful and descriptive representations of latent variable models," in *AAAI Spring Symposium on Knowledge Representation and Reasoning*, March 2015.

[32] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," *Journal of Big Data*, vol. 8, no. 1, Nov. 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00538-3

[33] M. Ruiz, J. Gómez-Romero, M. Molina-Solana, M. Ros, and M. Martin-Bautista, "Information fusion from multiple databases using meta-association rules," *International Journal of Approximate Reasoning*, vol. 80, pp. 185–198, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888613X1630158X

[34] K. Beedkar and R. Gemulla, "DESQ: Frequent sequence mining with subsequence constraints," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, Dec. 2016. [Online]. Available: https://doi.org/10.1109/icdm.2016.0092

[35] R. D. Huddleston, *English grammar*. Cambridge, England: Cambridge University Press, Jun. 1988.

[36] P. R. Kroeger, *Analyzing syntax*. Cambridge, England: Cambridge University Press, Jun. 2012.

[37] A. Mettinger, *Aspects of semantic opposition in English*, ser. Oxford Studies in Lexicography and Lexicology. Oxford, England: Clarendon Press, Feb. 1994.

[38] W. E. Zhang, Q. Z. Sheng, and A. Alhazmi, "Generating textual adversarial examples for deep learning models: A survey," *CoRR*, vol. abs/1901.06796, 2019. [Online]. Available: http://arxiv.org/abs/1901.06796

[39] N. P. Chue Hong, D. S. Katz, M. Barker, A.-L. Lamprecht, C. Martinez, F. E. Psomopoulos, J. Harrow, L. J. Castro, M. Gruenpeter, P. A. Martinez, and T. Honeyman. (2021) Fair principles for research software (fair4rs principles). Accessed on Jan. 28, 2023. [Online]. Available: https://rd-alliance.org/group/fair-research-software-fair4rs-wg/outcomes/fair-principles-research-software-fair4rs

A Appendix

A.1 Frequent Sequence Mining

Note that the Support means the percentage of occurring instances of a subsequence in the entirety of the dataset queries.

$s \in$ **'Neutral'**		$s \in$ **'Supports'**		$s \in$ **'Refutes'**	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
and	78%	acted	0.67%	exclusively	0.31%
the	70%	contains	0.29%	unable	0.19%
is	58%	birth	0.29%	yet	0.35%
a	57%	helped	0.05%	refused	0.36%
in	54%			incapable	0.61%
of	52%			zero	0.35%
was	37%			anything	0.22%
an	30%			except	0.19%
by	30%			failed	0.33%
to	28%			never	0.42%
as	27%			always	0.17%
American	26%			passed	0.13%
on	25%			outside	0.18%
for	23%			turned	0.14%
film	21%			does	0.40%
born	18%			appear	0.19%
s	18%			receive	0.14%
with	16%			every	0.15%
has	16%				

Table 2: selection of 20 **one**-item patterns with the highest coverage in the FEVER[19] dataset, grouped by query class

$s \in$ **'Neutral'**		$s \in$ **'Supports'**		$s \in$ **'Refutes'**	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
, and	35%	There is	0.69%	not a	0.70%
is a	28%	acted in	0.53%	incapable of	0.61%
in the	26%	her performance	0.44%	is only	0.55%
of the	25%	played the	0.42%	is incapable	0.41%
is an	18%	critically acclaimed	0.40%	only a	0.40%
an American	16%	also appeared	0.36%	was only	0.35%
and the	14%	has an	0.36%	has only	0.34%
, the	13%	worked with	0.35%	yet to	0.34%
on the	10%	also starred	0.35%	refused to	0.31%
is the	9%	her a	0.35%	has yet	0.31%
directed by	8%	Actress for	0.34%	failed to	0.30%
as the	8%	a nomination	0.33%	not an	0.27%
as a	7%	least one	0.33%	of being	0.26%
, is	7%	received the	0.33%	only ever	0.26%
the United	7%	a person	0.33%	not have	0.20%
United States	7%	acting career	0.33%	unable to	0.19%
for the	6%	received an	0.33%	was incapable	0.18%
to the	6%	a character	0.32%	to ever	0.17%
, which	6%	featured in	0.32%	has not	0.15%
at the	5%	his roles	0.32%	only one	0.15%

Table 3: selection of 20 **two**-item patterns with the highest coverage in the FEVER[19] dataset, grouped by query class

$s \in$ **'Neutral'**		$s \in$ **'Supports'**		$s \in$ **'Refutes'**	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
is an American	11%	There is a	0.74%	is not a	0.63%
, and the	5%	was in a	0.67%	is incapable of	0.43%
the United States	5%	was in the	0.61%	has yet to	0.32%
Award for Best	4%	is also known	0.54%	is only a	0.31%
one of the	4%	nominated for an	0.52%	incapable of being	0.27%
film directed by	3%	appeared in a	0.50%	is not an	0.23%
as well as	3%	nominated for a	0.50%	has only ever	0.20%
in the United	3%	starred in a	0.49%	was incapable of	0.18%
was an American	3%	the Golden Globe	0.48%	was unable to	0.15%
, is a	3%	for her roles	0.46%	does not have	0.15%
, as well	2%	also appeared in	0.46%	was not a	0.13%
, is an	2%	she won the	0.46%	only ever been	0.11%
an American actor	2%	film debut in	0.46%	to be a	0.17%
and directed by	2%	a nomination for	0.45%	is a 2015	0.17%
was nominated for	2%	at least one	0.45%	that premiered on	0.23%
Academy Award for	2%	for her performance	0.45%	most populous city	0.04%
best known for	2%	performance in the	0.45%	a population of	0.36%
, for which	2%	which she received	0.45%	the most populous	0.30%
known for his	2%	is in the	0.44%	is the capital	0.38%
Golden Globe Award	2%	she was nominated	0.43%	is a 2014	0.33%

Table 4: selection of 20 **three**-item patterns with the highest coverage in the FEVER[19] dataset, grouped by query class

$s \in$ **'Neutral'**		$s \in$ **'Supports'**		$s \in$ **'Refutes'**	
s	$Supp(s)$	s	$Supp(s)$	s	$Supp(s)$
in the United States	2.56%	for the Academy Award	0.36%	is incapable of being	0.20%
, as well as	2.37%	is an American film	0.34%	has only ever been	0.11%
is an American actor	1.82%	was nominated for a	0.34%	thriller film directed by	0.16%
Academy Award for Best	1.80%	is known for his	0.33%	, making it the	0.14%
, is an American	1.57%	for which she received	0.32%	an American rock band	0.13%
of the same name	1.55%	was nominated for an	0.31%	is a song by	0.14%
an American actor ,	1.37%	she was nominated for	0.31%		
is an American actress	1.36%	BAFTA Award for Best	0.28%		
written and directed by	1.26%	nominated for the Academy	0.28%		
the Academy Award for	1.24%	the Billboard Hot 100	0.27%		
Award for Best Actress	1.19%				
of the United States	1.17%				
, for which she	1.10%				
is best known for	1.05%				
film written and directed	1.05%				
singer , songwriter ,	0.99%				
, also known as	0.98%				
Golden Globe Award for	0.97%				
as one of the	0.95%				
Award for Best Supporting	0.95%				

Table 5: selection of 20 **four**-item patterns with the highest coverage in the FEVER[19] dataset, grouped by query class