Enhancing Trustworthiness in NLP Systems Through Explainability

Santiago González-Silot[1]

[1]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 imperative need for enhancing trustworthiness in NLP systems by integrating explainability mechanisms. 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 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

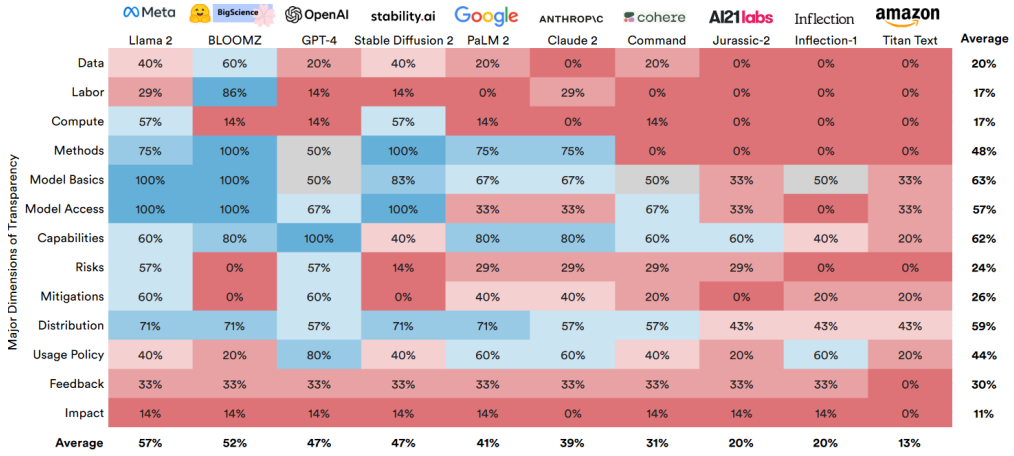
Since the introduction of Transformer-based models such as GPT and BERT, they have revolutionized most Natural Language Processing (NLP) tasks, such as machine translation, text summarization, and question answering among others. It is clear that Transformer-based models are the ones that obtain better 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 to ask how it works and why, but this is not really the case. The trustworthiness of Artificial

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

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

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



Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

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

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.

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 [7], 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 will be submitted in July.

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 Figure 2 from the document of ethics guidelines for trustworthy AI of the European Commision [8]. 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.

Figure 2: Trustworthy AI, key principles. Image from European Commision [8].

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.

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 specifics 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 [9], researchers have explored into various aspects to ensure trustworthy NLP models. For example, recent efforts have been made to enhance the trustworthiness of Graph Neural Networks (GNNs) through aspects like robustness, explainability, privacy, fairness, accountability, and environmental well-being [10]. Explainability in NLP has been a key focus, with research highlighting the importance of interpretability and the application of explainable AI (XAI) techniques to enhance understanding and trust in NLP models [11].

Model explainability in the Large Language Model's era of NLP has been a subject of interest, with discussions on how explainability analysis can help detect issues unique to NLP models post-training [12]. Additionally, rationalization for explainable NLP has been emphasized, recognizing the challenges faced in achieving explainability despite the practical applications of NLP [13]. Techniques such as feature attribution methods have been employed to visualize and understand the reasoning capabilities of NLP models, aiding experts in comprehending model outputs [14].

Moreover, the application of NLP in various domains, including social media data mining and knowledge discovery, has highlighted the significance of NLP in enabling machines to understand and generate meaningful content from diverse sources [15][16]. The development of ethical NLP models and the machine learning of ethical judgments from natural language have been explored, shedding light on the assumptions and methodologies involved in generating moral judgments through NLP [17]. Furthermore, in the healthcare domain, the use of NLP in structured real-world data for pharmacovigilance purposes has been investigated, emphasizing the importance of trustworthy AI criteria in AI model implementation due to the sensitivity of the task domain [18].

In conclusion, the synthesis of research in trustworthy and explainable NLP underscores the multidimensional efforts to enhance the reliability, interpretability, and ethical considerations within NLP models. By addressing issues such as bias, model explainability, and ethical judgments, researchers aim to advance the field of NLP towards more transparent and trustworthy applications and putting the focus on a key aspect for the application of AI in society, which is sometimes not given enough importance.

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. Analysis and development of datasets and pre-training methodologies that avoid the potential problems of inherent human language biases.
4. Creation of unbiased, more transparent, and explainable language models for a variety of particularly sensitive tasks such as fake news detection or sentiment analysis.
5. Design of frameworks to adapt current state-of-the-art models to the reliability needs required by their specific application domains.
6. Design of an evaluation framework that takes into account the different perspectives of trustworthiness, and in particular the level of explainability, unbiasedness, and robustness.

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 objective 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, trustworthiness, and the methodologies related to these concepts that may approach a more Trustworthy AI.
2. **Experimental design**: Development of techniques and methodologies to bring language models that act as black boxes closer to a more reliable AI. For this purpose, experiments will be carried out applying different NLP and Explainability techniques to obtain key insights that can guide the development of a Trustworthy AI.
3. **Trustworthy Data Creation and Curation**: Trustworthy data creation and curation in Natural Language Processing (NLP) is critical to ensuring the accuracy and reliability of information extracted from textual data and plays a key role in extracting valuable insights from unstructured text data. For this reason, datasets will be made to drive the explainable behavior of the language models. Additionally, data preprocessing techniques will also 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 in the different aspects involved (absence of biases, robustness, interpretability, etc). For this purpose, the key aspects and weaknesses of the current metrics will be analyzed to achieve a correct evaluation.

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. **Data Collection for Trustworthy AI**: The basis of Machine Learning is inductive learning, i.e., models learn from data. That is why one of the key points to make AI more trustworthy is to make both the datasets used and the collection process meet the necessary requirements for it. How can we collect data without biases of any kind? How do we evaluate if a dataset is biased?
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 [19]. Should we use Informal Examination, Comparison to Ground Truth or Human Evaluation? What are the advantages and disadvantages of using metrics such as BLEU [20], ROUGE [21], or Perplexity? Can we rely on what is relevant to attention mechanisms? [22]
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 shown the first steps of my doctoral thesis aimed at developing more explainable, interpretable, and unbiased models to bridge the gap between black-box models and Trustworthy Artificial Intelligence in the domain 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 put on the table.

References

[1] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, P. Liang, The foundation model transparency index, 2023. arXiv:2310.12941.

[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] S. González-Silot, Procesamiento de Lenguaje Natural Explicable para Análisis de Desinformación, Master's thesis, Universidad de Granada, 2023.

[8] European-Commision, Ethics guidelines for trustworthy AI., Technical Report, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[9] S. Prabhumoye, B. Boldt, R. Salakhutdinov, A. W. Black, Case study: Deontological ethics in nlp (2021). doi:10.18653/v1/2021.naacl-main.297.

[10] 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.

[11] M. Danilevsky, S. Dhanorkar, Y. Li, L. Popa, K. Qian, A. Xu, Explainability for natural language processing (2021). doi:10.1145/3447548.3470808.

[12] S. Gholizadeh, N. Zhou, Model explainability in deep learning based natural language processing (2021). doi:10.48550/arxiv.2106.07410.

[13] S. Gurrapu, Rationalization for explainable nlp: A survey, Frontiers in Artificial Intelligence (2023). doi:10.3389/frai.2023.1225093.

[14] X. Wang, R. Huang, Z. Jin, T. Fang, H. Qu, Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models, IEEE Transactions on Visualization and Computer Graphics (2023) 1–11. URL: http://dx.doi.org/10.1109/TVCG.2023.3327153. doi:10.1109/tvcg.2023.3327153.

[15] A. BOUCHEHAM, Natural language processing for social media data mining (2023). doi:10.47832/alanyacongress2-8.

[16] F. G. VanGessel, Natural language processing for knowledge discovery and information extraction from energetics corpora, Propellants Explosives Pyrotechnics (2023). doi:10.1002/prep.202300109.

[17] Z. Talat, H. Blix, J. Valvoda, M. I. Ganesh, R. Cotterell, A. Williams, On the machine learning of ethical judgments from natural language (2022). doi:10.18653/v1/2022.naacl-main.56.

[18] S. Dimitsaki, Applying artificial intelligence in structured real-world data for pharmacovigilance purposes: A systematic literature review (preprint) (2024). doi:10.2196/preprints.57824.

[19] 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.

[20] 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.

[21] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[22] 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.