COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks

Fanny Jourdan, Agustin Picard, Thomas Fel, Laurent Risser, Jean-Michel Loubes, Nicholas Asher

COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks

Fanny Jourdan*
IRIT, Université Paul-Sabatier
Toulouse, France
fanny.jourdan@irit.fr

Agustin Picard*†
IRT Saint-Exupéry
Toulouse, France
agustin-martin.picard@irt-saintexupery.com

Thomas Fel
Brown University, USA
SNCF, Toulouse, France

Laurent Risser
IMT, Université Paul-Sabatier
Toulouse, France

Jean-Michel Loubes
IMT, Université Paul-Sabatier
Toulouse, France

Nicholas Asher
IRIT, Université Paul-Sabatier
Toulouse, France

Abstract

Transformer architectures are complex and their use in NLP, while it has engendered many successes, makes their interpretability or explainability challenging. Recent debates have shown that attention maps and attribution methods are unreliable (Pruthi et al., 2019; Brunner et al., 2019). In this paper, we present some of their limitations and introduce COCKATIEL, which successfully addresses some of them. COCKATIEL is a novel, post-hoc, concept-based, model-agnostic XAI technique that generates meaningful explanations from the last layer of a neural net model trained on an NLP classification task by using Non-Negative Matrix Factorization (NMF) to discover the concepts the model leverages to make predictions and by exploiting a Sensitivity Analysis to estimate accurately the importance of each of these concepts for the model. It does so without compromising the accuracy of the underlying model or requiring a new one to be trained.

We conduct experiments in single and multi-aspect sentiment analysis tasks and we show COCKATIEL's superior ability to discover concepts that align with humans' on Transformer models without any supervision, we objectively verify the faithfulness of its explanations through fidelity metrics, and we showcase its ability to provide meaningful explanations in two different datasets.

Our code is freely available: https://github.com/fanny-jourdan/cockatiel

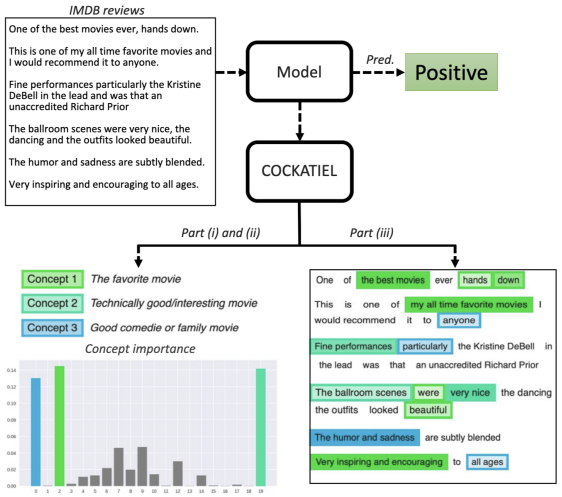

Figure 1: **An illustration of COCKATIEL**. Given some sentences of IMDB reviews, COCKATIEL *(i)* identifies concepts for prediction, *(ii)* ranks them, and *(iii)* gives the most important elements for each concept (to help us interpret the concept).

1 Introduction

NLP models have undeniably gotten increasingly more complex since the introduction of the transformer architecture (Vaswani et al., 2017; Devlin et al., 2018; Liu et al., 2019a). This trend, which is also occurring in the domain of Computer Vision, has brought about a need for understanding how these models make their predictions. The presence of bias in these models could indeed be prejudicial in applications where the user's lives are at stake (De-Arteaga et al., 2019). Humans should be able to comprehend the reasons behind the model's decisions if these models are to gain general acceptance. Also, companies need to ensure that they are deploying algorithms which are free of harmful biases and that the explanations that they are obligated to issue are easily understandable by employees and end-users alike (Kop, 2021).

Intelligibility by humans has then become a key topic in explainable AI systems. As AI systems become more sophisticated and are deployed in increasingly complex environments, the ability to

* Denotes equal contribution
†Work done as a Scalian employee, before April 2023 and joining IRT Saint-Exupéry.

Findings of the Association for Computational Linguistics: ACL 2023, pages 5120–5136
July 9-14, 2023

provide clear and concise explanations of their decisions becomes more pressing.

Researchers have proposed multiple solutions to address this challenge. The most straightforward approach analyzes how each part of the input influences the model's prediction. There are different ways of doing this, through perturbation (Ribeiro et al., 2016; Zeiler and Fergus, 2014) or by leveraging the gradients inside the neural network (Sundararajan et al., 2017a). However, these approaches suffer from being vulnerable to adversarial manipulation (Wang et al., 2020), from only performing partial input recovery (Adebayo et al., 2018) and from a general lack of stability with respect to the input (Ghorbani et al., 2019a). Another research path for transformer models harnesses the information in the attention maps of the transformers' layers to understand how the elements in the input relate to the output, implying that the attention mechanism is inherently interpretable. In spite of a number of supporters initially for this approach, there has been a recent wave of detractors of attention-based explanations (Jain and Wallace, 2019; Pruthi et al., 2019; Serrano and Smith, 2019).

More in line with our proposal work, researchers in the field of rationalization have proposed specific architectures to extract excerpts from whole inputs and predict a model's output based on these *rationales* (Lei et al., 2016; Jain et al., 2020; Chang et al., 2020; Yu et al., 2019; Bastings et al., 2019; Paranjape et al., 2020). These rationales can be seen as explanations that are sufficiently high-level to be easily understood by humans. However, they require to train an entirely new model. Only one rationale can also be found per input text, when there might intuitively be several predictions for a given prediction. Finally, these approaches use architectures that have mostly been left behind since the introduction of the transformer architecture, due to their inferior predictive capabilities.

In line with the project of generating explanations that are meaningful to humans, concept-based explainable AI (XAI) has lately advanced the sate of the art. The pioneer method TCAV (Kim et al., 2018) goes beyond widespread attribution methods to create high-level explanations based on hand-picked concepts. More recently, Fel et al. (2022) has extended this technique to discover automatically pertinent concepts inside the network's activation space and to find the parts of the input space that most align with each concept. Still, it has only been applied to convolutional architectures for image classification tasks.

In this paper, we present COCKATIEL, a novel technique for generating reliable and meaningful explanations for NLP neural architectures for classification problems. It extends CRAFT (Fel et al., 2022) and our contributions can be summarized as follows:

- We introduce a post-hoc explainability technique that is applicable to any neural network architecture containing non-negative activation functions. The technique is capable of explaining predictions of individual instances as well as providing insights of the model's general behavior.

- We measure COCKATIEL's ability to discover concepts that align with those that Humans would employ in a sentiment analysis application. Although we did not train the model on data annotated with these human concepts, COCKATIEL's explanations find them with high accuracy.

- We demonstrate that in addition to generating meaningful concepts for Humans, these explanations are faithful to the models: An explanation X provided by method C is faithful to a model M just in case if X is returned as a putative explanation of M's behavior by C, the X plays a causal role in M's behavior.

- We provide examples of explanations on fine-tuned RoBERTa models (Liu et al., 2019a) and bidirectional LSTMs trained from scratch to show how the concept decomposition can be used to understand the inner workings of complex models.

2 Related Work

2.1 Explaining through rationalization

Finding rationales in text refers to the process of identifying expressions that provide the key reasons or justifications that are provided for a particular claim or decision about that text. Lei et al. (2016) defined rationales as "a minimal set of text spans that are sufficient to support a given claim or decision". They should satisfy two desiderata: they should be interpretable, and they should reach nearly the same prediction as the original output.

To do so, they use a generator network that finds interesting excerpts and an encoder network that generates predictions based on them. However, their scheme requires the use of reinforcement learning (Williams, 1992) for the optimization procedure. Bastings et al. (2019) proposed to include a reparametrization trick to allow for better gradient estimations without the need for reinforcement learning techniques, and a sparsity constraint to encourage the retrieval of minimal excerpts.

Yu et al. (2019) and Paranjape et al. (2020) studied the problem of producing adequate rationales from a game-theoretic point of view. However, these models can be quite complex to train, as they either require a reparametrization trick or a reinforcement learning procedure. Jain et al. (2020) proposed to solve this problem by introducing a support model capable of producing continuous importance scores for instances of the input text, that the rationale extractor can use to decide whether an excerpt will make a good rationale or not.

All these rationales will serve as an explanation for single instances, but won't explain how models predict whole classes. Chang et al. (2019) introduced a rationalization technique that allows for the retrieval of rationales for factual and counterfactual scenarios using three players.

However, all these techniques are not model agnostic and require specific architectures, in particular rather simple architectures or LSTMs, and training procedures. But these architectures have been shown to not produce optimal results.

2.2 Concept-based explanations

Concept-based explainability is a growing area of research in AI, focused on generating human-understandable explanations for the decisions made by machine learning models. One popular approach for generating concepts is TCAV (Kim et al., 2018). It uses gradient-based techniques to identify the important features of a model. However, TCAV relies on Human inputs, as it requires the user to manually specify the concepts to be tested. This can be time-consuming and may not always produce the most comprehensive explanations (Ghorbani et al., 2019b).

Another approach, ACE (Ghorbani et al., 2019b), aims to automate the concept extraction process. ACE uses a clustering algorithm to identify interpretable concepts in the model's activations, without the need for Human input. While this approach has the potential to greatly reduce the time and effort required for concepts extraction, the authors criticize their own reliance on pre-defined clustering algorithms, which may not always produce the most relevant or useful concepts.

An alternative uses matrix factorization techniques, such as non-negative matrix factorization (NMF) (Lee and Seung, 1999), to identify interpretable factors in the data (Zhang et al., 2021; Fel et al., 2022). As presented in Section 3, or strategy is inspired by (Fel et al., 2022) and is therefore a concept-based explanations XAI method. In (Fel et al., 2022), the authors developed a framework for generating global and local explanations. They successfully tested the meaningfulness and the capacity of these explanations to help Humans to understand the model's behavior through psychological experiments. However, this approach has only been applied to convolutional neural networks for image classification tasks so far.

For NLP applications, Bouchacourt and Denoyer (2019) proposed a self-interpretable neural architecture capable of simultaneously generating a prediction on classification tasks and its concept-based explanation. These concepts are learned without supervision from excerpts using a bidirectional LSTM during the training phase of the model, and the predictions are only based on the presence or absence of the individual concepts in the input sentences. Despite of its capacity to generate interesting concepts, its low prediction accuracy for the classification task is a serious limitation (see Table 1). Going further, Antognini and Faltings (2021) introduced ConRAT, a technique that includes orthogonality, cosine similarity and knowledge distillation constraints, as well as a concept pruning procedure to improve on both the quality of the extracted concepts and the model's accuracy.

3 COCKATIEL

In this section, we describe COCKATIEL, our concept-based XAI technique for NLP models to generate human-understandable explanations. It has three main components: *(i)* it uses Non-Negative Matrix Factorization (NMF) to discover the concepts that the neural network under study leverages to make predictions; *(ii)* it exploiting Sensitivity Analysis to estimate accurately the importance of each of these concepts for the model; and *(iii)* it uses a black-box explainability technique to generate instance-wise explanations at a per-word

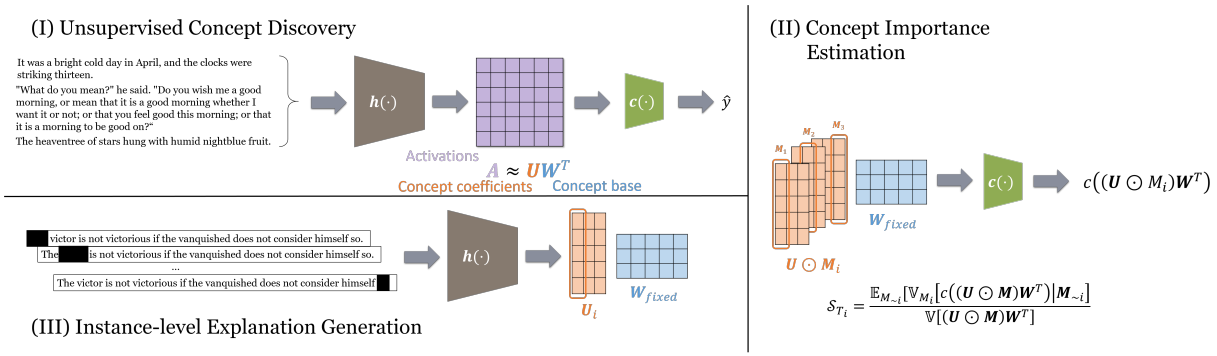

Figure 2: **Overview of our method:** COCKATIEL can be divided into three phases. *(i)* The first step is assembling the concepts base. We propose to do this by constituting a database of whole or excerpts of input texts, projecting each one of these elements into the embedding of the model of our choice $h(\boldsymbol{x})$ and using the NMF algorithm to decompose the resulting non-negative matrix into two low-rank, non-negative matrices: $\boldsymbol{U}$ and $\boldsymbol{W}$. *(ii)* Once $\boldsymbol{U}$ and $\boldsymbol{W}$ have been computed, we can compute the Total Sobol indices for the concept base's columns by masking the coefficients and by looking at their effect on the classifier's output: $c((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T)$. *(iii)* Finally, we propose to retrieve the influence of each word of the instance under study in each concept through Occlusion, that is, by applying masks to each word (or clause) in the input and quantifying the changes in each of the concept coefficients.

and per-clause level. Fig. 2 presents a schematic outline of COCKATIEL.

Notation In a supervised learning framework, we assume that a neural network model $\boldsymbol{f}: \mathcal{X}^n \to \mathcal{Y}^n$ has already been trained for some classification task. We denote by $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathcal{X}^n$ a set of n input texts and $(y_1, ..., y_n) \in \mathcal{Y}^n$ their associated labels. We consider $\boldsymbol{f}$ to be a composition of h, the last embedding of $\boldsymbol{x}$ (*i.e.* the last layer of the feature extractor model), and c, the classification function, $\boldsymbol{f}(\boldsymbol{x}) = c \circ h(\boldsymbol{x})$ with $h(\boldsymbol{x}) \subseteq \mathbb{R}^p$.

COCKATIEL will factorize h through NMF, so we require h to be non-negative – i.e. $h(\boldsymbol{x}) \geq 0 \ \forall \boldsymbol{x} \in \mathcal{X}$. This constraint is typically verified when the last layer has an activation function such that $\boldsymbol{\sigma}(\boldsymbol{x}) \geq 0$, which is the case in (but it's not limited to) layers or blocks using *ReLU*.

3.1 Unsupervised concept discovery - "Concept part"

COCKATIEL discovers concepts without supervision by factorizing the neural network's intermediary activations by using a NMF algorithm. Because we are factorizing h, we can generate explanations on embeddings without needing to deal with the complexities of attention layers (Pruthi et al., 2019); nor do we have to deal with the non-identifiability of transformer models (Brunner et al., 2019). Thus, the concept extraction phase of our method does not depend on the specificities of attention. We will address this later on in Section 3.3 to be able to generate our instance-level explanations.

NMF algorithm: We choose an excerpt-extraction function $\boldsymbol{\tau}_1$ to generate a database of excerpts coming from texts that the model places in the desired class d_c – i.e. $\boldsymbol{X}_i = \boldsymbol{\tau}_1(x_i)$ such that $\boldsymbol{f}(x_i) = d_c$. Then, we place ourselves at the model's last layer and we extract the activations $\boldsymbol{A} = h(\boldsymbol{X}_i)$ for each of the excerpts $\boldsymbol{X}_i$ in the database. With this information, we solve the constrained optimization problem engendered by the NMF algorithm:

$$(\boldsymbol{U}, \boldsymbol{W}) = \underset{\boldsymbol{U} \geq 0, \boldsymbol{W} \geq 0}{\arg\min} \frac{1}{2} \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}^T\|_F^2, \quad (1)$$

where $||\cdot||_F$ is the Frobenius norm.

This allows us to decompose the high-rank matrix containing all activations $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ into two low-rank matrices $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ and $\boldsymbol{W} \in \mathbb{R}^{p \times r}$. Intuitively, this corresponds to $\boldsymbol{W}$ being a matrix whose columns represent the concepts that we will use to generate explanations, and $\boldsymbol{U}$ is a matrix containing the coefficients quantifying the presence of each concept. These matrices are built so as to minimize the reconstruction error $\frac{1}{2}\|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{W}\|_F^2$, enforcing the relevance of the concepts, and with a non-negative constraint for each matrix, thus encouraging sparsity in their elements.

It is important to note that these coefficients $u_{ij} \in \mathbb{R}_+$, so the presence of a concept can be determined by where its value stands in the concept's coefficients distribution. In practice, we have found that fixing a threshold at the quantile repre-

senting the 10% highest values leads to accurate and easy to interpret explanations.

Choice of τ_1: As we want the concepts to be descriptive enough to convey an abstraction but short enough to only contain one, we work with excerpts chosen by an excerpt-extraction function τ_1. The choice of τ_1, which should depend on the dataset and the text's format, heavily impacts the type of explanations that we are able to generate.

We have identified 3 possible τ_1 functions: *(i)* take all the full text ; *(ii)* split the text into sentences (of at least 6 words) ; *(iii)* split the text into clauses. Linguistically, it doesn't make sense to take smaller tokens like one or two words since their meaning is typically too unfocused to provide a real explanation.

We therefore chose τ_1 to respond specifically to each use-case. If we want to capture the mood of whole inputs, we can designate the inputs as the excerpts, and then interpret them by leveraging the local part of our method. If we instead wish to extract more simple but structured concepts, we can choose τ_1 to pick sentences of at least 6 words and ending in a full-stop. The first condition is necessary in the case of the *beer review* dataset, which is composed of short sentences containing very simple descriptions. For this dataset, using only very short excerpts would fail to convey the complexity of the ideas conveyed by the concepts. In this paper, we present results using these two excerpt-extraction functions.

3.2 Concept importance estimation - "Ranking part"

A common issue when utilizing concept extraction methods is the discrepancy between concepts deemed relevant by humans and those utilized by the model for classification. To mitigate the potential for confirmation bias during the concept analysis phase, we estimate the overall importance of the extracted concepts.

To determine which concept has the most significant impact on the model output, we use a counterfactual reasoning (Peters et al., 2017; Pearl et al., 2016), and then use sensitivity analysis (Cukier et al., 1973; Iooss and Lemaître, 2015). A classic strategy in this area is the use of total Sobol indices (Sobol, 1993). This method captures the importance of a concept, along with its interactions with other concepts, on the model output by calculating the expected variance that would remain if all the

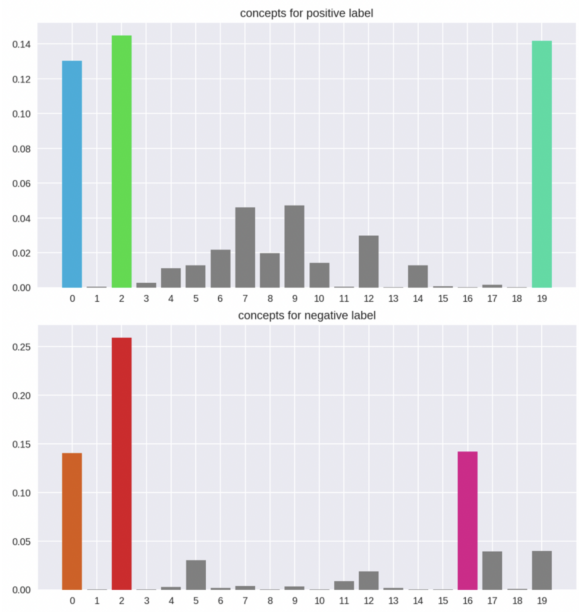

Figure 3: **Concept importance:** The global influence of the NMF concepts on the predictions on RoBERTa model is measured using Sobol indices. There are different concepts for each class (positive and negative label).

indices of the masks except $\boldsymbol{M}_i$ were fixed.

Definition 3.1 (**Total Sobol indices**). *The total Sobol index $\mathcal{ST}_i$, which measures the contribution of a concept $\boldsymbol{U}_i$ as well as its interactions of any order with any other concepts to the model output variance, is given by:*

$$\mathcal{S}_{T_i} = \frac{\mathbb{E}_{\boldsymbol{M}_{\sim i}}(\mathbb{V}_{M_i}(\boldsymbol{Y}|\boldsymbol{M}_{\sim i}))}{\mathbb{V}(\boldsymbol{Y})} \tag{2}$$

$$= \frac{\mathbb{E}_{\boldsymbol{M}_{\sim i}}(\mathbb{V}_{M_i}(c((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T)|\boldsymbol{M}_{\sim i}))}{\mathbb{V}(c((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T))}. \tag{3}$$

To estimate the importance of a concept $\boldsymbol{U}_i$, we measure the fluctuations of the model output $c(\boldsymbol{U}\boldsymbol{W}^T)$ in response to perturbations of the concept coefficient $\boldsymbol{U}_i$. Specifically, we use a sequence of random variables $\boldsymbol{M}$ to introduce concept fluctuations and reconstruct a perturbed activation $\tilde{\boldsymbol{A}} = (\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T$. We then propagate this perturbed activation to the model output $\boldsymbol{Y} = c(\tilde{\boldsymbol{A}})$. An important concept will have a large variance in the model output, while an unused concept will barely change it.

The method for calculating (2) and (3) exploits the Sobol-Hoeffding decomposition and is in the supplementary materials (appendix A).

There are already a plethora of different techniques that allow us to compute this index efficiently (Saltelli et al., 2010; Marrel et al., 2009;

Janon et al., 2014; Owen, 2013; Tarantola et al., 2006). But concretely, we estimate the total Sobol indices using the Jansen estimator (Janon et al., 2014), a widely recognized efficient method (Puy et al., 2022). The Jansen estimator is commonly utilized in conjunction with a Monte Carlo sampling strategy, but we improve over Monte Carlo by using a Quasi-Monte Carlo sampling strategy. This technique generates sample sequences with low discrepancy, resulting in a more rapid and stable convergence rate (Gerber, 2015).

3.3 Instance-level explanation generation - "Interpretable elements part"

In this part, we interpret the concepts found previously. To do this, we find which words and clauses are associated with each concept.

We adapt Occlusion (Zeiler and Fergus, 2014): a black-box attribution method that works by masking each word looking at the impact on the model output. In this case, to get an idea of the importance of each word for a given concept, we mask words in a sentence and measure the effect of the new sentence (without the words) on the concept. This operation can be performed at word or clause level – i.e. mask words or whole clauses – to obtain explanations that are more or less fine-grained depending on the application.

Motivations: This choice has been shown to perform particularly well on NLP models (Fel et al., 2021a) and doesn't suffer from the inefficiency of having to sample a considerable amount of masks for each explanation. Indeed, in (Fel et al., 2021a), they compared Occlusion to other explainability techniques that are commonly used in NLP, and they showed that it is more faithful to the model than Saliency (Simonyan et al., 2014), Grad-Input, SmoothGrad (Smilkov et al., 2017), Integrated Gradients (Sundararajan et al., 2017b), and their own Sobol method on both LSTM and BERT models.

In addition, in the case of transformer models, using a black-box method such as Occlusion avoids manipulating the attention layers between the input and the activation matrix A, where our concepts are located. In doing so, we avoid the non-identifiability problem of transformer models (Pruthi et al., 2019).

Application: Empirically, we perform the following operations:

For a sentence X_i, $A_i = h(X_i)$. We have a fixed W calculate with the NMF and W_k, the k concept of W. As before, we get the importance of the sentence X_i for the concept k:

$$U_i^k = \underset{U \geq 0}{\arg\min} \ \frac{1}{2} \|A_i - U W_k^T\|_F^2.$$

Then, we remove the element j from the sentence i: $\tilde{X}_{i-j}$ (i.e. we replace the (tokenized) feature by a zero). So we have $\tilde{A}_{i-j} = h(\tilde{X}_{i-j})$, and:

$$\tilde{U}_{i-j}^k = \underset{U \geq 0}{\arg\min} \ \frac{1}{2} \|\tilde{A}_{i-j} - U W_k^T\|_F^2,$$

So, $\phi(k, i, j)$ quantifies the influence of the element j in the sentence i for the concept k:

$$\phi(k, i, j) = U_i^k - \tilde{U}_{i-j}^k,$$

For the visualisations (see e.g. Fig. 6), we color the element with the color of the concept for which it is most important. In addition, the darker the color, the more important the element is for the concept.

Choice of τ_2: Just like in the case of the NMF, the choice of the form of the elements of the input to occlude will have an impact on the understandability of the explanations. This can be generalized via another excerpt extraction function τ_2, whose optimal shape will depend on the dataset, the text's format and the learned concepts (i.e. Occlusion shouldn't be applied at a per-clause level if the concepts were learned using a τ_1 providing single words, so this first exceprt extraction function must be taken into consideration). There is a certain trade-off between the granularity and the interpretability of the explanations, as illustrated in Figure 11 in the appendix which contains some examples with different choices of τ_2. In general, we advise to try different combinations of τ_i to find the desired level of granularity in the explanations for each use-case.

4 Experimental evaluation

For all of our results, we fine-tuned RoBERTa (Liu et al., 2019a) based models on each dataset. We ensured the non-negativity of at least one layer of the model by adding a ReLU activation after the first layer of the 1-hidden-layer, dense MLP of the classification head. For the qualitative analysis, we also tested COCKATIEL's performance on bidirectional LSTM models trained from scratch. More details about the implementations are left in appendix B.

		Average				*Appearance*			*Aroma*			*Palate*			*Taste*		
	Model	**Acc.**	**Prec.**	**Rec.**	**Fsc.**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
$l = 20$	RNP	81.1	24.7	21.3	24.9	28.6	23.2	26.5	22.1	21.0	21.5	17.7	24.1	20.4	28.1	16.7	20.9
	RNP-3P	80.5	26	21.8	23.3	30.4	25.6	27.8	19.3	20.4	19.8	10.3	12.0	11.1	43.9	28.4	34.5
	Intro-3P	85.6	21	18.0	19.1	28.7	24.8	26.6	14.3	14.4	14.3	16.6	19.3	17.9	24.2	13.6	17.4
	InvRAT	82.9	37.5	31.6	33.8	54.5	45.5	49.6	26.1	27.6	26.9	22.6	25.9	24.1	46.6	27.4	34.5
	ConRAT	91.4	**43.8**	39.7	40.9	57.8	53.0	55.3	31.9	35.5	33.6	**29.0**	36.3	32.3	**56.5**	33.9	42.4
	Ours	**95.2**	40.6	**58.4**	**47**	**67.5**	**71.4**	**69.4**	**34.1**	**42.3**	**37.7**	24.8	**46.7**	**32.4**	36.1	**73.3**	**48.4**
$l = 10$	RNP	84.4	32.7	14.5	19.5	40.1	12.0	18.5	33.3	18.7	24.0	25.1	17.4	20.6	32.3	9.8	15.07
	RNP-3P	83.1	28.4	13.2	17.8	41.8	19.2	26.3	22.2	12.4	15.9	16.5	10.4	12.7	33.2	10.6	16.1
	Intro-3P	80.9	24	12.2	16.1	51.0	26.0	34.4	18.8	9.7	12.8	16.5	10.6	12.9	9.7	2.6	4.1
	InvRAT	81.9	36.6	15.7	21.8	59.4	26.1	36.3	31.3	15.5	20.8	16.4	9.6	12.1	39.1	11.6	17.9
	ConRAT	91.3	38.2	17.6	23.8	51.7	26.2	34.8	**32.6**	17.4	22.7	23.0	13.8	17.3	**45.3**	13.1	20.3
	Ours	**95.2**	**39.5**	**58.4**	**45.5**	**63.3**	**56.4**	**59.7**	27.3	**67.4**	**38.9**	**26**	**43.5**	**32.5**	41.4	**66.1**	**50.9**

Table 1: Objective performance of rationales for the multi-aspect beer reviews. All baselines are trained separately on each aspect rating, except for ConRAT (Antognini and Faltings, 2021), which is trained on the *Overall* label just like our method. Bold and underline denote the best and second-best results, respectively.

We will first analyze the meaningfulness of the discovered concepts by measuring their alignment with human annotations on the different aspects of a multi or single-aspect sentiment analysis task. Then, we will ensure that our explanations are faithful to the model through an adaptation of the insertion and deletion metrics to concept-based XAI. Finally, we will showcase some examples of explanations and of applications for our method.

4.1 Alignment with human concepts

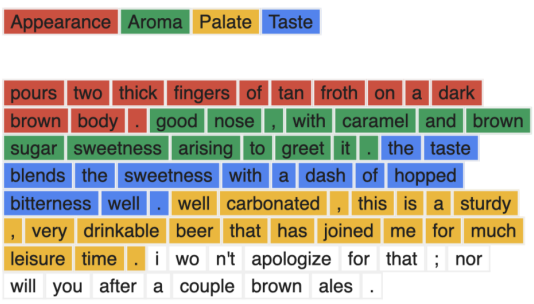

Figure 4: Concepts generated with $l = 20$ for a beer review. The colors depict the aspects for each annotate concept. COCKATIEL is trained only on the label and we use the NMF part of the method to find annotate concepts. *For other examples of review, see appendix D*

Following the human-alignment evaluation in (Antognini and Faltings, 2021), we perform beer task:

Beer Task We will measure the extent to which our concepts overlap the human annotations for the 4 different aspects of the multi-aspect *beer reviews* dataset (McAuley et al., 2012). This dataset contains reviews for beers with commentary and marks (from 0 to 5) on 5 different aspects: Appearance, Aroma, Palate, Taste and Overall. The model will be trained to predict whether the overall score is greater than 3 – i.e. a positive review on the beer – and will not have access to the labels for the other aspects. Additionally, it includes 994 reviews with annotations indicating the position of these aspects in the text. The objective of this evaluation is to look for concepts that align with these annotations and measure their capacity to predict the location of each different aspect. In particular, we searched across the whole annotated dataset for the concepts whose F1 score for the prediction of each aspect was maximal. It is important to note that this does not take into account to which extent they are important for the model to predict, but this only serves as an automatized test for determining whether the explainability technique is capable of generating understandable concepts.

We calculate the precision, recall and F1 scores for each aspect, and we do so with $l = 10$ and $l = 20$ concepts. We remind the reader that, unlike the baselines, our method is a post-hoc technique, and thus, the model does not need to be re-trained, and that changing the number of concepts takes only a few minutes of compute on GPU.

In Table 1, we present a comparison of our results to those obtained with some rationalization techniques: RNP (Lei et al., 2016), RNP-3P (Yu

et al., 2019), InvRAT (Chang et al., 2020) and CoRAT (Antognini and Faltings, 2021) for the task on *Beer*. We demonstrate that not only our model achieves the highest accuracy, but also that it outperforms all the other methods in its ability to accurately recognize the human annotations, be it by its precision, recall or F1 score.

4.2 Evaluation of Explanation Faithfulness

We have demonstrated that we can generate concepts that greatly align with humans', but to legitimately serve as an explainability technique, we must also guarantee its faithfulness. This element is key, as the concepts leveraged by the model may not perfectly align with humans in every task, but we still want the explanation to reflect what the model is doing. An XAI method is said to be faithful if its explanations faithfully convey the information that the model is using to generate its predictions. In (Ghorbani et al., 2019b; Zhang et al., 2021), they proposed to use an adaptation of the deletion and insertion explainability metrics to concept-based methods. In essence, they proposed to gradually mask/add the concepts (following their importance) and seeing the impact on the logits. If the concepts are indeed important for the model to predict, they should drastically decrease/increase as vital information for the prediction is progressively being erased/added.

To evaluate the explanation Faithfulness and present qualitative results, we used the IMDB dataset (Maas et al., 2011). The IMDB dataset is a collection of 50K movie reviews from the Internet Movie Database (IMDB) website. For each review, IMDB specifies whether it is positive or negative (the label). The dataset is balanced, with 25K positive and 25K negative reviews. We used a RoBERTa model to predict the label from the reviews.

In Fig. 5, we showcase the plots for these two fidelity metrics on the *IMDB Reviews* dataset. We observe that the concepts are indeed important for the model's predictions. In the both plots, the curve corresponding to the concept ranked in order of importance according to our Sobol method is better than a random ranking of these concepts, and much better than if we had taken the order of Sobol importance in reverse. In particular, to obtain statistically significant results, we took 10 sets of 10k reviews, and computed the mean and standard deviation values for both of the metrics.

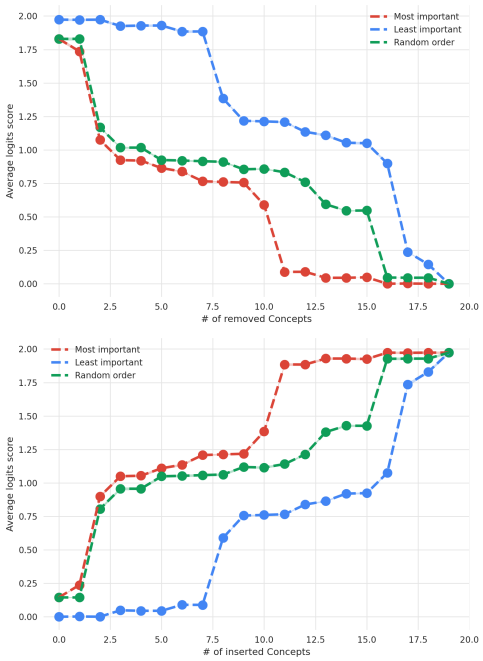

Figure 5: (Upper) Deletion curve for RoBERTa on *IMDB Reviews* (lower is better). (Lower) Insertion curve for RoBERTa on *IMDB Reviews* (higher is better).

4.3 Qualitative evaluation

A model with a good accuracy like RoBERTa gives very good explanations. Others like LSTM (see appendix C) do not do so well and do not yield good explanations. This is not a surprise; if the model predicts badly, necessarily the concepts it uses to predict will be bad. Similarly, if the model is very basic, it uses simple concepts to predict. The reviews in IMDB are also well written, so it is more comfortable to analyse sentences and words to properly call the concepts found by the NMF.

In Fig. 6, we can see the 3 most important concepts for each label class. Each of its concepts "*the favorite movie*", "*technically good/interesting movie*", "*good comedie of family movie*" for the positive class or "*the worst movie*", "*middling movie*", "*boring/stupid movie*" for the negative class are ideas that seem natural and which structures our vision of why a film would be positive or negative.

5 Conclusion

In this paper, we revisited concept-based explainability techniques and presented COCKATIEL, a post-hoc, model agnostic method capable of generating meaningful and faithful explanations for NLP models trained on classification tasks. The method has three parts: *(i) a concept part*, using Non-Negative Matrix Factorization to discover the

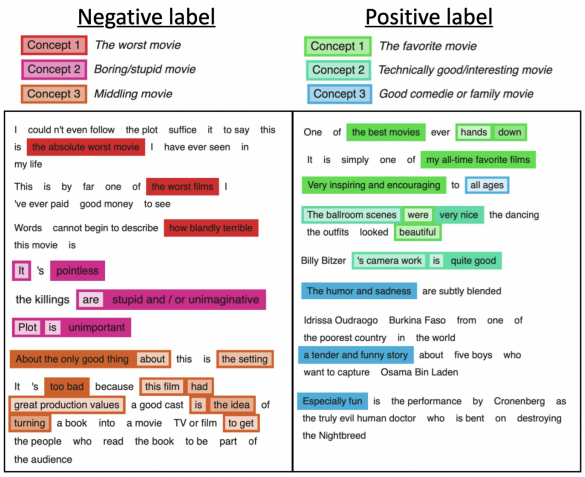

Figure 6: Concepts generated with $l = 20$ for a few sentences taken from IMDB reviews. The colored elements are those important for the concept of the corresponding color (calculated with part *(iii)* of our method). The more colorful the element, the more important it is for the concept (continuously). We have selected the 3 most important concepts for each label (see Fig. 3). The name of the concept is chosen manually in view of the important elements corresponding to the concepts.

concept, *(ii) a ranking part*, using Total Sobol indices to measure the influence of each concept, and *(iii) an interpretable elements part*, using a black-box attribution method to quantify the impact of each element out of each concept.

We measured COCKATIEL's ability to discover concepts that align with those humans and obtained better scores than state-of-the-art methods. We demonstrated that in addition to generating meaningful concepts for humans, these explanations are faithful to the models. Finally, we gave some qualitative examples of explanations for different models to understand the method "in practice".

Limitations

We have demonstrated that COCKATIEL is capable of generating meaningful explanations that align with human concepts, and that they tend to explain rather faithfully the model.

The concepts extracted of NMF are abstract and we interpret them using part 3 of the method. However, for the interpretation, we rely on our own understanding of the concept linked to the examples of words or clauses associated with the concept. This part therefore requires human supervision and will not be identical depending on who is looking. One way to add some objectivity to this concept labeling task would be to leverage topic modeling models to find a common theme to each concept.

In addition, τ_1 and τ_2 were chosen empirically to allow for an adequate concept complexity/human understandability trade-off in our examples. We recognize that this choice might not be optimal in every situation, as more complex concept may be advantageous in some cases, and more easily understandable ones, in others. We surmise that this choice might also depend on the amount of concepts and on the model's expressivity.

Finally, we have studied the meaningfulness and fidelity of our generated concepts, but ideally, the simulatability should also be tested. This property measures the explanation's capacity to help humans predict the model's behavior, and has recently caught the attention of the XAI community (Fel et al., 2021b; Shen and Huang, 2020; Nguyen, 2018; Hase and Bansal, 2020). We leave this analysis for future works.

Ethics Statement

This work contributes to the field of explainability. This field has strong links with the field of fairness, because explaining a model makes it possible to understand its biases. Transformers are a type of model that are little studied in explainability and yet it is widely used. COCKATIEL is a tool to explain transformers and therefore avoid using biased models against the minority.

It is important to remark that this need for understanding automatic decisions start being enforced by Law, as for instance by the so-called *AI act*[1] of the European Union. As a consequence, companies need to ensure that they are deploying algorithms which are free of harmful biases and that the explanations that they're obligated to issue are easily understandable by employees and end-users alike.

Acknowledgements

We thank the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) funded by the ANR-19-PI3A-0004 grant for research support. We also thank the reviewers for their insightful comments. This work was conducted as part of the DEEL[2] project.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[2] https://www.deel.ai

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Diego Antognini and Boi Faltings. 2021. Rationalization through concepts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 761–775, Online. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. *arXiv preprint arXiv:1905.08160*.

Diane Bouchacourt and Ludovic Denoyer. 2019. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2019. A game theoretic approach to class-wise selective rationalization. *Advances in neural information processing systems*, 32.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR.

RI Cukier, CM Fortuin, Kurt E Shuler, AG Petschek, and JH Schaibly. 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of chemical physics*, 59(8):3873–3878.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. 2021a. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in Neural Information Processing Systems*, 34.

Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. 2021b. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *arXiv preprint arXiv:2112.04417*.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. 2022. Craft: Concept recursive activation factorization for explainability. *arXiv preprint arXiv:2211.10154*.

Mathieu Gerber. 2015. On integration methods based on scrambled nets of arbitrary size. *Journal of Complexity*, 31(6):798–816.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019a. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019b. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Bertrand Iooss and Paul Lemaître. 2015. A review on global sensitivity analysis methods. *Uncertainty management in simulation-optimization of complex systems*, pages 101–122.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. 2014. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Mauritz Kop. 2021. Eu artificial intelligence act: The european approach to ai. Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust ....

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. Cc-news-en: A large english news corpus. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 3077–3084, New York, NY, USA. Association for Computing Machinery.

Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant. 2009. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

Art B Owen. 2013. Better estimation of small sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. *arXiv preprint arXiv:2005.00652*.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

Arnald Puy, William Becker, Samuele Lo Piano, and Andrea Saltelli. 2022. A comprehensive comparison of total-order estimators for global sensitivity analysis. *International Journal for Uncertainty Quantification*, 12(2).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. 2010. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer physics communications*, 181(2):259–270.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Hua Shen and Ting-Hao Huang. 2020. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Ilya M Sobol. 1993. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*, 1:407–414.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017a. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017b. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Stefano Tarantola, Debora Gatelli, and Thierry Alex Mara. 2006. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of nlp models is manipulable. *arXiv preprint arXiv:2010.05419*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. 2021. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11682–11690.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.

A Sobol technique in details

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space of possible concept perturbations. To build these concept perturbations, we use $\boldsymbol{M} = (M_1, \dots, M_r) \in \mathcal{M} \subseteq [0,1]^r$, i.i.d. stochastic masks on the original vector of concept coefficients $\widehat{U} \in \mathbb{R}^r$. We define concept perturbation $U = \boldsymbol{\pi}(\widehat{\boldsymbol{U}}, \boldsymbol{M})$ with the perturbation operator $\pi(\tilde{\boldsymbol{U}}, \boldsymbol{M}) = \tilde{\boldsymbol{U}} \odot \boldsymbol{M} + (\mathbf{1} - \boldsymbol{M})\mu$ with $\odot$ the Hadamard product and $\mu \in \mathbb{R}$ a baseline value, here zero.

We denote the set $\mathcal{U} = \{1, \dots, r\}$, $\boldsymbol{u}$ a subset of $\mathcal{U}$, its complementary $\sim \boldsymbol{u}$ and $\mathbb{E}(\cdot)$ the expectation over the perturbation space. We define $\boldsymbol{c} : \mathcal{A} \to \mathbb{R}$, the classification function and we assume that $\boldsymbol{c} \in \mathbb{L}^2(\mathcal{A}, \mathbb{P})$ i.e. $|\mathbb{E}(\boldsymbol{c}(\boldsymbol{U}))| < +\infty$.

The Hoeffding decomposition gives $\boldsymbol{c}$ in function of summands of increasing dimension, denoting $\boldsymbol{c_u}$ the partial contribution of the concepts $\boldsymbol{U_u} = (U_i)_{i \in \boldsymbol{u}}$ to the score $\boldsymbol{c}(\boldsymbol{U})$:

$$\begin{aligned}\boldsymbol{c}(\boldsymbol{U}) &= \boldsymbol{c}_{\varnothing} \\ &+ \sum_{i}^{r} \boldsymbol{c}_i\left(U_i\right) \\ &+ \sum_{1 \leqslant i < j \leqslant r} \boldsymbol{c}_{i,j}\left(U_i, U_j\right) + \cdots \\ &+ \boldsymbol{c}_{1,\dots,r}\left(U_1, \dots, U_r\right) \\ &= \sum_{\boldsymbol{u} \subseteq \mathcal{U}} \boldsymbol{c_u}\left(\boldsymbol{U_u}\right)\end{aligned} \tag{4}$$

Eq. 4 consists of 2^r terms and is unique under the orthogonality constraint:

$$\begin{aligned}&\mathbb{E}\left(\boldsymbol{c_u}\left(\boldsymbol{U_u}\right)\boldsymbol{c_v}\left(\boldsymbol{U_v}\right)\right) = 0, \\ &\forall (\boldsymbol{u}, \boldsymbol{v}) \subseteq \mathcal{U}^2 \text{ s.t. } \boldsymbol{u} \neq \boldsymbol{v}\end{aligned}$$

Moreover, thanks to orthogonality, we have $\boldsymbol{c_u}\left(\boldsymbol{U_u}\right) = \mathbb{E}\left(\boldsymbol{c}(\boldsymbol{U}) \mid \boldsymbol{U_u}\right) - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \boldsymbol{c_v}\left(\boldsymbol{U_v}\right)$ and we can write model variance as:

$$\begin{aligned}\mathbb{V}(\boldsymbol{c}(\boldsymbol{U})) &= \sum_{i}^{r} \mathbb{V}\left(\boldsymbol{c}_i\left(U_i\right)\right) \\ &+ \sum_{1 \leqslant i < j \leqslant r} \mathbb{V}\left(\boldsymbol{c}_{i,j}\left(U_i, U_j\right)\right) \\ &+ \ldots + \mathbb{V}\left(\boldsymbol{c}_{1,\dots,r}\left(U_1, \dots, U_r\right)\right) \\ &= \sum_{\boldsymbol{u} \subseteq \mathcal{U}} \mathbb{V}\left(\boldsymbol{c_u}\left(\boldsymbol{U_u}\right)\right)\end{aligned} \tag{5}$$

Eq. 5 allows us to write the influence of any subset of concepts $\boldsymbol{u}$ as its own variance. This yields, after normalization by $\mathbb{V}(\boldsymbol{c}(\boldsymbol{U}))$, the general definition of Sobol' indices.

Definition A.1. *Sobol indices (Sobol, 1993).* The sensitivity index $\mathcal{S}_{\boldsymbol{u}}$ which measures the contribution of the concept set $\boldsymbol{U_u}$ to the model response $f(\boldsymbol{U})$ in terms of fluctuation is given by:

$$\mathcal{S}_{\boldsymbol{u}} = \frac{\mathbb{V}\left(\boldsymbol{c_u}\left(\boldsymbol{U_u}\right)\right)}{\mathbb{V}(\boldsymbol{c}(\boldsymbol{U}))} = \tag{6}$$

$$\frac{\mathbb{V}\left(\mathbb{E}\left(\boldsymbol{c}(\boldsymbol{U}) \mid \boldsymbol{U_u}\right)\right) - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} \mathbb{V}\left(\mathbb{E}\left(\boldsymbol{c}(\boldsymbol{U}) \mid \boldsymbol{U_v}\right)\right)}{\mathbb{V}(\boldsymbol{c}(\boldsymbol{U}))}$$

Sobol indices provide a numerical assessment of the importance of various subsets of concepts in relation to the model's decision-making process. Thus, we have: $\sum_{\boldsymbol{u} \subseteq \mathcal{U}} \mathcal{S}_{\boldsymbol{u}} = 1$.

Additionally, the use of Sobol' indices allows for the efficient identification of higher-order interactions between features. Thus, we can view the Total Sobol indices defined in 2 as the sum of of all the Sobol indices containing the concept $i : \mathcal{S}_{T_i} = \sum_{\boldsymbol{u} \subseteq \mathcal{U}, i \in \boldsymbol{u}} \mathcal{S}_{\boldsymbol{u}}$.

B Implementation Details

We trained 3 different models. For each model, we performed a single run and we split datasets in 70% for train, 10% for validation and 20% for test.

B.1 Trained RoBERTa on Beer dataset

We used a RoBERTa base pretrained on hugging face by Liu et al. (2019b) (all the information on the pretrain can be found in the paper). The model was pretrained on the reunion of five datasets:

- BookCorpus (Zhu et al., 2015), a dataset containing 11,038 unpublished books;
- English Wikipedia (excluding lists, tables and headers) ;
- CC-News (Mackenzie et al., 2020), a dataset containing 63 millions English news articles crawled between September 2016 and February 2019 ;
- OpenWebText (Radford et al., 2019), an open-source recreation of the WebText dataset used to train GPT-2 ;
- Stories (Trinh and Le, 2018) a dataset containing a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas.

We then trained the model on Beer dataset. The model was trained on 2 GPUs for 10 epochs with a batch size of 32 and a sequence length of 512. The optimizer was AdamW with a learning rate of 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e6$

B.2 Trained RoBERTa on IMDB dataset

We used a RoBERTa model already fine-tuned on IMDB from hugging face. This model used the pretraining presented above, we fine-tuned it with 2 epochs, a batch size of 16, and an Adam optimizer with a learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 8$.

B.3 Trained LSTM on IMDB dataset

We created our LSTM with:

```
SentimentRNN(
  (embedding): Embedding(1001, 512)
  (lstm): LSTM(512, 128,
  num_layers=4, batch_first=True,
  bidirectional=True)
  (dropout): Dropout(p=0.3,
  inplace=False)
  (fc_1): Linear(in_features=128,
  out_features=128, bias=True)
  (relu): ReLU()
  (fc_2): Linear(in_features=128,
  out_features=2, bias=True)
  (sig): Softmax(dim=1)
)
```

Then, we trained it on the IMDB dataset. The model was trained on 2 GPUs for 5 epochs with a batch size of 128 and a sequence length of 512. The optimizer was Adam with a learning rate of 1e-4.

C LSTM example

LSTMs are much less complex than RoBERTa, and as such, we can expect them to leverage less and much simpler concepts for their predictions.

In particular, COCKATIEL identified 3 concepts that monopolized the importance score for each class on the RoBERTa model. For the positive class, we had "*the favorite movie*", "*technically good/interesting movie*" and "*good comedie or family movie*". For the negative class, we also had "*the worst movie*", "*middling movie*" and "*boring movie*".

In contrast, in the case of the LSTM (see figure 8), COCKATIEL detected a single important concept per predicted class. For the positive class, this concept encompasses *the positive language elements* mostly, and for the negative class, *the negative elements*. This is a much more basic view of the review classification problem, and COCKATIEL allows us to confirm our intuitions about the richness of the embedding learned by the LSTM.

D Other examples of COCKATIEL explanations for RoBERTa

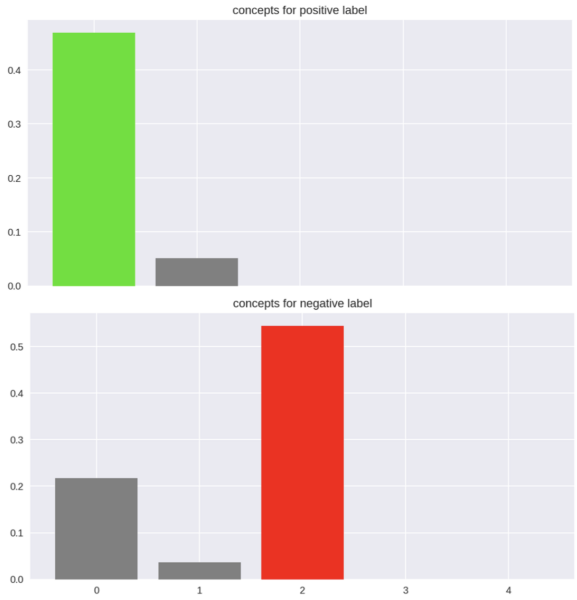

Figure 7: **Concept importance:** The global influence of the NMF concepts on the predictions on LSTM Model is then measured using Sobol indices. We have different concepts for each class (positive and negative label).

Figure 9: Concepts generated with $l = 20$ for some beer reviews with RoBERTa model. The color depicts the aspects for each annotate concept. COCKATIEL is trained only on the label and we use the NMF part of the method to find annotate concepts.

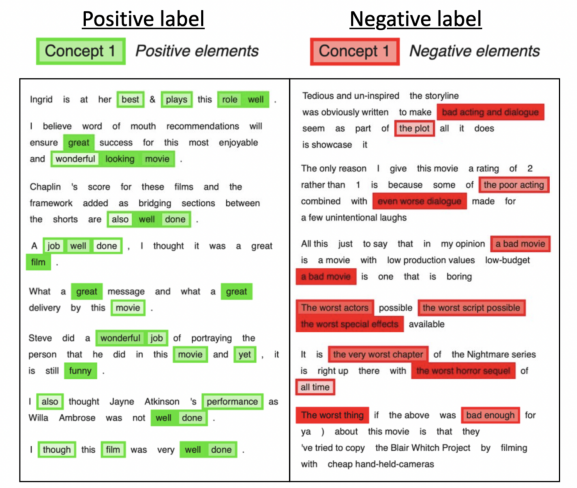

Figure 8: Concepts generated for a LSTM model with $l = 5$ for a few sentences out of IMDB reviews. The colored elements are those important for the concept of the corresponding color (calculated with part *(iii)* of our method). The more colorful the element, the more important it is for the concept (continuously). We have selected the most important concept for each label (see Fig. 7). The name of the concept is chosen manually in view of the important elements corresponding to the concepts.

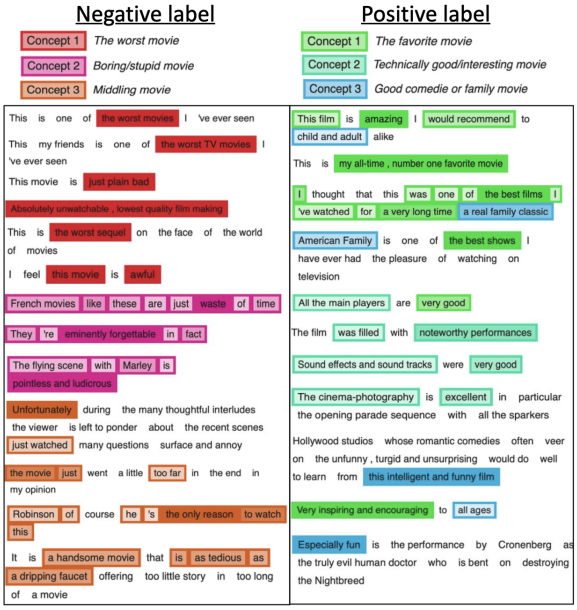

Figure 10: Concepts generated for a RoBERTa model with $l = 20$ for a few sentences taken out of IMDB reviews. The colored elements are those important for the concept of the corresponding color (calculated with part *(iii)* of our method). The more colorful the element, the more important it is for the concept (continuously). We have selected the 3 most important concepts for each label (see Fig. 3). The name of the concept is chosen manually in view of the important elements corresponding to the concepts.

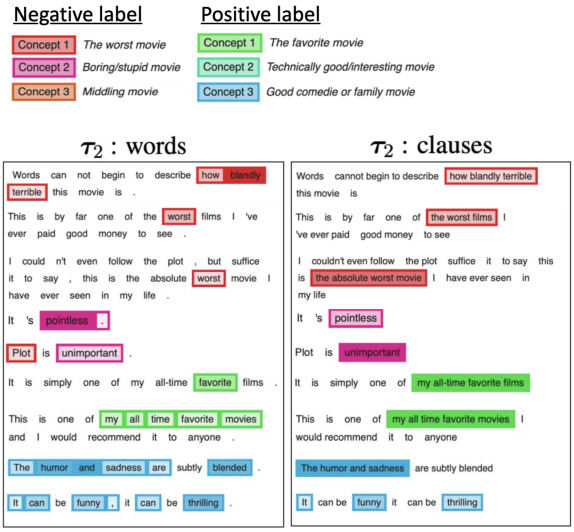

Figure 11: Concepts generated for a RoBERTa model with $l = 20$ for a few sentences taken out of IMDB reviews. The excerpts chosen by an excerpt-extraction function τ_1 are sentences for both (so, we have same concepts). The colored elements are those that are considered to be the most important for the concept of the corresponding color (calculated with part *(iii)* of our method). We compare the visualisations of the same sentences with two different excerpt-extraction functions τ_2: words (on the left) and clauses (on the right). We split the text into clauses for occlusion using the fair library's SequenceTagger implementation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section, after conclusion.

☑ A2. Did you discuss any potential risks of your work?
In limitations section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We did use available and standard datasets.

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The two datasets used are well known and public domain. Their intended use is known.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The two datasets used are well known and public domain.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix 2

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix 2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix 2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix 2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix 2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

We use human annotations but they are only in a used dataset. We did not collect human annotations.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.