HiTZ@Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine

Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, Iker Garcia-Ferrero, Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau and Anar Yeginbergenova

HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain

Abstract

Providing high quality explanations for AI predictions based on machine learning is a challenging and complex task. To work well it requires, among other factors: selecting a proper level of generality/specificity of the explanation; considering assumptions about the familiarity of the explanation beneficiary with the AI task under consideration; referring to specific elements that have contributed to the decision; making use of additional knowledge (e.g. expert evidence) which might not be part of the prediction process; and providing evidence supporting negative hypothesis. Finally, the system needs to formulate the explanation in a clearly interpretable, and possibly convincing, way. Given these considerations, ANTIDOTE fosters an integrated vision of explainable AI, where low-level characteristics of the deep learning process are combined with higher level schemes proper of the human argumentation capacity. ANTIDOTE will exploit cross-disciplinary competences in deep learning and argumentation to support a broader and innovative view of explainable AI, where the need for high-quality explanations for clinical cases deliberation is critical. As a first result of the project, we publish the *Antidote CasiMedicos* dataset to facilitate research on explainable AI in general, and argumentation in the medical domain in particular.

Keywords

Explainable AI, Digital Medicine, Question Answering, Argumentation, Natural Language Processing

1. Introduction

ANTIDOTE[1] is a European CHIST-ERA project where each partner is funded by their national Science Agencies. As the Spanish partner in the Consortium is the HiTZ Center - Ixa, from the University of the Basque Country UPV/EHU, the project was funded by the *Proyectos de Colaboración Internacional* (PCI 2020) program of the Spanish Ministry of Science and Innovation. The other European partners are the following: Université Côte d'Azur (UCA) from France and coordinators of the international consortium, Fondazione Bruno Kessler (FBK) from Italy, KU Leuven/Computer Science, in Belgium and Universidade Nova de Lisboa (NOVA) in Portugal.

The aim of ANTIDOTE is to exploit cross-disciplinary competences in three areas, namely, deep learning, argumentation and interactivity, to support a broader and innovative view of explainable AI.

Providing high quality explanations for AI predictions based on machine learning is a challenging and complex task. To work well it requires, among other aspects: (i) selecting a proper level of generality/specificity of the explanation, (ii) considering assumptions about the familiarity of the explanation beneficiary with the AI task

under consideration, (iii) referring to specific elements that have contributed to the decision, (iv) making use of additional knowledge (e.g. metadata) which might not be part of the prediction process, (v) selecting appropriate examples and, (vi) providing evidence supporting negative hypotheses. Finally, the system needs to formulate the explanation in a clearly interpretable, and possibly convincing, way.

Taking into account these considerations, ANTIDOTE fosters an integrated vision of Explainable AI (XAI), where the low-level characteristics of the deep learning process are combined with higher level schemes proper of human argumentation. Following this, the ANTIDOTE integrated vision is supported by three considerations. First, in neural architectures the correlation between internal states of the network (e.g., weights assumed by single nodes) and the justification of the network classification outcome is not well studied. Second, high quality explanations are crucially based on argumentation mechanisms (e.g., provide supporting examples and rejected alternatives). Finally, in real settings, providing explanations is inherently an interactive process involving the system and the user.

Thus, ANTIDOTE will exploit cross-disciplinary competences in three areas, namely, deep learning, argumentation and interactivity, to support a broader and innovative view of explainable AI. There are several research challenges that ANTIDOTE will address to advance the state-of-the-art in explainable AI.

The first challenge is to take advantage of the huge

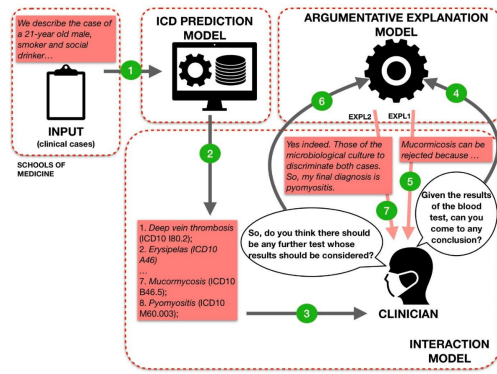[1]https://univ-cotedazur.eu/antidote

Figure 1: ANTIDOTE use-case scenario.

body of past research on argumentation to complement state-of-the-art approaches on explainability. In addition, the recent resurgence of AI highlights the idea that low-level system behavior not only needs to be interpretable (e.g., showing those elements that most contributed to the system decision), but that also needs to be joined by high level human argumentation schemes. The second challenge is to automatically learn explanatory argumentation schemas in Natural Language (NL) and to effectively combine evidence-based decision making with high level explanations. The third challenge for ANTIDOTE is that a task-specific prediction model and a general argumentation model need to be combined to produce explanatory argumentations.

While neural networks for medical diagnosis have become exceedingly accurate in many areas, their ability to explain how they achieve their outcome remains problematic. Herein lies the main novelty of the ANTIDOTE project: it focuses on elaborating argumentative explanations to diagnosis predictions in order to assist student clinicians to learn making informed decisions.

The explanatory argumentative scenario envisaged by ANTIDOTE will involve a student clinician, who will need to hypothesize about the clinical case of a patient and will have to provide argumentative explanations about them. The focus of the experimental setting is set on the capacity of the ANTIDOTE Explanatory AI system to provide correct predictions and consistent arguments, without forgetting also the linguistic quality of the dialogues (e.g., naturalness of the utterances, etc.).

In our scenario depicted in Figure 1, the clinician queries the ANTIDOTE XAI for explanations (arguments) on its diagnosis of the clinical case. The ANTIDOTE XAI provides hypotheses (differential diagnosis) about the clinical case, as well as arguments to support its prediction and arguments discarding alternative predictions.

The student clinician has the possibility to take the initiative to ask additional questions and clarifications. The goal of the explanatory argumentation in a differential diagnosis is to validate the correctness of the diagnosis and the ANTIDOTE XAI capacity to argue in favour of the correct hypothesis and to counter-argue against alternative hypotheses.

2. Related Work

In this section we review the most relevant previous work focusing on argumentation and explainable AI for the medical domain.

2.1. Argumentation mining and generation

Argumentation mining is a research area that moves between natural language processing, argumentation theory and information retrieval. The aim of argumentation mining is to automatically detect the argumentation of a document and its structure. This implies the detection of all the arguments involved in the argumentation process, their individual or local structure (rhetorical or argumentative relationships between their propositions), and the interactions between them, namely, the global argumentation structure.

Argumentation mining in Natural Language Processing has been applied to various domains such as persuasive essays, legal documents, political debates and social media data [1]. For instance, Stab and Gurevych [2] built an annotated dataset of persuasive essays with corresponding argument components and relations. Using this corpus, Eger et al. [3] developed an end-to-end neural method for argument structure identification. Furthermore, Nguyen and Litman [4] also applied an end-to-end method to parse argument structure and used the argument structure features to improve automated persuasive essay scoring. Other approaches studied context-dependent claim detection by collecting annotations for Wikipedia articles [5]. Using this corpus, the task of automatically identifying the corresponding pieces of evidence given a claim has also been investigated [6].

Argumentation generation remains a research area in which there is still a long way to go. Recent work has made progress towards this goal through the automated generation of argumentative text [7, 8, 9, 10]. Thus, Alshomary et al. [11] proposed a Bayesian argument generation system to generate arguments given the corresponding argumentation strategies. Sato et al. [9] presented a sentence-retrieval-based end-to-end argument generation system that can participate in English debating games.
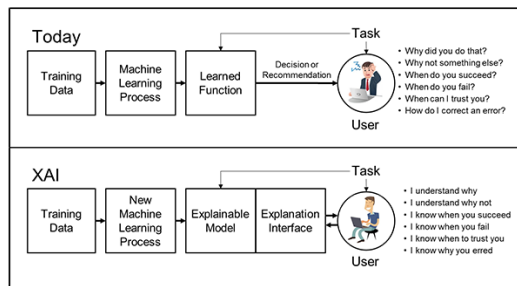
Figure 2: Explainable AI with Human in the Loop.

There have also been some works exploring counter-argument generation to select the main talking points to generate a counter-argument [12]. In this line of research, Hidey and McKeown [13] proposed a neural model that edited the original claim semantically to produce a claim with an opposing stance. They also incorporated external knowledge into the encoder-decoder architecture showing that their model generated arguments that were more likely to be on topic.

Finally, an autonomous debating system (Project Debater) able to engage in competitive debates with humans was developed. The system consisted of a pipeline of four main modules: argument mining, an argument knowledge base, argument rebuttal, and debate construction [14].

2.2. Explainable AI

Explainable artificial intelligence (XAI) aims to address the needs of users wanting to understand how a program's artificial intelligence works and how to evaluate the results obtained. Otherwise, there is no basis for real confidence in the work of the AI system, as illustrated by Figure 2[2] The transparency offered by explainable AI is therefore essential for the acceptance of artificial intelligence.

There has been a surge of interest in explainable artificial intelligence (XAI) in recent years. This has produced a myriad of algorithmic and mathematical methods to explain the inner workings of machine learning models [15]. However, despite their mathematical rigor, these works suffer from a lack of usability and practical interpretability for real users. Although the concepts of interpretability and explainability are hard to rigorously define, multiple attempts have been made towards that goal [16, 17].

Adadi and Berranda [18] presented an extensive literature review, collecting and analyzing 381 different scientific papers between 2004 and 2018. They arranged all of the scientific work in the field of explainable AI along four main axes and stressed the need for more formalism to be introduced in the field of XAI and for more interaction between humans and machines.

In a more recent study [19] introduced a different type of arrangement that initially distinguishes transparent and post-hoc methods and subsequently created sub-categories.

Taking into account argumentation principles, ANTIDOTE will explain machine decisions based on four modes of explanations to be auditable by humans: (i) analytic statements in NL that describe the elements and context that support a choice, (ii) visualizations that highlight portions of the raw data that support a choice, (iii) cases that invoke specific examples, and (iv) rejections of alternative choices that argue against less preferred answers based on analytics, cases, and data.

3. Methology and Work Plan

The main scientific challenge for the project is the combination of three models depicted in Figure 1: (1) The Prediction Model has to predict appropriate International Classification of Diseases (ICD) codes given a clinical case; (2) the Argumentative Model selects proper arguments (i.e., entity and relations) to support or attack a given topic. It may use both information included in the clinical cases used by the prediction model and additional sources of knowledge; (3) the Interaction Model provides argumentative explanations about a certain prediction. An integrated approach is proposed to both predict the outcome of a clinical course of action and justify a medical diagnosis by a language model. A starting point will be using current large language models [20] to generate appropriate explanations guided by the activated view on a textual snippet that contributed to the decision, namely, the argument for the decision.

3.1. Work Plan

The Work Plan is structured in six Work Packages of which three are focused on the scientific contributions of the project.

WP2: Methodology and Design (Leader: FBK). Participants: UCA, UPV/EHU, KU, NOVA. The purpose of WP2 is to define, adapt and integrate the modules, resources, data structures, data formats and module APIs of the ANTIDOTE architecture. This includes designing the experiments, datasets, standard protocols, information flow and main architecture of ANTIDOTE.

WP3: Machine Learning (ML) for predicting clinical outcomes (Leader: KU). Participants: UPV/EHU, FBK, UCA, NOVA. WP3 targets (1) the development of a multitask learning model to jointly predict and justify a medical diagnosis by a deep learning model; (2) surfacing and making explicit the underlying aspects (identification of the most relevant/informative terms, identification of relations among terms) driving neural network decisions during the diagnosis prediction process; (3) retrieve external information to support the explanation.

WP4: Explanatory arguments in natural language (Leader: UCA). Participants: UPV/EHU, FBK, UCA, NOVA. WP4 relies on the textual arguments that form the basis for the decisions generated in WP3. WP4 targets (1) the definition and analysis of explanatory argumentative patterns to be used to construct natural language explanatory arguments of predictions; (2) the creation of a resource of annotated natural language explanatory arguments; (3) the development of explanatory arguments in natural language by mining and collecting them from trusted textual resources in the medical domain.

WP5: Evaluation (use cases in healthcare) (Leader: UPV/EHU). Participants: KU, UCA, FBK, NOVA. WP5 aims to (1) evaluate the effectiveness and quality of the prediction and the plausible alternatives (2) the quality of the generated explanatory arguments regarding the supporting evidence found in the clinical case in favor of the prediction and the positive or negative evidence found to discard other plausible alternatives, (3) the intrinsic quality of the generated arguments.

3.2. Evaluation

The generation of arguments will be quantitatively evaluated by computing metrics used in text generation to measure their overlap with ground truth arguments [21]. Moreover, the argumentative model will be evaluated following the criteria of coherence, simplicity, and generality [22]: explanations with structural simplicity, coherence, or minimality are preferred. With respect to argument mining, standard metrics such as F1 and accuracy will be used.

Generation of explanatory arguments will be also qualitative evaluated by medical students. Given the objectives and context of the project, ANTIDOTE will be based on previous work by Johnson [23], whereby the arguments will be evaluated for their (informal) inferential structure in terms of acceptability, relevance, and sufficiency of reasons provided, as well as their answerability to human agents' doubts and objections.

4. Ongoing Work

There are a number of tasks currently being undertaken within the project. In this section we provide details of the most central ones with respect to the objectives and motivation provided in the introduction.

4.1. ANTIDOTE Datasets

In order to carry out the tasks related to the main use-case presented in Figure 1, we need to identify, collect and annotate the most suitable corpus with which to train different models. In this regard, we have identified two possible data sources that will help us meet our objectives and that will constitute an important contribution of the ANTIDOTE project: SAEI and CasiMedicos.

The SAEI Corpus is a collection of differential diagnosis in Spanish collected by *La Sociedad Andaluza de Enfermedades Infecciosas* (The Andalusian Society of Infectious Diseases)[3]. This society is a non-profit association formed almost entirely by physicians specializing in Internal Medicine with special dedication to the management of infectious diseases, whose general purpose is the promotion and development of this medical discipline (training, care and research). We have selected the books that are of interest to us in order to carry out our objectives, namely, those that include clinical cases of infectious diseases for residents that are available, for the years 2011, 2015, 2016, 2017 and 2020. Among all these books we have extracted cleaned and pre-processed a total of 244 clinical cases with differential diagnosis.

CasiMedicos is a community and collaborative medical project run by volunteer medical doctors[4]. Among all the information created and made publicly available by this collaborative project, we have identified as an adequate data source the MIR exams commented by voluntary medical doctors with the aim of providing answers and explanations[5] to the MIR exams annually published by the Spanish Ministry of Health. In this data source we have extracted and pre-processed 622 commented questions from the MIR exams held between the years 2005, 2014, 2016, 2018, 2019, 2020, 2021 and 2022. The cleaned corpus, named the *Antidote Casimedicos* dataset, is publicly available to encourage research on explainable AI in the medical domain in general, and argumentation in particular[6].

Unlike popular Question Anwering (QA) datasets for English based on medical exams [24], both SAEI and CasiMedicos include not only the explanations for the

---

[3]https://www.saei.org
[4]https://www.casimedicos.com/
[5]https://www.casimedicos.com/mir-2-0/
[6]https://github.com/ixa-ehu/antidote-casimedicos

corrent answer (diagnosis or treatment), but also explanatory arguments written by medical doctors explaining why the rest of the possible answers are incorrect.

After pre-processing, these datasets have been translated from Spanish to English with the objective of starting various annotation tasks at various levels of complexity: (i) linking the explanatory sequences with respect to each possible answer; (ii) labeling of hierarchical argumentative structures; (iii) discourse markers. The resulting corpus will be the first corpus (multilingual or otherwise) with this type of annotations for the medical domain. Whenever ready, the corpus will be distributed under a free license to promote further research and to ensure reproducibility or results.

4.2. Question Answering in the Medical Domain

While there are several QA datasets for English based on medical exams [24], none of the previously published works contain two features which are unique of both SAEI and CasiMedicos: (i) the presence of explanations for both correct and incorrect answers; (ii) an argumentative structure arguing and counter-arguing about the possible answers. These features make it possible to define new Question Answering tasks, both from an extractive and generative point of view. In extractive QA, the objective would consist of identifying, in a given context, the explanation to the correct answer. In terms of generative QA, it will also allow us to leverage large language models [20] to learn generating the explanatory arguments with respect to both correct and incorrect possible answers.

4.3. Crosslingual Knowledge Transfer

The only corpus annotated with argumentative structure currently available for the medical domain is the AbstRCT dataset, which consists of English clinical trials [25]. In order to investigate the different strategies of transferring knowledge from English to other languages, especially those applying model- and data-transfer techniques previously discussed for other application domains [26], ongoing work is focused on adapting such knowledge transfer techniques for argumentation in the medical domain. As a result, we are undertaking novel experimental work on argument mining in Spanish for the medical domain [27]. This also involves the generation of the first Spanish dataset annotated with argumentative structures for the medical domain. Finally, the plan is to apply the developed technique to other languages of interest for the ANTIDOTE project (French and Italian).

In this line of research, and taking as starting point the ongoing work mentioned in the previous section, we plan to investigate also crosslingual and multilingual approaches to Question Answering techniques in the medical domain.

5. Concluding Remarks

In this paper we provide a description of the ANTIDOTE project, mostly focusing on identifying and generating high-quality argumentative explanations for AI predictions in the medical domain. So far, ongoing work has been focused on dataset collection and annotation and novel experimental work on Question Answering and Crosslingual Argument Mining. This work has leveraged multilingual encoder and decoder large language models [28, 20] for both extractive and generative experimentation.

Still, providing high-quality explanations for AI predictions based on machine learning is a challenging and complex task [24]. To work well, it requires, among other factors, making use of additional knowledge (e.g. medical evidence) which might not be part of the prediction process, and providing evidence supporting negative hypotheses. With these issues in mind, ANTIDOTE aims to address the challenge of providing an integrated vision of explainable AI, where low-level characteristics of the deep learning process are combined with higher level schemes proper of the human argumentation capacity. In order to do so, ANTIDOTE will be focused on a number of deep learning tasks for the medical domain, where the need for high quality explanations for clinical cases deliberation is critical.

References

[1] M. Dusmanu, E. Cabrio, S. Villata, Argument mining on twitter: Arguments, facts and sources, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2317−2322.

[2] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics 43 (2017) 619–659.

[3] S. Eger, J. Daxenberger, I. Gurevych, Neural end-to-end learning for computational argumentation mining, arXiv preprint arXiv:1704.06104 (2017).

[4] H. Nguyen, D. Litman, Argument mining for improving the automated scoring of persuasive essays, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[5] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, N. Slonim, Context dependent claim detection, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1489–1500.

[6] R. Rinott, L. Dankin, C. Alzate, M. M. Khapra, E. Aharoni, N. Slonim, Show me your evidence-an automatic method for context dependent evidence detection, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 440–450.

[7] R. Bar-Haim, L. Eden, R. Friedman, Y. Kantor, D. Lahav, N. Slonim, From arguments to key points: Towards automatic argument summarization, arXiv preprint arXiv:2005.01619 (2020).

[8] X. Hua, L. Wang, Neural argument generation augmented with externally retrieved evidence, arXiv preprint arXiv:1805.10254 (2018).

[9] M. Sato, K. Yanai, T. Miyoshi, T. Yanase, M. Iwayama, Q. Sun, Y. Niwa, End-to-end argument generation system in debating, in: Proceedings of ACL-IJCNLP 2015 System Demonstrations, 2015, pp. 109–114.

[10] M. Alshomary, S. Syed, A. Dhar, M. Potthast, H. Wachsmuth, Argument Undermining: Counter-Argument Generation by Attacking Weak Premises, arXiv preprint arXiv:2105.11752 (2021).

[11] M. Alshomary, S. Syed, M. Potthast, H. Wachsmuth, Target inference in argument conclusion generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4334–4345.

[12] X. Hua, Z. Hu, L. Wang, Argument generation with retrieval, planning, and realization, arXiv preprint arXiv:1906.03717 (2019).

[13] C. Hidey, K. McKeown, Fixed that for you: Generating contrastive claims with semantic edits, 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. 1756–1767.

[14] N. Slonim, Y. Bilu, C. Alzate, R. Bar-Haim, B. Bogin, F. Bonin, L. Choshen, E. Cohen-Karlik, L. Dankin, L. Edelstein, et al., An autonomous debating system, Nature 591 (2021) 379–384.

[15] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), volume 8, 2017, pp. 8–13.

[16] Z. C. Lipton, The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.

[17] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[18] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[19] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information fusion 58 (2020) 82–115.

[20] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: NAACL, 2021.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: International Conference on Learning Representations (ICLR), 2020.

[22] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[23] R. H. Johnson, Manifest rationality: A pragmatic theory of argument, Routledge, 2012.

[24] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large language models encode clinical knowledge, in: arXiv 2212.13138, 2022.

[25] T. Mayer, S. Marro, E. Cabrio, S. Villata, Enhancing Evidence-Based Medicine with Natural Language Argumentative Analysis of Clinical Trials, Artificial Intelligence in Medicine (2021) 102098.

[26] I. García-Ferrero, R. Agerri, G. Rigau, Model and data transfer for cross-lingual sequence labelling in zero-resource settings, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022.

[27] A. Yeginbergenova, R. Agerri, Cross-lingual argument mining in the medical domain, in: arXiv 2301.10527, 2023.

[28] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT:

pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019, pp. 4171–4186.