International Journal of User-System Interaction 14(2) 2021, 68-84 

# Explainable Artificial Intelligence in Natural Language Processing

Marian Gabriel Sandu[1], Ştefan Trăușan-Matu[1,2,3]

[1]University Politehnica of Bucharest
313 Splaiul Independenței, Bucharest, Romania

[2]Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăganescu"
Calea 13 Septembrie nr. 13, Bucharest, Romania

[3]Academy of Romanian Scientists
54 Splaiul Independenței, 050094, Bucharest, Romania

*E-mail: sandugabriel97@gmail.com, stefan.trausan@upb.ro*

**Abstract.** Explainable Artificial Intelligence has received a lot of interest and has been growing steadily in the last few years. This is because Machine Learning and Deep Learning domains overgrew creating more complex models that are highly accurate but lack explainability and interpretability. The aim of this paper is to present two most used model-agnostic explanation methods and experiment with them on a conversational dataset.

**Keywords**: artificial intelligence; machine learning; deep learning; neural networks; human-AI interaction

DOI: 10.37789/ijusi.2021.14.2.2

## 1. Introduction

Human-Computer Interaction is changing these years towards Human-Artificial Intelligence Interaction in an ever-growing degree. This fact is empowered by the remarkable achievements of Artificial Intelligence (AI) in using Machine Learning (ML) and its subdomain of Deep Learning (DL) (Krzysztof, 2018) , which became now the predominant paradigm, replacing the Knowledge-Based, Symbolic AI. However, these achievements come with a cost: many ML algorithms and especially DL (based on Deep Neural Networks - DNN), do not provide means for explaining their decisions or, in the case of Natural Language Processing (NLP), artificial dialogue systems,

cannot justify their generated utterances. Therefore, it can be said that such AI systems behave like "black-boxes", for example, in the case of conversational agents with NLP, they are not answerable in the sense of Mikhail Bakhtin (1993), as emphasized in several studies and official documents (Banavar, 2016; European Commission, 2019; AI HLEG, 2019). Moreover, the problem of lacking the possibility of providing explanations may drive to the impossibility of growing a real dialog with an artificial agent (Trausan-Matu, 2019). This major concern gave birth to a domain of research and development called *eXplainable Artificial Intelligence* (XAI), which aims at finding algorithms and methods able to interpret and explain the results provided by machine learning methods. In addition to XAI, another concern regarding AI is related to ethical problems (Trausan-Matu, 2019; Trausan-Matu, 2020; O'Neil, 2016), which are also generated by the opacity of how the results were obtained by many ML approaches.

The world of XAI is rapidly growing when looking at the number of scientific papers and conferences over the recent years. The problem with this field is the fact that there is a lot of new information, which needs to be structured and reviewed. The goal of this paper is to study all the state-of-the-art explainability techniques and present them in an informative manner.

Three reasons may be identified, which enforce the idea of creating explainability:

- The need for transparent models.
- The development of techniques that allow humans to interact with a model and grow a real dialog (Trausan-Matu, 2019).
- The building of trust between machine learning models and humans.

There are several approaches for XAI, which can be grouped in the following four classes:

- Interpretable models
- Model-agnostic methods
- Example-based explanations
- Neural network interpretation

In this paper, we will focus on model-agnostic methods for text classification tasks, and also test them on a dataset that has never been used before for this task. We have considered all the above-mentioned classes, but

we have taken the decision to go with model-agnostic methods due to the fact that they can be used to explain any algorithm, there are not dedicated to a specific approach, and this is what we consider that Explainable AI should be.

## 2. The concept of Interpretability

Interpretability has no formal definition, but we have found two informal definitions that we consider illustrative. The first definition states that interpretability is the degree to which a human can understand the cause of a decision (Miller, 2019). Other definition focuses on the degree to which a human can consistently predict the model's result (Kim, Khanna, & Koyejo, 2016).

### 2.1. Importance of Interpretability

Nowadays, in the Machine Learning community, the focus was directed towards the performance of models instead of explainability and interpretability. The question that is the center of this subject is "Why should we not just trust the model?". As Doshi-Velez and Kim (2016) state, a single metric (accuracy) cannot fully describe the performance of a model.

Unfortunately, there is a trade-off between the performance of a model and its transparency (interpretability). One must weigh the importance of a model's predictions to be explained, such that it will be a drop in the model's performance if this path is chosen. Furthermore, knowing "why" a result was generated may help with learning more about the problem, or about the data, and how it behaves in certain situations. The most important thing is for the designer of the model to figure out if the task at hand is a low-risk or a high-risk task, if it may be biased and generate unethical utterances. 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 (Doshi-Velez & Kim, 2017).

Regarding the importance of interpretability, there are more reasons, which are listed below:

- *Human curiosity and learning*: Humans are usually trying to learn pairs of cause and effect. Therefore, in order to facilitate learning and satisfy the curiosity of human beings, the explainability and

interpretability are especially important regarding Artificial Intelligence models.

- *Safety measures:* This is valid for example in automotive, self-driving cars, where the safety of a passenger is based on the performance of machine learning models, so there needs to be an explanation for the decision it makes (as a form of model debugging).
- *Detecting bias:* Machine learning pick up bias from training from a biased dataset. A good example is a model which automatically accepts or rejects people for a credit. This model may discriminate based on the background dataset, which is not allowed due to UE regulations. By introducing interpretability, we can detect that bias and modify the data in order to remove it with certain specialized algorithms.

## 2.2 Taxonomy of Interpretability Methods

In this section, a classification of Machine Learning interpretability methods will be described based on some relevant criteria. The first criterion of classification of these methods is *whether interpretability is achieved by reducing the complexity of the model (intrinsic) or by applying the methods of interpretability after the model was trained (post hoc)*. Regarding intrinsic interpretability, machine learning models that enter this category are interpretable by nature due to their simple structure, such as sparse linear models or decision tree-based models.

On the other hand, post-hoc interpretability can only be achieved after the training of the model occurs. An example of an interpretation method that achieves post-hoc interpretability Explainable AI in Natural Language Processing is SHAP (SHapley Additive exPlanations - Lundberg and Lee, 2017), which is derived from game theory and tries to compute feature importances based on predictions from the trained model. Another method that is not so ambiguous is permutation feature importance, which can be computed for decision trees. This paper will be organized based on the criteria we have briefly explained here.

The second criterion these methods can be classified with is according to *the result of the interpretation method*. The result can be classified into:

- *Feature summary statistics:* The methods included here should return summary statistics for each feature in the data set, such as feature

importance, or more complex results, such as the pairwise feature interaction strength.

- *Feature summary visualization:* Based on the methods from the category above, the problem of some results may be the fact that simple numbers may not present an actual information, and because of this fact a visualization method for those results is required. An example of such a method is the Partial Dependent Plot. This method shows curves that show a certain feature together with the average predicted outcome.
- *Model internals* (e.g., model weights): As a clear example, intrinsically interpretable models can be classified into this category. Model internals can be weights in linear models, or decision tree splits and thresholds. Regarding a highly successful branch of Artificial Intelligence, Computer Vision, the method of visualization of feature detectors learned in convolutional neural networks falls into this category.
- *Data point:* This category might be ambiguous, since it includes all the interpretability methods that return data points as explanations, such as counterfactual examples. This method takes as an input a data point from the original data set, and try to change certain features in order to change the prediction of the model for that data point (e.g. change the predicted class of a point).
- *Intrinsically interpretable model:* This category has not had that much development as the others and represents the interpretation of black box models by approximating them with an interpretable model, which can be easily explained.

Another meaningful criterion is *the usefulness of the interpretability method (model-specific or model-agnostic)*. Model-specific methods are limited to a specific class of models. Instead, model-agnostic methods can be used on any machine learning model, and also can be only applied after the training step. These methods are useful when we do not have access to the model's architecture or internal parameters, meaning they focus on input-output data pairs.

The last criterion of the interpretability methods is *the scope (local or global)*. This criterion refers to the question of whether the method explains an individual prediction or the entire model.

## 2.3. Scope of Interpretability

In this subsection we will present the evaluation of the level of transparency of each step in a model creation (training, inference, etc.)

*2.3.1. Algorithm Transparency*

Algorithm transparency can be defined as the method the algorithm uses in order to create the model. The algorithm learns a model through data, and this metric describes the method and the relationships the model can learn. Thus, the transparency of an algorithm is described only by the knowledge we have about it. For example, algorithms that have a linear characteristic are well studied and are characterized as being very transparent. On the other hand, deep learning methods, which imply learning weights that are propagated through thousands of neurons cannot be well explained and thus are less transparent.

*2.3.2 Global, Holistic Model Interpretability*

A model can be fully interpretable if one can explain the entire model at once (Lipton, 2018). The problem with the global interpretability is the fact that you need a lot of information to achieve it, such as the trained model, knowledge of the algorithm and the data used for training. Global model interpretability can be also described as finding out the distribution of the outcomes (predictions) based on the input features. This raises the problem that if a model has a substantial number of parameters or weights the chance for a human to understand that model is not good. Also, the data used today has more than three features, meaning that a hyperplane cannot be visualized in order to see feature relationships.

*2.3.3 Global Model Interpretability on a Modular Level*

As we have stated in the previous subsection, generating global interpretability is untouchable for certain types of models. But, if a model can be sectioned in smaller parts, such as single weights, it can become interpretable. The linear models' parts that can be interpretable are the weights, for trees are the splits and leaf node predictions. Regarding linear models, the problem is that the interpretation of a single weight is linked with all the other weights due to connections between the distributions of the

features. Even though this is not a good thing, the weights in a linear model are much more interpretable than the weights of a deep neural network.

*2.3.4 Local Interpretability for a Single Prediction*
As said before, global explanations are harder to obtain for complex models. Instead, trying to explain a single example from the dataset might give an insight of the behavior of the model. The problem with these methods is the fact that you do not have a big picture on the model, and by explaining some predictions you may not find the complex dependences between features.

# 3. Explanations and their properties

## 3.1 Definition and metrics

An explanation can be defined as the feature values of an instance to its model prediction in an understandable way (Robnik-Šikonja & Bohanec, 2018). Further, properties of these explanation will be detailed (Robnik-Šikonja & Bohanec, 2018). The first described properties are about the explanation methods:

- *Expressive power*: Describes the structure of an explanation, or how the explanation looks to the end user. They can be rules, decision trees, a numerical metric, natural language, etc.
- *Translucency*: Refers to how tied the explanation is to the model's parameters and architecture. For example, methods of interpretability that rely on intrinsically explainable models such as linear regression are highly translucent, but methods that only rely on observing the inputs and the predictions are not translucent. There is no good value of translucency, because it is dependent on the use case. For instance, a high translucency relies on more information to give an explanation, but the low translucency may make the explanation method more portable.
- *Portability*: This measures the diversity of the explanation method, and it depends on the number of different machine learning architectures can be used from.
- *Algorithmic complexity*: It describes the complexity of the algorithm for generating explanations.

A second group of properties are related to the individual explanations:

- *Accuracy*: This metric describes how good the explanation method predicts on unseen data. It is not a restrictive metric because of the fact that it is normal if both the explanation method and the machine learning model have low accuracy.
- *Fidelity*: This is one of the most important metrics for individual explanations, and it describes how well the explanation approximates the black box prediction. An explanation is invalid if it has low fidelity. The problem is that, because of local explanations, the method may only have good local fidelity on a subset of data, and not globally.
- *Consistency*: The degree to which similar explanations are generated from different models trained on a similar task. For examples, explanations may vary because of the explanation method even though the models are trained on the same data and the predictions being similar.
- *Stability*: The degree to which similar explanations are generated for similar instances. This metric is different from consistency because of the fact that it is based on a single model.
- *Comprehensibility*: This depends on how readable the explanations are (depends on the audience and the size of the explanation).
- *Certainty*: It is described as the degree to which the explanations reflect the model's prediction.
- *Degree of importance*: The degree to which the explanations are reflecting the importance for each returned item.
- *Novelty*: It is related to certainty, and describes if explanations would reflect the fact that the explained instance is from a new region, meaning that it is out of distribution.
- *Representativeness*: It describes how the model is represented by the explanations (local or global explanations).

## 2.2 Human-friendly explanations

Explanations and their classification is a very sensitive subject. Miller (2019)

has conducted a huge survey of publications about this topic, and this subsection will review this publication in detail, and put an emphasis to what is important for Interpretable Machine Learning.

Many times, an explanation is an answer to a “why” question (Miller, 2019). Regarding Interpretable Machine Learning, everyday-type explanations are of interest to us. There are a couple of criteria that needs to be taken into consideration when thinking if an explanation is proper for human beings, and these are:

- Explanations should be selected, not in full form, because of how human attention work.
- The social environment and the target audience of the machine learning model need to be taken into consideration when some predictions are explained.
- "If one of the input features for a prediction was abnormal in any sense (like a rare category of a categorical feature) and the feature influenced the prediction, it should be included in an explanation, even if other ’normal’ features have the same influence on the prediction as the abnormal one" (Molnar, 2019).
- The explanation should predict the event with a high fidelity score, meaning it should be as truthful as possible.
- Better explanations may be provided by giving the machine learning algorithm a set of apriori knowledge, such as rules, but this would damage the performance of the model too much so it is not feasible.
- A good explanation can be considered to be one that is more general and can explain different events of prediction. The problem with this statement is that it is in contradiction with the third item in the list, the "abnormality" clause.

## 3. Model-Agnostic Methods

The model-agnostic methods can be used with any machine learning model. There are a couple of desirable aspects of such methods (Ribeiro, Singh, & Guestrin, 2016):

- *Model flexibility*: The interpretation method can work with any machine learning method.
- *Explanation flexibility*: The form of the explanation is not limited or

fixed, may be a graph or a formula.

- *Representation flexibility*: The explanation system may use different feature representation than the trained model.

Next, we will describe a couple of popular and state-of-the art agnostic methods.

## 3.1 Local Surrogate (LIME)

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) is a concise way of implementing an algorithm to approximate predictions of a black-box model. Instead of training a global surrogate model, LIME focuses on local explanations (single examples).

The first step in applying this algorithm is to consider that we only have a blackbox model that has a prediction function. The goal is to understand why this model makes certain predictions. LIME tests what happens when you try to predict data derived from the original data by performing perturbations. There are five steps in training a local surrogate model:

1. Select the example from the dataset which it is wanted to be explained based on a black-box model.
2. Perturb the dataset and extract predictions with the black-box model.
3. Weight the new samples based on proximity.
4. Train an interpretable surrogate model with the perturbed data.
5. Explain the prediction.

The LIME method has several advantages:

- The same interpretable model can be used regarding what the underlying black-box model is. This statement empowers the fact that the explanations are human-friendly.
- If a decision tree or Lasso regression (Tibshirani, 1996) is used, the explanations are short and possibly contrastive.
- Works for tabular, image, and text data.
- The fidelity measurement, meaning the R-squared score, gives a very good interpretation on how accurate the surrogate model mimics the black-box model.
- Actual LIME implementations are very easy to use (Python).

However, LIME has also several disadvantages:

- The neighborhood of a point is very hard to correctly define, and it is an unsolved issue, meaning that we need to be very careful when using this algorithm such that we have good explanations.
- The sampling steps can be improved from the actual version, because of the fact that this step can yield unlikely data points.
- Instability of the explanation because of the sampling step.
- The algorithm can be tempered with in order to hide biases in the dataset or model, and this decreases the trust in LIME.

## 3.2 SHAP

SHAP is a method proposed by Lundberg and Lee (2017) which explains individual predictions. The goal of this method is to explain an instance $x$ by calculating the contributions made by all the involved features. By using Shapley values (Shapley, 1952), we find out about how to fairly distribute the prediction among features. Furthermore, there is an innovation, which suggests that Shapley values should be represented as an additive feature attribution method, similar to LIME (Lundberg & Lee, 2017).

Shapley values have several properties (Shapley, 1952): Efficiency, Symmetry, Dummy and Additivity. SHAP also satisfies these properties, the three most desirable ones being the following (Molnar, 2019):

- Local accuracy
- Missingness. This property enforces all missing features' Shapley
- values to 0.
- Consistency: If a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features), the Shapley value also increases or stays the same.

There are two derived methods from SHAP: KernelSHAP and TreeSHAP (also proposed by Lundberg et. al (2017), a variant of the SHAP for tree-based models).

- The disadvantages of SHAP and its variants are:
- KernelSHAP is slow because it requires a lot of computations.
- SHAP considers that there are no dependences between features, which is not true and may yield unreliable results in some cases. This happens when random values for features are sampled which are out

of distribution.

- TreeSHAP may yield unintuitive feature attributions, because of the fact also stated above (different than 0 feature importances).

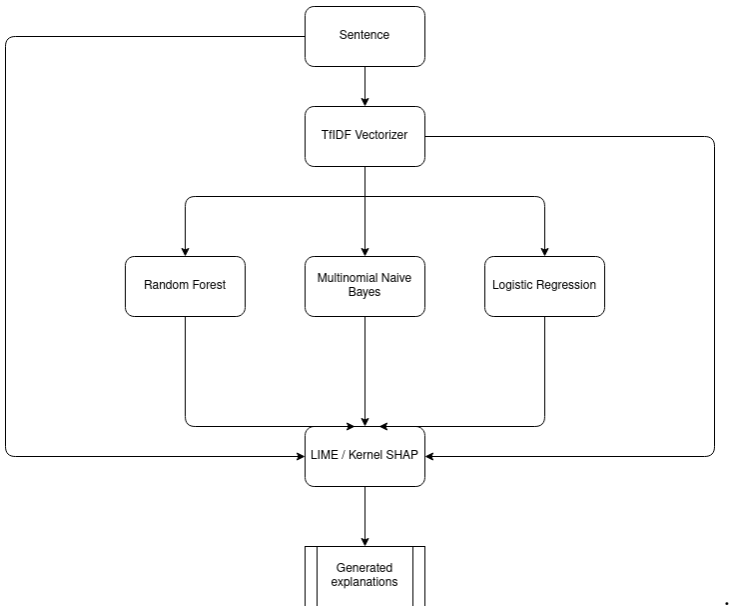



Figure 1. Experimental setup.

## 4. Experiments

Several experiments were performed with the LIME and SHAP model-agnostic interpretable methods for Natural Language Processing. The experimental setup is shown in Figure 1 and it has four stages:

1. Initial preprocessing of data (removing stop words, removing special characters, emojis and punctuation).
2. Transform text utterances into text embeddings.
3. Train three Machine Learning models (Logistic Regression, Multinomial Naive Bayes, and Random Forest) on the embeddings.
4. Generate explanations using LIME and SHAP for each model.

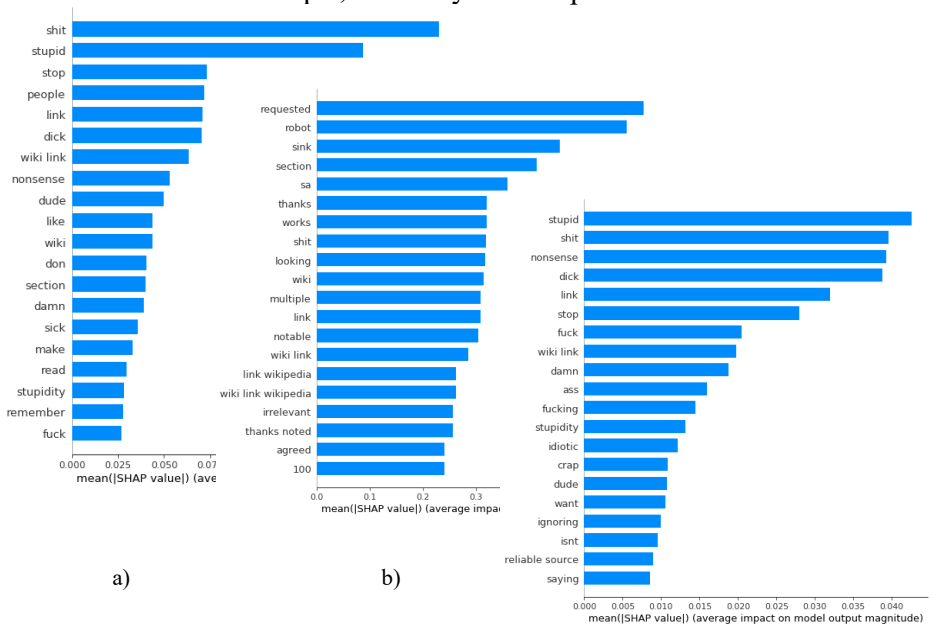
The dataset used to make the experiments is "Conversations Gone Awry"

from Cornell University (Zhang, 2018), a corpus of conversations having utterances that may be characterized as personal attacks. The three ML models were used for classifying utterances as attacks (“class 1”) and not attacks (“class 0”).


Figure 2. Lime individual explanation for a single example belonging to class 0 (a, b, c) and class 1 (d, e, f), for Logistic Regression (a, d), Multinomial Naive Bayes (b, e), and Random Forest (c, f).

The original dataset was very unbalanced between the two classes. The consequence is 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. Furthermore, by applying the preprocessing steps we have leveled the character and word.

Comparative results of the LIME and SHAP approaches to each of the three ML models are presented in Figures 2 and 3. The explanations of the classifications of the utterances performed by the ML algorithms consist in emphasizing how words in the texts contributed to the provided results. For example, in Figure 2, LIME emphasizes the importance of each word by different colors and intensity. SHAP offers several ways of illustrating the results, in Figure 3 only a bar chart is used for the representation of the contribution of the concepts, sorted by their importance.




a) b)

Figure 3. SHAP Summary Plot for Logistic Regression (a), Multinomial Naive Bayes (b), and Random Forest (c).

In order to build the explanations, we have chosen two instances that are from different classes in order to illustrate the functionalities of SHAP and LIME. By looking at both explanations from these algorithms, we can clearly see differences because of the significant differences between the algorithms.

Due to the Shapley values calculations, this algorithm guarantees local accuracy and robustness in explanations, as faced with LIME, which builds surrogate models that locally mimic the original explained model. The main advantage of SHAP over LIME is the fact that it can explain a model globally, by computing an average over all the importances on a partition of examples, while LIME can only explain single instances.

Furthermore, we can also look at how explanations are presented to the user, at which LIME has a much better visualization technique due to its nature. Unfortunately, both methods have certain issues. One such issue is the fact that both methods are based on sampling, meaning that the absence of features is simulated by sampling examples from a background dataset and replacing those features with the sampled features. The problem is the fact that the sampling methods used so far do not generate examples on the data manifold (not probable examples). Also, due to this, two experiments with the same experimental setup may yield slightly different explanations.

There are many algorithms that explain text classifiers, but we have chosen to focus on the model-agnostic explanations because of the fact that model-specific explainability methods are much harder to understand and implement, and also are not that robust mathematically. Examples of different methods that are not model-agnostic are variants of SHAP (LinearExplainer, DeepExplainer), DeepLift, and other experimental theoretic methods of manipulating the neural networks' learnt parameters.

## 4. Conclusions

In this paper, we have focused on introducing and classifying the main concepts and features of XAI, and on presenting an experiment that emphasizes the explanations generated for three machine learning models for text classification with the help of model-agnostic methods, such as SHAP and LIME, and state the differences and their problems. As further research, we will try to come with something new regarding the explanation of conversations as a whole, not only single utterances.

## References

AI HLEG (2019) Ethics guidelines for trustworthy AI. Downloaded from https://data.europa.eu/doi/10.2759/177365

Bakhtin, M.M. (1993) *Toward a Philosophy of the Act*, University of Texas Press

Banavar, G. (2016) Learning to trust artificial intelligence systems: Accountability, compliance, and ethics in the age of smart machines. Armonk, NY: IBM Research.

Doshi-Velez, F. and B. Kim (2017) Towards a rigorous science of interpretable machine learning, https://arxiv.org/abs/1702.08608

European Commission (2019) Building Trust in Human-Centric Artificial Intelligence, COM 168, Available at https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF, (Last accessed: 25 July 2020)

Kim, B. Khanna, R., and Koyejo, O. (2016) Examples are not enough, learn to criticize! criticism for interpretability, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., p. 2288–2296.

Krzysztof, C. (2018). Deep Neural Networks - A Brief History. *Advances in Data Analysis with Computational Intelligence Methods*, pp.183-200

Lipton, Z. C. (2018) The mythos of model interpretability, *ACM Queue*, | may-june, https://dl.acm.org/doi/10.1145/3236386.3241340

Lundberg, S. M., and S.-I. Lee (2017) A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., p. 4768–4777.

Lundberg, S. M., G. G. Erion, G. G., and Lee, S.-I. (2019) Consistent individualized feature attribution for tree ensembles, https://arxiv.org/abs/1802.03888, (Last accessed: 25 July 2020)

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence*, Volume 267, Pages 1-38, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2018.07.007, (Last accessed: 25 July 2020)

Molnar, C. (2019) Interpretable Machine Learning https://christophm.github.io/ interpretable-ml-book/, (Last accessed: 25 July 2020)

O'Neil, C (2016) Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy, Crown Books

Ribeiro, M.T., S. Singh, and C. Guestrin (2016) why should i trust you?: Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, p. 1135–1144. Available: https: //doi.org/10.1145/2939672.2939778

Ribeiro, M.T., S. Singh, and C. Guestrin (2016) Model-agnostic interpretability of machine learning, https://arxiv.org/abs/1606.05386, (Last accessed: 25 July 2020)

Robnik-Šikonja, M and M. Bohanec (2018) Perturbation-Based Explanations of Prediction Models, J. Zhou and F. Chen, Eds. Cham: Springer International Publishing. Available: https://doi.org/10.1007/978-3-319-90403-0_9

Shapley, L. S. (1952) A Value for n-Person Games. Santa Monica, CA: RAND Corporation.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the lasso". Journal of the

Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88

Trausan-Matu, S. (2019) Is it possible to grow an I–Thou relation with an artificial agent? A dialogistic perspective, *AI & Society, Journal of Knowledge, Culture and Communication*, ISSN 0951-5666, Vol. 34, Nr. 1, Special Issue: Ethics of AI and Robotics, Springer-Verlag London, pp. 9-17, https://dl.acm.org/doi/10.1007/s00146-017-0696-5

Trausan-Matu, S. (2020) Ethics in Artificial Intelligence. *International Journal of User-System Interaction* 13(3), pp. 136-148, doi: 10.37789/ijusi.2020.13.3.2

Zhang, J., Chang, J., C. Danescu-Niculescu-Mizil, C., L. Dixon, L., Y. Hua, Y., D. Taraborelli, D., and N. Thain, N. (2018) Conversations gone awry: Detecting early signs of conversational failure, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, pp. 1350–1361. Available: https://www.aclweb.org/anthology/P18-1125, doi:10.18653/v1/P18-1125.