ECAI 2024
U. Endriss et al. (Eds.)

© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA241028

ANTIDOTE: ArgumeNtaTIon-Driven explainable artificial intelligence fOr digiTal mEdicine

Cristian Cardellino[a], Theo Collias[a], Benjamin Molinet[a], Erwan Hain[a], Wei Sun[b], Rodrigo Agerri[c], Serena Villata[a] and Elena Cabrio[a]

[a]Université Côte d'Azur, I3S, CNRS, Inria
[b]Department of Computer Science, KU Leuven
[c]HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract. The need for transparent AI systems in sensitive domains like medicine has become key. In this paper we present ANTIDOTE, a software suite proposing different tools for argumentation-driven explainable Artificial Intelligence for digital medicine. Our system offers the following functionalities: multilingual argumentative analysis for the medical domain, explanation extraction and generation of clinical diagnoses, multilingual large language models for the medical domain, and the first multilingual benchmark for medical question-answering. Experimental results demonstrate the efficacy of ANTIDOTE across different tasks, highlighting its potential as an asset in medical research and practice and fostering transparency, which is crucial for informed decision-making in healthcare.

1 Introduction

Argument Mining (AM) is a research topic of increasing interest within Artificial Intelligence (AI). AM has been applied to different domains, such as persuasive essays [28], scientific articles [30], web debating platforms [13], and political speeches [21], aiming to automatically extract and analyze argumentative content in natural language. AM has also been employed in the medical domain [38, 12, 17], looking to support the decision-making process, like diagnosis. Different models and datasets have been proposed to aid in this process [18, 20, 16]. Recently, Large Language Models (LLM) have favored the predominance of black box methods for classification, leaving the issue of explainability and transparency of the decision-making process open [1, 8]. This is a key challenge, in particular for systems that support decision-making in sensible scenarios like medicine, where it is important to deliver explanations that are understandable and significant to the user [27, 9].

In this work, we present a publicly available online demo system[1], with an accompanying demo video[2], and open source code[3] to aid students, professionals, and researchers in the medical field in exploring and evaluating medical data. It provides access to different tools resulting from the ANTIDOTE project[4]: *(i) argument mining for the medical domain* includes the demo of ACTA [22], access to a mBERT [7] model fine-tuned for multilingual argument component detection [37], and access to a dual-tower multiscale convolution neural network model for argument structure learning [29]; *(ii) explanation extraction and generation* includes the SYMEXP demo, and access to a fine-tuned mDeBERTa [14] model and the dataset for the extraction of correct explanations in the medical domain [2, 11]; *(iii) Medical mT5* includes the first multilingual text-to-text LLM for the medical domain [10] plus all the resources generated to train and evaluate the models; and finally, (iv) *MedExpQA* presents the only available multilingual benchmark for Medical Question Answering, plus a Mistral 7B fine-tuned model evaluated on the benchmark [3].

2 Argument Mining in the Medical Domain

AM aims to aid evidence-based decision-making in medicine. More precisely, AM aims to identify argument components (i.e., claims and premises) and argument relations (i.e., attack or support) to then reason upon the extracted arguments to deliberate.

2.1 ACTA 3.0

ACTA is an online tool[5] for the detection of argument components and their relations, the detection of PICO (Patient, Intervention, Comparison, Outcome) elements, and the effects of interventions on outcomes (i.e., Increased, Decreased, Improved, No Occurrences, No Difference). We refactored and revisited the ACTA pipeline [18], improving and updating technical aspects of the original code to increase its overall stability, documentation, and compatibility, especially with newer models available on Hugging Face [33]. The module was released as an open-source library[6]. We updated and relaunched the ACTA 2.0 demo [22], keeping its core functionality but also adding an open and documented REST API. ACTA 3.0 includes:

Search on PubMed PubMed[7] is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. In ACTA, we included the possibility of searching for a set of abstracts directly on the PubMed catalog through their API, integrated as a search bar to enter queries in the PubMed format. The abstracts of the queried publications can be subject to argument mining analysis with our models.

[1] http://antidote.i3s.unice.fr
[2] https://youtu.be/ZJFWAwuimVA
[3] https://gitlab.com/wimmics-antidote/antidote-server
[4] https://univ-cotedazur.eu/antidote
[5] http://antidote.i3s.unice.fr/acta/
[6] https://gitlab.com/wimmics-antidote/antidote-acta
[7] https://pubmed.ncbi.nlm.nih.gov/

Argumentative Analysis From a medical argumentative text, the system identifies the argumentative components and relations. The extracted argumentation graph is browsable in the demo, with the highlighted components and the graph of relations among them. The component detection model is DEBERTa-v3 [14], achieving a macro f1-score of 0.81, 0.82, and 0.82 for the AbstRCT-Neoplasm, AbstRCT-Glaucoma, and AbstRCT-Mixed tests sets respectively [20]. The relation classification model is SciBERT [4] uncased base model, which achieves a macro f1-score of 0.70 for the AbstRCT-Neoplasm test set [20]. Both model are fine-tuned on the AbstRCT dataset [19].

PICO Element Detection The system also extracts the PICO elements which are highlighted in the browsable demo. The model used here is trained on the EBM-NLP dataset [23] with coarse labels following the set-up in [20]. The f1-score on the test set is 0.69.

Effects on Outcome The last module of ACTA aims to identify the effects that interventions have on outcomes. The identified effects are also highlighted in the browsable demo. The module was trained on the part of AbstRCT containing outcomes. The task is addressed as a two-step pipeline: *(i)* outcome detection, and *(ii)* effect classification [20]. The outcome detection and effect classification tasks together reach a macro f1-score of 0.80.

REST API To foster versatility and re-usability, we also enhance the ACTA tool so that the full pipeline and each of the processing steps can be executed as independent units via our publicly available REST API. The API is documented and provides examples to run it.

Browsing Data The processed argumentative text is available for visualization via our web user interface. The browser shows the different elements (i.e., argument components, PICO elements, and effects on outcomes) highlighted in the analyzed text with different colors and displayed in a table with the text and their corresponding label. The argument relations are displayed in a graph that differentiates premises from claims by color, visualizing relations (i.e., attack or support) as directed edges. It is possible to download the analyzed data as a JSON file that has the metadata and the annotated text, both via the REST API and the browser UI. We also provide the user with pre-computed examples to visualize the outcome of ACTA.

2.2 Multilingual Argument Component Detection

Transfer learning and pre-trained language models are closely related, as the knowledge learned for one or more tasks in one specific language can be applied to other tasks or languages [31]. Given that the only existing dataset manually annotated with argumentative structures is currently available only for English, exploring this idea is particularly interesting for AM in the medical domain. A detailed study of a variety of techniques for cross-lingual transfer learning [37] has allowed us to conclude that *data-transfer*, i.e., translating the data and projecting the required annotations to a given target language, outperforms other alternatives. Furthermore, it also helps to improve results over training with the original English data only [20].

We provide the best performing mBERT model fine-tuned on the *data-transfer* approach to address the task of argument component detection in four languages: English, French, Italian, and Spanish[8]. The model can be easily integrated into applications via the HuggingFace API. Furthermore, the model card comes with an inference API that allows to test it with various examples in the four languages. The dataset used to train and evaluate the model is publicly available[9].

2.3 Dual-tower Multi-scale Convolution Neural Network (DMON)

DMON [29] is a model for Argument Structure Learning (ASL), i.e., the classification of relations between arguments. The model has four components that work as follows: first, an encoder extracts pairwise argument representations. The relationships are represented as an asymmetric relationship tensor. During training, a cropping strategy selects sub-tensors from the relationship tensor. Then, a bidirectional learning mechanism is applied to the cropped relationship tensors to capture contextual arguments and their relationships. Finally, the model employs label fusion to merge two predicted label matrices into one label matrix. During evaluation, the full relationship tensor is fed into the model. DMON was obtained by fine-tuning BioLinkBERT [36] for the AbstRCT dataset [19] and LinkBERT [36] for the SciDTB [35] and CDCP [25] datasets. DMON achieved macro F1-scores of 0.763, 0.742, 0.741, 0.484, and 0.681 on AbstRCT-Neoplasm, AbstRCT-Glaucoma AbstRCT-Mixed, SciDTB, and CDCP datasets, respectively.

3 Explanation Extraction and Generation

In this section, we present the two tools for the extraction and generation of argument-based natural language explanations.

3.1 SymExp

SymExp, i.e., Symptomatic Explanation, is an online tool[10] that provides natural language explanations for clinical cases and is used to train medical residents. The tool provides both a user interface and an REST API for consultation.

Explanation Generation The UI asks for a clinical case, the correct diagnosis for that clinical case, and up to 4 incorrect diagnoses. The result is a template-based explanation of why the correct diagnosis is the correct one and why the other diagnoses cannot be good for this precise clinical case. The tool provides a set of functionalities:

Entities detection Patient's symptoms in clinical cases are often described in layperson terms and can be missed by traditional medical Named Entity Recognition (NER) systems [26]. For example, occurrences of the symptom "Dyspnea" can be expressed as "shortness of breath" or "difficulty breathing". To tackle this, we provided a system to detect instances of these layperson terms (symptoms) as well as medical findings (blood tests, vital signs, ...). The model was trained using the MEDQA-USMLE-Symp dataset [16], and the best results were obtained using the biomedical domain-specific SciBERT [4] uncased transformer model initialized with its respective pre-trained weights, obtaining a macro f1-score of 0.86.

Findings to symptom conversion In addition to symptoms, medical explanations often rely on the results of additional tests known as medical findings, such as blood tests and vital sign measurements. The system employs the findings extracted through entity detection and converts them into symptoms. An example of this would be replacing the measure "platelets count is 50000 mcL" with "Thrombocytopenia", which is the term (i.e., concept) more commonly used in medical ontologies. A database of findings matched to symptoms, manually curated by experts, is used to make the conversion, and an option allows the user to use an LLM in an Input-Output Zero-Shot Prompting [32] setting if no corresponding symptom is found in the database. ChatGPT-4 [24] was the best performing for the findings to

[8] https://huggingface.co/HiTZ/mbert-argmining-abstrct-multilingual

[9] https://huggingface.co/datasets/HiTZ/multilingual-abstrct

[10] http://antidote.i3s.unice.fr/symexp/

symptoms conversion task. Under a few-shot setting [5], an accuracy of 0.64 was obtained.

Ontology alignment The alignment module associates, whenever possible, the detected symptoms and converted findings mentioned in the clinical case description with a term (concept) found in the Human Phenotype Ontology (HPO) [15].

Template-based explanations The system also incorporates a template-based explanation generation module. The goal is to explain why a patient is given a diagnosis based on her symptoms, and to discard alternative diagnoses by explaining why they are not viable. To support these explanations, we employ the previously converted findings and detected symptoms, as well as statistical information present in HPO, such as the frequency of these symptoms' incidence. Additionally, the module is able to identify symptoms that are key to a diagnosis but are not invoked in the given explanation. Finally, references to the original terms used in the clinical case description (i.e., layperson terms and findings) are kept as a reference in the produced explanations.

REST API Like in ACTA, we provide a publicly available documented REST API. This API provides examples for each of the modules in the tool.

3.2 Correct Answer Explanation Extraction

We provide a mDeBERTa model[11] fine-tuned for a novel extractive task consisting of *identifying the explanation of the correct answers* in medical exams commented with gold reference explanations [11]. In other words, the model is trained to answer medical questions regarding the correct answer in a multiple-choice setting *by providing the piece of text in the context that explains why a given answer is correct*. The model is fine-tuned using the CasiMedicos dataset [2, 11] for English, French, Italian, and Spanish[12], and it can be easily accessed using the HuggingFace API. The model card setup for this model in HuggingFace provides a demo to test it in inference mode. The model obtained a 0.746 f1-score (partial match) averaged across the four languages, showing the feasibility of our approach.

4 Medical mT5

Medical mT5 [10] is an encoder-decoder model developed by continuing the training of publicly available mT5 [34] checkpoints on medical domain data for English, Spanish, French, and Italian. The model was tested on different medical tasks, including two new multilingual sequence labeling (argument component detection[13]) and generative question answering[14] datasets for the evaluation of multilingual LLMs in the medical domain.

We distribute the base[15] and a fine-tuned version[16] of the models for multitask and multilingual sequence labeling in the medical domain. The models can be used via Hugging Face and tested using the available demo in their respective model cards.

A comprehensive experimental evaluation showed that Medical mT5 outperforms similarly-sized text-to-text models for the Spanish, French, and Italian benchmarks while being competitive in English with respect to the current state-of-the-art text-to-text [34, 6] and encoder-only models [14, 20].

5 MedExpQA - Multilingual Medical QA

MedExpQA [3] is the first multilingual benchmark (EN, ES, FR, IT) based on medical exams to evaluate LLMs in Medical Question Answering. To the best of our knowledge, MedExpQA includes, for the first time, reference gold explanations written by medical doctors[17]. Comprehensive experimentation using both gold reference explanations and Retrieval Augmented Generation (RAG) methods demonstrate that high performance in LLMs in this benchmark remains a challenge, especially for Spanish, French, and Italian. In addition, we also release the best-performing model in MedExpQA, a Mistral 7B model fine-tuned with maximum RAG, which obtained 0.6 in accuracy averaged across the four languages[18], through Hugging Face.

6 Concluding remarks

In this paper, we presented ANTIDOTE, a suite of tools and services for Argument Mining (AM) and explainable Artificial Intelligence (XAI) in the medical domain. With the growing interest in AM, our work addresses the critical need for transparent and understandable AI systems in a sensitive domain. Among the different functionalities offered by ANTIDOTE, we present tools for *(i)* AM on medical data: the updated ACTA demo, a full pipeline for AM in the medical domain, support for multilingual component detection, and a state-of-the-art model for argument structure learning; *(ii)* explanation extraction and generation: the SymExp demo that provides explanations, in natural language, based on diagnoses from clinical cases, as well as a multilingual model for the novel task of identifying the explanation of the correct answers; *(iii)* a Large Language Model fine-tuned for multiple multilingual medical tasks; and *(iv)* a new multilingual benchmark for evaluation of LLMs in medical question answering. All these modules aim at supporting evidence-based clinical decision-making and enhancing the interpretability of AI systems through natural language argument-based explanations. Additionally, experimental results showcase the efficacy of the tools presented across all of these tasks, highlighting the potential of ANTIDOTE as a valuable asset in medical research and training, offering transparency and comprehensibility that are crucial for facilitating informed decision-making in healthcare.

Acknowledgements

This work has been partially supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002. This work was also partially supported by the CHIST-ERA grant of the Call XAI 2019 of the ANR with the grant number Project-ANR-21-CHR4-0002. Rodrigo Agerri acknowledges the support of the following MCIN/AEI/10.13039/501100011033 projects: (i) Antidote (PCI2020-120717-2) and by European Union NextGenerationEU/PRTR; (ii) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of making Europe; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR. We also thank the European High Performance Computing Joint Undertaking (EuroHPC Joint Undertaking, EXT-2023E01-013) for the GPUs.

[11] https://huggingface.co/HiTZ/mdeberta-expl-extraction-multi

[12] https://huggingface.co/datasets/HiTZ/casimedicos-squad

[13] https://huggingface.co/datasets/HiTZ/multilingual-abstrct

[14] https://huggingface.co/datasets/HiTZ/Multilingual-BioASQ-6B

[15] https://huggingface.co/HiTZ/Medical-mT5-large

[16] https://huggingface.co/HiTZ/Medical-mT5-large-multitask

[17] https://huggingface.co/datasets/HiTZ/MedExpQA

[18] https://huggingface.co/HiTZ/Mistral-7B-MedExpQA-EN

Ethical Statement

We acknowledge the development of ANTIDOTE has ethical implications. The broader impact of this work lies in its potential to improve the medical and linguistic research communities. However, it also raises ethical considerations in the risks that come with applying LLMs, given the likelihood of incorrect inferences as the information is automatically generated. Furthermore, we are committed to transparency and fairness in our models' development and evaluation to reduce biases; and commit to open source of our models, data and code, promoting collaboration within the research community.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

[2] R. Agerri, I. Alonso, A. Atutxa, A. Berrondo, A. Estarrona, I. Garcia-Ferrero, I. Goenaga, K. Gojenola, M. Oronoz, I. Perez-Tejedor, G. Rigau, and A. Yeginbergenova. HiTZ@Antidote: Argumentation-driven Explainable Artificial Intelligence for Digital Medicine. In *SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing*, 2023.

[3] I. Alonso, M. Oronoz, and R. Agerri. MedExpQA: Multilingual Benchmarking of Large Language Models for Medical Question Answering. *arXiv 2404.05590*, 2024.

[4] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[8] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[9] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, and V. Patkar. Argumentation-based inference and decision making–a medical perspective. *IEEE intelligent systems*, 22(6):34–41, 2007.

[10] I. García-Ferrero, R. Agerri, A. A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, et al. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain. In *Proceedings of LREC-COLING*, 2024.

[11] I. Goenaga, A. Atutxa, K. Gojenola, M. Oronoz, and R. Agerri. Explanatory Argument Extraction of Correct Answers in Resident Medical Exams. *arXiv preprint arXiv:2312.00567*, 2023.

[12] N. L. Green, E. Cabrio, S. Villata, and A. Wyner. Argumentation for scientific claims in a biomedical research article. In *ArgNLP*, pages 21–25, 2014.

[13] I. Habernal and I. Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, 2016.

[14] P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv 2111.09543*, 2021.

[15] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, et al. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217, 2021.

[16] S. Marro, B. Molinet, E. Cabrio, and S. Villata. Natural language explanatory arguments for correct and incorrect diagnoses of clinical cases. In *ICAART 2023-15th International Conference on Agents and Artificial Intelligence*, volume 1, pages 438–449, 2023.

[17] T. Mayer, E. Cabrio, M. Lippi, P. Torroni, S. Villata, et al. Argument mining on clinical trials. In *COMMA*, pages 137–148, 2018.

[18] T. Mayer, E. Cabrio, and S. Villata. Acta: A tool for argumentative clinical trial analysis. In *IJCAI 2019-Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6551–6553, 2019.

[19] T. Mayer, E. Cabrio, and S. Villata. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press, 2020.

[20] T. Mayer, S. Marro, E. Cabrio, and S. Villata. Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*, 118:102098, 2021.

[21] S. Menini, E. Cabrio, S. Tonelli, and S. Villata. Never retreat, never retract: Argumentation analysis for political speeches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[22] B. Molinet, S. Marro, E. Cabrio, S. Villata, and T. Mayer. Acta 2.0: A modular architecture for multi-layer argumentative analysis of clinical trials. In *IJCAI 2022-Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

[23] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, and B. C. Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[24] R. OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

[25] J. Park and C. Cardie. A corpus of eRulemaking user comments for measuring evaluability of arguments. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1257.

[26] S. Raza, D. J. Reji, F. Shajan, and S. R. Bashir. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152, 2022.

[27] E. Reiter. Natural language generation challenges for explainable ai. *arXiv preprint arXiv:1911.08794*, 2019.

[28] C. Stab and I. Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

[29] W. Sun, M. Li, J. Sun, J. Davis, and M.-F. Moens. Dmon: A simple yet effective approach for argument structure learning. *arXiv preprint arXiv:2405.01216*, 2024.

[30] S. Teufel, A. Siddharthan, and C. Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502, 2009.

[31] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun. Pre-trained language models and their applications. *Engineering*, 25:51–65, 2023.

[32] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[34] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

[35] A. Yang and S. Li. Scidtb: Discourse dependency treebank for scientific abstracts, 2018.

[36] M. Yasunaga, J. Leskovec, and P. Liang. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022.

[37] A. Yeginbergen, M. Oronoz, and R. Agerri. Argument Mining in Data Scarce Settings: Cross-lingual Transfer and Few-shot Techniques. In *ACL*, 2024.

[38] J. Žabkar, M. Možina, J. Videcnik, and I. Bratko. Argument based machine learning in a medical domain. *Frontiers in Artificial Intelligence and Applications*, page 59, 2006.