Comparing model-agnostic and model-specific XAI methods in Natural Language Processing

Marian Gabriel Sandu
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
Sandugabriel97@gmail.com

Ștefan Trăușan-Matu
University Politehnica of Bucharest
313 Splaiul Independentei,
Bucharest, Romania
and
Research Institute for Artificial Intelligence
stefan.trausan@upb.ro

ABSTRACT

Explainable Artificial Intelligence is very important from a user-computer interaction perspective and has been growing steadily in the last few years. This is because of the fact that ML and Deep Learning overgrew creating more complex models that are highly accurate but lack explainability and interpretability. The aim of this paper is to present the concept of explainability in ML and to compare state-of-the-art model agnostic and model-specific explanation methods in order to determine which type is the most concise based on different evaluation metrics fitted for text classification tasks.

Author Keywords
NLP; XAI; Deep Learning; Transformers

ACM Classification Keywords
I.2.7 Natural Language Processing: Text analysis.
General Terms
Human Factors; Design; Measurement.

DOI: 10.37789/rochi.2022.1.1.19

INTRODUCTION

The domain of eXplainable Artificial Intelligence (XAI) is rapidly growing because many artificial intelligence (AI) applications based on machine learning (ML), even if they perform excellently, they cannot provide explanations in order that users can interpret their proposed solutions. This fact caused the rapid development of several algorithms and methods to interpret and explain ML models. The goal of this paper is to study the state-of-the-art of explainability techniques, present them in an informative manner and analyse them from the performance standpoint.

Interpretability is defined in literature as ”the degree to which a human can understand the cause of a decision” [1] and as ”the degree to which a human can consistently predict the model’s result” [2]. Unfortunately, there is a trade-off between the performance of a model and its transparency (interpretability). Furthermore, knowing ”why” may help with learning more about the problem, or about the data, and how it behaves in certain situations. The most important thing for the designer of the model is to figure out if the task at hand is a low-risk or a high-risk task. ”The need for interpretability arises from an incompleteness in problem formalization, which means that for many problems it is not enough to get the prediction, but also the explanation” [3]. Regarding the importance of interpretability, there are more reasons for this which will be listed below: human curiosity and learning, safety measures, and detecting biases.

STATE OF THE ART

SHAP

SHAP is a method which explains individual predictions [4]. The goal of this method is to explain an instance by calculating the contributions made by all implied features. By using Shapley values, we find out how to correctly attribute the prediction among features. Furthermore, there is an innovation [4] which suggests that Shapley values should be represented as an additive feature attribution method, like LIME. Shapley values have the following properties: Efficiency, Symmetry, Dummy and Additivity [5].

LIME

LIME (Local Interpretable Model-agnostic Explanations) is a model-agnostic explanation method [6], which explains predictions locally by the use of a surrogate model and local perturbations. The goal of this explanation method is to give attributions for each input feature by performing local permutations and training a surrogate interpretable model that is weighted according to the distance between the original prediction and the latter one. Due to the local nature of LIME, the learned model is a good approximation of the original model locally, thus the accuracy may be interpreted as fidelity of the prediction

Integrated Gradients

Integrated gradients is an attribution method, which requires no modifications to the networks [7] and is a combination of the Implementation Invariance from Gradients and Sensitivity from LRP or DeepLift. Let us consider a neural network, that is trained on a certain data set. In order to retrieve the attributions for a certain prediction, the gradients are computed in all of the points along the path between the example and a baseline input, which is a zero embedding vector. One of the most important benefit of this method is the fact that it satisfies the completeness axiom, which states that ”the sum of attributions is equal to the difference between the output of the network at the input x and the

baseline" [8]. An example of model-agnostic explanation method that also satisfies this axiom is SHAP [4].

Expected Gradients

Expected Gradients is a method derived from Integrated Gradients but with fewer hyperparameters [9]. The authors of the paper that introduced this method argued that it is a hard task of choosing the baseline embedding, and some choosing may be wrong, yielding wrong attributions. They have proposed a non-arbitrary selection of baseline, modelling the value of a not-known feature by integrating over a distribution of background data. The only problem this method has when comparing it to integrated gradients is the fact that the complexity of the algorithm increases, leading to a more processing power needed for generating attributions for a single example. We will try to conclude an experiment regarding the time required for each algorithm to generate local attributions based on the example's size.

Evaluation metrics

A very thorough review on the evaluation metrics for XAI was performed on approximately 600 papers, and concluded that there are 12 quality properties that should be further studied [8]. Unfortunately, few research studies focused on applying these properties in explanation methods for textual data. For that reason, we will try to alter these metrics such that they will fit NLP models. These properties are categorized into two primary areas: with or without user studies. Hence the previous experiments, we will focus on the properties that do not depend on peer review or user studies.

Further, we dived deeper into each of the four categories [8] in order to find suitable functional evaluation methods for NLP explanations. The types with which we will try to evaluate the methods we have chosen are: Incremental Deletion / Addition; Single Deletion; Data Randomization Check; Covariate Regularity

Incremental Addition is a method that tries to evaluate output-completeness by incrementally adding features for an example that are important for that prediction. If a method completely satisfies the output-completeness, then a wrong decision should be made by the model when all the important features are removed. Another method of measuring this property by using incremental addition is to count how many important features need to added / deleted in order to change the prediction of the model.

The Single Deletion method consists of deleting one feature from the initial example, measuring the difference in output, and finally comparing it with the difference in explanation. In the ideal case, "the explanation's feature importance score should be proportional to the output shift" [10]. Simply put, if a feature with a high importance score is deleted from the input example, then the output should have also a big change. This method can also be used to test the "null hypothesis", which states that if a feature has an almost 0 feature importance, the change in model output should not exist.

The Data Randomization Check is a method which acts as a sanity check for the "sensitivity of an explanation method to the relationship between instances and targets" [11]. This method states that if a model is trained on a dataset with shuffled labels, then since the model will learn a different target distribution, the explanations should be different. We will try to measure the difference by using a function that computes cosine distance between pairs of explanations, and then averaging them.

Yu and Varshney [12] suggest that a decision rule should be easier to remember if it is less entropic. By following this argument, we can calculate the Shannon entropy of the importance scores, and then computing the average for the entire test set. Theoretically, this score would indicate how noisy the explanations given by a method are.

IMPLEMENTATION

Dataset

The dataset used to make the experiments is "Conversations Gone Awry" from Cornell University [13], which is pre-processed by removing stop words and special characters, emojis and punctuation.

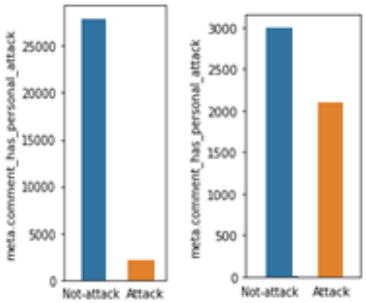

Figure 1. Label distribution for the original dataset (left) and for the sampled dataset (right)

It can be seen in Figure 1 (left) that the dataset is very unbalanced between not attacking and attacking utterances in conversations, implying that the model's performance on the small class is very poor and the explanation method's performance will be affected. Therefore, we sampled an equal number of examples from both classes, as seen in Figure 1 (right).

Model setup

Regarding the model choice, we have chosen a model that would work with both types of methods, the pretrained DistilBERT transformer network, from the HuggingFace library [14]. Table 1 shows the performance of the model we have fine-tuned on the dataset.

	Precision	Recall	F1 Score
Train	0.967	0.980	0.973
Test	0.971	0.966	0.968

Table 1. Model performance

As we can see, the performance on the test set is very good, so we will not need to investigate this matter, thus focus on the explanation methods and how we can compare the two categories.

Evaluation metrics

For these experiments, we have chosen four evaluation methods: Faithfulness, Monotonicity, Shannon Entropy, and Data Randomization Check

Comparison Setup

A couple of experiments were performed regarding the overall performance of the four explanation methods, and compared in terms of how different they are to one another. As shown in Figure 2, there are hree layers of abstractization in our experiments: deep learning model, explanation method and lastly the evaluation method.

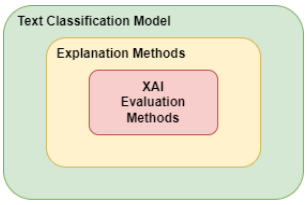

Figure 2. Layers of abstractization.

Figure 3 shows the steps taken to evaluate the explanation methods. The first step required to evaluate these methods was to train a model that would have an incredibly good performance, such that we would exclude the prediction errors from our analysis. After that, we have used the explanation methods to extract attributions for each word in each test example. By using these attribution vectors, we were able to compute the evaluation methods and interpret the results.

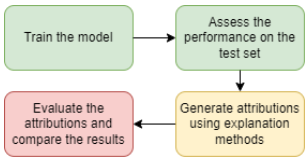

Figure 3. Comparison experiments' methodology.

RESULTS AND DISCUSSION

Table 2 shows the comparison experiment we have conducted in order to see the differences between model-agnostic and model-specific explanation methods.

Faithfulness

Figure 4 shows, from the faithfulness perspective, that SHAP achieved the highest score. Both SHAP and Faithfulness metric assume the fact that features are independent, not taking into account the context. On the other hand, gradient-based methods should in theory consider the context. As a conclusion, SHAP and these performance metrics have more in common, which could be a motive for the higher score.

From the monotonicity standpoint, all three methods have a very low performance, with an average of 3 percent of all the explanations being monotonic. Another assumption that we might take is the fact that, since we are using a small sample of baseline examples for Expected Gradient, the manifold of the data might be under-representative for our training dataset, hence the low score. Initially, we have measured only faithfulness and monotonicity metrics, but hence the assumption that these metrics are biased we have also computed two more evaluation metrics, Data Randomization Check and Mean Shannon Entropy. Figure 5 shows the distribution of faithfulness scores for each example in the sampled dataset. LIME and SHAP have better singular results on the explanations, where integrated gradients has worse results than expected gradients, as we have had predicted.

	Faithfulness	**Monotonicity**	**Data Randomization Check**	**Mean Shannon Entropy**
SHAP	0.3578	0.03%	0.0729	3.2063
LIME	0.3315	0.03%	-0.0138	3.2108
Integrated Gradients	0.0749	0.03%	-0.0399	5.2985
Expected Gradients	-0.1028	0.02%	-0.0285	4.8441

Table 2. Faithfulness and monotonicity metrics calculated for the four methods.

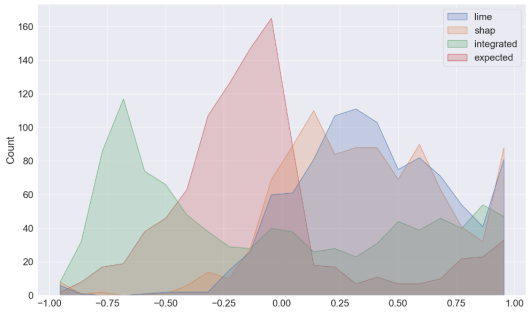

Figure 4. Faithfulness Scores' distribution on the test examples.

Data Randomization Check

From an intuitive standpoint, Data Randomization Check measures the ability of explanation methods to be able to distinguish between models that are trained on examples with initial and randomized labels. As further the random explanations are from the initial explanations, the better the explanation methods is, because it computes Spearman Correlation on feature attribution vectors. From our measurements found in Table 2, all four explanation methods have the same performance, which is particularly good, since the values hover around 0. Figure 5 shows the distribution of correlations measured on all the sampled examples. All the methods' correlations have a distribution that is close to a

normal distribution, which is seen in the Mean Spearman Correlations as well, which indicates that all the methods correctly assign the good attributions to the model that is not trained on randomized labels.

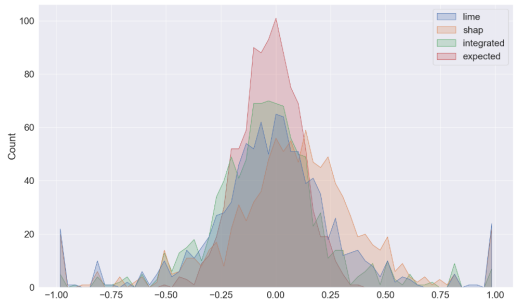

Figure 5. Data Randomization Check distribution of correlations.

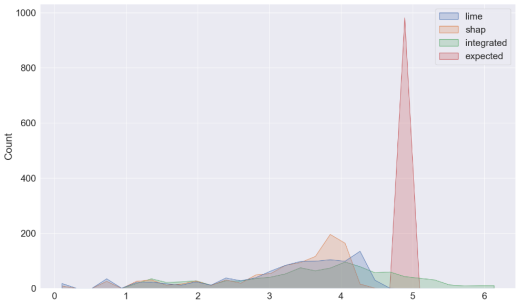

Figure 6. Shannon Entropies' distribution

Shannon Entropy

The Shannon Entropy metric tries to describe the readability of the explanations. Since we use a task of text classification, each word having an attribution score to the prediction, the metric will evaluate if an explanation is relying on few words or more, in which case the Shannon Entropy will be higher. As it can be seen in Table 2, model-agnostic explanation methods clearly performed better than model-agnostic methods, but the difference is not large. This result was expected since the granularity of methods that use gradients is much higher, hence the higher entropy. In this case, since this metric is not scaled between certain values, we do not know the minimum value that can be achieved by an explanation method, and so we will only compare the mean values of the four explanation methods between each other. Figure 6 shows the calculated entropies' distribution for all four methods. It can clearly be seen that Expected Gradients has a much higher mean, which is expected since it calculates mean values over a lot of samples. Attribution vectors given by this methods tend to have more non-zero values, and this is not desirable since we want to have concise explanations.

CONCLUSIONS

This paper had the purpose of comparing the two types of explainability algorithms, model-agnostic and model-specific. We have partially achieved this task by finding different evaluation metrics for this explainable AI algorithms, such that we could have a valid comparison. We firstly trained a text classification model on the dataset we have presented, and then by validating that this model has a particularly superior performance, we were able to use it further in our experiments. After this step was achieved, we have extracted explanations with the four presented algorithms such that we would be able to analyze the results, and then calculated the metrics mentioned in the article to decide which type of explanation algorithm is better for this task. Despite our expectation that the model-specific algorithms would perform better due to their access to the model's architecture, model-agnostic methods performed better taking into consideration all the metrics. Further work will try to fundament these findings and produce more evaluation methods for XAI methods.

REFERENCES

1.Miller T., Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Vol. 267, 2019, pp. 1-38,

2.Been K., Rajiv K., and Koyejo O., Examples are not enough, learn to criticize! criticism for interpretability. In Procs. of the NIPS'16, pp. 2288–2296, 2016.

3. Doshi-Velez F. and Been K., Towards a rigorous science of interpretable machine learning, arXiv, 2017.

4. Lundberg S. M., Lee S., A unified approach to interpreting model predictions. Procs. NIPS'17, pp. 4768–4777, 2017.

5. Shapley L., A Value for n-Person Games. RAND Corporation, Santa Monica, CA, 1952.

6. Ribeiro M. T., Singh S., and Guestrin C. " Why should I trust you?":Explaining the predictions of any classifier, 2016.

7. Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks. CoRR, abs/1703.01365, 2017.

8. Nauta M. et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. CoRR, abs/2201.08164, 2022

9. Erion, G. et al. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat Mach Intell **3,** 620–631 (2021).

10. Labreuche C., Fossier S. Explaining multi-criteria decision aiding models with an extended Shapley value. In Proceedings of IJCAI-18, pp. 331–339, 2018.

11. Adebayo J. et al. Sanity checks for saliency maps. CoRR, abs/1810.03292, 2018

12. Yu H., Varshney L. R. Towards deep interpretability: Learning hierarchical representations of tonal music, ICLR 2017.

13. Zhang J. et al., Conversations gone awry: Detecting early signs of conversational failure. In Proceedings of the 56th CACL, Vol. 1, pp. 1350–1361, 2018.

14.Wolf, T. et al., Huggingface's transformers: State-of-the-art natural language processing. CoRR, abs/1910.03771, 201