Detecting disinformation through computational argumentation techniques and large language models

Ana Gutiérrez[1], Stella Heras[2] and Javier Palanca[2]

[1]Universitat Politècnica de València (UPV), 46022 Valencia, Spain
[2]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV), 46022 Valencia, Spain

Abstract

Nowadays, the spread of disinformation poses a major challenge for society. Citizens find themselves immersed in a complex and data-saturated digital context that hinders their ability to critically discern between truthful information and disinformation. The task is complex because disinformation often appears in the form of rhetorical manipulations, logical fallacies, or distortions of the truth. This project aims to develop a system for detecting disinformation based on computational argumentation techniques and large language models, which promotes critical thinking and media literacy in society.

The implemented web tool analyzes the patterns of human reasoning used in argumentation, classifying arguments into argumentation schemes defined by argumentation theory. After identifying the argumentation scheme, the system thoroughly examines the reasoning presented in the argument and uses a set of critical questions to question its validity. Using a large language model, enhanced with external contextualization from various information sources, the system is guided in the process of evaluating the truthfulness of the argument. The final response includes both a qualitative and quantitative justification of the level of truthfulness, providing links and references to the information sources used in the evaluation.

Keywords

Computational Argumentation, Large Language Models (LLM), Natural Language Processing (NLP)

1. Introduction

This project presents a practical application of computational argumentation in a specific domain of natural argumentation, such as the dissemination and exchange of information in social networks, media, and other communication channels, where the fight against disinformation is a complex and crucial task. In this context and recognizing that disinformation often involves logical fallacies and rhetorical manipulation, this work investigates how Artificial Intelligence, particularly computational argumentation and large language models, can be leveraged to develop a system that enhances critical thinking and media literacy.

2. State of the art

An extensive review of the current state of the art reveals that disinformation detection is highly complex due to natural language's inherent ambiguities, vagueness, enthymemes[1], and dialectical variations, which can easily lead to deception or fallacies. As a fundamental natural language processing (NLP) task, disinformation detection involves analyzing text and speech. Before applying specialized Argument Mining techniques, it is crucial to first address key NLP aspects such as text preprocessing, feature extraction, and numerical representation, as discussed in [1].

However, the primary challenge in performing NLP tasks is preserving the order of words within their original context, as losing this information can degrade model performance. Capturing long-term dependencies is challenging, but sequential models like **Recurrent Neural Networks (RNNs)**[2] and **Transformers**[3] address this issue by processing data sequentially. Among RNNs, Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are particularly adept at long-term

Computational Models of Natural Argument (CMNA24)
agutman@upv.edu.es (A. Gutiérrez); stehebar@upv.es (S. Heras); jpalanca@dsic.upv.es (J. Palanca)

[1]An argument in which one premise is not explicitly stated.

learning. Transformers, which have gained significant popularity, also excel in various NLP tasks with encoder-based models for classification (e.g., BERT, DistilBERT, RoBERTa) [4] and decoder-based models for generation (e.g., GPT [5]).

Transformer-based models outperform traditional architectures like RNNs due to their attention mechanisms, which effectively capture long-range relationships without being constrained by word dependency length or complexity. Consequently, large language models (LLMs) based on Transformers represent the state-of-the-art in NLP tasks for disinformation detection. However, it has been observed that LLMs have certain limitations when applied to this area of study.

2.1. Critical review: limitations of LLMs

LLMs consist of millions of parameters, leading to high computational costs, substantial economic investment, and a significant environmental impact due to their carbon footprint. Although techniques like fine-tuning can refine these models, they still require powerful GPUs. Therefore, finding methods to reduce computational expenses without sacrificing performance is essential, making prompt engineering a promising approach.

These architectures often have difficulty retrieving specific information if it wasn't part of their training data, leading to "*hallucinations*," where the model generates information that seems plausible but is actually incorrect. Furthermore, it becomes evident that these models lack a logical layer. In other words, they do not consider the underlying logic that makes a text or argument truthful; instead, they simply detect and classify content based on its structure without paying attention to the reasoning involved, as seen in [6]. While LLMs are very adept at generating informative text, they often struggle with extensive reasoning and may even end up contradicting themselves.

As outlined in [7], relying solely on LLMs for disinformation detection is insufficient, highlighting the need to incorporate a logical layer into these models. The proposed solution in this project is to introduce this logical layer by applying techniques from computational argumentation.

2.2. Computational Argumentation

A key concept used throughout this project is argumentation schemes [8], as abstract patterns of human reasoning. The system outlined in [8] has been followed, which categorizes different argumentation schemes into three major groups: Source-Based Arguments, Applying Rules to Cases Arguments and Reasoning Arguments. For example, Figure 1 illustrates the argumentation scheme for the "Argument from Position to Know/Authority" which belongs to the Source-Based Arguments category.

Argument from Position to Know/Authority		
Argumentation Scheme		**Critical Questions**
Major Premise	Source a is in position to know certain facts related to a subject of domain S contained in proposition A.	CQ1: Is a in position to know if A is true (false)?
Minor Premise	a states that A is true (false).	CQ2: Is a a reliable source?
Conclusion	A is true (false).	CQ3: Did a really affirm that A is true (false)?
Example	Medical professionals and scientists are in position to know about the spread of infectious diseases in public spaces. Medical professionals and scientists claim that wearing facemasks in public spaces can significantly reduce the spread of infectious diseases. Therefore, wearing facemasks in public spaces is an effective measure to prevent the spread of infectious diseases.	

Figure 1: Argumentation Scheme from Authority including its premises, conclusions and set of critical questions

3. Proposed solution

The disinformation detection system[2] comprises two interconnected modules (Figure 2), as follows:

[2]Link to the implemented web tool: http://desinformacion.gti-ia.upv.es/

- **Module 1: "Classifier System in Argumentation Schemes"**. This module takes a sentence (written in English) representing an argument as input and classifies it according to an argumentation scheme from argumentation theory. This module uses RASA[3], an open-source framework for developing conversational systems.
- **Module 2: "Veracity Level Generator and Evaluator System"**. This module receives the output from Module 1. Based on the argumentation scheme and a set of critical questions, the system uses a LLM with external contextualization to finally provide a qualitative justification and a quantitative value indicating the argument's level of truthfulness. The LLAMA[4] family was investigated, with the LLAMA 3 70B model used as the LLM for this module.

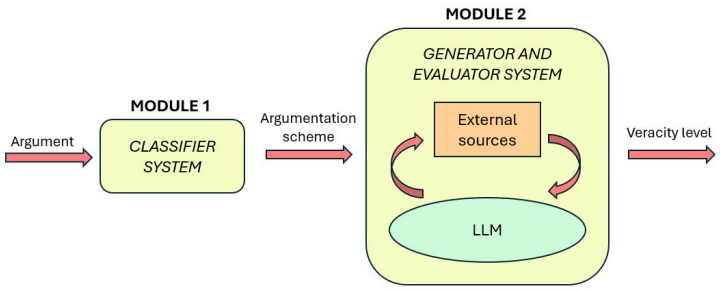

Figure 2: General architectural design

4. Solution design and implementation

4.1. Module 1: Classifier System in Argumentation Schemes

The NLAS-MULTI corpus [9], developed by the Centre for Argument Technology at the University of Dundee, has been utilized for this project. This dataset was modified to create a specific version in YAML (YAML Ain't Markup Language) format [10], compatible with RASA. It includes 2,188 arguments in English covering both sides of 50 topics and associated with 19 argumentation schemes. A significant modification was adding a "*no scheme*" class to handle arguments that don't fit any specific scheme, enhancing the system's adaptability to real-world scenarios.

Initially, a single-layer classifier was implemented to identify the argumentation scheme of an input sentence from 19 possible classes. However, due to sub-optimal results (discussed in section 5.1), a two-layer classifier was adopted, as Figure 3 illustrates. This new approach, based on Walton's theory, divides the 19 schemes into four groups, creating four classifiers in RASA, distributed across two layers. The first layer classifies the sentence into one of these groups, and the second layer activates the corresponding classifier to determine the specific argumentation scheme.

To implement these classifiers, the RASA NLU module based on intent detection was developed, with each class representing an intent defined by multiple examples. A consistent configuration was used across all four classifiers, including the default pipeline: the *WhitespaceTokenizer* [11], the *RegexFeaturizer* [12], the *LexicalSyntacticFeaturizer* [13], the *CountVectorsFeaturizer* [14], the *DIETClassifier* [15], the *ResponseSelector* [16] and the *FallbackClassifier* [17]. The dataset was split using the holdout method, allocating 80% for training across 100 epochs and 20% for evaluation.

4.2. Module 2: Veracity Level Generator and Evaluator System

In this module, the quantized LLAMA 3 model with 70 billion parameters is employed using an on-premise deployment of the model. The process begins with the creation of a client using the OpenAI

[3]https://rasa.com/
[4]https://llama.meta.com/

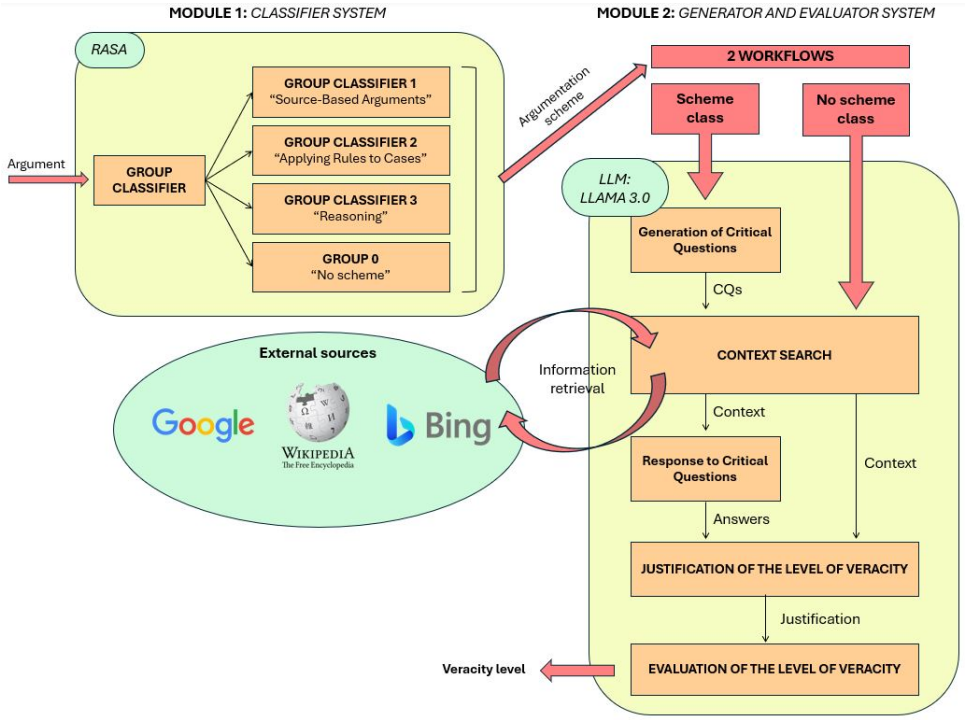

Figure 3: Detailed design of the complete system architecture

Python library, enabling interaction with the model through chat completions. Prompt engineering is utilized to craft requests in JSON format, which instructs the model on the specific tasks to be performed.

As shown in Figure 3, two workflows are defined depending on the previous classification—either an argumentation scheme or a non-schema type. If the input sentence does not correspond to any argumentation scheme, a context search is conducted using the sentence's inherent information. On the contrary, if the sentence aligns with an argumentation scheme, critical questions are generated and tailored to the content of the sentence. A parallel context search then retrieves information from three external sources (Google, Wikipedia, and Bing) for each critical question. Finally, with the gathered context, the model addresses the defined critical questions and synthesizes all collected information to provide a structured evaluation of the input argument's veracity, acting as an expert assistant in computational argumentation.

5. Experimentation and results

5.1. Module 1: Classifier System in Argumentation Schemes

In Module 1, a comparison of the unilayer and bilayer architectures reveals that the two-layer model consistently outperforms the single-layer model across all metrics, including accuracy, recall, precision, and F1-score, achieving values between 85% and 87% (Figure 4). The primary issue with the unilayer approach was the error introduced by the non-schema class. This is evident from the confusion matrix (Figure 5), where a prominent column in the non-schema class area indicates a tendency of the classifier to misclassify arguments into this category. The two-layer classifier addresses this problem effectively, reducing the noise associated with the non-schema class and improving the precision for this class ("*group0*") from 20.19% to 72.22%, as shown in Figure 4.

5.2. Module 2: Veracity Level Generator and Evaluator System

To evaluate the performance of module 2, a survey was conducted with the answers generated by the system for 20 examples of arguments, covering the 19 classes of argumentation schemes. Responses from

	ACCURACY		RECALL		PRECISION		F1-SCORE	
	UNILAYER	BILAYER	UNILAYER	BILAYER	UNILAYER	BILAYER	UNILAYER	BILAYER
GROUP 0	100,00%	100,00%	100,00%	100,00%	20,19%	72,22%	33,59%	83,87%
GROUP 1	51,53%	86,03%	51,53%	86,03%	99,16%	94,26%	67,82%	89,95%
GROUP 2	48,63%	91,78%	48,63%	91,78%	98,61%	83,75%	65,14%	87,58%
GROUP 3	46,85%	69,93%	46,85%	69,93%	100,00%	98,04%	63,81%	81,63%
OVERALL PERFORMANCE	61,75%	86,93%	61,75%	86,93%	79,49%	87,07%	57,59%	85,76%

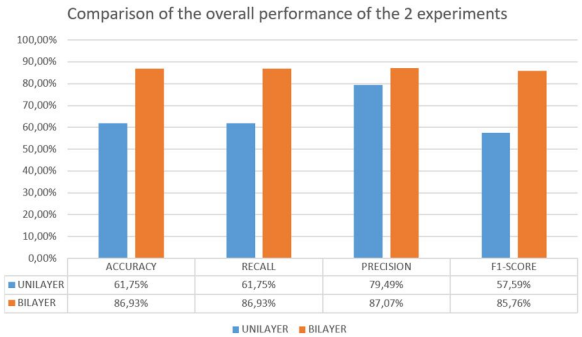

Figure 4: Comparison of the unilayer and bilayer experiments

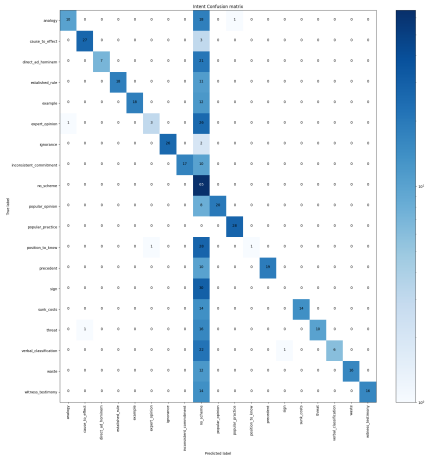

Figure 5: Confusion matrix unilayer experiment

a total of 80 adults between the ages of 18 and 65 were collected and analyzed. Each respondent evaluated three specific aspects for each example: the quantitative justification of the level of truthfulness, the qualitative justification of the level of truthfulness, and the adequacy of the sources.

Figure 6 illustrates that all aspects of the system received similarly positive ratings. In terms of overall satisfaction (Figure 6), a significant majority (83.8%) rated the system's responses as satisfactory with a high degree of contentment, indicating consistent coherence across all areas.

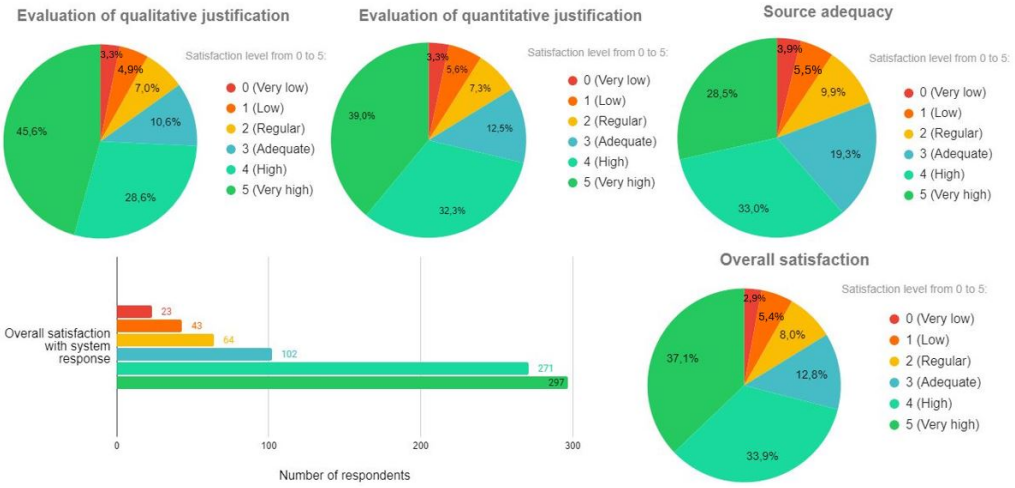

Figure 6: Evaluation of system responses according to several aspects and overall satisfaction

6. Conclusions and future work

In conclusion, this work successfully developed a tool for detecting disinformation in societal arguments. This research has met its objectives through the creation of a web-based tool comprising two interconnected modules that integrate computational argumentation techniques with LLMs.

Looking ahead, for scaling the system, future work could involve integrating intelligent agents. Implementing domain-specific expert agents or agents with access to diverse information sources could enhance the system's final responses through the use of advanced technologies and collaborative approaches. Additionally, the veracity of the sources used to generate context should also be studied.

References

[1] A. Montoro-Montarroso, P. Rosso, Á. Panizo-Lledot, B. Calvo-Figueras, F. J. Cantón-Correa, B. Chulvi, J. Huertas-Tato, M. J. Rementeria, J. Gómez-Romero, Inteligencia artificial contra la desinformación: fundamentos, avances y retos (2023).

[2] S. Yang, X. Yu, Y. Zhou, Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example, in: 2020 International workshop on electronic communication and artificial intelligence (IWECAI), IEEE, 2020, pp. 98–101.

[3] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, Overview of the transformer-based models for nlp tasks, in: 2020 15th Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2020, pp. 179–183.

[4] A. F. Adoma, N.-M. Henry, W. Chen, Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition, in: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2020, pp. 117–121.

[5] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, et al., Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, IEEE Access (2024).

[6] P. Goffredo, M. Espinoza, S. Villata, E. Cabrio, Argument-based detection and classification of fallacies in political debates, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 11101–11112.

[7] R. Ruiz-Dolz, J. Lawrence, 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, 2023.

[8] D. Walton, C. Reed, F. Macagno, Argumentation schemes, Cambridge University Press, 2008.

[9] R. Ruiz-Dolz, J. Taverner, J. Lawrence, C. Reed, Nlas-multi: A multilingual corpus of automatically generated natural language argumentation schemes, arXiv preprint arXiv:2402.14458 (2024).

[10] O. Ben-Kiki, C. Evans, B. Ingerson, Yaml ain't markup language (yaml™) version 1.1, Working Draft 2008 5 (2009).

[11] S. Vijayarani, R. Janani, et al., Text mining: open source tokenization tools-an analysis, Advanced Computational Intelligence: An International Journal (ACII) 3 (2016) 37–47.

[12] M.-H. Hwang, J. Shin, H. Seo, J.-S. Im, H. Cho, Korasa: Pipeline optimization for open-source korean natural language understanding framework based on deep learning, Mobile Information Systems 2021 (2021) 1–9.

[13] D. S. Mishra, A. Agarwal, B. Swathi, K. C. Akshay, Natural language query formalization to sparql for querying knowledge bases using rasa, Progress in Artificial Intelligence 11 (2022) 193–206.

[14] T. T. Nguyen, A. D. Le, H. T. Hoang, T. Nguyen, Neu-chatbot: Chatbot for admission of national economics university, Computers and Education: Artificial Intelligence 2 (2021) 100036.

[15] W. Astuti, D. P. I. Putri, A. P. Wibawa, Y. Salim, A. Ghosh, et al., Predicting frequently asked questions (faqs) on the covid-19 chatbot using the diet classifier, in: 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), IEEE, 2021, pp. 25–29.

[16] X. Kong, G. Wang, A. Nichol, Conversational AI with Rasa: Build, test, and deploy AI-powered, enterprise-grade virtual assistants and chatbots, Packt Publishing Ltd, 2021.

[17] F. S. Khan, M. Al Mushabbir, M. S. Irbaz, M. A. Al Nasim, End-to-end natural language understanding pipeline for bangla conversational agents, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 205–210.