Enhancing Trustworthiness in NLP Systems Through Comprehensive Explainability Approaches

Santiago González-Silot

Centro de Estudios Avanzados en TIC, Universidad de Jaén, Campus Las Lagunillas s/n, 23007, Jaén, Spain

Abstract

Natural Language Processing (NLP) systems have made significant strides in recent years, achieving remarkable success in various applications such as machine translation, sentiment analysis, and question answering. However, the black-box nature of many advanced NLP models raises concerns about their trustworthiness and reliability, especially in critical domains like healthcare, legal, and disinformation. This doctoral thesis addresses the need for enhancing trustworthiness in NLP systems by integrating explainability through three main approaches: Feature Importance Methods, Natural Language Generation (NLG) Explanations, and Probing Techniques. The research presented here aims to bridge the gap between complex NLP models and their end-users by developing and evaluating methods that provide transparent and interpretable insights throughout the Machine Learning production cycle: data acquisition, preprocessing, training, and inference. This doctoral thesis hypothesizes that achieving reliable, explainable, and unbiased language models through these three complementary approaches will lead to more human-friendly and usable Artificial Intelligence.

Keywords

LLM, Language Models, XAI, Trustworthy AI, Explainable AI, Interpretability

1. Justification of the proposed research

Since the introduction of Transformer-based models such as GPT and BERT, they have revolutionized most Natural Language Processing (NLP) tasks, such as machine translation, text summarization, and question answering among others. It is clear that Transformer-based models are the ones that obtain better results than others, even more so if we talk about Large Language Models (LLM), but due to their complex and non-linear structure, these learning models are often black-boxes that obtain results in a totally opaque way. This is a major problem, especially for the application of these models in sectors such as medicine, psychology, or social sciences which need high reliability, robustness, and safety. Unfortunately, as can be seen in Figure 1, most of the most widely used models have major reliability problems from several points of view [1].

All of this is aggravated if we take into account that research in Artificial Intelligence (AI) and more specifically in NLP has been marked by a SOTA-Chasing trend by the entire scientific community [2], which is more focused on obtaining better metrics or scores in a leaderboard of questionable relevance rather than obtaining real insights and their explanation. It would seem that machine learning has become so powerful (and opaque) that it is no longer important

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

sgs00034@red.ujaen.es (S. González-Silot)

0000-0001-8378-5840 (S. González-Silot)

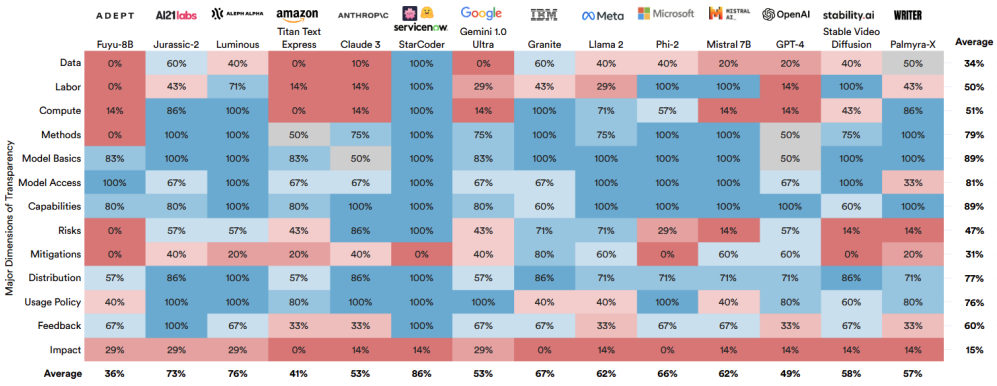

Figure 1: Foundation Models Transparency Index. Image from [1].

to ask how it works and why, but this is not really the case. The trustworthiness of Artificial Intelligence is key for it to have a good impact on society and the acceptance of users to use it correctly without fears and prejudices. For example, people are more open to use AI if they know how it works and why they make certain decisions [3].

If we do not know why the AI makes a decision, produces a response, or acts in a certain way, we will not know if that decision is really correct, since in many cases this AI response is highly subjective, variable, and multifactorial. Many papers [4, 5, 6] have shown that AI is plagued by biases of all kinds, e.g., gender, ethnicity, and religion, which are inherent in the data used for training and can condition it to make decisions that are dangerous to humans. That is, sometimes these biases come from humans themselves. The issue of opacity and limitations in explanations is determined by the typology of neural networks. For example, rule-based expert systems do achieve an acceptable ability to explain their decisions [7].

In addition, explainability is not only a goal to see why a model makes a decision and to see the model's behavior, it also serves to justify that decision and to help users to investigate uncertain or inconsistent predictions. For example, in my previous work [8], I applied SHAP and observed that the state-of-the-art models of fake news detection took into consideration spurious features and named entities, which is a violation of impartiality. Thanks to this application of explainability, I was able to develop a methodology of working to reduce biases in this task and make the model less biased, more robust to adversarial attacks, more generalizable, and generally more trustworthy. It is worth mentioning that a paper on the application of this methodology has been written and is currently under review in a journal.

Trustworthy AI has become increasingly crucial due to the growing landscape of regulations designed to ensure ethical, transparent, and accountable use of Artificial Intelligence, as can be seen in the document of ethics guidelines for trustworthy AI of the European Commision [9]. As governments and international bodies establish guidelines to protect individual rights and societal interests, AI researchers and organizations must prioritize trustworthiness to comply with these standards. Trustworthy AI not only helps in avoiding legal repercussions and financial penalties but also fosters public confidence and adoption of AI technologies. It encompasses principles such as fairness, privacy, security, robustness, and explainability, which

are essential to mitigate biases, prevent misuse, and promote transparency. Adhering to these regulations ensures that AI systems operate responsibly and equitably, reinforcing their positive impact on society while maintaining public trust and safeguarding against potential harm.

For these reasons, the objective of this doctoral thesis is to bridge the gap between black-box, biased, and opaque models to a more secure, transparent, unbiased, and generally more trustworthy Artificial Intelligence in the Natural Language Processing domain, focusing on three distinct yet complementary explainability approaches: Feature Importance Methods, Natural Language Generation Explanations, and Probing Techniques.

These three explainability approaches—Feature Importance Methods, NLG Explanations, and Probing Techniques—offer complementary perspectives on model behavior. While feature importance methods highlight influential input elements, NLG explanations provide accessible rationales, and probing reveals internal representations and linguistic capabilities. Together, they form a comprehensive toolkit for enhancing the trustworthiness and interpretability of NLP systems.

The remaining sections of this paper are organized as follows: Section 2 covers the background and related work of Trustworthy and Explanability in NLP; Section 3 the main hypothesis and objectives of the doctoral thesis; Section 4 the research methodology and experiments for this thesis; Section 5 the specific research elements proposed for discussion; Finally, Section 6 depicts the conclusions.

2. Background and related work

Trustworthy and explainable natural language processing (NLP) has become a critical area of research in recent years. With the increasing focus on ethical challenges within NLP, such as bias mitigation, identifying objectionable content, and enhancing system design and data handling practices [10], researchers have delved into various aspects to ensure trustworthy NLP models. Recent efforts have been made to enhance the trustworthiness of models through aspects like robustness, explainability, privacy, fairness, accountability, and environmental well-being [11].

The field of explainable NLP has evolved to encompass various methodologies and techniques aimed at enhancing model interpretability. Based on the categorization presented in [12], we identify three principal approaches to explainability in NLP: **Feature Importance Methods**, **Natural Language Generation (NLG) Explanations**, and **Probing Techniques**. Each of these approaches offers unique insights into model behavior and decision-making processes, contributing to the broader goal of trustworthy AI.

2.1. Feature Importance Methods

Feature importance methods focus on identifying and quantifying the contribution of input features to model predictions. These techniques aim to answer the question "*Which parts of the input were most influential for the model's decision*" by generating attribution scores for individual tokens, words, or phrases. Several prominent approaches have emerged in this category:

- **Gradient-based** methods such as Integrated Gradients [13] and SmoothGrad [14] utilize the gradient of the model output with respect to input features to determine feature importance. These methods provide fine-grained explanations but can be computationally intensive and may produce noisy attributions.
- **Perturbation-based** methods like LIME (Local Interpretable Model-agnostic Explanations) [15] and SHAP (SHapley Additive exPlanations) [16] observe changes in model predictions when input features are perturbed or removed. LIME approximates complex models locally with interpretable surrogates, while SHAP draws from cooperative game theory to assign contribution values to features.
- **Attention-based** interpretations leverage the attention mechanisms inherent in Transformer models, providing visualization of which parts of the input the model "focuses" on during prediction [17]. However, research by Serrano and Smith [18] has questioned whether attention weights directly translate to feature importance.

Recent research has explored how these methods can be adapted specifically for NLP tasks. For instance, Jin et al. [19] proposed hierarchical explanations for text classification that account for both word-level and phrase-level contributions. Similarly, Wallace et al. [20] introduced AllenNLP Interpret, which integrates various feature attribution methods for NLP models.

2.2. Natural Language Generation (NLG) Explanations

Natural Language Generation approaches produce textual explanations that describe the reasoning process or decision factors of a model. These explanations are often more accessible to non-technical users compared to numerical scores or visualizations. The key advantage of NLG explanations is their ability to communicate complex decision processes in a familiar format—natural language.

Self-explanatory models incorporate explanation generation as an intrinsic component of their architecture. Models like ExplanationLP [21] and CoS-E [22] are trained to generate both predictions and explanations simultaneously. These approaches often use multitask learning frameworks where explanation generation is an auxiliary task alongside the primary NLP task.

Post-hoc explanation generators, on the other hand, produce explanations after the model has made its prediction. Such systems may be trained on human-authored explanations to mimic human reasoning patterns [23], or they may utilize large language models to generate plausible rationales for predictions [24].

Rationalization techniques aim to extract segments of the input text that justify the model's prediction [25, 26]. These methods typically employ selective or extractive approaches to identify crucial portions of the input that influence the output decision.

Recent advancements in this area include the development of faithfulness metrics to evaluate how accurately natural language explanations reflect the model's true decision process [27]. Additionally, researchers have explored generating contrastive explanations that highlight why one prediction was made over another potential outcome [28].

2.3. Probing Techniques

Probing techniques, also known as diagnostic classifiers or linguistic probing, investigate what linguistic properties or structures are captured by different components of a model. These methods help researchers understand the internal representations learned by NLP models and analyze what information is encoded at different processing stages.

Structural probing methods examine how well models capture syntactic and hierarchical linguistic structures. For instance, [29] demonstrated that BERT's representations encode parse tree distances, suggesting the model implicitly learns syntactic information during pretraining.

Semantic probing assesses a model's understanding of meaning-related properties. This includes probing for semantic roles, lexical relations, entity types, and compositional semantics. [30] showed how different layers in BERT capture different levels of linguistic information, from surface features in early layers to semantic information in later layers.

Behavioral probing examines how models respond to specific challenges or manipulations of the input [31]. Methods like CheckList [31] provide a framework for testing specific linguistic capabilities through carefully crafted test cases.

Advanced probing techniques include controlled interventions [32], where specific neurons or attention heads are manipulated to observe their impact on model behavior, and cross-architectural comparisons [33], which analyze how different model architectures represent similar linguistic phenomena.

3. Main Hypothesis and Objectives

3.1. Main Hypothesis

The hypothesis behind this line of research is that if we develop explainable, interpretable, and less-biased models, we can create a more Trustworthy AI which is more usable, human-friendly, and responsible.

This doctoral thesis aims to bridge the gap between black-box, biased, and opaque models to a more secure, transparent, unbiased, robust, and generally more trustworthy Artificial Intelligence in the Natural Language Processing domain.

3.2. Objectives

1. Analyze the state of the art of Explainability and Trustworthiness in AI and specifically in NLP
2. Analyze the possible regulations that exist and will exist in AI to adapt the line of research and application to these regulations.
3. Develop and evaluate Feature Importance Methods for NLP models that provide transparent insights into which input features influence model predictions, with particular focus on addressing biases and improving model robustness.
4. Design and implement Natural Language Generation approaches that produce accessible and faithful explanations of model behavior, enabling users to understand model decisions in human-readable format.

5. Apply and extend Probing Techniques to systematically investigate what linguistic properties are captured by NLP models and how these representations relate to model performance and biases.
6. Design of an evaluation framework that takes into account the different perspectives of trustworthiness, comparing and integrating insights from all three explainability approaches to provide a comprehensive understanding of model behavior.
7. Create a methodology for applying appropriate explainability techniques based on specific domain requirements and user needs, particularly for sensitive applications such as fake news detection, medical text analysis, and legal document processing.

4. Research Methodology and Proposed Experiments

To achieve the objectives and validate the hypothesis, the research will proceed in four stages:

1. **Analysis of relevant literature sources:** To achieve the objectives of the thesis, an exhaustive analysis of relevant sources has to be performed. This includes the review of scientific literature related to language models, explainability techniques (Feature Importance Methods, NLG Explanations, and Probing), trustworthiness, and the methodologies that may approach a more Trustworthy AI.
2. **Experimental design:** Development of techniques and methodologies across the three explainability approaches to bring language models closer to a more reliable AI. The experimental design includes:
    - Application of Feature Importance Methods to identify influential features in model decisions and detect biases
    - Development of Natural Language Generation techniques for explaining model decisions in human-readable format
    - Implementation of Probing methods to understand what linguistic information is encoded in model representations
3. **Trustworthy Data Creation and Curation:** Development of datasets specifically designed to drive explainable behavior of language models and to evaluate the effectiveness of different explainability approaches. Additionally, data preprocessing techniques will be developed to ensure privacy and unbiasedness throughout the data lifecycle.
4. **Evaluation of results:** Application and development of different evaluation metrics that measure how reliable an AI model is across different aspects of trustworthiness (absence of biases, robustness, interpretability, etc.). The evaluation will focus on comparing the insights gained from each explainability approach and assessing their complementarity.

5. Research Elements for Discussión

In a field as broad and incipient as trustworthy AI, there is a discussion on a wide range of issues, but in particular, I show below the 3 elements of the discussion that I am debating in the current state of the doctoral thesis.

1. **Integration of Multiple Explainability Approaches** While each explainability approach (Feature Importance, NLG Explanations, and Probing) offers valuable insights, how can we effectively integrate these diverse perspectives into a cohesive understanding of model behavior? Do these approaches sometimes provide contradictory explanations, and if so, how should such contradictions be resolved? Furthermore, how can we determine which explainability method is most appropriate for specific stakeholders, domains, or use cases?
2. **Evaluation Techniques for Measuring the Quality of an Explanation**: A model's quality should be evaluated not only by its accuracy and performance but also by how well it provides explanations for its predictions [12]. Should we use Informal Examination, Comparison to Ground Truth or Human Evaluation? What are the advantages and disadvantages of using metrics such as BLEU [34], ROUGE [35], or Perplexity? Can we rely on what is relevant to attention mechanisms? [36].
3. **Effective evaluation of the degree of bias of a language model**. The degree of trustworthiness of a language model depends on several factors such as its robustness, interpretability, or absence of bias among others. How can we effectively measure the degree of bias of a language model? How can we know if there is a real bias in the model output? How can we identify from which part of the model development cycle the bias comes?

6. Conclusions

This paper has outlined the initial phase of my doctoral research, which focuses on creating models that are more explainable, interpretable, and fair, with the goal of narrowing the gap between opaque black-box systems and the principles of Trustworthy Artificial Intelligence within the field of Natural Language Processing.

For this purpose, the state of the art has been analyzed, the objectives to be achieved have been presented, the methodology to achieve them has been described, and finally, different elements for discussion have been introduced.

Declaration on Generative AI

Declaration on Generative AI. During the preparation of this work, the author used DeepL and, occasionally, ChatGPT in order to check grammar, spelling, and translation. After using these tools, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Bommasani, K. Klyman, S. Kapoor, S. Longpre, B. Xiong, N. Maslej, P. Liang, The foundation model transparency index v1. 1: May 2024, arXiv preprint arXiv:2407.12929 (2024).

[2] K. W. Church, V. Kordoni, Emerging trends: Sota-chasing, Natural Language Engineering 28 (2022) 249–269. doi:10.1017/S1351324922000043.
[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.
[4] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, CoRR abs/1607.06520 (2016). URL: http://arxiv.org/abs/1607.06520. arXiv:1607.06520.
[5] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences 115 (2018) E3635–E3644.
[6] I. Garrido-Muñoz, F. Martínez-Santiago, A. Montejo-Ráez, Maria and beto are sexist: evaluating gender bias in large language models for spanish, Language Resources and Evaluation (2023) 1–31.
[7] C. Yáñez-Márquez, Toward the bleaching of the black boxes: minimalist machine learning, IT Professional 22 (2020) 51–56.
[8] S. González-Silot, Procesamiento de Lenguaje Natural Explicable para Análisis de Desinformación, Master's thesis, Universidad de Granada, 2023.
[9] European-Commision, Ethics guidelines for trustworthy AI., Technical Report, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[10] S. Prabhumoye, B. Boldt, R. Salakhutdinov, A. W. Black, Case study: Deontological ethics in NLP, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3784–3798. URL: https://aclanthology.org/2021.naacl-main.297/. doi:10.18653/v1/2021.naacl-main.297.
[11] H. Zhang, B. Y. Wu, X. Yuan, S. Pan, H. Tong, J. Pei, Trustworthy graph neural networks: Aspects, methods and trends (2022). doi:10.48550/arxiv.2205.07424.
[12] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A Survey of the State of Explainable AI for Natural Language Processing, in: K.-F. Wong, K. Knight, H. Wu (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.
[13] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, CoRR abs/1703.01365 (2017). URL: http://arxiv.org/abs/1703.01365. arXiv:1703.01365.
[14] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, arXiv preprint arXiv:1706.03825 (2017).
[15] M. T. Ribeiro, S. Singh, C. Guestrin, " 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, 2016, pp. 1135–1144.
[16] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances

in neural information processing systems 30 (2017).
[17] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does bert look at? an analysis of bert's attention, arXiv preprint arXiv:1906.04341 (2019).
[18] S. Serrano, N. A. Smith, Is attention interpretable?, arXiv preprint arXiv:1906.03731 (2019).
[19] X. Jin, Z. Wei, J. Du, X. Xue, X. Ren, Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models, arXiv preprint arXiv:1911.06194 (2019).
[20] E. Wallace, J. Tuyls, J. Wang, S. Subramanian, M. Gardner, S. Singh, Allennlp interpret: A framework for explaining predictions of nlp models, arXiv preprint arXiv:1909.09251 (2019).
[21] B. Hancock, M. Bringmann, P. Varma, P. Liang, S. Wang, C. Ré, Training classifiers with natural language explanations, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2018, 2018, p. 1884.
[22] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).
[23] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).
[24] S. Wiegreffe, A. Marasović, N. A. Smith, Measuring association between labels and free-text rationales, arXiv preprint arXiv:2010.12762 (2020).
[25] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, arXiv preprint arXiv:1606.04155 (2016).
[26] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, L. Freeman, F. Batarseh, Rationalization for explainable nlp: A survey. arxiv 2023, arXiv preprint arXiv:2301.08912 (????).
[27] A. Jacovi, Y. Goldberg, Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, arXiv preprint arXiv:2004.03685 (2020).
[28] T. Wu, M. T. Ribeiro, J. Heer, D. S. Weld, Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models, arXiv preprint arXiv:2101.00288 (2021).
[29] J. Hewitt, C. D. Manning, A structural probe for finding syntax in word representations, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4129–4138.
[30] I. Tenney, D. Das, E. Pavlick, Bert rediscovers the classical nlp pipeline, arXiv preprint arXiv:1905.05950 (2019).
[31] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of nlp models with checklist, arXiv preprint arXiv:2005.04118 (2020).
[32] R. Rudinger, A. Teichert, R. Culkin, S. Zhang, B. Van Durme, Neural-davidsonian semantic proto-role labeling, arXiv preprint arXiv:1804.07976 (2018).
[33] Z. Wu, Y. Chen, B. Kao, Q. Liu, Perturbed masking: Parameter-free probing for analyzing and interpreting bert, arXiv preprint arXiv:2004.14786 (2020).
[34] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[35] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization

branches out, 2004, pp. 74–81.
[36] S. Serrano, N. A. Smith, Is attention interpretable?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951. URL: https://aclanthology.org/P19-1282. doi:10.18653/v1/P19-1282.