
EXPLORING NATURAL EXAMPLE-BASED LOCAL AND GLOBAL EXPLAINABILITY ACROSS DIFFERENT DATA TYPES

Claudia Maiolino

claudia.maiolino@studio.unibo.it

Tian Cheng Xia

tiancheng.xia@studio.unibo.it

ABSTRACT

Natural example-based explainability methods aim at finding representative samples drawn from the training set of the model. The rationale behind these methods is that human intuition and reasoning is heavily based on concepts built upon examples, making these approaches an ideal candidate for enhancing explainability. In this project, we experiment with this class of methods and apply them on image, text, and tabular data in a post-hoc data-agnostic classification setup. We experiment with different models and search hyperparameters to analyze the outcomes in different embedding spaces. From the experimental results, we found out that examples provide insights on how the model behaves. In particular, our main findings are that: examples can provide information on how the model is classifying a sample by giving an idea of the form of the embedding space and the decision boundaries; each class has preferred recurrent axes in the embeddings which allow identifying patterns and biases in the dataset; and examples provide information that can be useful to analyze misclassified samples and to detect simple adversarial attacks.

1 Introduction

Natural examples are a way to explain a given query by looking in the training data for a set of samples that follows some properties. Compared to a more general generative approach, these methods ground examples into existing data which give more control over the output and avoid cases of non-interpretable results. Example-based explainability methods themselves are inspired from human reasoning as it usually builds new knowledge on top of past experience. In this project, we explore methods to search for similar, semi-factual, and counterfactual examples, as well as prototypes. We work in a post-hoc data-agnostic setup for classification models with the common architecture composed of a backbone and classification layer, allowing to work with embeddings. We experiment with image, text, and tabular datasets trained on different models to analyze the results when working in different embedding spaces. We also experiment with different hyperparameters for the search methods and ultimately apply these methods for explainability, concepts extraction, error analysis, and adversarial attack detection. In particular, we use as datasets EuroSAT [7], brain tumor MRIs, COVID-19 fake news [19], textual diagnosis, COMPAS [10], and weather classification; and as models MobileNet [9], MobileViT [16], BERT [4], BioBERT [13], RoBERTa [14], BioMed-RoBERTa [6], and fully-connected feedforward networks. As distances, we experiment with the traditional L1, L2, and cosine distance, as well as a weighted distance we propose based on LIME or SHAP.

The rest of the report is structured as follows: in Section 2, we formalize the problem of natural example search and describe each type of example; in Section 3, we detail the four applications we mentioned for explainability, concepts extraction, error analysis, and adversarial attack detection; in Section 4, we describe our experimental setup; in Section 5, we present our results and findings; in Section 6, we wrap-up with the final observations and future work, while in Section 7, we list the contributions of each group member.

2 Natural example search

In this section, we present the main methods to select examples we employed in this project. We focus on natural examples, which are those that are extracted from the training set. Moreover, we work in a post-hoc data-agnostic setup for classification and consider models using the traditional architecture of a feature extractor $e(\cdot) : \mathbb{R}^{\text{in}} \rightarrow \mathbb{R}^E$ and a classification layer $f(\cdot) : \mathbb{R}^E \rightarrow \mathbb{R}^C$, allowing to work with embedding spaces regardless of the type of input data. More formally, given the embeddings $e(x_q)$ of a query sample x_q and the training set $\mathcal{D}^{\text{train}} = \{(x_i^{\text{train}}, c_i^{\text{train}})\}_{i=1}^{N_{\text{train}}}$, we

are interested in extracting some examples $\{x_j^{\text{train}}\}_{j \in S}$ with $S \subseteq \{1, \dots, N_{\text{train}}\}$ such that they satisfy some conditions with the prediction $\arg \max f(e(x_q))$ and some distance criteria dist to the embedding $e(x_q)$ of the query.

2.1 Similar examples

Similar (or factual) examples are those of the training set that are the closest to the query. Their application in explainability is inspired on how humans reason based on past experience. This class of examples are the most direct in explaining a query as they provide other instances that would have been classified in the same way as the query. However, a drawback is that the connection between the query and the examples a model considers as similars is not always clear and understandable [17].

In our case, we are interested in examples whose embeddings are the most similar to the one of the query. In the literature, most of the approaches for similar examples are based on k -nearest neighbors (k -NN) [20], which can be easily applied with embeddings. Therefore, the design choices for retrieving similar examples revolve around deciding the embedding space to work with and the distance metric to use. The problem of finding k similar examples can be formulated as:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{\text{train}}\}, \\ |S|=k}} \sum_{j \in S} \text{dist}(e(x_j^{\text{train}}), e(x_q)) \\ & \text{subject to } c_j^{\text{train}} = \arg \max f(e(x_q)) \quad \forall j \in S. \end{aligned}$$

We also consider a variation of this method, which will be useful for analyzing some instances, that we denote as “most similar” examples and are computed as the similar examples we mentioned without the constraint of extracting samples of the same class as the query prediction. This allows to further study the embeddings to have an idea of where they are positioned in the latent space and extract insights on how the model is making the prediction.

2.2 Counterfactual examples

Counterfactual examples are those closest to the query but belonging to a different class. Intuitively, they provide some information about what would happen if some parts of the input are modified [1]. Counterfactuals have grounding both in psychology and philosophy, and they are mostly related to the concept of cause and effect [20]. Applied to explainability, they are a way to highlight which parts of the input query are the most sensitive in predicting a different class. In our case, working with embedding spaces, they can also provide some insights about the decision boundaries of the classifier.

In the literature, the most straightforward method for finding counterfactuals is by using nearest unlike neighbors (NUN) [5]. The problem can be formulated as:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{\text{train}}\}, \\ |S|=k}} \sum_{j \in S} \text{dist}(e(x_j^{\text{train}}), e(x_q)) \\ & \text{subject to } c_j^{\text{train}} \neq \arg \max f(e(x_q)) \quad \forall j \in S. \end{aligned}$$

However, this approach is sensitive to outlier counterfactuals that end up close to the query in the latent space. Therefore, we follow the method proposed in KLEOR [2] which they call as *the nearest miss that lies over the boundary* (NMOTB) and aims at counteracting this issue. The counterfactual examples with this approach are computed as:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{\text{train}}\}, \\ |S|=k}} \sum_{j \in S} \text{dist}(e(x_j^{\text{train}}), e(x_q)) \\ & \text{subject to } c_j^{\text{train}} \neq \arg \max f(e(x_q)) \wedge \\ & \quad \text{dist}(e(x_q), e(x_{\text{sim}}^{\text{train}})) < \text{dist}(e(x_j^{\text{train}}), e(x_{\text{sim}}^{\text{train}})) \quad \forall j \in S, \end{aligned}$$

where $x_{\text{sim}}^{\text{train}}$ is a similar example of x_q .

2.3 Semi-factual examples

Semi-factual examples are those farthest from the query but belonging to the same class. They give an idea of which input features can be modified while keeping the prediction unchanged, making them typically paired alongside counterfactuals in the literature due to their semantic similarity [20]. For explainability, they can be used to detect the boundaries that the features should maintain to stay within the same class or, in some cases, locate which features

are less impactful in the prediction. Also, analyzed together with counterfactuals, semi-factuals allow obtaining more precise information on the decision boundaries.

In practice, a reference counterfactual example x_{cf}^{train} is required to find semi-factuals. A naive way to compute them is to consider the samples of the same class as the query and closest to x_{cf}^{train} . The problem is therefore defined as:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{train}\}, \\ |S|=k}} \sum_{j \in S} \text{dist}(e(x_j^{train}), e(x_{cf}^{train})) \\ & \text{subject to } c_j^{train} = \arg \max f(e(x_q)) \quad \forall j \in S. \end{aligned}$$

As the case of counterfactuals, there might be instances where the boundaries do not necessarily produce well-behaved regions, which would lead to outliers to be selected as examples. Therefore, we follow the *global similarity* approach of KLEOR [2] and constrain the problem to select examples whose distance to the query is lower than the distance between the query and the reference counterfactual:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{train}\}, \\ |S|=k}} \sum_{j \in S} \text{dist}(e(x_j^{train}), e(x_{cf}^{train})) \\ & \text{subject to } c_j^{train} = \arg \max f(e(x_q)) \wedge \\ & \quad \text{dist}(e(x_q), e(x_j^{train})) < \text{dist}(e(x_q), e(x_{cf}^{train})) \quad \forall j \in S. \end{aligned}$$

2.4 Prototypes

Prototypes are data samples that aim to be representative of the data distribution. They provide some insights on which data points are most representative of the training set and can be used to determine how the embedding space is partitioning the samples.

In practical terms, a post-hoc method to compute prototypes in our case is to use clustering. We consider the simplest case where we want to search for a fixed number of prototypes using k -means and leave exploring other methods for future work. Therefore, we can formulate the problem of finding the prototypes for class c as follows:

$$\begin{aligned} & \arg \min_{\substack{S \subseteq \{1, \dots, N_{train}\}, \\ |S|=k}} \sum_{\substack{i=1, \dots, N_{train} \\ c_i^{train}=c}} \min_{j \in S} \{ \text{dist}(e(x_i^{train}), e(x_j^{train})) \} \\ & \text{subject to } c_j^{train} = c \quad \forall j \in S. \end{aligned}$$

3 Applications

In this section, we present the applications that can be built on top of the methods we presented in Section 2. We think that examples are relatively flexible and can be employed in various scenarios. In this project, we experiment with examples applied for (i) explainability, (ii) concepts extraction, (iii) error analysis, and (iv) adversarial attack detection.

3.1 Example-based explainability

Given the examples of a query point, the most straightforward application is to use them as an explanation of the prediction which can be useful both for the end-user and the model developer. Similar examples provide information on which data points would have been classified in the same way as the query, allowing to extrapolate common features that contributed to the prediction and understand what the embedding space is encoding as similar samples; counterfactuals provide an idea of which changes would push the current prediction to another class and can provide insights on which features are the most sensitive; semi-factuals, with the same idea of counterfactuals, also provide knowledge on which features are moving the prediction towards another class and can be useful to have an idea of where the decision boundaries are.

3.2 Concepts extraction

It is well known that the embeddings produced by a feature extractor are semantically rich and compress many features of the input. Due to the effectiveness of example search methods, we argue that related inputs also share common patterns in embedding space, which can be a possible way to achieve post-hoc global explainability. To test this

hypothesis, we consider the two components of a traditional classification model – the feature extractor and the classifier – separately and apply a local explainability method (e.g., SHAP) in a two-step method depicted in Figure 1 that works as follows:

1. Explain the output of the classifier from the embedding space to determine the importance of each of its axes.
2. Explain the path from the input sample to the embeddings, considering only the relevant axes filtered by, for instance, thresholding.

We expect that the explanation obtained using this method highlights the most relevant parts of the input that contributed to the most important axes of the embeddings, which ideally might also correspond to some sort of human-understandable concept or pattern.

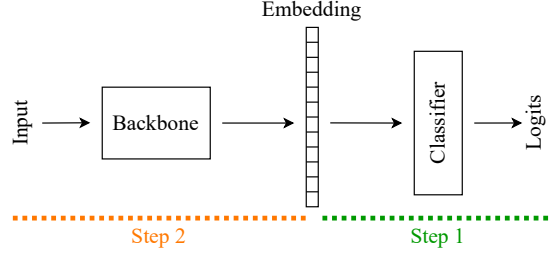


Figure 1: Workflow for concepts extraction. We first apply a local explanation method on the classifier to score the axes of the embedding. Then, we explain the backbone to score the input and only consider the values associated to the highest scoring axes of the embedding.

3.3 Error analysis

Examples can also be used to have an idea of why a model misclassifies a sample, particularly when there is uncertainty in the confidence of the prediction. Given a misclassified query point, we argue that its embedding might contain information of the correct class which can be visualized through examples. We expect some form of intersection between the query and the semi-factuals, which indicates that the query itself is positioned near the decision boundaries. On the same idea, we can also expect that, by considering the most similar examples (i.e., those without the constraint of being of the same class as the query), nearby samples are not necessarily all belonging to the predicted class. Finally, we would expect that counterfactuals provide some information on the nearest classes which can give insights on which other possible classes the model is considering possible for the query. Overall, we think that all these observations provide some reference points for the developers to refine the model by, for instance, enhancing the dataset.

3.4 Adversarial attack detection

Adversarial attacks based on perturbation are types of attack that introduce slight modifications to the input that are mostly imperceptible to humans but are able to hijack the prediction to another class. We think that, in some cases, the adversarial input will be encoded into an embedding that still preserves some of its original information which will emerge in the examples found for it. We experiment with this hypothesis by inspecting counterfactuals and most similar examples, expecting to obtain samples of the original class. In practice, we consider the simplest case of a white-box perturbation attack on images and leave the study of more complex attacks for future development.

4 Experimental setup

4.1 Datasets

In this project, we experiment with example search methods using different types of data and work in their embedding spaces. In particular, we experiment with image, text, and tabular data classification tasks. Preliminary experiments were performed with images using toy datasets such as MNIST [12] and CIFAR10 [11], then, each data type has been experimented with using two datasets each.

For images, we use EuroSAT¹ [7] and brain tumor MRIs². The former contains satellite images taken from ESA’s Sentinel-2: it is composed by ten different classes of land types, collected over 34 European countries. The latter contains brain MRIs taken from different angles. It is the composition of different Kaggle challenges and its goal is to classify whether the scan contains a glioma, meningioma, pituitary, or no tumor.

For text, we use COVID-19 fake news³ [19] and diagnosis text⁴. The former consists of social media posts and articles of real and fake news on COVID-19. The latter is a synthetic dataset based on DDXPlus [22] from which age, sex, and symptoms have been parsed through a text-generation model to create first-person descriptions of the symptoms to simulate patients. The goal given the text is to classify it into one disease among ten possible classes.

For tabular data, we use COMPAS⁵ [10] and weather classification⁶. The former is the well-know dataset used by ProPublica to evaluate the fairness of the homonymous decision-making system whose task was to settle the level of recidivism of a real offender. The latter is a synthetic dataset containing weather information with the goal of determining whether they represent a rainy, sunny, cloudy, or snowy day.

4.2 Embedding spaces

A key element for the example search methods we selected is the embedding space. Therefore, different architectures might vary in results and, to test this, we experiment with various models.

For images, the comparison revolves around convolutional neural networks and vision transformer models, which are known to have their own advantages and disadvantages [15]. In particular, convolutions build a hierarchy of features based on the local receptive field of the kernels while the multi-head attention mechanism of transformers forms the embeddings by considering global relations. In practice, we experiment with MobileNetV3 [9] and MobileViT [16], both starting from the pre-trained weights on ImageNet [3].

For textual data, the current state-of-the-art is the transformer architecture, usually employed after a pre-training phase on large corpora of data. In particular, as we are tackling classification tasks using embeddings, we are interested in encoder-only transformers and a key point of comparison of these models is the pre-training corpora that have been used as there can be differences in performance when using a general-purpose and domain-specific model [23]. For our experiments, we use BERT [4] and BioBERT [13] for the diagnosis text dataset and RoBERTa [14] and BioMed-RoBERTa [6] for COVID-19 fake news, which allows us to try both different pre-training routines and corpora.

For tabular data, we take the traditional approach of using fully-connected feedforward networks. In this case, we consider as embeddings for computing examples the activations at the last layer before the last linearity that maps into classes. This setup allows testing example search on a model that is architecturally small and that is trained from scratch without any form of pre-training.

4.3 Distances

All methods to search examples we use in this project rely on some concept of distance between embeddings. Therefore, the choice of the distance metric becomes a topic of our analysis. We select and experiment with traditional metrics such as the Manhattan (L1), Euclidean (L2), and cosine distance. For the specific case of this project, as we are working with classification models, we also propose a weighted distance based on the scores of local explainability techniques similar to the concepts extraction method described in Section 3.2. Given a query x_q , we consider the classifier stage of the model and apply the explainability method to obtain the importance scores $w_q \in \mathbb{R}^E$ of each axis of the input embeddings $e(x_q) \in \mathbb{R}^E$. The distance between the query and another sample x can be computed as the distance of the axis-wise weighted embeddings:

$$\text{dist}((e(x_q) * w_q), (e(x) * w_q)),$$

where $*$ is the element-wise product. The rationale behind this method is that, as embeddings are information-rich, weighing with respect to the importance scores of the query could lead to examples that better capture its main content. However, we must also observe that a downside of this method is that the distance metric is query-specific, therefore requiring to be recomputed for each different query. Also, we experiment with different setups of these weights by considering w_q as-is, in absolute value, and by using only its positive components.

¹https://huggingface.co/datasets/blanchon/EuroSAT_RGB

²<https://huggingface.co/datasets/PranomVignesh/MRI-Images-of-Brain-Tumor>

³https://huggingface.co/datasets/nany1025/covid_fake_news

⁴<https://huggingface.co/datasets/ninaa510/diagnosis-text>

⁵<https://huggingface.co/datasets/imodels/compas-recidivism>

⁶<https://www.kaggle.com/datasets/nikhil7280/weather-type-classification>

4.4 Evaluation metrics

Evaluating an explainability technique is notoriously cumbersome in terms of automatic quantitative metrics [21]. However, for the case of example-based explainability, some metrics have been proposed in the literature to capture the meaningfulness of the examples. In particular, to provide a preliminary evaluation on the effectiveness of the methods, we use as metrics: non-representativeness and diversity [18], which are computed as follows:

$$\text{non-repres} = \frac{\sum_{j \in S} l(f(e(x_j^{\text{train}})), c_q)}{k}$$

$$\text{diversity} = \sum_{\substack{j_1, j_2 \in S, \\ j_1 \neq j_2}} \frac{\text{dist}(e(x_{j_1}^{\text{train}}), e(x_{j_2}^{\text{train}}))}{2k},$$

where l is the training loss function (i.e., cross-entropy in this project), c_q is the predicted class of the query, and k is the number of examples. In terms of interpretation, non-representativeness models the fidelity of the examples in terms of prediction accuracy and lower values indicate examples that are in agreement with the class of the query. By definition, we prefer similar and counterfactual examples to have a lower non-representativeness, while semi-factuals should maximize it. Instead, diversity measures the examples in terms of sparsity in the latent space, which ultimately gives an indicator on redundancy and coverage, making it a metric we want to maximize.

Alongside the metrics, we perform a manual qualitative analysis on the results. To obtain more structured observations, we define a checklist to follow for the evaluation process. Firstly, when using examples to explain a given query, we assess which features the similar examples have in common with the query and what type of correlation, if any, is present between semi-factual and counterfactual examples. Instead, for concept analysis, we first check whether there are patterns in the activations considering multiple sample of the same class and whether the attribution maps provide insights on the embedding space and the prediction. Lastly, for error analysis and adversarial attack detection, we check if most similar and counterfactual examples provide some information of the correct outcome and whether these observations are in line with the confusion matrices of the model predictions.

5 Results

5.1 Preliminaries

Due to the high number of possible combinations of hyperparameters we can use, we first conduct a preliminary study on CIFAR10 to filter out redundant setups. First, we observed that different strategies for the weighted distance result in negligible differences (for compactness, we refer to Appendix A.1 for more details). Therefore, we perform all the experiments with weighted distances by only considering positive scores. Instead, in terms of metrics, we present the results in Table 1. It can be seen that there is an expected increase in term of non-representativeness from similar to counterfactual examples and that the numbers are roughly similar, indicating that the distance metric might not be too relevant. Also, in Table 2, we report the average overlap for the examples found using the plain L1, L2, and cosine distance. It can be seen that there is a strong consensus in the results with similars and counterfactuals, and a slightly weaker result for semi-factuals. Therefore, with these observations, for this project we select the L2 distance as the metric of choice and test it in its plain and weighted versions.

Table 1: Quantitative evaluation on CIFAR10 using MobileViT with 5 examples and different distance metrics

Distance	Similars		Semi-factuals		Counterfactuals	
	Non-repres.	Diversity	Non-repres.	Diversity	Non-repres.	Diversity
L1	0.03	3.73	0.10	4.63	0.86	7.08
L2	0.03	3.73	0.09	4.68	0.84	7.08
cosine	0.02	3.78	0.08	4.58	0.96	7.28
SHAP weighted L1	0.02	3.78	0.09	4.76	1.02	7.58
SHAP weighted L2	0.02	3.86	0.07	4.89	1.28	7.82
SHAP weighted cosine	0.02	3.92	0.05	4.60	1.13	7.58
LIME weighted L1	0.03	3.82	0.10	4.85	0.85	7.44
LIME weighted L2	0.04	3.88	0.09	4.81	0.96	7.61
LIME weighted cosine	0.02	3.84	0.06	4.50	0.86	7.54

Table 2: Overlap evaluation on CIFAR10 with MobileViT

Example type	Overlap ($k = 1$)	Overlap ($k = 3$)	Overlap ($k = 5$)
Similars	65.47	70.36	72.68
Semi-factuals	47.00	51.40	52.51
Counterfactuals	82.27	79.49	81.49

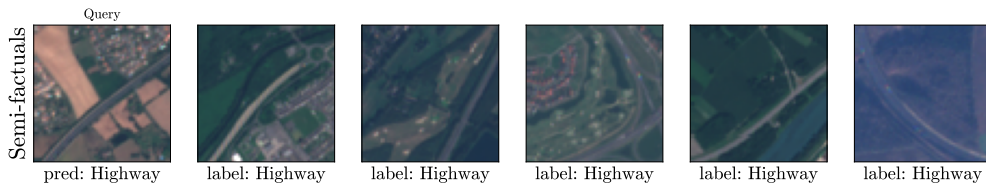
5.2 EuroSAT

For the EuroSAT dataset, we present the metrics in Table 3. The results show that similars and semi-factuals reach a very low non-representativeness, while counterfactuals are less representative, as expected. Furthermore the two models behave similarly, probably due to the fact they almost perfectly fit the dataset.

Table 3: Quantitative evaluation on EuroSAT with 5 examples

Example type	MobileNet		MobileViT	
	Non-repres.	Diversity	Non-repres.	Diversity
Similars [L2]	0.01	116.74	0.01	20.93
Similars [pos. LIME weighted L2]	0.00	119.07	0.01	21.33
Similars [pos. SHAP weighted L2]	0.01	121.33	0.00	21.33
Semi-factuals [L2]	0.03	135.04	0.11	31.10
Semi-factuals [pos. LIME weighted L2]	0.03	130.84	0.10	31.18
Semi-factuals [pos. SHAP weighted L2]	0.02	127.45	0.10	31.29
Counterfactuals [L2]	7.01	134.75	3.31	42.22
Counterfactuals [pos. LIME weighted L2]	6.94	140.30	2.96	44.16
Counterfactuals [pos. SHAP weighted L2]	7.05	144.90	3.14	44.12

The similarity of the behavior of the two models is also confirmed by the qualitative results, for which we show some selected samples. We can see that for the *Highway* class, the similar examples in Figure 2 maintain the color and the pattern of the query on the neighborhood of the street, while the semi-factuals in Figure 3 do not. In line with the confusion matrix in Figure 8, the class of the counterfactuals in Figure 4 is prevalently the *River* one. The same considerations could be done for the *Industrial* class, where the similar examples in Figure 5 maintain the dimension and the color of the buildings in the query compared to the semi-factuals in Figure 6, while the counterfactuals in Figure 7 are prevalently classified as *Residential*, always according to the confusion matrix and our intuition.

Figure 2: Five similar examples considering a query of class *Highway* using L2 distance on EuroSAT with MobileViTFigure 3: Five semi-factuals considering a query of class *Highway* using L2 distance on EuroSAT with MobileViT

For what concerns concepts extraction, we can see from Figure 9 that both models share some recurring axes that are mostly active for a specific class, enforcing the observation that a small subset of axes is responsible for the prediction of a certain class and encodes some specific information for it. Indeed, if we compute the attribution maps of the image

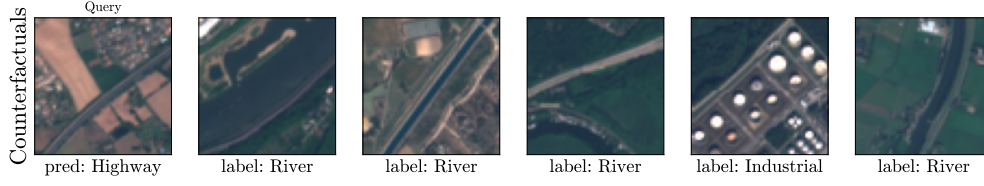


Figure 4: Five counterfactuals considering a query of class *Highway* using L2 distance on EuroSAT with MobileViT

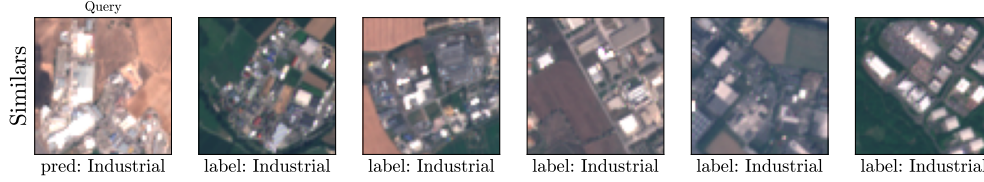


Figure 5: Five similar examples considering a query of class *Industrial* using L2 distance on EuroSAT with MobileViT

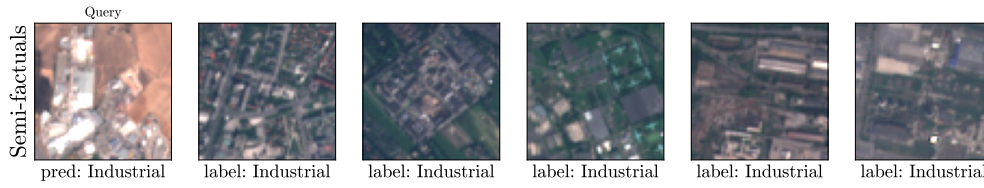


Figure 6: Five semi-factuals considering a query of class *Industrial* using L2 distance on EuroSAT with MobileViT

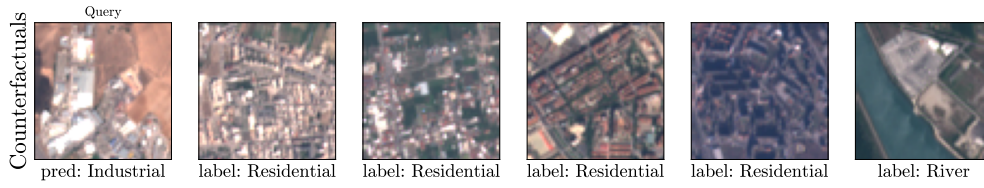


Figure 7: Five counterfactuals considering a query of class *Industrial* using L2 distance on EuroSAT with MobileViT

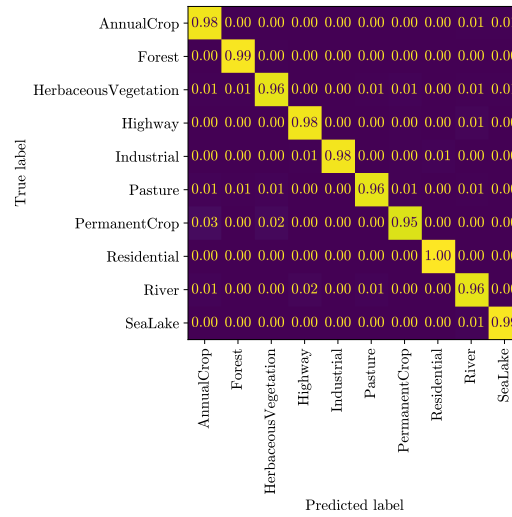


Figure 8: Confusion matrix of the MobileViT model on EuroSAT

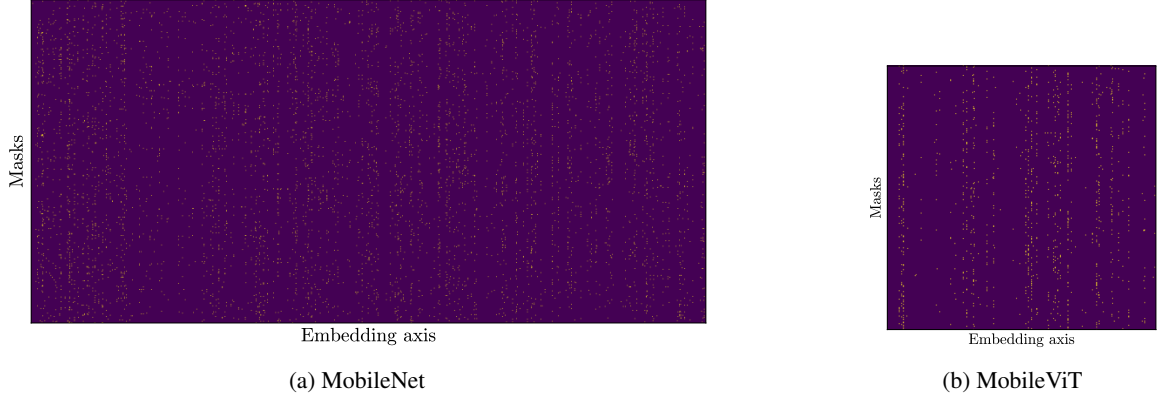


Figure 9: Top relevant embedding axes of correctly classified samples with class *Highway*

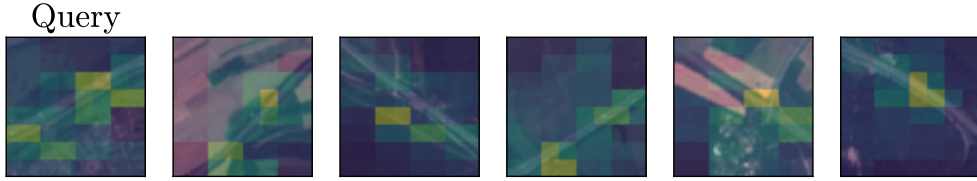


Figure 10: Attribution maps on EuroSAT with MobileViT applied on a query of class *Highway* and its similar examples

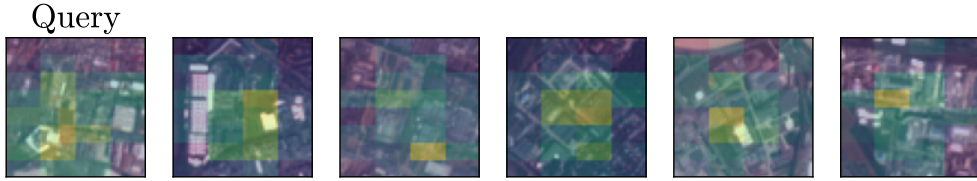


Figure 11: Attribution maps on EuroSAT with MobileViT applied on a query of class *Industrial* and its similar examples

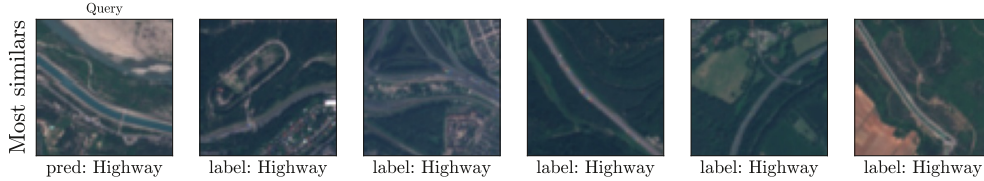


Figure 12: Five most similar examples using L2 distance on EuroSAT with MobileViT on a misclassified query. The correct class is *River*

by only considering the high-scoring axes, we can see from Figure 10 that the relevant patches of the image are the ones that follow the path of the street. However, the attribution maps are not always clearly interpretable with human-like concepts as we can see in Figure 11.

In terms of error analysis, we choose, as before, the more interpretable sample consisting of a query of the *River* class misclassified as *Highway*. We can see in Figure 13 that the counterfactual examples are of the correct class, as expected. Instead, in Figure 12 is shown that the most similar examples are still of the wrong predicted class. One possible explanation could be that the *River* class is identified by a larger path than the *Highway* one, which is more narrow, and therefore the model is more confident in the prediction of the *Highway* class, even if the river is present in the image. We can see this insight by analyzing the attribution maps of the correct (i.e., the *River*) and the wrong (i.e., the *Highway*) prediction in Figure 14. The attribution maps of the correct prediction indicate a larger path, while the attribution maps of the wrong prediction are more focused on a smaller path. This is also confirmed if we check the prototypes of the two classes in Figure 15, where the *Highway* class has a more narrow path than the *River* one.

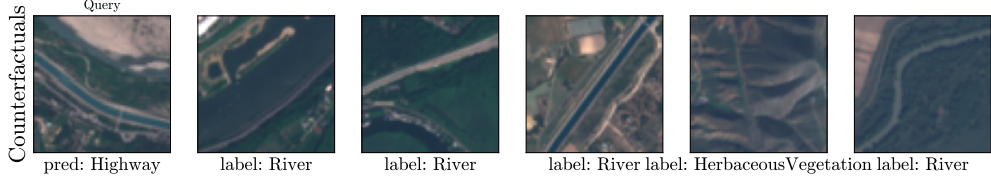
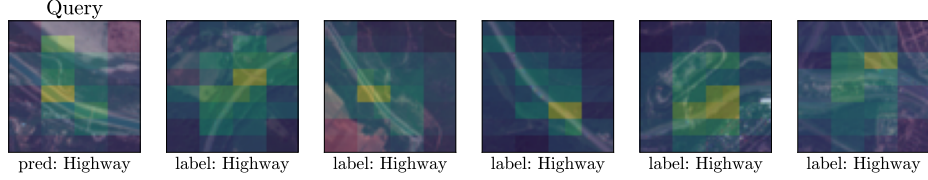
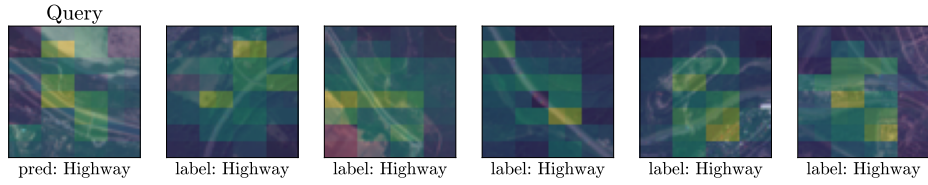


Figure 13: Five counterfactuals using L2 distance on EuroSAT with MobileViT on a misclassified query. The correct class is *River*

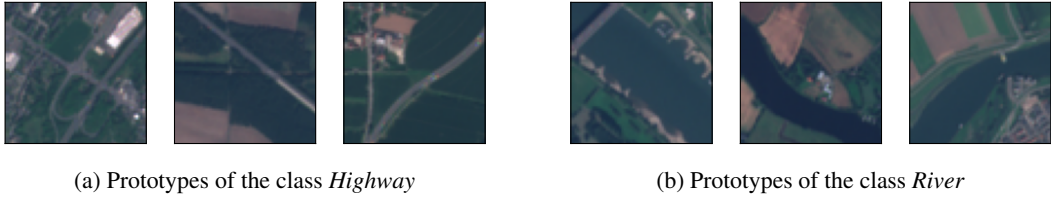


(a) Attribution maps considering the predicted class *Highway*



(b) Attribution maps considering the true class *River*

Figure 14: Attribution maps on EuroSAT with MobileViT applied on a query and its similar examples



(a) Prototypes of the class *Highway*

(b) Prototypes of the class *River*

Figure 15: Prototypes of the classes *Highway* and *River* on EuroSAT with MobileViT.

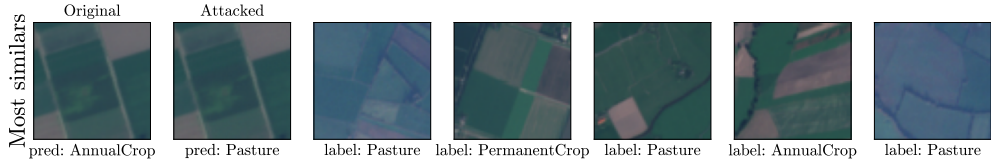


Figure 16: Five most similar examples using L2 distance on EuroSAT with MobileViT on an adversarial query. The original query is of the class *AnnualCrop*, once attacked it is classified as *Pasture*.

Lastly, we present the results of the adversarial attack detection on EuroSAT. We can see in Figure 16 that the most similar examples are coming from different classes and do not show an unique and simple behavior. Instead, for the counterfactuals in Figure 17, we can observe that most of the examples are of the class of the query before the attack, indicating that some information has been preserved and is still recoverable.

5.3 Brain tumor MRI

For the brain tumor MRI dataset, we present the metrics in Table 4. Due to the fact that both models fit the dataset nearly perfectly, the results in this case show that similars and semi-factuals reach a very low non-representativeness, which does not provide any relevant information. Instead, counterfactuals are correctly the ones with the highest

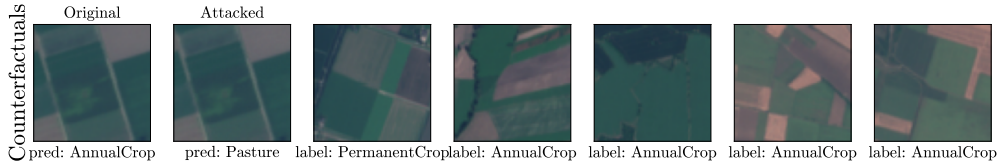


Figure 17: Five counterfactuals using L2 distance on EuroSAT with MobileViT on an adversarial query. The original query is of the class *AnnualCrop*, once attacked it is classified as *Pasture*.

Table 4: Quantitative evaluation on brain tumor MRI with 5 examples

Example type	MobileNet		MobileViT	
	Non-repres.	Diversity	Non-repres.	Diversity
Similar [L2]	0.00	50.03	0.00	1.61
Similar [SHAP weighted L2]	0.00	52.09	0.00	1.65
Similar [LIME weighted L2]	0.00	51.17	0.00	1.62
Semi-factuals [L2]	0.00	61.70	0.05	3.44
Semi-factuals [SHAP weighted L2]	0.00	63.99	0.05	3.58
Semi-factuals [LIME weighted L2]	0.00	63.50	0.05	3.53
Counterfactuals [L2]	11.21	70.77	3.57	4.44
Counterfactuals [SHAP weighted L2]	12.82	72.90	4.53	6.99
Counterfactuals [LIME weighted L2]	11.79	71.41	4.50	7.17

non-representativeness. In terms of diversity, there is an increase from similars to counterfactuals, with the former being the lowest, indicating that similar examples are less spread across the latent space.

Qualitatively, we observed similar results for both MobileNet and MobileViT. Firstly, similar examples are the most coherent across all samples we evaluated, allowing to conclude that, as shown for instance in Figure 18, the embedding spaces of both models are able to capture features such as type of scan and position of the tumor. Instead, by considering semi-factuals and counterfactuals, we observed some less well-behaved results but nevertheless encoding some interpretable information. In particular, we observed instances where the differences are either in terms of type of scan (as shown in Figure 19), position of the tumor (as shown in Figure 20), or size of the patient’s head in the scan. Finally, in terms of distance metrics, we observed that the same behavior is noticeable in all the three we tested, indicating that, even if weighted, most of the information is contained in specific axes of the embedding. By moving to prototypes, we observed, as seen in Figure 21, that the results are scans from different angles, indicating that in the embedding space this information is accounted for, making them clustered accordingly.

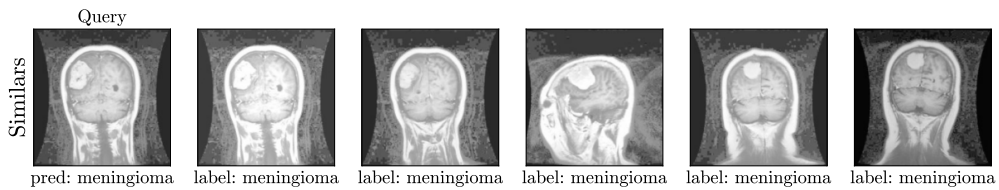


Figure 18: Five similar examples using L2 distance on brain tumor MRI with MobileViT. All examples have in common the position of the tumor and, in all but one case, the same type of scan.

By analyzing the axes of the embeddings more in depth, we can observe from Figure 22 that both MobileNet and MobileViT present some recurring axes that are mostly active for a specific class. This supports the observation we previously made on the fact that a small subset of axes are responsible for the prediction of a class and it might encode some specific information for it. Indeed, if we compute the attribution maps of the image by only considering the high-scoring axes, we can see from Figure 23 that the relevant portions of the image are detected as the most important. However, this does not always result in interpretable attribution maps as shown in Figure 24, which indicates that the most important axes of the embedding might encode the relevant information for classifying the class, but that it is not necessarily what intuition would suggest and be directly interpretable.

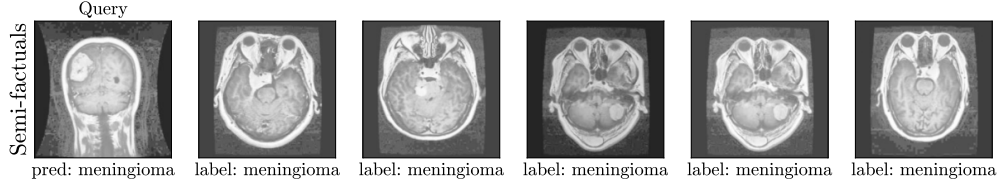


Figure 19: Five semi-factuals using L2 distance on brain tumor MRI with MobileViT. All examples are from a different type of scan.

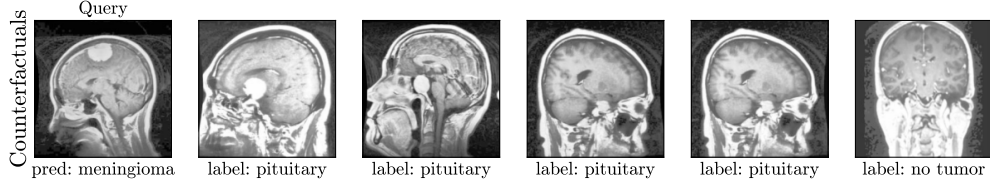


Figure 20: Five counterfactuals using L2 distance on brain tumor MRI with MobileViT. Almost all examples share the same type of scan, but the tumor is of a different type.

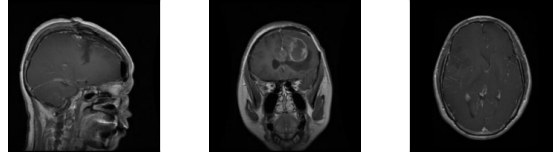


Figure 21: Three prototypes on brain tumor MRI with MobileViT for the class *glioma*. Each sample represents a scan from a different angle.

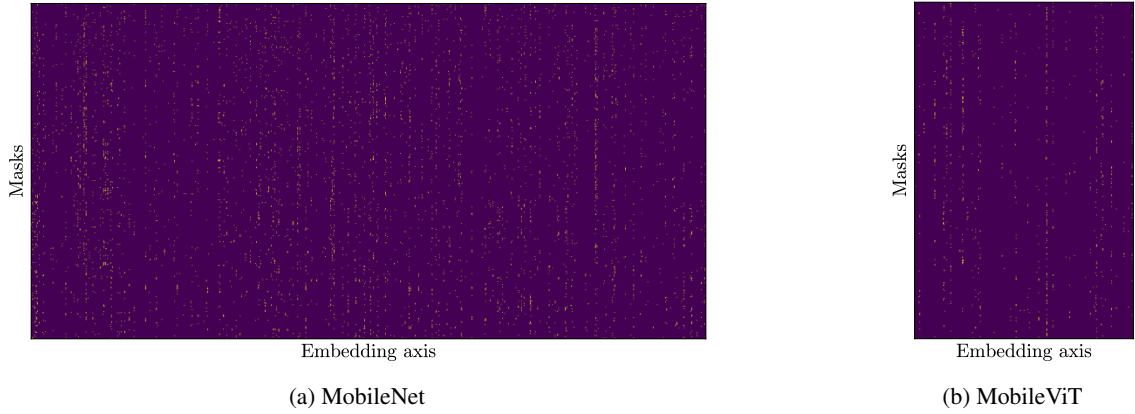


Figure 22: Top relevant embedding axes of correctly classified samples with class *glioma*

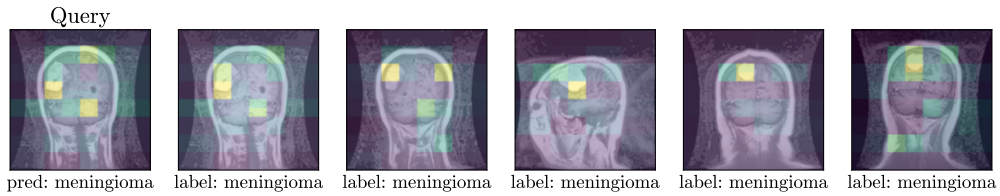


Figure 23: Attribution maps on brain tumor MRI with MobileViT applied on a query and its similar examples

In terms of error analysis, as we already mentioned, on this dataset both models performs nearly perfectly and have a high prediction confidence for all samples. This includes misclassified images which are indistinguishable in embedding

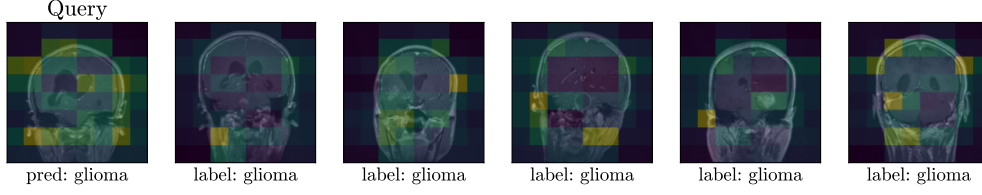


Figure 24: Attribution maps on brain tumor MRI with MobileViT applied on a query and its similar examples

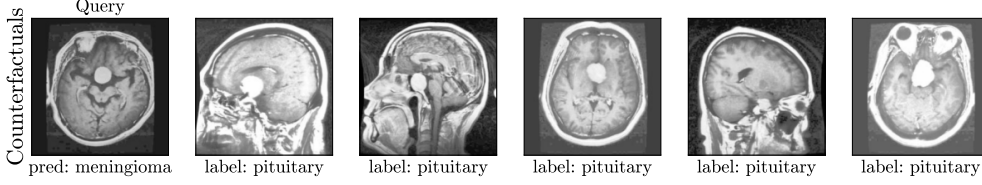


Figure 25: Five counterfactuals using L2 distance on brain tumor MRI with MobileViT on a misclassified query. The correct class is *pituitary*.

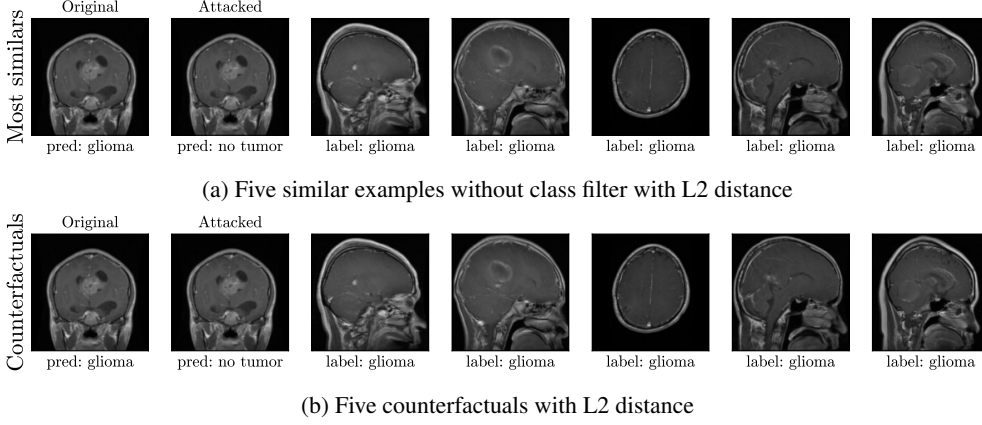


Figure 26: Adversarial attack on brain tumor MRI with MobileViT

space from the ground-truth, making the analysis of the closest embeddings not fruitful. Nevertheless, by inspecting the counterfactuals, it is still possible to extract some insights. In particular, we observed that, in most of the cases, the closest counterfactuals are those of the correct class indicating that embeddings, although leading to the wrong prediction, are still nearby the correct ones and share some of their information. This provides some knowledge on how the training data was encoded by the model and can be useful to determine possible directions to improve the results by enhancing the dataset.

Lastly, we observed that examples can also provide some information in case of adversarial attacks similarly to misclassified samples. In this case, as shown in Figure 26, both the most similar samples and counterfactuals are able to capture the original class, which indicates that attacks that are not too disruptive still leave relevant information in the embeddings.

5.4 COVID-19 fake news

We report the results of the quantitative evaluation on the COVID-19 fake news dataset in Table 5. The results of non-representativeness are consistent with our expectations; furthermore, considering that the two models are the same version of RoBERTa pre-trained on different corpus, the values are quite the same for both models.

Considering a sample classified as *fake*, in Table 6 we highlight that the similar examples are consistent with our intuition of similarity (i.e., the subject of all the tweets is Donald Trump). Instead, we can observe that the semi-factual and counterfactual examples, in Table 7 and Table 8 respectively, do not show a particular evidence of a link with the query. Nevertheless, due to the fact that the classification task is binary, counterfactual and semi-factual examples could

Table 5: Quantitative evaluation on COVID-19 fake news with 5 examples

Example type	RoBERTa		BioMed-RoBERTa	
	Non-repres.	Diversity	Non-repres.	Diversity
Similar [L2]	0.01	3.08	0.02	2.96
Similar [pos. LIME weighted L2]	0.00	2.99	0.01	2.90
Similar [pos. SHAP weighted L2]	0.00	2.95	0.01	2.91
Semi-factuals [L2]	0.04	4.64	0.04	3.89
Semi-factuals [pos. LIME weighted L2]	0.03	4.82	0.02	4.11
Semi-factuals [pos. SHAP weighted L2]	0.02	5.12	0.01	4.07
Counterfactuals [L2]	0.25	7.25	0.29	7.03
Counterfactuals [pos. LIME weighted L2]	0.23	7.48	0.26	7.37
Counterfactuals [pos. SHAP weighted L2]	0.23	7.48	0.26	7.32

Table 6: Five similar examples using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	.@realDonaldTrump says US is at the top globally in COVID-19 fight. It isn't. https://t.co/JY6Ojdjhwy #coronavirus #Covid19 https://t.co/Gf8j2nLjZb
Label: fake	.@realDonaldTrump says U.S. is at the <i>top globally</i> in COVID-19 <i>fight</i> . It isn't. https://t.co/JY6Ojd1G7Y https://t.co/15iUPwpH64
Label: fake	.@realDonaldTrump has claimed that the U.S. has "one of the lowest mortality rates in the world" from COVID-19. That's False. https://t.co/GdFjrXl032
Label: fake	.@realDonaldTrump said COVID-19 is totally harmless in 99% of the cases. From deaths to hospitalizations to a growing list of ailments that hit people with even mild cases the numbers say that's wrong. https://t.co/rnqVXpEZEx #coronavirus https://t.co/68pH28q7YE
Label: fake	.@PressSec says @realDonaldTrump never downplayed the virus. He expressly said that he did. https://t.co/ZFGtnXMadB #coronavirus#WoodwardTapes https://t.co/gfMa7e4qkf
Label: fake	.@realDonaldTrump has claimed that the U.S. has "one of the lowest mortality rates in the world" from COVID-19. That's False. https://t.co/GdFjrXl032 https://t.co/k9KYDPJkvA
Top examples 1-grams	co (5), realdonaldtrump (5), 19 (4), covid (4), u (3), says (2), one (2), rates (2), lowest (2), gdfjrxl032 (2), world (2), claimed (2), mortality (2), false (2), coronavirus (2), said (2)
Top examples 2-grams	covid 19 (4), mortality rates (2), claimed u (2), one lowest (2), u one (2), lowest mortality (2), rates world (2), 19 false (2), co gdfjrxl032 (2), false https (2), world covid (2), realdonaldtrump claimed (2)
Top examples 3-grams	claimed u one (2), covid 19 false (2), world covid 19 (2), rates world covid (2), false https co (2), 19 false https (2), https co gdfjrxl032 (2), u one lowest (2), realdonaldtrump claimed u (2), mortality rates world (2), lowest mortality rates (2), one lowest mortality (2)

Table 7: Five semi-factuals using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	.@realDonaldTrump says US is at the top globally in COVID-19 fight. It isn't. https://t.co/JY6Ojdjhwy #coronavirus #Covid19 https://t.co/Gf8j2nLjZb
Label: fake	Donald Trump fails to distinguish between younger and older students and <i>coronavirus</i> risks. Early studies suggest older teens can catch and spread COVID-19 much like adults. https://t.co/hDLIVcUzSY
Label: fake	<i>Coronavirus</i> 'still not as stressful as explaining Windows 10 upgrade to your parents' https://t.co/c6JiLnWADx
Label: fake	Pharmacological and cardiovascular perspectives on the treatment of COVID-19 with chloroquine derivatives. https://t.co/3Ei1hY2SpA
Label: fake	No the CDC did not reduce its COVID-19 Florida case count. https://t.co/BBbCvi6m76 https://t.co/5tCAx7MBao
Label: fake	Confused about mortality rate for <i>coronavirus</i> and case-fatality rate? We explain the difference here. https://t.co/wv1o4b4EfF
Top examples 1-grams	co (5), 19 (3), covid (3), coronavirus (3), case (2)
Top examples 2-grams	covid 19 (3)

be used to understand the model boundary between the two classes, rather than a further insight of the query. This observation is also supported by the fact that semi-factuals and counterfactuals are the same for different queries of the same class as shown in Table 9 and in Table 10

For what concerns concepts extraction, as we have seen in the other datasets, both for RoBERTa and BioMed-RoBERTa, the computed masks in the embedding space in Figure 27 highlight a pattern of the activations used by the model for predicting a certain class (i.e., *fake*, for this case). Instead, in terms of attribution maps on the dataset, we can observe that one of the most important words for the prediction of the *fake* class for a query and its similars is *realdonaldtrump* as shown in Table 11. Instead, for the *real* class, two of the most recurring words are *covid* and *India* as shown in Table 12.

In terms of error analysis, we present a tweet of the *real* class classified as *fake* by the model. We can observe that the most similar examples in Table 13 belong to both the *real* and *fake* classes; furthermore, in the similar and semi-factual

Table 8: Five counterfactuals using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	.@realDonaldTrump says US is at the top globally in COVID-19 fight. It isn't. https://t.co/JY6Ojdjhwy #coronavirus #Covid19 https://t.co/Gf8j2nLjZb
Label: real	India fifth country to isolate <i>Coronavirus</i> at NIV Pune lab https://t.co/xCQE8uFULC
Label: real	The FDA has revoked its emergency use authorization for the drugs hydroxychloroquine and chloroquine for the treatment of <i>COVID</i> .
Label: real	ICMR set to start multi-centric study to evaluate efficacy of BCG vaccine against <i>COVID19</i> in elderly https://t.co/GS9BiGJpov
Label: real	1000 COVID-19 testing labs in India: ICMR https://t.co/1c027BsM2y
Label: real	India crosses nine million COVID-19 tests https://t.co/tqIMHdoaxH
Top examples 1-grams	co (4), india (3), covid (3), icmr (2), 19 (2)
Top examples 2-grams	covid 19 (2)

Table 9: Five semi-factuals using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	Asia's richest man Mukesh Ambani's wife Nita Ambani received Rs.500 as welfare fund from Central Government.
Label: fake	By revoking the permission to use chloroquine in hospitalized patients the FDA expanded the use of the drug in the USA.
Label: fake	Pharmacological and cardiovascular perspectives on the treatment of COVID-19 with chloroquine derivatives. https://t.co/3Ei1hY2SpA
Label: fake	There is no known cure for COVID-19 hydroxychloroquine is not a proven treatment and public health officials advise everyone to wear face masks in public. https://t.co/vIH89d6GZJ
Label: fake	This US study of 821 people focused on the drug as a way to avoid getting COVID19. The hydroxychloroquine group was slightly less likely to test positive after two weeks 11.8% compared to 14.3% in the placebo group but researchers said # were too small. https://t.co/Nhh5MkkIR9
Label: fake	This UK study randomly assigned over 4600 patients with COVID-19 to hydroxychloroquine treatment and a standard care. Among the hydroxy treatment group 25.7% died compared with 23.5% for the standard care group. That was not statistically different. https://t.co/2i0FGZrjvZ
Top examples 1-grams	co (4), treatment (3), 19 (3), covid (3), hydroxychloroquine (3), drug (2), patients (2), chloroquine (2), compared (2), group (2), study (2)
Top examples 2-grams	covid 19 (3), 19 hydroxychloroquine (2)
Top examples 3-grams	covid 19 hydroxychloroquine (2)

Table 10: Five counterfactuals examples using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	Asia's richest man Mukesh Ambani's wife Nita Ambani received Rs.500 as welfare fund from Central Government.
Label: real	The FDA has revoked its emergency use authorization for the drugs hydroxychloroquine and chloroquine for the treatment of COVID.
Label: real	India fifth country to isolate Coronavirus at NIV Pune lab https://t.co/xCQE8uFULC
Label: real	ICMR set to start multi-centric study to evaluate efficacy of BCG vaccine against COVID19 in elderly https://t.co/GS9BiGJpov
Label: real	7000 healthcare workers have died of COVID-19 led by Mexico and the U.S. https://t.co/LbopwsHu2d #medtwitter
Label: real	1000 COVID-19 testing labs in India: ICMR https://t.co/1c027BsM2y
Top examples 1-grams	co (4), covid (3), india (2), icmr (2), 19 (2)
Top examples 2-grams	covid 19 (2)

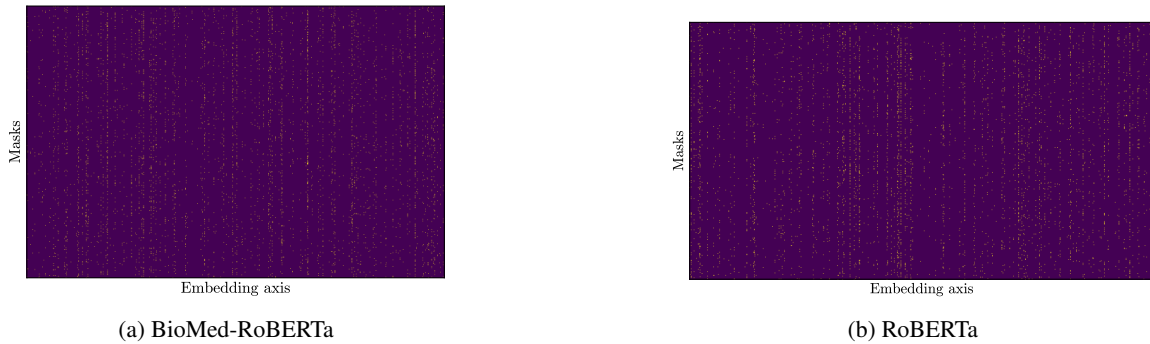


Figure 27: Top relevant embedding axes of correctly classified samples with class *fake*

examples in Tables 14 and 15, respectively, we can find some overlaps. This is a clear indicator that the embedding of the misclassified query lies at the boundary of the two classes.

Table 11: Tokens importance extracted on COVID-19 fake news with BioMed-RoBERTa applied on a query of class *fake* and its similar examples. Scores are shown below tokens. Tokens with a score lower than 0.5 are in gray.

Query	<s>	.	@	realDonaldTrump	Gsays	GUS	Gis	Gat	Gthe	Gtop	Gglobally	Gin	GCO	VID	-	19	Gfight	Glt	Gisn			
Pred: fake	0.343	0.343	0.617	0.617	0.617	0.617	0.297	0.297	0.297	0.297	0.142	0.142	0.142	0.142	0.131	0.131	0.131	0.131	0.915	0.915		
	âG	L	t	.	Ghttps	//	t	.	co	/	J	Y	6	O	j	dj	h	wy	G-	cor	on	av
	0.490	0.490	0.490	0.733	0.259	0.259	0.259	0.259	0.120	0.120	0.120	0.241	0.241	0.241	0.241	0.241	0.241	0.865	1.000	1.000	1.000	
	irus	G-	C	ov	id	19	Ghttps	//	t	.	co	/	G	f	8	j	2	n	L	j	Z	b
	1.000	0.273	0.273	0.273	0.273	0.062	0.062	0.062	0.062	0.062	0.154	0.154	0.154	0.154	0.308	0.308	0.200	0.200	0.200	0.200	0.200	
	</s>																					
	0.000																					
Label: fake	<s>	.	@	realDonaldTrump	Gsays	GU	.	S	.	Gis	Gat	Gthe	Gtop	Gglobally	Gin	GCO	VID	-	19	Gfight		
	0.136	0.136	0.461	0.461	0.461	0.461	0.309	0.070	0.070	0.009	0.009	0.009	0.009	0.016	0.016	0.016	0.016	0.101	0.101	0.265		
	.	Glt	Gisn	âG	L	t	.	Ghttps	//	t	.	co	/	J	Y	6	O	j	d	l	G	7
	0.038	1.000	0.793	0.439	0.439	0.140	0.084	0.084	0.084	0.045	0.042	0.042	0.042	0.162	0.162	0.162	0.162	0.041	0.041	0.041		
	Y	Ghttps	//	t	.	co	/	15	i	UP	wp	H	64	</s>								
	0.041	0.000	0.000	0.000	0.000	0.031	0.031	0.046	0.046	0.046	0.046	0.033	0.033									
Label: fake	<s>	.	@	realDonaldTrump	Gsaid	GCO	VID	-	19	Gis	Gtotally	Gharmless	Gin	G99	%	Gof	Gthe	Gcases	.			
	0.181	0.181	1.000	1.000	1.000	0.259	0.259	0.259	0.259	0.006	0.006	0.006	0.006	0.018	0.018	0.018	0.018	0.018	0.018	0.018		
	GFrom	Gdeaths	Gto	Ghospital	izations	Gto	Ga	Ggrowing	Glist	Gof	Gailments	Gthat	Ghit	Gpeople	Gwith	Geven	Gmild	Gcases				
	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.075	0.075	0.075	0.075	0.127	0.127	0.127	0.127	0.223	0.223	0.223				
	Gthe	Gnumbers	Gsay	Gthat	âG	L	s	Gwrong	.	Ghttps	//	t	.	co	/	m	q	V	X	p	E	
	0.223	0.240	0.240	0.240	0.240	0.345	0.345	0.345	0.319	0.512	0.512	0.512	0.313	0.046	0.046	0.033	0.033	0.033	0.033	0.033		
	Z	Ex	G-	cor	on	av	irus	Ghttps	//	t	.	co	/	68	p	H	28	q	7	YE	</s>	
	0.033	0.033	0.729	0.729	0.729	0.729	0.008	0.008	0.008	0.008	0.021	0.021	0.063	0.063	0.063	0.138	0.138	0.000	0.000	0.000		
Label: fake	<s>	.	@	realDonaldTrump	Ghas	Gclaimed	Gthat	Gthe	GU	.	S	.	Ghas	G"	one	Gof	Gthe	Glowest	Gmortality			
	0.040	0.040	0.128	0.128	0.128	0.099	0.099	0.099	0.099	0.051	0.060	0.033	0.033	0.000	0.000	0.019	0.019	0.019				
	Grates	Gin	Gthe	Gworld	âG	L	Gfrom	GCO	VID	-	19	.	Gthat	's	Gfalse	.	Ghttps	//	t	.	co	
	0.019	0.015	0.015	0.050	0.050	0.010	0.010	0.010	0.010	0.006	0.006	0.405	0.405	1.000	0.216	0.076	0.076	0.089	0.061	0.061		
	/	G	d	F	j	r	X	l	032	</s>												
	0.061	0.053	0.053	0.053	0.026	0.026	0.026	0.043	0.043													
Label: fake	<s>	.	@	realDonaldTrump	Ghas	Gclaimed	Gthat	Gthe	GU	.	S	.	Ghas	G"	one	Gof	Gthe	Glowest	Gmortality			
	0.135	0.135	0.300	0.300	0.300	0.300	0.221	0.221	0.221	0.221	0.134	0.134	0.095	0.095	0.081	0.081	0.095	0.095	0.095			
	Grates	Gin	Gthe	Gworld	âG	L	Gfrom	GCO	VID	-	19	.	Gthat	's	Gfalse	.	Ghttps	//	t	.	co	
	0.095	0.093	0.093	0.153	0.153	0.082	0.082	0.082	0.082	0.072	0.072	0.072	0.928	0.928	1.000	0.266	0.124	0.124	0.116	0.060		
	/	G	d	F	j	r	X	l	032	Ghttps	//	t	.	co	/	k	9	KY	DP	J	k	
	0.060	0.093	0.093	0.093	0.093	0.093	0.093	0.059	0.059	0.059	0.059	0.081	0.081	0.131	0.062	0.062	0.062	0.062	0.062	0.062		
	A	</s>																				
	0.062	0.000																				
Label: fake	<s>	.	@	Press	Sec	Gsays	G@	realDonaldTrump	Gnever	Gdown	played	Gthe	Gvirus	.	GHe	Gexpressly	Gsaid	Gthat	Ghe			
	0.259	0.259	0.539	0.539	0.539	0.959	0.959	0.959	0.959	0.959	0.353	0.353	0.353	0.353	0.185	0.506	0.506	0.574	0.574	0.251		
	Gdid	.	Ghttps	//	t	.	co	/	Z	FG	tn	X	Mad	B	G-	cor	on	av	irus	G-	Wood	
	0.251	0.251	0.396	0.396	0.396	0.204	0.047	0.047	0.034	0.034	0.034	0.034	0.034	0.442	1.000	1.000	1.000	1.000	0.431	0.574	0.574	
	T	apes	Ghttps	//	t	.	co	/	g	f	Ma	7	e	4	q	k	f	</s>				
	0.574	0.574	0.012	0.012	0.012	0.012	0.332	0.332	0.332	0.332	0.332	0.171	0.171	0.171	0.171	0.171	0.171	0.000				

Table 12: Tokens importance on COVID-19 fake news with BioMed-RoBERTa applied on a query of class *real* and its similar examples. Scores are shown below tokens. Tokens with a score lower than 0.5 are in gray.

Query	<s>	-	India	F	ights	Cor	ona	GAs	GIndia	Grecords	Gmore	Grecover	ies	Gthan	Gthe	Gnew	Gcases	Gthe	Ggap	Gbetween		
Pred: real	0.331	0.331	0.592	0.592	0.800	0.800	0.800	0.603	0.603	0.464	0.464	0.156	0.156	0.156	0.156	0.451	0.451	0.426	0.426	0.426		
	GRec	overed	GCases	Gand	Gthe	GActive	GCases	Gis	Gcontinuously	Gwidening	.	GThe	GRecovered	ies	Gare	Gmore	Gthan	G5				
	0.159	0.159	0.159	0.159	0.223	0.223	0.223	0.223	0.875	0.875	1.000	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158	0.158		
	Gtimes	GActive	GCases	.	Ghttps	//	t	.	co	/	f	D	ml	E	w	Y	i	3	D	</s>		
	0.136	0.136	0.136	0.136	0.284	0.284	0.284	0.095	0.227	0.227	0.076	0.076	0.076	0.076	0.076	0.076	0.076	0.000	0.000	0.000		
Label: real	<s>	-	India	F	ights	Cor	ona	G-	Cor	ona	V	irus	Update	GAs	G-	CO	VID	19	GRecover	ies	Gin	GIndia
	0.269	0.269	0.595	0.595	0.595	0.595	0.595	0.595	0.469	0.469	0.469	0.469	0.469	0.620	0.620	0.620	0.774	0.774	1.000	1.000		
	Ggrow	Gdifference	Gwith	GActive	GCases	Gwid	ens	GRec	overed	GPatients	Gare	G3	5	Gtimes	Gthe	GActive	GCases					
	0.847	0.847	0.281	0.281	0.281	0.285	0.285	0.285	0.285	0.515	0.515	0.515	0.407	0.321	0.321	0.321	0.321	0.321	0.321	0.321		
	Ghttps	//	t	.	co	/	da	9	UV	yg	D	5	Ghttps	//	t	.	co	/	b	M	FH	
	0.293	0.293	0.293	0.189	0.469	0.469	0.469	0.391	0.391	0.391	0.391	0.391	0.148	0.148	0.148	0.148	0.148	0.000	0.000	0.307	0.307	
	s	IU	8	N	c	</s>																
	0.307	0.307	0.049	0.049	0.049	0.049																
Label: real	<s>	-	India	F	ights	Cor	ona	GIndia	âG	L	s	Gtotal	GRecover	ies	Gcross	G23	Glakh	.	GPatients	GRec	overed	
	0.726	0.726	0.539	0.539	0.620	0.620	0.620	0.384	0.384	0.272	0.272	0.272	0.000	0.000	0.000	0.000	0.000	0.122	0.201	0.201	0.201	
	Gare	Gmore	Gthan	Gthree	Gtimes	Gof	Gthe	GActive	GCases	.	GRecover	ies	Gexceed	Gactive	GCases	Gby	Gmore	Gthan	G16			
	0.201	0.201	0.201	0.201	0.354	0.354	0.354	0.354	0.354	1.000	0.178	0.178	0.178	0.178	0.148	0.148	0.148	0.148	0.148	0.148		
	Glakh	.	Ghttps	//	t	.	co	/	D	v	D	Gy	F	tek	A	Ghttps	//	t	.	co	/	
	0.148	0.148	0.270	0.270	0.270	0.270	0.114	0.114	0.094	0.094	0.094	0.094	0.069	0.069	0.069	0.099	0.099	0.099	0.099	0.083	0.083	
	Q	x	Z	VG	w	LD	9	</s>														
	0.026	0.026	0.026	0.026	0.026	0.026	0.030	0.030														
Label: real	<s>	-	India	F	ights	Cor	ona	GAs	GIndia	Grecords	Gmore	Grecover	ies	Gthan	Gthe	Gnew	Gcases	Gthe	Ggap	Gbetween		
	0.567	0.567	0.592	0.592	0.593	0.593	0.593	0.689	0.689	0.516	0.516	0.046	0.046	0.046	0.046	0.428	0.428	0.382	0.382	0.382		
	GRec	overed	GCases	Gand	Gthe	GActive	GCases	Gis	Gcontinuously	Gwidening	.	GThe	GRecovered	GCases	G(	467	49	87				
	0.170	0.170	0.170	0.170	0.218	0.218	0.218	0.218	0.789	0.789	1.000	0.530	0.530	0.530	0.530	0.110	0.110	0.110	0.110	0.110		
	)	Gexceed	Gactive	GCases	G(	9	66	382	)	Gby	Gmore	Gthan	G37	Glakh	.	Ghttps	//	t	.	co	/	
	0.110	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.254	0.254	0.254	0.254	0.254	0.254	0.126	0.126	0.126	0.150	0.150		
	0	L	nm	4	H	TC	ci	</s>														
	0.150	0.150	0.150	0.002	0.002	0.002	0.002	0.002														
Label: real	<s>	-	India	F	ights	Cor	ona	GIndia	's	G-	CO	VID	GRecover	ies	Ghave	Gwitnessed	Ga	Gsteep	Gexponential			
	0.681	0.681	1.000	1.000	1.000	1.000	1.000	0.808	0.808	0.319	0.319	0.319	0.319	0.425	0.425	0.425	0.425	0.425	0.246			
	Grise	G-	Gfrom	G5	0000	Gin	GMay	Gto	Gover	G36	Glakh	Gin	GSept	.	GA	Ghigh	Glevel	Gof	Gmore	Gthan	G7	
	0.246	0.246	0.246	0.138	0.138	0.138	0.138	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.000	0.000	0.000	0.000	0.000	0.000		
	GRecover	ies	Greported	Gevery	Gday	.	GRecover	ies	Gare	Gnearly	G3	8	Gtimes	Gthe	GActive	GCases	G(	under				
	0.372	0.372	0.372	0.372	0.372	0.372	0.606	0.274	0.274	0.274	0.274	0.274	0.274	0.399	0.399	0.399	0.399	0.399	0.687	0.197		
	G(	/	4	Gtotal	GCases	)	Ghttps	//	t	.	co	/	n	f	Q	T	k	A	s	2	u	
	0.197	0.197	0.197	0.311	0.311	0.311	0.311	0.282	0.282	0.282	0.272	0.272	0.251	0.251	0.251	0.251	0.251	0.251	0.068	0.068		

Table 13: Five most similar examples using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	The gallery like many arts institutions across the country has been hit hard by the COVID-19 pandemic and is reportedly considering lay-offs thanks to a 75% drop in revenue https://t.co/Mmb1D0rPXg
Label: fake	Kylie Minogue admits she'd find it quite funny if Glastonbury 2020 was cancelled due to coronavirus after people worked so <i>hard</i> to get tickets https://t.co/hTZ96Ulzq5
Label: fake	Scientists are expressing cautious optimism that a COVID-19 vaccine can be ready to go by the late spring of 2021 although it's unclear how much longer it would take to distribute the vaccine widely. https://t.co/srNfKUra8c
Label: real	7000 healthcare workers have died of COVID-19 led by Mexico and the U.S. #medtwitter
Label: fake	We fact-checked this false comparison of daily coronavirus deaths with other diseases and included our own graph comparing the international coronavirus death counts for two different days. https://t.co/2XkUhB5Z9W
Label: real	The NHS is partnering with Facebook to enlist new blood donors in England Wales and Northern Ireland ahead of a potential second wave of COVID-19 https://t.co/8m5UEUYXf
Top examples 1-grams	co (5), covid (3), 19 (3), coronavirus (2)
Top examples 2-grams	covid 19 (3)

Table 14: Five similar examples using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	The gallery like many arts institutions across the country has been hit hard by the COVID-19 pandemic and is reportedly considering lay-offs thanks to a 75% drop in revenue https://t.co/Mmb1D0rPXg
Label: fake	Kylie Minogue admits she'd find it quite funny if Glastonbury 2020 was cancelled due to coronavirus after people worked so <i>hard</i> to get tickets https://t.co/hTZ96Ulzq5
Label: fake	Scientists are expressing cautious optimism that a COVID-19 vaccine can be ready to go by the late spring of 2021 although it's unclear how much longer it would take to distribute the vaccine widely. https://t.co/srNfKUra8c
Label: fake	We fact-checked this false comparison of daily coronavirus deaths with other diseases and included our own graph comparing the international coronavirus death counts for two different days. https://t.co/2XkUhB5Z9W
Label: fake	NEW: A coronavirus vaccine: Where does it stand? Experts we checked with remain optimistic about a vaccine in the first quarter of 2021. https://t.co/alji0QepqT https://t.co/UGgf3mALp7
Label: fake	Biden said feds have "no game plan" to distribute COVID-19 vaccine. A plan is in the works. Committees are expected to make recommendations on which groups should get the vaccine first. https://t.co/XxeraXmw6i https://t.co/F292jogKJ9
Top examples 1-grams	co (5), coronavirus (3), vaccine (3), get (2), covid (2), 19 (2), 2021 (2), distribute (2), checked (2), first (2)
Top examples 2-grams	covid 19 (2), 19 vaccine (2), vaccine first (2)
Top examples 3-grams	covid 19 vaccine (2)

Table 15: Five semi-factuals using L2 distance on COVID-19 fake news with BioMed-RoBERTa. *Italicized* words in the examples also appear in the query.

Query Pred: fake	The gallery like many arts institutions across the country has been hit hard by the COVID-19 pandemic and is reportedly considering lay-offs thanks to a 75% drop in revenue https://t.co/Mmb1D0rPXg
Label: fake	Scientists are expressing cautious optimism that a COVID-19 vaccine can be ready to go by the late spring of 2021 although it's unclear how much longer it would take to distribute the vaccine widely. https://t.co/srNfKUra8c
Label: fake	NEW: A coronavirus vaccine: Where does it stand? Experts we checked with remain optimistic about a vaccine in the first quarter of 2021. https://t.co/alji0QepqT https://t.co/UGgf3mALp7
Label: fake	Any recent manufacturing gains were wiped out by the COVID-19 lockdown. https://t.co/3wEiEpUtaO https://t.co/K7lucok4N7
Label: fake	#Coronavirus in which <i>country</i> of the world patients treated <i>like</i> prisoner. https://t.co/WyRBBfL76Y
Label: fake	We fact-checked this false comparison of daily coronavirus deaths with other diseases and included our own graph comparing the international coronavirus death counts for two different days. https://t.co/2XkUhB5Z9W
Top examples 1-grams	co (5), coronavirus (3), covid (2), 19 (2), 2021 (2), vaccine (2), checked (2)
Top examples 2-grams	covid 19 (2)

5.5 Diagnosis text

For the transformer models on diagnosis text, we report the metrics in Table 16. As before, for both models, non-representativeness correctly increases from similars to counterfactuals while similars and semi-factuals tend to be close. This can be explained by the fact that samples of the same class all tend to result into similar confidence scores. In terms of diversity, the highest values are reached by semi-factuals, which can be an indicator that points at the boundaries are less compact.

In terms of visual results, we observed similar patterns for both BERT and BioBERT using all three distance metrics. For similar examples, the overall behavior is that the selected samples share most of the symptoms of the query as shown in Table 17. Semi-factuals and counterfactuals instead, as shown in Tables 18 and 19, have mostly instances that partially match the symptoms of the query while containing additional or different ones. This indicates that samples at the decision boundary tend to have a few symptoms that are considered relevant for the prediction. In terms of

Table 16: Quantitative evaluation on diagnosis text with 5 example

Example type	BERT		BioBERT	
	Non-repres.	Diversity	Non-repres.	Diversity
Similar [L2]	0.56	2.29	0.61	2.64
Similar [SHAP weighted L2]	0.53	2.31	0.59	2.67
Similar [LIME weighted L2]	0.54	2.32	0.59	2.68
Semi-factuals [L2]	0.57	9.13	0.57	8.43
Semi-factuals [SHAP weighted L2]	0.55	8.59	0.58	8.68
Semi-factuals [LIME weighted L2]	0.56	9.18	0.59	8.73
Counterfactuals [L2]	0.71	4.81	0.84	4.90
Counterfactuals [SHAP weighted L2]	0.69	4.74	0.82	5.03
Counterfactuals [LIME weighted L2]	0.70	4.72	0.83	4.96

Table 17: Five similars on diagnosis text with BERT. *Italicized* words in the examples also appear in the query.

Query	I have symptoms of vomiting while coughing, along with pain and swelling, with sharp pain in my left sole and swelling on the top of my left foot.
Pred: Whooping cough	
Label: Whooping cough	A 2-year-old female with <i>symptoms</i> of <i>vomiting</i> during <i>coughing</i> , <i>pain</i> , <i>swelling</i> , and recent weight gain, experiencing <i>sharp pain</i> in the right <i>top</i> of the <i>foot</i> .
Label: Whooping cough	I am a 79-year-old female suffering from <i>vomiting</i> during <i>coughing</i> , swollen lymph nodes, sensitive <i>pain</i> in the right knee, shortness of breath, muscle contractions or absence episodes, and redness in the eyes.
Label: Whooping cough	I am a 54-year-old male with <i>vomiting</i> during <i>coughing</i> , <i>sharp pain</i> in the <i>left top</i> of the <i>foot</i> , and <i>swelling</i> on the right <i>sole</i> .
Label: Whooping cough	I am a 33-year-old female with <i>vomiting</i> during <i>coughing</i> , <i>sharp pain</i> on the <i>left</i> shoulder, skin lesions or rashes, mouth ulcers, constant fatigue, and a red rash on the nose.
Label: Whooping cough	I am a 60-year-old female experiencing <i>vomiting</i> while <i>coughing</i> , heavy <i>pain</i> in my <i>left sole</i> , and <i>swelling</i> on the right <i>top</i> of my <i>foot</i> .
Top examples 1-grams	coughing (5), vomiting (5), year (5), old (5), pain (5), right (4), female (4), swelling (3), top (3), foot (3), sharp (3), left (3), experiencing (2), sole (2)
Top examples 2-grams	year old (5), vomiting coughing (5), old female (4), sharp pain (3), top foot (3), pain left (3), right top (2), pain right (2), coughing sharp (2), swelling right (2)
Top examples 3-grams	year old female (4), right top foot (2), vomiting coughing sharp (2), sharp pain left (2), coughing sharp pain (2)

Table 18: Five semi-factuals on diagnosis text with BERT. *Italicized* words in the examples also appear in the query.

Query	I have symptoms of vomiting while coughing, along with pain and swelling, with sharp pain in my left sole and swelling on the top of my left foot.
Pred: Whooping cough	
Label: Whooping cough	At 36 years old and female, I am experiencing wheezing on inhale, with <i>vomiting</i> after <i>coughing</i> and intense <i>coughing</i> fits.
Label: Whooping cough	Having <i>coughing</i> fits with <i>pain</i> , extreme fatigue, skin lesions, muscle <i>pain</i> , diarrhea, nausea, swollen lymph nodes, weight loss, and a sore throat. The sensitive <i>pain</i> is in the <i>left</i> temple with pale rashes on the <i>left</i> cheek.
Label: Whooping cough	As a 25-year-old female, I am having <i>coughing</i> fits <i>along</i> with <i>pain</i> , shortness of breath, muscle contractions, redness in the eyes, and vaginal discharge. The burning <i>pain</i> is in my <i>left</i> palm.
Label: Whooping cough	Having <i>coughing</i> fits with <i>sharp pain</i> in the upper abdomen, <i>along</i> with diarrhea, skin lesions, nausea, allergic reaction, shortness of breath, and <i>swelling</i> .
Label: Whooping cough	Being a 67-year-old male, I am facing inhale wheezing, heavy <i>pain</i> in my <i>left top</i> of the <i>foot</i> , and <i>swelling</i> .
Top examples 1-grams	coughing (4), fits (4), pain (4), old (3), left (3), wheezing (2), female (2), inhale (2), lesions (2), diarrhea (2), skin (2), muscle (2), nausea (2), year (2), breath (2), shortness (2), along (2), swelling (2)
Top examples 2-grams	coughing fits (4), pain left (3), old female (2), skin lesions (2), year old (2), shortness breath (2)
Top examples 3-grams	–

prototypes, BERT and BioBERT again perform similarly and the overall observation we can make, as it can be seen in Table 20, is that prototypes of the same class tend to vary based on the symptoms, which is reasonable for this dataset.

Moving onto analyzing the embeddings, we observed the usual recurrent behavior where only a small number of axes of the embeddings are relevant to make the prediction as shown in Figure 28. This is true for both models, although, even with the same architecture, BERT and BioBERT do not share the same relevant axes. More concretely, by computing the importance scores of the tokens, we can observe from Tables 21 and 22 that BERT focuses on a common symptom across different examples while, similarly, BioBERT tends to focus on multiple ones.

Table 19: Five counterfactuals on diagnosis text with BERT. *Italicized* words in the examples also appear in the query.

Query	I have symptoms of vomiting while coughing, along with pain and swelling, with sharp pain in my left sole and swelling on the top of my left foot.
Pred: Whooping cough	
Label: Influenza	I have been wheezing on inhaling and also experienced <i>vomiting</i> after <i>coughing</i> episodes.
Label: Tuberculosis	I am an 18-year-old female with <i>coughing</i> fits, suffering from diarrhea, <i>pain</i> in the upper abdomen characterized as <i>sharp</i> , skin lesions or rashes on the lower back that are pink, had an allergic reaction, shortness of breath, <i>swelling</i> in the nose, loss of consciousness or fainting, high-pitched breathing sound, and wheezing on exhale.
Label: Influenza	<i>Vomiting</i> after <i>coughing</i> and intense <i>coughing</i> fits are my <i>symptoms</i> , along with muscle <i>pain</i> .
Label: Tuberculosis	I am a 20-year-old female experiencing <i>coughing</i> , <i>pain</i> described as burning in the right middle finger.
Label: Tuberculosis	I am a 7-year-old male with repetitive <i>coughing</i> , increased sweating, diarrhea, painful pulse-like sensation at the back of my head, pale rash on the <i>left</i> cheek, nausea, <i>vomiting</i> , and muscle <i>pain</i> .
Top examples 1-grams	coughing (5), pain (4), vomiting (3), year (3), old (3), wheezing (2), diarrhea (2), fits (2), female (2), back (2), muscle (2)
Top examples 2-grams	year old (3), vomiting coughing (2), coughing fits (2), old female (2), muscle pain (2)
Top examples 3-grams	year old female (2)

Table 20: Three prototypes on diagnosis text with BERT for the class *Allergic sinusitis*

My main symptom is eye itching, along with pain, skin lesions or rashes, nausea or vomiting, allergic reactions, shortness of breath, and swelling. The pain feels sharp in the upper abdomen, with a pink rash on my right ankle and swelling on my nose.
I am experiencing a runny nose along with swollen lymph nodes, increased sweating, diarrhea, pain, fever, skin lesions, weight loss, extreme fatigue affecting activities, a sore throat, and diffuse muscle pain. The exhausting pain I feel is in my left temple, with a pale rash on my palate.
I am a 53-year-old female with a cough, increased sweating, pain, presence of fever, skin lesions or rashes, nasal congestion, extreme fatigue, diffuse muscle pain, loss of appetite, chills, and a sensitive pain in the forehead with a pink rash on the left side of the neck.

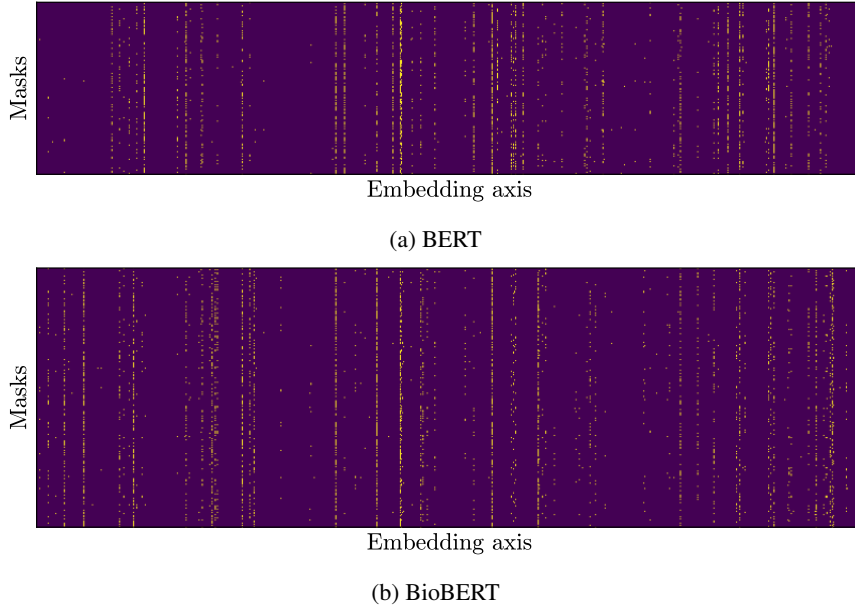


Figure 28: Top relevant embedding axes of correctly classified samples with class *influenza*

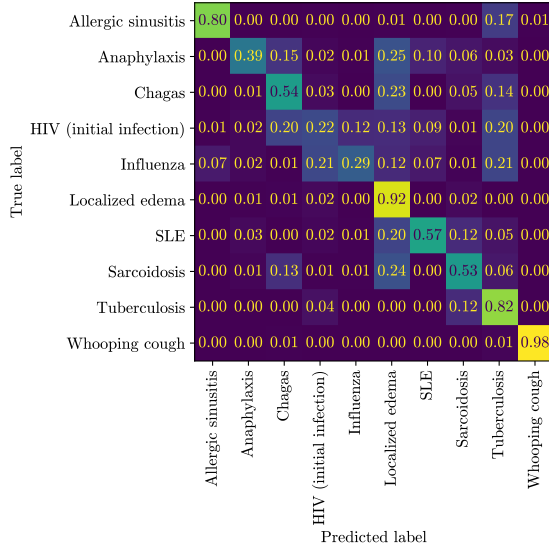
In terms of misclassified samples, the results this dataset show do not provide insightful information on the embedding space. In fact, differently from the previous experiments, counterfactuals, as shown for reference in Table 23, do not encode clear information of the correct class, indicating that both models do not produce well partitioned embedding spaces. This is somewhat expected due to the low accuracy the models reach as shown in Figure 29 and the composition of the dataset that contains phrases structured similarly and with similar symptoms associated to different diseases. Ultimately, this can be used as a starting point for further refinement of the model.

Table 21: Tokens importance on diagnosis text with BERT for the class *influenza*. Scores are shown below tokens. Tokens with a score lower than 0.5 are in gray.

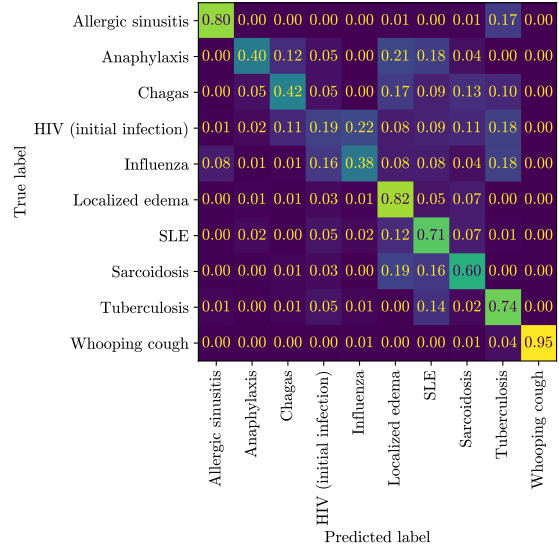
Query Pred: Influenza	[CLS] I have lost my appetite and experiencing pain , skin lesions or r 0.244 0.244 0.255 0.255 1.000 1.000 0.206 0.277 0.285 0.229 0.145 0.145 0.244 0.214 -ash -es , short -ness of breath , a sensitive left shoulder pain , and 0.214 0.214 0.236 0.121 0.121 0.121 0.121 0.057 0.092 0.092 0.086 0.086 0.086 0.189 0.147 a red r -ash on my right cheek . [SEP] 0.069 0.069 0.069 0.069 0.105 0.105 0.105 0.105 0.105 0.105 0.000
Label: Influenza	[CLS] I have lost my appetite and experiencing pain and swelling (heavy pain 0.187 0.187 0.245 0.245 1.000 1.000 0.327 0.293 0.261 0.465 0.571 0.411 0.160 0.160 in my right sole and swelling on the left top of the foot) . 0.160 0.160 0.105 0.105 0.105 0.105 0.105 0.105 0.105 0.099 0.099 0.099 0.099 0.214 [SEP] 0.000
Label: Influenza	[CLS] I have lost my appetite , alongside symptoms of swollen l -ymph nodes 0.172 0.172 0.234 0.234 1.000 1.000 0.208 0.246 0.246 0.221 0.221 0.117 0.117 0.117 di -ar -r -hea , pain , fever , and new fatigue . An exhaust 0.099 0.113 0.113 0.113 0.113 0.060 0.065 0.065 0.065 0.059 0.092 0.092 0.092 0.186 0.307 -ing pain is felt at the back of my head . [SEP] 0.307 0.087 0.087 0.069 0.069 0.091 0.091 0.091 0.091 0.091 0.078 0.000
Label: Influenza	[CLS] I have lost my appetite with increased sweating , pain , fever , 0.023 0.023 0.655 0.661 1.000 1.000 0.159 0.159 0.159 0.114 0.156 0.152 0.190 0.207 nasal congestion , sore throat , di -ff -use muscle pain , loss of appetite 0.248 0.248 0.234 0.122 0.122 0.061 0.079 0.079 0.079 0.021 0.021 0.021 0.158 0.158 exhaust -ing pain at the back of my head , and chill -s . [SEP] 0.081 0.060 0.060 0.054 0.054 0.066 0.066 0.066 0.066 0.030 0.071 0.071 0.071 0.000
Label: Influenza	[CLS] I have lost my appetite and am dealing with increased sweating , pain 0.332 0.332 0.332 0.332 1.000 1.000 1.000 0.384 0.384 0.384 0.384 0.251 0.177 fever , skin lesions , nasal congestion , extreme fatigue , sore throat 0.177 0.220 0.220 0.248 0.248 0.303 0.303 0.303 0.143 0.143 0.143 0.109 0.109 and di -ff -use muscle pain . The pain is exhaust -ing and located 0.109 0.117 0.117 0.117 0.117 0.171 0.171 0.171 0.166 0.166 0.166 0.166 0.401 0.222 in my throat , with pink skin lesions on my forehead . [SEP] 0.222 0.222 0.222 0.321 0.226 0.226 0.444 0.444 0.251 0.251 0.251 0.000
Label: Influenza	[CLS] I have lost my appetite and experiencing di -ar -r -hea , pain 0.135 0.135 0.135 0.135 0.150 1.000 0.080 0.080 0.080 0.080 0.080 0.080 0.058 0.024 skin lesions or r -ash -es , nausea or vomit -ing , swelling , 0.024 0.024 0.024 0.024 0.045 0.045 0.045 0.045 0.097 0.097 0.118 0.118 0.063 0.063 light -headed -ness , with c -ram -p - like pain in my upper abdomen 0.052 0.052 0.052 0.052 0.087 0.049 0.049 0.049 0.043 0.043 0.045 0.103 0.103 0.115 pink r -ash on my right ankle , and swelling on my nose . 0.036 0.065 0.065 0.065 0.057 0.057 0.057 0.057 0.049 0.035 0.035 0.035 0.035 0.035 [SEP] 0.000

Table 22: Tokens importance on diagnosis text with BioBERT for the class *influenza*. Scores are shown below tokens. Tokens with a score lower than 0.5 are in gray.

Query Pred: Influenza	[CLS] I have lost my appetite and experiencing pain , skin lesions or r 0.191 0.191 0.222 0.222 1.000 1.000 0.325 0.433 0.546 0.456 0.414 0.390 0.368 0.405 -ash -es , short -ness of breath , a sensitive left shoulder pain , and 0.405 0.350 0.364 0.236 0.236 0.209 0.209 0.190 0.170 0.170 0.181 0.181 0.181 0.486 0.126 a red r -ash on my right cheek . [SEP] 0.126 0.126 0.126 0.126 0.149 0.149 0.149 0.149 0.149 0.149 0.000
Label: Influenza	[CLS] I have lost my appetite and experiencing di -ar -r -hea , pain 0.153 0.153 0.171 0.171 0.193 1.000 0.104 0.104 0.146 0.146 0.146 0.146 0.134 0.099 skin lesions or r -ash -es , nausea or vomit -ing , swelling , 0.099 0.139 0.139 0.139 0.155 0.155 0.155 0.155 0.139 0.139 0.131 0.131 0.131 0.082 0.082 light -headed -ness , with c -ram -p - like pain in my upper abdomen 0.130 0.130 0.130 0.130 0.136 0.067 0.067 0.067 0.065 0.065 0.062 0.062 0.094 0.094 0.144 pink r -ash on my right ankle , and swelling on my nose . 0.115 0.119 0.119 0.119 0.069 0.069 0.069 0.069 0.069 0.089 0.079 0.079 0.079 0.079 [SEP] 0.000
Label: Influenza	[CLS] I have lost my appetite and am dealing with pain , swelling , w 0.106 0.106 0.182 0.182 0.105 1.000 0.134 0.275 0.238 0.130 0.278 0.043 0.103 0.104 0.088 -hee -zing on ex -hale , swollen l -ymph nodes , muscle contraction -s 0.088 0.088 0.063 0.063 0.063 0.063 0.092 0.092 0.092 0.092 0.092 0.123 0.123 0.123 and red -ness in my eyes . [SEP] 0.119 0.202 0.130 0.130 0.092 0.092 0.092 0.092 0.133 0.000
Label: Influenza	[CLS] I have lost my appetite and experiencing pain and swelling (heavy pain 0.218 0.218 0.334 0.334 1.000 1.000 0.353 0.727 0.623 0.293 0.293 0.342 0.229 0.229 in my right sole and swelling on the left top of the foot) . 0.229 0.229 0.228 0.228 0.228 0.228 0.228 0.228 0.212 0.212 0.212 0.212 0.604 [SEP] 0.000
Label: Influenza	[CLS] I have lost my appetite and also have swollen l -ymph nodes , pain 0.266 0.266 0.574 0.574 1.000 1.000 0.476 0.316 0.316 0.383 0.341 0.341 0.341 0.357 0.318 short -ness of breath , red -ness in my eyes , v -agi -nal 0.316 0.383 0.383 0.383 0.383 0.490 0.197 0.197 0.197 0.197 0.197 0.251 0.280 0.280 0.280 discharge , and a burning type of pain specifically felt in my right shoulder 0.292 0.292 0.600 0.119 0.119 0.119 0.119 0.195 0.195 0.195 0.195 0.159 0.159 0.159 [SEP] 0.159 0.000
Label: Influenza	[CLS] I have lost my appetite and am experiencing pain in my left palm 0.204 0.204 0.299 0.299 1.000 1.000 0.337 0.511 0.511 0.465 0.465 0.230 0.230 0.230 along with skin r -ash -es . The pain is sharp , and there is 0.291 0.297 0.297 0.576 0.311 0.311 0.311 0.355 0.320 0.320 0.321 0.321 0.250 0.543 0.241 0.241 a red r -ash on my nose . [SEP] 0.241 0.241 0.180 0.180 0.180 0.180 0.180 0.314 0.000



(a) BERT



(b) BioBERT

Figure 29: Confusion matrix for the diagnosis text dataset

Table 23: Five counterfactuals on diagnosis text with BioBERT on a misclassified query. *Italicized* words in the examples also appear in the query. The correct class is *Influenza*.

Query Pred: Allergic sinusitis	I am a 57-year-old female with a cough, pain, swollen lymph nodes, diarrhea, shortness of breath, unintentional weight loss, and a pulse-like pain on my left temple.
Label: Influenza	I am a 38-year-old male with a <i>cough</i> , <i>swollen lymph nodes</i> , increased sweating, <i>diarrhea</i> , <i>pain</i> , fever, and sensitive <i>pain</i> in my right <i>temple</i>
Label: Tuberculosis	I am a 50-year-old <i>female</i> with a <i>cough</i> , <i>swollen lymph nodes</i> , sharp <i>pain</i> in the right shoulder, <i>shortness of breath</i> , and redness in the eyes.
Label: Tuberculosis	I am a 78-year-old <i>female</i> with a <i>cough</i> , <i>swollen lymph nodes</i> , <i>pain</i> in the right middle finger, <i>shortness of breath</i> , and muscle contractions/absence episodes.
Label: Tuberculosis	A 43-year-old male with a <i>cough</i> , skin lesions, allergic reactions, <i>shortness of breath</i> , swelling, <i>loss</i> of consciousness, high-pitched breathing, sharp <i>pain</i> in my upper abdomen, and pink skin rash on my right ankle.
Label: Tuberculosis	I, a 21-year-old <i>female</i> with a <i>cough</i> , have <i>pain</i> in my upper abdomen, <i>swollen lymph nodes</i> , <i>diarrhea</i> , nausea or vomiting, <i>shortness of breath</i> , <i>unintentional weight loss</i> , new fatigue or muscle aches, and a pulse-like <i>pain</i> .
Top examples 1-grams	cough (5), old (5), pain (5), year (5), swollen (4), right (4), lymph (4), nodes (4), shortness (4), breath (4), female (3), diarrhea (2), male (2), sharp (2), muscle (2), loss (2), upper (2), abdomen (2)
Top examples 2-grams	year old (5), swollen lymph (4), lymph nodes (4), shortness breath (4), cough swollen (3), pain right (3), female cough (3), old female (3), old male (2), male cough (2), sharp pain (2), upper abdomen (2), pain upper (2)
Top examples 3-grams	swollen lymph nodes (4), cough swollen lymph (3), old female cough (3), year old female (3), old male cough (2), year old male (2), female cough swollen (2), pain upper abdomen (2)

5.6 COMPAS

We proceed to analyze the COMPAS dataset. As we can see for a query of the *Not Recidivism* class, the similar examples in Table 25 show the same values in terms of one-hot encoded features, such as gender and ethnicity, but also in terms of age and priors count. The semi-factuals in Table 26 instead show a more diverse set of values as we can expect, while for the counterfactuals in Table 27 we can observe that the query share some common attributes with the examples (i.e., age, gender, ethnicity) changing only the label. The same consideration could be done for a query of the *Recidivism* class as we can see in Table 28 for the similars, in Table 29 for the semi-factuals, and in Table 30 for the counterfactuals. This behavior underlines that the separation of the two classes is encoded in very few changes in the attributes values.

This insight is confirmed by the metrics for the COMPAS dataset in Table 24. In fact, we highlight that, compared to the other experiments, the non-representativeness does not follow the trend that we should expect (i.e., in an increasing way starting from the similars, to the semi-factuals, and finally to the counterfactuals). This is probably due to the fact that in the embedding space the two classes are very close to each other and one possible reason consists of the fact that

Table 24: Quantitative evaluation on COMPAS with 5 examples

Example type	Non-repres.	Diversity
Similar [L2]	0.41	0.26
Similar [pos. LIME weighted L2]	0.40	0.30
Similar [pos. SHAP weighted L2]	0.40	0.30
Semi-factuals (glob. sim. + CF NMOTB) [L2]	0.46	2.91
Semi-factuals (glob. sim. + CF NMOTB) [pos. LIME weighted L2]	0.45	2.93
Semi-factuals (glob. sim. + CF NMOTB) [pos. SHAP weighted L2]	0.45	2.83
Counterfactuals (NMOTB) [L2]	0.42	0.44
Counterfactuals (NMOTB) [pos. LIME weighted L2]	0.41	0.50
Counterfactuals (NMOTB) [pos. SHAP weighted L2]	0.41	0.49

Table 25: Five similar examples using L2 distance with a fully-connected network on COMPAS of a *Not Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Pred: Not Recidivism																				
Label: Not Recidivism	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	41.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	43.00	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	40.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	40.00	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0

Table 26: Five semi-factuals using L2 distance with a fully-connected network on COMPAS of a *Not Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Pred: Not Recidivism																				
Label: Not Recidivism	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	57.00	0	0	27	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	26.00	0	-1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
Label: Not Recidivism	37.00	2	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0
Label: Not Recidivism	30.00	0	-1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	1

the final linear classifier of the model does not enforce a clear separation of the two classes and their inter-class distance is very thin [8].

In the prototypes for the *Not Recidivism* class in Table 31 and in the one for the *Recidivism* class in Table 32, we can observe that the dataset is prone to group bias: two out of three of the prototypes for *Not Recidivism* are Caucasian subjects, while for the *Recidivism* class, two out of three prototypes are Afro-American subjects. Also, for what concerns the gender, we can see no females in the prototypes of the *Recidivism* class, while in the *Not Recidivism* class we have. Finally, as we can expect, the priors count is higher in the prototypes of the *Recidivism* class.

Table 27: Five counterfactuals using L2 distance with a fully-connected network on COMPAS of a *Not Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Pred: Not Recidivism																				
Label: Recidivism	42.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Recidivism	43.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Recidivism	44.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Recidivism	44.00	1	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Recidivism	39.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0

Table 28: Five similar examples using L2 distance with a fully-connected network on COMPAS of a *Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	24.00	0	-1	6	0	2	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Pred: Recidivism																				
Label: Recidivism	22.00	0	-1	13	0	2	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	23.00	1	-1	6	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	22.00	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	24.00	1	-1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	22.00	1	-1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1

Table 29: Five semi-factual examples using L2 distance with a fully-connected network on COMPAS of a *Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	24.00	0	-1	6	0	2	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Pred: Recidivism																				
Label: Recidivism	22.00	0	-1	13	0	2	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	25.00	0	-1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0
Label: Recidivism	43.00	3	-3	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1
Label: Recidivism	20.00	0	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	35.00	1	-1	49	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

In line with the experiments of the previous datasets, also in the embedding space of COMPAS in Figure 30, we can see a pattern in the activation of axes related to a specific class. Translating this information in the features space, we can see in Table 33 that the gender plays a predominant role to classify a subject as *Not Recidivism*, instead, as we can see

Table 30: Five counterfactual examples using L2 distance with a fully-connected network on COMPAS of a *Recidivism* query

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25 - 45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	24.00	0	-1	6	0	2	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Pred: Recidivism																				
Label: Not Recidivism	23.00	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Not Recidivism	23.00	1	-1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Not Recidivism	21.00	1	-1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Not Recidivism	21.00	0	-1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Not Recidivism	24.00	0	0	-1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1

Table 31: Prototypes of the *Not Recidivism* class on COMPAS

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25 - 45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
45.00	2	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1
22.00	1	12	42	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1
32.00	0	-1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0

Table 32: Prototypes of the *Recidivism* class on COMPAS

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25 - 45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
32.00	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
31.00	4	-11	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1
39.00	13	-1	0	0	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1

in Table 34, the ethnicity, age, and priors count are relevant for what concerns the classification of *Recidivism*. This insight underlines and confirms the presence of group bias in the dataset: probably the model focuses on the female gender to classify a subject as *Not Recidivism*, instead, besides the priors count, being Caucasian or Afro-American is a determining factor on whether the model predicts a subject as *Recidivism*.

In terms of error analysis, we take a query of the *Recidivism* class predicted by the model as *Not Recidivism*. By analyzing the concept mask of the wrong predicted class (i.e., *Not Recidivism*) in Table 35 and the one of the correct class (i.e., *Recidivism*) in Table 36, we can say that, again, the ethnicity plays a fundamental role in the classification and probably in this case the fact that the subject was a Caucasian man lead the model to fail.



Figure 30: Top relevant embedding axes of correctly classified samples of the COMPAS dataset

Table 33: Feature importance on COMPAS with a fully-connected network applied on a query of class *Not Recidivism* and its similar examples

	age	priors_count	days_b_screening_arrest	e_jail_time	juv_fel_count	juv_other_count	juv_misd_count	e_charge_degree:F	e_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	42.00	0	-3	0	0	0	0	1	0	0	0	1	0.086	0	0	1	0	0	1	0
Pred: Not Recidivism	0.340	0.322	0.002	0.000	0.000	0.000	0.000	0.323	0.638	0.358	0.000	0.305	0.086	0.000	0.003	0.441	0.410	0.142	1.000	0.542
Label: Not Recidivism	42.00	0	-3	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	0.216	0.413	0.000	0.000	0.000	0.000	0.000	0.337	0.154	0.267	0.000	0.449	0.068	0.000	0.006	0.131	0.166	0.213	1.000	0.554
Label: Not Recidivism	41.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	0.174	0.425	0.000	0.000	0.000	0.000	0.000	0.342	0.163	0.266	0.000	0.446	0.092	0.000	0.005	0.129	0.168	0.191	1.000	0.558
Label: Not Recidivism	43.00	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	0.314	0.441	0.000	0.000	0.000	0.000	0.000	0.344	0.155	0.276	0.000	0.468	0.084	0.000	0.000	0.128	0.172	0.190	1.000	0.576
Label: Not Recidivism	40.00	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	0.149	0.433	0.000	0.000	0.004	0.000	0.000	0.333	0.169	0.266	0.000	0.457	0.077	0.000	0.000	0.126	0.162	0.194	1.000	0.556
Label: Not Recidivism	40.00	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Label: Not Recidivism	0.119	0.423	0.000	0.000	0.000	0.000	0.003	0.353	0.173	0.268	0.000	0.450	0.102	0.000	0.005	0.114	0.173	0.201	1.000	0.562

Table 34: Feature importance on COMPAS with a fully-connected network applied on a query of class *Recidivism* and its similar examples

	age	priors_count	days_b_screening_arrest	e_jail_time	juv_fel_count	juv_other_count	juv_misd_count	e_charge_degree:F	e_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25_-45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	21.00	1	-1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	1
Pred: Recidivism	0.220	0.092	0.007	0.033	0.000	0.007	0.000	0.754	0.944	0.724	0.000	1.000	0.078	0.000	0.015	0.453	0.043	0.390	0.262	0.175
Label: Recidivism	21.00	1	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	0.911	0.545	0.000	0.031	0.000	0.000	0.000	0.344	0.367	0.623	0.000	0.513	0.125	0.000	0.038	0.798	0.139	1.000	0.242	0.220
Label: Recidivism	21.00	1	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	0.868	0.533	0.000	0.042	0.000	0.000	0.000	0.333	0.375	0.619	0.000	0.522	0.118	0.000	0.003	0.789	0.154	1.000	0.252	0.229
Label: Recidivism	21.00	1	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	0.854	0.553	0.000	0.033	0.000	0.000	0.000	0.350	0.388	0.619	0.000	0.537	0.134	0.000	0.000	0.794	0.144	1.000	0.302	0.225
Label: Recidivism	21.00	1	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	0.888	0.561	0.000	0.030	0.000	0.000	0.000	0.354	0.412	0.638	0.000	0.540	0.127	0.000	0.017	0.818	0.131	1.000	0.316	0.194
Label: Recidivism	21.00	1	-1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1
Label: Recidivism	0.869	0.541	0.000	0.046	0.000	0.010	0.000	0.351	0.393	0.617	0.000	0.518	0.126	0.000	0.002	0.784	0.145	1.000	0.290	0.211

Table 35: Feature importance on COMPAS of a misclassified query and its similars using the mask of the predicted class *Not Recidivism*

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25 - 45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	29.00	1	-23	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Pred: Not Recidivism	0.097	0.131	0.000	0.000	0.000	0.001	0.000	0.317	0.158	0.357	0.000	0.247	0.063	0.000	0.000	0.375	0.102	0.190	1.000	0.648
Label: Not Recidivism	29.00	1	-17	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.153	0.258	0.232	0.000	0.000	0.000	0.000	0.607	0.437	0.648	0.008	1.000	0.077	0.000	0.033	0.219	0.204	0.390	0.180	0.381
Label: Not Recidivism	30.00	2	-21	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.128	0.000	0.292	0.013	0.000	0.000	0.000	0.557	0.450	0.620	0.000	1.000	0.097	0.000	0.028	0.207	0.199	0.336	0.131	0.403
Label: Not Recidivism	25.00	1	-22	-1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.228	0.238	0.311	0.000	0.000	0.000	0.000	0.595	0.418	0.639	0.000	1.000	0.124	0.000	0.015	0.212	0.200	0.342	0.175	0.427
Label: Not Recidivism	38.00	2	-20	19	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.090	0.000	0.282	0.000	0.000	0.000	0.001	0.617	0.449	0.638	0.000	1.000	0.181	0.000	0.019	0.227	0.196	0.366	0.180	0.435
Label: Not Recidivism	29.00	2	-23	22	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
	0.147	0.000	0.313	0.030	0.000	0.000	0.000	0.553	0.441	0.597	0.000	1.000	0.095	0.000	0.000	0.201	0.191	0.357	0.135	0.392

Table 36: Feature importance on COMPAS of a misclassified query and its similars using the mask of the correct class *Recidivism*

	age	priors_count	days_b_screening_arrest	c_jail_time	juv_fel_count	juv_other_count	juv_misd_count	c_charge_degree:F	c_charge_degree:M	race:African-American	race:Asian	race:Caucasian	race:Hispanic	race:Native_American	race:Other	age_cat:25 - 45	age_cat:Greater_than_45	age_cat:Less_than_25	sex:Female	sex:Male
Query	29.00	1	-23	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Pred: Not Recidivism	0.683	0.062	0.008	0.000	0.000	0.000	0.000	0.367	0.359	0.560	0.000	0.362	0.075	0.000	0.029	0.521	0.103	0.674	0.604	1.000
Label: Not Recidivism	29.00	1	-17	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.484	0.171	0.305	0.000	0.000	0.000	0.000	0.560	0.442	0.683	0.000	1.000	0.080	0.000	0.000	0.522	0.142	0.308	0.525	0.303
Label: Not Recidivism	30.00	2	-21	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.403	0.000	0.391	0.022	0.000	0.000	0.000	0.562	0.456	0.729	0.000	1.000	0.161	0.000	0.025	0.564	0.178	0.265	0.512	0.328
Label: Not Recidivism	25.00	1	-22	-1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.811	0.160	0.390	0.000	0.000	0.000	0.000	0.610	0.447	0.694	0.000	1.000	0.134	0.000	0.000	0.545	0.138	0.275	0.525	0.337
Label: Not Recidivism	38.00	2	-20	19	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
Label: Not Recidivism	0.084	0.000	0.341	0.006	0.000	0.000	0.000	0.599	0.443	0.708	0.000	1.000	0.177	0.000	0.057	0.545	0.152	0.272	0.527	0.330
Label: Not Recidivism	29.00	2	-23	22	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
	0.483	0.000	0.433	0.047	0.000	0.000	0.000	0.583	0.451	0.737	0.000	1.000	0.156	0.000	0.050	0.564	0.157	0.313	0.520	0.330

5.7 Weather classification

For the weather classification dataset, we report the results in Table 37. As in the previous cases, there is a correct increase in non-representativeness moving from similars to counterfactuals. Diversity is also higher for counterfactuals, which indicates that samples are spread along the decision boundary.

As for the examples, we observed, as in all the previous cases, that similars tend reflect the query and only differ by few numerical features. Semi-factuals differ from the query more visibly both in categorical and numerical features. Counterfactuals are instead closer to semi-factuals in terms of features, which might indicate that samples of different classes at the decision boundary are closely related. For prototypes, as in all previous cases, we can observe that all samples are different among themselves in nearly all features as shown for instance in Table 41, indicating that the embeddings are able to model these differences.

By viewing the embedding spaces, as shown in Figure 31, it is again possible to locate preferred embedding axes for each specific class. However, by computing the attribution scores, we were unable to detect any particular patterns, which might indicate again that what the model considers relevant features that are not necessarily human interpretable.

Table 37: Quantitative evaluation on weather classification with 5 example

Example type	Feed-forward network	
	Non-repres.	Diversity
Similar [L2]	0.12	2.73
Similar [SHAP weighted L2]	0.10	3.07
Similar [LIME weighted L2]	0.10	3.17
Semi-factuals [L2]	0.10	5.18
Semi-factuals [SHAP weighted L2]	0.07	5.88
Semi-factuals [LIME weighted L2]	0.07	5.68
Counterfactuals [L2]	0.88	7.69
Counterfactuals [SHAP weighted L2]	1.09	8.23
Counterfactuals [LIME weighted L2]	1.14	7.98

Table 38: Five similar examples on weather classification

	Temperature	Humidity	Wind Speed	Precipitation (%)	Atmospheric Pressure	UV Index	Visibility (km)	Cloud Cover_clear	Cloud Cover_cloudy	Cloud Cover_overcast	Cloud Cover_partly cloudy	Season_Autumn	Season_Spring	Season_Summer	Season_Winter	Location_coastal	Location_inland	Location_mountain
Query	38.00	27.00	7.50	14.00	1017.69	6.00	6.50	1	0	0	0	0	0	0	1	0	1	0
Pred: Sunny																		
Label: Sunny	35.00	36.00	4.00	11.00	1023.49	6.00	7.00	1	0	0	0	0	0	0	1	0	1	0
Label: Sunny	35.00	21.00	9.00	17.00	1014.91	5.00	8.00	1	0	0	0	0	0	0	1	0	1	0
Label: Sunny	35.00	29.00	9.50	9.00	1022.77	5.00	7.00	1	0	0	0	0	0	0	1	0	1	0
Label: Sunny	34.00	29.00	2.50	4.00	1029.35	7.00	7.00	1	0	0	0	0	0	0	1	0	1	0
Label: Sunny	32.00	29.00	2.00	16.00	1025.82	6.00	8.50	1	0	0	0	0	0	0	1	0	1	0

Table 39: Five semi-factuals on weather classification

	Temperature	Humidity	Wind Speed	Precipitation (%)	Atmospheric Pressure	UV Index	Visibility (km)	Cloud Cover_clear	Cloud Cover_cloudy	Cloud Cover_overcast	Cloud Cover_partly cloudy	Season_Autumn	Season_Spring	Season_Summer	Season_Winter	Location_coastal	Location_inland	Location_mountain
Query	38.00	27.00	7.50	14.00	1017.69	6.00	6.50	1	0	0	0	0	0	0	1	0	1	0
Pred: Sunny																		
Label: Sunny	20.00	48.00	6.00	8.00	1016.99	7.00	10.00	0	0	0	1	0	0	0	1	0	0	1
Label: Sunny	40.00	36.00	7.50	15.00	1016.03	7.00	10.00	0	0	0	1	0	0	0	1	0	1	0
Label: Sunny	41.00	63.00	6.50	18.00	1022.65	9.00	8.50	0	0	0	1	0	1	0	0	0	1	0
Label: Sunny	20.00	60.00	0.50	18.00	1015.43	9.00	9.00	0	0	0	1	0	0	0	1	0	0	1
Label: Sunny	21.00	41.00	2.50	18.00	1012.67	8.00	9.00	0	0	0	1	0	0	1	0	0	1	0

In terms of error analysis as well, examples were unable to provide insights for misclassified samples. This can provide a signal on the fact that the samples are mapped into the embedding space with little inter-class spacing, as mentioned for COMPAS, which results in producing counterfactuals that span across different classes.

Table 40: Five counterfactuals on weather classification

	Temperature	Humidity	Wind Speed	Precipitation (%)	Atmospheric Pressure	UV Index	Visibility (km)	Cloud Cover_clear	Cloud Cover_cloudy	Cloud Cover_overcast	Cloud Cover_partly cloudy	Season_Autumn	Season_Spring	Season_Summer	Season_Winter	Location_coastal	Location_inland	Location_mountain
Query Pred: Sunny	38.00	27.00	7.50	14.00	1017.69	6.00	6.50	1	0	0	0	0	0	0	1	0	1	0
Label: Snowy	43.00	48.00	5.00	26.00	936.13	7.00	9.00	0	0	0	1	0	0	1	0	0	1	0
Label: Snowy	18.00	37.00	3.00	11.00	888.19	11.00	1.00	0	0	0	1	1	0	0	0	0	1	0
Label: Rainy	14.00	44.00	9.00	11.00	853.01	14.00	3.50	0	0	0	1	0	0	0	1	1	0	0
Label: Cloudy	34.00	50.00	11.50	16.00	1015.68	4.00	7.00	0	0	0	1	0	1	0	0	0	1	0
Label: Cloudy	33.00	55.00	0.00	18.00	1006.11	3.00	7.00	0	0	0	1	1	0	0	0	0	0	1

Table 41: Three prototypes on weather classification for the class cloudy

Temperature	Humidity	Wind Speed	Precipitation (%)	Atmospheric Pressure	UV Index	Visibility (km)	Cloud Cover_clear	Cloud Cover_cloudy	Cloud Cover_overcast	Cloud Cover_partly cloudy	Season_Autumn	Season_Spring	Season_Summer	Season_Winter	Location_coastal	Location_inland	Location_mountain
30.00	52.00	6.00	27.00	1013.51	1.00	8.50	0	0	1	0	0	0	1	0	1	0	0
33.00	83.00	27.50	102.00	1003.99	4.00	5.00	0	0	0	1	1	0	0	0	1	0	0
46.00	69.00	3.50	43.00	842.22	12.00	5.50	0	1	0	0	0	1	0	0	0	1	0

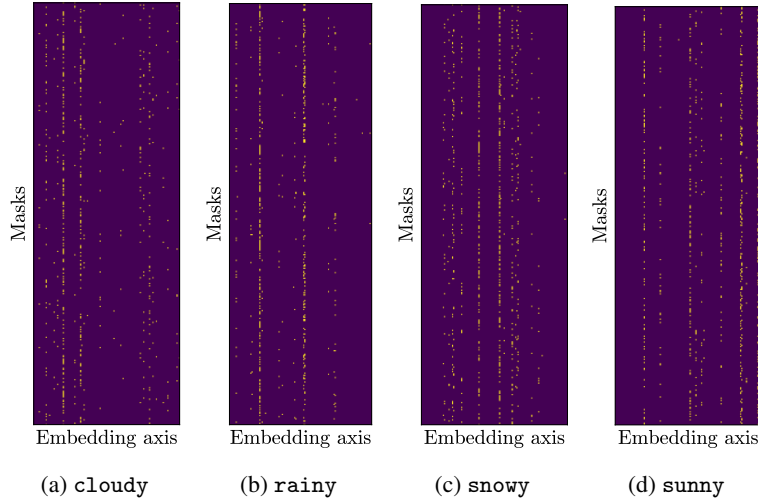


Figure 31: Top relevant embedding axes of weather classification of correctly classified samples of each class

6 Conclusion

In this project, we explored methods to search examples in a post-hoc model-agnostic setup for classification models. We experimented with similar, semi-factual, counterfactual, and prototypical examples search methods applied to explainability, concepts extraction, error analysis, and adversarial attack detection using images, text, and tabular data with different models. We found out that similar examples are consistent in all data types and are able to well represent

the query. Instead, semi-factuais and counterfactuals vary depending on the dataset and do not always show a clear pattern. Also, with all data types, there are patterns in the embeddings where only a small subset of activations is relevant for the final prediction. Also, by inspecting attribution maps, we observed that they vary depending on the instance and do not always show a human-interpretable result. In terms of error analysis, again, results vary depending on the dataset. In some instances, most similar and counterfactual examples provide some insights on how the model is misclassifying. However, in others, this behavior is not always clear which still gives some information on how the latent space can be improved. Also, we observed, for the case of images, that simple adversarial attacks can be detected using examples similarly as for error analysis. In conclusion, we argue that example search methods are an effective way for understanding how the model is interpreting and representing the dataset to support explainability for both end-user and developer, and can also provide a way to analyze the data itself to extrapolate patterns and biases.

Future directions we think can be explored are to (i) experiment with different distance metrics; (ii) explore other methods to search for examples and prototypes, and using different embedding spaces; (iii) use other ways of finding and analyzing the relevant axes of the embeddings; (iv) experiment with models trained using contrastive losses that produce clearer class separation; (v) analyze the results with other types of adversarial attacks, also on textual and tabular data.

7 Contributions and artifacts

In terms of work distribution, the initial study of the literature was performed together. The idea of analyzing the axes of the embeddings was proposed by Xia while using examples for error analysis and adversarial attack detection was devised by Maiolino. In practical terms, most of the implementation of the basic structure and methods for the project was done in pair programming. For the datasets and experimental results, Maiolino worked on EuroSAT, COVID-19 fake news, and COMPAS. Instead, Xia worked on brain tumor MRIs, diagnosis text, and weather classification.

The trained models and outputs of the experiments are available on OneDrive at https://liveunibo-my.sharepoint.com/:f:/g/personal/tiancheng_xia_studio_unibo_it/EpQglIa0S5x0vW3ZyTUIDgIBt5Gt8rKzGgk80vEDHK08MQ?e=U9Dey5.

References

- [1] Ruth M. J. Byrne. “Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJ-CAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, 2019, pp. 6276–6282. DOI: 10.24963/IJCAI.2019/876.
- [2] Lisa Cummins and Derek G. Bridge. “KLEOR: A Knowledge Lite Approach to Explanation Oriented Retrieval”. In: *Comput. Artif. Intell.* 25.2-3 (2006), pp. 173–193. URL: <http://www.cai.sk/ojs/index.php/cai/article/view/338>.
- [3] Jia Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/V1/N19-1423.
- [5] Dónal Doyle et al. “Explanation Oriented Retrieval”. In: *Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, Madrid, Spain, August 30 - September 2, 2004, Proceedings*. Ed. by Peter Funk and Pedro A. González-Calero. Vol. 3155. Lecture Notes in Computer Science. Springer, 2004, pp. 157–168. DOI: 10.1007/978-3-540-28631-8_13.
- [6] Suchin Gururangan et al. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 2020, pp. 8342–8360. DOI: 10.18653/V1/2020.ACL-MAIN.740.
- [7] Patrick Helber et al. “EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification”. In: *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 12.7 (2019), pp. 2217–2226. DOI: 10.1109/JSTARS.2019.2918242.

- [8] Shota Horiguchi, Daiki Ikami, and Kiyoharu Aizawa. “Significance of Softmax-Based Features in Comparison to Distance Metric Learning-Based Features”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 42.5 (2020), pp. 1279–1285. DOI: 10.1109/TPAMI.2019.2911075.
- [9] Andrew Howard et al. “Searching for MobileNetV3”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.
- [10] Larson Jeff et al. *How We Analyzed the COMPAS Recidivism Algorithm*. 2016. URL: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [11] Alex Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: (2009), pp. 32–33. URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [12] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proc. IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.
- [13] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinform.* 36.4 (2020), pp. 1234–1240. DOI: 10.1093/BIOINFORMATICS/BTZ682.
- [14] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* (2019). arXiv: 1907.11692.
- [15] José Maurício, Inês Domingues, and Jorge Bernardino. “Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review”. In: *Applied Sciences* 13.9 (2023). ISSN: 2076-3417. DOI: 10.3390/app13095521.
- [16] Sachin Mehta and Mohammad Rastegari. “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer”. In: *CoRR* (2021). arXiv: 2110.02178.
- [17] Meike Nauta et al. “This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition”. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*. Vol. 1524. Communications in Computer and Information Science. Springer, 2021, pp. 441–456. DOI: 10.1007/978-3-030-93736-2_34.
- [18] An-phi Nguyen and María Rodríguez Martínez. “On quantitative aspects of model interpretability”. In: *CoRR* (2020). arXiv: 2007.07584.
- [19] Parth Patwa et al. *Fighting an Infodemic: COVID-19 Fake News Dataset*. 2020. arXiv: 2011.03327 [cs.CL].
- [20] Antonin Poché, Lucas Hervier, and Mohamed Chafik Bakkay. “Natural Example-Based Explainability: A Survey”. In: *Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part II*. Ed. by Luca Longo. Vol. 1902. Communications in Computer and Information Science. Springer, 2023, pp. 24–47. DOI: 10.1007/978-3-031-44067-0_2.
- [21] Francesco Sovrano et al. “A Survey on Methods and Metrics for the Assessment of Explainability Under the Proposed AI Act”. In: *Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021*. Ed. by Erich Schweighofer. Vol. 346. Frontiers in Artificial Intelligence and Applications. IOS Press, 2021, pp. 235–242. DOI: 10.3233/FAIA210342.
- [22] Arsène Fansi Tchango et al. “DDXPlus: A new Dataset for Medical Automatic Diagnosis”. In: *CoRR* (2022). arXiv: 2205.09148.
- [23] Amalie Trewartha et al. “Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science”. In: *Patterns* 3.4 (2022), p. 100488. DOI: 10.1016/J.PATTER.2022.100488.

A Quantitative results

A.1 Comparison of strategies for weighted distances

Table 42: Quantitative evaluation of similar examples on CIFAR10 using MobileViT with 5 examples and weighted distance

Example type	Non-repres.	Diversity
Similar [pos. SHAP weighted L1]	0.02	3.78
Similar [abs. SHAP weighted L1]	0.02	3.80
Similar [pos. SHAP weighted L2]	0.02	3.86
Similar [abs. SHAP weighted L2]	0.02	3.87
Similar [pos. SHAP weighted cos.]	0.02	3.92
Similar [abs. SHAP weighted cos.]	0.02	3.92
Similar [pos. LIME weighted L1]	0.03	3.82
Similar [abs. LIME weighted L1]	0.03	3.81
Similar [pos. LIME weighted L2]	0.04	3.88
Similar [abs. LIME weighted L2]	0.04	3.87
Similar [pos. LIME weighted cos.]	0.02	3.84
Similar [abs. LIME weighted cos.]	0.02	3.85

Table 43: Quantitative evaluation of semi-factual examples on CIFAR10 using MobileViT with 5 examples and weighted distance

Example type	Non-repres.	Diversity
Semi-factuals [pos. SHAP weighted L1]	0.07	6.26
Semi-factuals [abs. SHAP weighted L1]	0.07	6.26
Semi-factuals [pos. SHAP weighted L2]	0.14	5.04
Semi-factuals [abs. SHAP weighted L2]	0.14	5.04
Semi-factuals [pos. SHAP weighted cos.]	0.09	4.81
Semi-factuals [abs. SHAP weighted cos.]	0.10	4.82
Semi-factuals [pos. LIME weighted L1]	0.07	6.26
Semi-factuals [abs. LIME weighted L1]	0.07	6.26
Semi-factuals [pos. LIME weighted L2]	0.18	5.01
Semi-factuals [abs. LIME weighted L2]	0.19	5.01
Semi-factuals [pos. LIME weighted cos.]	0.10	4.65
Semi-factuals [abs. LIME weighted cos.]	0.10	4.66

Table 44: Quantitative evaluation of counterfactual examples on CIFAR10 using MobileViT with 5 examples and weighted distance

Example type	Non-repres.	Diversity
Counterfactuals [pos. SHAP weighted L1]	1.02	7.58
Counterfactuals [abs. SHAP weighted L1]	1.03	7.54
Counterfactuals [pos. SHAP weighted L2]	1.28	7.82
Counterfactuals [abs. SHAP weighted L2]	1.28	7.81
Counterfactuals [pos. SHAP weighted cos.]	1.13	7.58
Counterfactuals [abs. SHAP weighted cos.]	1.14	7.57
Counterfactuals [pos. LIME weighted L1]	0.85	7.44
Counterfactuals [abs. LIME weighted L1]	0.84	7.36
Counterfactuals [pos. LIME weighted L2]	0.96	7.61
Counterfactuals [abs. LIME weighted L2]	0.93	7.60
Counterfactuals [pos. LIME weighted cos.]	0.86	7.54
Counterfactuals [abs. LIME weighted cos.]	0.86	7.53