Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

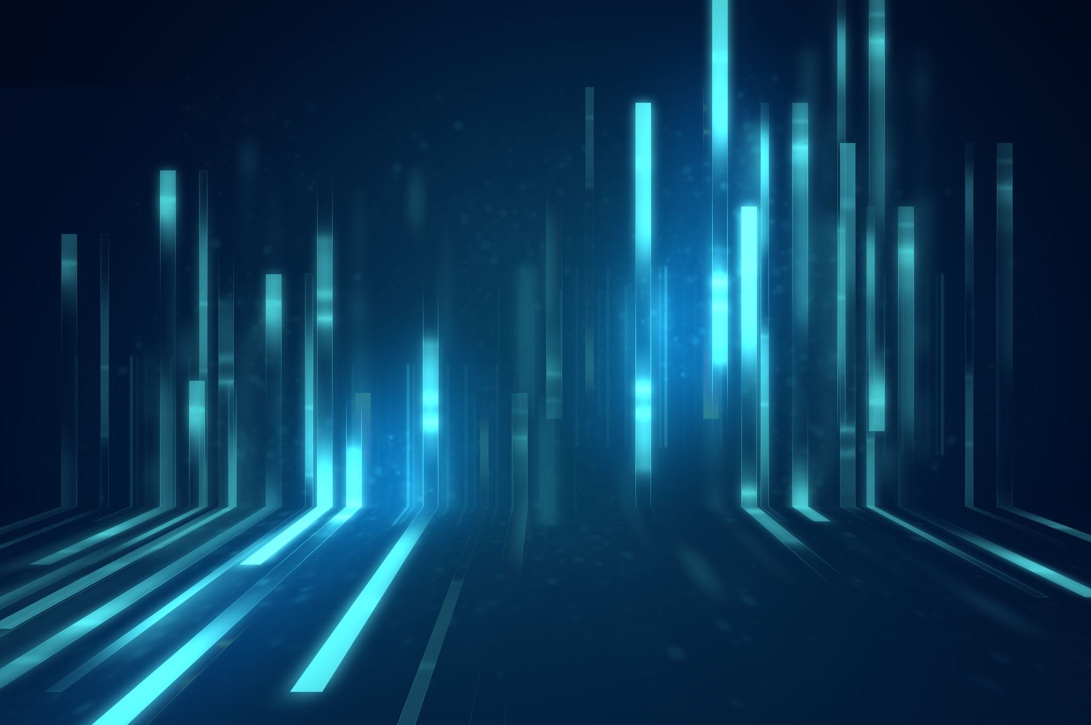
­­­

Transparency

Expert

Module 1

Explainability in NLP – lesson 2



# 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**

* Evaluating Generated Explanations

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

* Saliency Heatmaps in NLP

Arras, Leila, et al. "Explaining recurrent neural network predictions in sentiment analysis." arXiv preprint arXiv:1706.07206 (2017)

* Explainable Text Classification Dashboard- OmniXAI

<https://sfr-omnixai-demo.herokuapp.com/nlp>

* Explainable Text Classification Dashboard- AllenNLP

Gardner, Matt, et al. "Allennlp: A deep semantic natural language processing platform." arXiv preprint arXiv:1803.07640 (2018)

* Interpretations of Explainability Models
  + OmniXAI – <https://github.com/salesforce/OmniXAI>
  + Alibi – <https://github.com/SeldonIO/alibi>
  + SHAP – <https://github.com/slundberg/shap>
  + LIME – <https://lime-ml.readthedocs.io/en/latest>
  + InterpretML – <https://github.com/interpretml/interpret>
  + Captum – <https://captum.ai/tutorials/IMDB_TorchText_Interpret>
  + Anchor – <https://github.com/marcotcr/anchor>
  + Transformers Interpret <https://github.com/cdpierse/transformers-interpret>
  + Ferret XAI <https://github.com/g8a9/ferret>
  + TextAttack – <https://github.com/QData/TextAttack>

Real World Case Studies - Why Explainability Matters in NLP?

* GPT-3 makes racist jokes, condones terrorism, and accuses people of being rapists

<https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible>

* Amazon scraps secret AI recruiting tool that showed bias against women – <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>

**Self-assessment pass/fail questions**

1. Which of the following evaluation methods can have multiple explanations for one text instance? (Select all that apply)

a. Human grounded evaluation

b. Quantifying evaluations by comparing to ground truths

c. Informal investigation of explanations by looking at desirable properties of explainability methods

d. Counterfactuals

e. None of these

2. Which of the following are valid ways to evaluate the effectiveness of generated explanations?

a. Informal examination of explanations

b. Using some quantitative metrics for faithfulness of explanations

c. Human based evaluations by giving them a combination of predictions or explanations or both and asking them to assess model behaviour.

d. Compare generated explanations to ground truth data in order to quantify the performance of explainability techniques

e. All of the above

3. Which of the following methods achieve global explanability? (Tick all that apply)

a. Collection of local explanations

b. SHAP waterfall plots

c. Model distillations

d. SHAP summary plots

e. SHAP force plots

4. Which of libraries contains implementation of Anchors?

a. Omnixai

b. Alibi

c. SHAP

d. InterpretML

e. Option A and Option B

5. Which of the following SHAP explainers are model agnostic? (Tick all that apply)

a. Gradient Explainer

b. Deep Explainer

c. Tree Explainer

d. Partition Explainer

e. Kernel Explainer

6. Which of the following libraries contains explainability methods especially built for pytorch users?

a. Captum AI

b. OmniXai

c. Alibi

d. SHAP

e. None of the above

7. Which of the following is the TRUE about the following force plot?

Timeline

Description automatically generated

a. The model predicts in favour of the positive class

b. The model would predict the positive class with 0.4936 probability given a prediction neutral input

c. Words honest, good contribute positively towards prediction of the positive class

d. The model predicts the negative class with 0.43 probability

e. All of the above

8. Which of the following explainability methods requires NOT access to a background/reference dataset?

a. Deep SHAP

b. Shapley

c. Layer wise relevance propagation

d. Deep LIFT

e. Integrated Gradients

9. Which of the following methods is notable when talking about fairness in NLP?

a. Counterfactuals

b. SHAP

c. LIME

d. Anchors

e. Integrated Gradients

10. Which of the following sampling methods in Alibi for the anchors algorithm replaces words by sample words from a corpus with the same Part-of-speech tag and similarity in embedding space?

a. Sampling by UNK token.

b. Sampling by using probability distribution of a masked language model.

c. Sampling by similarity.

d. All of the above

e. None of the above

11. Which of the following methods are available to approximate integral in IG in the Alibi library?

a. riemann\_left

b. riemann\_right

c. riemann\_trapezoid

d. gausslegendre

e. All of the above

**Answers**

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

# Pass/fail questions – further reading

Question 2

Read more here:

<https://aclanthology.org/2020.aacl-main.46.pdf>

# References

* Interpretability in Natural language processing (NLP)

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.

<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

Molnar, C. (2020). Interpretable machine learning. Lulu. Com

* Shapley value

<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

* 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

* Anchors vs LIME

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>

* Interpreting Model Predictions

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.
* Molnar, C. (2020). Interpretable machine learning. Lulu. com
* Lertvittayakumjorn, P. and Toni, F. (2019). Human-grounded Evaluations of Explanation Methods for Text Classification. [online] ACLWeb. doi:10.18653/v1/D19-1523.
* ‌ Nguyen, D. (2018). Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. [online] ACLWeb. doi:10.18653/v1/N18-1097.
* ‌ Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In International conference on machine learning (pp. 1885-1894). PMLR.
* Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi Explain: Algorithms for Explaining Machine Learning Models. J. Mach. Learn. Res., 22, 181-1.
* Yang, W., Le, H., Savarese, S., & Hoi, S. C. (2022). OmniXAI: A Library for Explainable AI. arXiv preprint arXiv:2206.01612.
* Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., ... & Lee, S. I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering, 2(10), 749-760.
* Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896.
* Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. arXiv preprint arXiv:2005.05909.
* Attanasio, G., Pastor, E., Di Bonaventura, C., & Nozza, D. (2022). ferret: a Framework for Benchmarking Explainers on Transformers. arXiv preprint arXiv:2208.01575.
* Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., ... & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640.
* Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., & Singh, S. (2019). Allennlp interpret: A framework for explaining predictions of nlp models. arXiv preprint arXiv:1909.09251
* Arras, Leila, et al. "Explaining recurrent neural network predictions in sentiment analysis." arXiv preprint arXiv:1706.07206 (2017).



­­­