Title:
Argumentation Technology for Explainable Misinformation Identification

Description:
The dissemination of misinformation has become one of the major concerns of the decade, being classified by the World Economic Forum as one of the main global economic risks due to its rapid propagation [1]. With the popularisation of generative Artificial Intelligence (AI) and Large Language Models (LLMs), this problem has become even worse, producing high quality natural language difficult to distinguish from the human-generated one and setting a new challenge in the identification of misinformation [2]. Automated systems for misinformation identification have become a promising effective countermeasure to the fast spread of online misinformation. Most of the previous work has focused, however, on sequence classification approaches, considering a text sequence (e.g., a news or a post) as the input and making a prediction based on the language distribution [3, 4]. This approach relying on finding linguistic patterns can work for factual misinformation (i.e., fake news), but presents important limitations when used for identifying rational misinformation (i.e., fallacies) [5]. This is mainly due to the fact that sequence classification approaches fail to model the complex dimension of natural language reasoning and inference.

Argumentation technology enables a richer analysis of the natural language inputs, by incorporating concepts from argumentation theory and reasoning to the modelling of natural language. Furthermore, by adding this argumentation-informed layer to the misinformation identification process, it will be possible not only to point out potential pieces of misinformation, but also the reasons behind this process. Other aspects beyond system performance can also be improved by this approach, including the trust placed in the system, its persuasiveness, as well as its educational capacity. This PhD project aims at integrating argumentation technology with automated misinformation identification systems, making its predictions more transparent and explainable.

References
[1] Tong, A., Du, D.Z. and Wu, W., 2018. On misinformation containment in online social networks. *Advances in neural information processing systems*, *31*.

*[2] C*hen, C. and Shu, K., Can LLM-Generated Misinformation Be Detected?. *In The Twelfth International Conference on Learning Representations*.

[3] Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R. and Schoelkopf, B., 2022, December. Logical Fallacy Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 7180-7198).

[4] Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 266–277, Bangkok, Thailand. Association for Computational Linguistics.

[5] Ruiz-Dolz, R. and Lawrence, J., 2023, December. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

Supervisory Team:
1st Supervisor: Dr. Ramon Ruiz-Dolz (rruizdolz001@dundee.ac.uk)
2nd Supervisor: Prof. Chris Reed (c.a.reed@dundee.ac.uk)