P4 - Explanation evaluation in text classification

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Salvatore Greco

Project. Evaluating the quality of explanations generated by explainability methods is critical for ensuring the reliability of explanations in real-world applications. This project focuses on enhancing the evaluation of explanation quality in the context of text classification.

Overview.

Evaluating the quality of explanations of the prediction of machine learning models is crucial for ensuring the reliability and trust of explanations. Various evaluation methods focused on assessing different aspects of explanation quality such as faithfulness, plausibility, robustness, and compactness [1, 3, 5, 6]. Considering this relevance, different libraries and tools have been proposed for evaluating explanations [2, 4]. Despite the efforts, many explainability libraries only cover a subset of these methods.

Goal.

The task of the project is first to systematically review existing evaluation methods to assess the quality of the explanation. Then, the project aims to improve the evaluation capabilities of a specific package, ferret [4], which focuses on explainability methods tailored for transformers. The package ferret includes only a few faithfulness measures and plausibility measures. The project involves studying other possible metrics suitable for the task of text classification. Once identified, the task of the project is to implement them and assess them.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing evaluation methods for explanation quality assessment in text classification.
- **Identification of Metrics.** Identify a set of metrics suitable for evaluating explanations for text classifiers not yet included in the *ferret* package.

- **Implementation.** Select and implement 2-3 promising metrics within the *ferret* package.
- **Evaluation.** Assess the effectiveness and applicability of the newly implemented metrics by comparing them across different attribution-based explanation methods.

References

[1] Julius Adebayo et al. "Sanity checks for saliency maps". In: vol. 31. 2018.

[2] Chirag Agarwal et al. "OpenXAI: Towards a Transparent Evaluation of Model Explanations". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 15784–15799. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/65398a0eba88c9b4a1c38ae405b125ef-Paper-Datasets_and_Benchmarks.pdf.

[3] Pepa Atanasova et al. "A Diagnostic Study of Explainability Techniques for Text Classification". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 3256–3274. DOI: 10.18653/v1/2020.emnlp-main.263. URL: https://aclanthology.org/2020.emnlp-main.263.

[4] Giuseppe Attanasio et al. "ferret: a Framework for Benchmarking Explainers on Transformers". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Danilo Croce and Luca Soldaini. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 256–266. DOI: 10.18653/v1/2023.eacl-demo.29. URL: https://aclanthology.org/2023.eacl-demo.29.

[5] Jay DeYoung et al. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://aclanthology.org/2020.acl-main.408.

[6] Yang Liu et al. "Synthetic Benchmarks for Scientific Research in Explainable Machine Learning". In: *Advances in Neural Information Processing Systems - Datasets and Benchmarks*. 2021.