Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

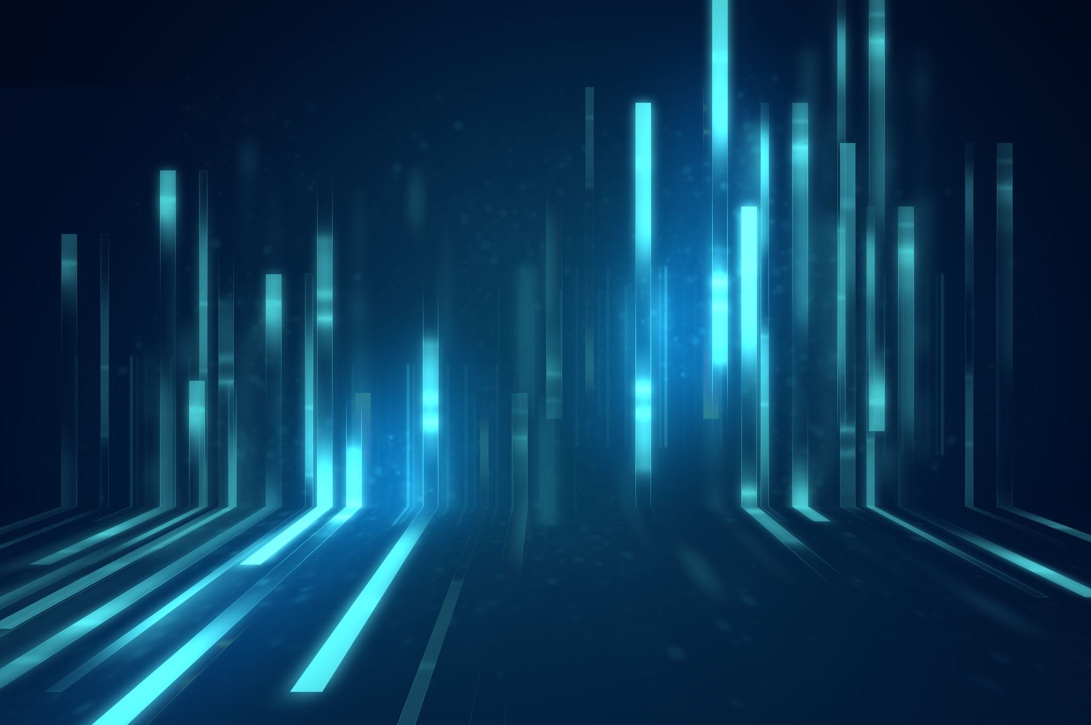
­­­

Transparency

Expert

Module 1

Explainability in NLP – lesson 1



# Introduction

This module focuses on explainability In the NLP domain, discussing methods to generate explanation of text models:

* Methods to generate post-hoc explanations of the prediction of a text classification model
* Approaches to generate explanations of recognition models
* Visualisations as explanations of NLP models

­­­

**Reading**

* Belle, Vaishak, and Ioannis Papantonis. "Principles and practice of explainable machine learning." Frontiers in big Data (2021): 39.
* Danilevsky, Marina, et al. "A survey of the state of explainable AI for natural language processing." arXiv preprint arXiv:2010.00711 (2020).
* Balkir, Esma, et al. "Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models." arXiv preprint arXiv:2206.03945 (2022).
* Interpretability in Natural language processing (NLP)

Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2022).

* Slides from: <https://explainml-tutorial.github.io/neurips20>
* Desirable properties of Post-Hoc Explainability methods

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in big Data, 39

* Survey on explainability approaches in text classification:
  + Atanasova, Pepa, et al. "A diagnostic study of explainability techniques for text classification." arXiv preprint arXiv:2009.13295 (2020).
* Fundamental axioms in cooperative game theory from:
  + Molnar, C. (2020). Interpretable machine learning. Lulu. Com
* Shapley value:
  + Shapely value slides: <https://icml.cc/media/icml-2019/Slides/4776.pdf>
  + <https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1>
* Local Interpretable Model agnostic Explanations (LIME)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016

* Anchors: High-Precision Model-Agnostic Explanations

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018

* SHapley Additive exPlanations (SHAP)

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions

* Influence functions

Han, Xiaochuang, Byron C. Wallace, and Yulia Tsvetkov. "Explaining black box predictions and unveiling data artifacts through influence functions." arXiv preprint arXiv:2005.06676(2020)

* Layer-wise Relevance Propagation (LRP)

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, 193-209

* Integrated gradients (IG)

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR

* DeepLIFT (Deep Learning Important FeaTures)

Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMLR

* Gradient based methods in SHAP

Detailed Implementation differences on how gradients are computed between Deep LIFT and Deep SHAP can be found here: <https://github.com/kundajelab/deeplift#what-are-the-similarities-and-differences-between-the-deeplift-like-implementations-in-deepexplain-from-ancona-et-al-iclr-2018-and-deepshapdeepexplainer-from-the-shap-repository>

* Model Predictions

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions

<https://github.com/slundberg/shap>

**Self-assessment pass/fail questions**

1. Which of the following approaches is not model agnostic?

a. Anchors

b. LIME

c. Kernel SHAP

d. Partition SHAP

e. Layer wise relevance propagation

2. Which of the following offers consistent global and local explainability:

a. SHAP

b. LIME

c. Integrated Gradients

d. All of the above

e. None of the above

3. Which of these axioms are NOT satisfied by Integrated Gradients? (Tick all that apply)

a. Symmetry

b. Additivity

c. Implementation Invariance

d. Linearity

e. Violation of null effects

4. Which of the following methods uses the concept of a baseline to calculate feature attributions? (Tick all that apply)

a. Integrated Gradients (IG)

b. Layer wise relevance propagation (LRP)

c. Deep Lift

d. Deep Shap

e. All of the above

5. Which of the following axioms are satisfied by Shapley values?

a. Linearity

b. Completeness

c. Symmetry

d. Dummy

e. All of the above

6. Which of the following statements about Shapley values is NOT true?

a. Exact computation of Shapley values can take a long time as feature set and data set size increase

b. The Shapley value of a feature value is the difference of the predicted value after removing the feature from the model training

c. Explanations created with the Shapley value method always use all the features

d. The Shapley value is the wrong explanation method if you seek sparse explanations (explanations that contain few features)

e. Shapley values are a model agnostic method

7. Which of the following is a limitation of explainability methods in NLP?

a. Gaps in what benchmarks show and real-world data applications

b. Evaluation criteria are often itself very simple

c. User Interfaces for explainability vary across studies

d. Perturbation based methods can likely produce unrealistic samples with limited coverage

e. All of the above

8. Which of the following is a limitation of LIME compared to Anchors?

a. Computing perturbation distribution can be time-consuming.

b. Coverage is sometimes not clear when investigating explanations.

c. LIME explanations are more intuitive.

d. Explanations computed for local instances of interest are generally accurate.

e. All of the above

9. Which of the following axioms is/are NOT satisfied by LRP and DeepLIFT?

a. Sensitivity

b. Completeness

c. Implementation Invariance

d. Additivity

e. All of these are satisfied

10. Which of the following computations are intractable?

a. Integral of the gradients of a neural networks output with respect to the inputs in Integrated gradients explainability method

b. For an arbitrary perturbation distribution and black-box model f, computing precision directly in the Anchors explainability method

c. Computing exact Shapley values for a large set of features

d. A, b, c

e. None of the above

11. Which of the following methods is often used as reference when evaluating explanations generated by explainability methods in isolation?

a. LIME

b. SHAP

c. Anchors

d. Integrated Gradients

e. Influence functions

**Answers**

1) e, 2) a, 3) a, 4) a/c/d, 5) e, 6) b, 7) e, 8) b, 9) c, 10 a/b/c, 11) a

# Pass/fail questions – further reading

Question 6

“*The Shapley value can be misinterpreted. The Shapley value of a feature value is not the difference of the predicted value after removing the feature from the model training. The interpretation of the Shapley value is: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.”* <https://christophm.github.io/interpretable-ml-book/shapley.html#the-shapley-value-in-detail>

# References

Shapley, L.S. (1952). A Value for N-Person Games. [online] www.rand.org. Available at: <https://www.rand.org/pubs/papers/P295.html> [Accessed 22 Sep. 2022].

Balkir, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2022). Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. arXiv preprint arXiv:2206.03945.​

NeurIPS 2020 Tutorial

<https://explainml-tutorial.github.io/neurips20>​

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in big Data, 39.​

Molnar, C. (2020). Interpretable machine learning. Lulu. com​

The Shapley Value for ML Models

<https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1>​

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.​

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.​

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. ​

Han, Xiaochuang, Byron C. Wallace, and Yulia Tsvetkov. "Explaining black box predictions and unveiling data artifacts through influence functions." arXiv preprint arXiv:2005.06676(2020).​

​

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, 193-209.​

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.​

Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In International conference on machine learning (pp. 3145-3153). PMLR.​

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. ​

<https://github.com/slundberg/shap​>

Masís, Serg. Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples. Packt Publishing Ltd, 2021.

Štrumbelj, E. and Kononenko, I. (2013). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), pp.647–665. doi:10.1007/s10115-013-0679-x.

Wu, T., Ribeiro, M., Heer, J. and Weld, D. (2021). Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. [online] pp.6707–6723. Available at: <https://www.cs.cmu.edu/~sherryw/assets/pubs/2021-polyjuice.pdf> [Accessed 22 Sep. 2022].

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. arXiv preprint arXiv:2010.00711.

Atanasova, P., Simonsen, J.G., Lioma, C. and Augenstein, I. (2020). A Diagnostic Study of Explainability Techniques for Text Classification. [online] ACLWeb. doi:10.18653/v1/2020.emnlp-main.263.

Ancona, M., Ceolini, E., Öztireli, C. and Gross, M. (2022). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. [online] openreview.net. Available at: <https://openreview.net/forum?id=Sy21R9JAW>