Dr hab. inż. Maciej Piasecki
Katedra Sztucznej Inteligencji
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl

21st June 2026

Reviewer's opinion
on Ph.D. dissertation authored by
Arkadiusz Modzelewski
entitled:
Leveraging Persuasion and Intent for Analysis and Reasoning-based Detection of Disinformation with Large Language Models

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?
The thesis addresses the problem of disinformation detection and proposes an approach that enhances it by leveraging contextual analysis informed by the recognition of signals of persuasion and malicious intent. The problem of disinformation detection and the correlated subproblems are well defined by novel, unique textual datasets built as a result of a carefully planned and conducted manual annotation process.

Is it a scientific one?
The core problem is a scientific challenge that remains unresolved. In addition, the problem of the development of a new multilingual annotated dataset is an important scientific problem that is fundamental for fostering further development in the area of disinformation detection and mitigation.

Does it have a practical meaning?

There are several aspects of practical importance in the work done for the thesis: several unique research datasets, a large number of implemented recognition methods, and many reported and carefully documented experimental results. As all the results have been published on open licences, the thesis expresses a high level of practical importance.

2. Contribution

What is the main, original contribution of the dissertation?
The main contribution of the thesis is a complex approach to disinformation detection in which the core task is preceded by solving one of the correlated subtasks, namely recognition of persuasive strategies or malicious intent. From the Machine Learning perspective, such an approach is a kind of multitask learning, but it is an original, novel approach in relation to disinformation detection. However, it should be emphasised that this similarity to multitask learning was neither noticed nor explored in the thesis.

A very important contribution of the thesis is several original manually annotated datasets that define both the core problem and the considered correlated subproblems. The datasets, all but one, are based on original data and a well-defined process of manual annotation and quality control. For instance, all the datasets planned to be used in the evaluation of the methods based on LLMs were built from the

texts published after the source cut-off dates (i.e. the dates declared by the LLM developers as the final dates for collecting texts for pretraining) for all the LLMs used in the given round of evaluation experiments.

All the language resources – training-testing text datasets – developed as a result of the thesis are a very valuable contribution to the development of NLP and its applications.

"Persuasion-Augmented Chain of Thought", which is proposed in the thesis, together with a very similar method utilising malicious intent recognition, are presented as "novel zero-shot methods" that improve disinformation detection. However, they represent a variant of the well-known Chain of Thought technique or simplified reasoning, and the proposed methods are not discussed from this perspective and compared to the appropriate works from outside of disinformation detection.

An important contribution of the thesis is the exploration of the idea that recognition of an aspect of intentionality is critical for disinformation definition and detection. It is a pity that recognition of persuasion and intent signals was not combined into one joint approach.

In addition to the three main contributions characterised briefly above, the author defined the fourth main objective of his thesis, namely, analysis and comparison of "AI-generated and human-authored persuasive content based on linguistic characteristics and detection difficulty for automatic disinformation systems." The main purpose of this task seems to be unintentionally disguised, as the generation is just an exercise in LLM-based augmentation of a text dataset (this task is not named correctly in the thesis). In addition, the range of LLMs used is limited and a little attention was given to the fact that LLMs form a kind of different families due to their origin. However, the presented in-depth analysis on the basis of comparison of the generated texts to human-written texts brought many interesting findings, especially with respect to the stylometric techniques applied. Nevertheless, the similarity of these experiments to the task of the recognition of human-written text from LLM-generated texts was not noticed or explored in the thesis, nor were they compared to the approaches from the literature.

As the thesis is, in fact, based on a body of publications (see the next section for the discussion), and the key chapters are slightly rewritten conference papers, it is necessary to raise the question about the personal contribution of the thesis author to each of the publications. Here we come to a difficulty, as the author's personal contribution to the four publications included in the body of works (presented in the thesis as bases for the four key chapters) is not clarified or declared in the thesis. All the works are multi-author, while the obtained results are presented as the results of the thesis. In the case of the work of resources, we can guess the role of the thesis author, but not in the case of the methods. A strong signal is the fact that Mr Arkadiusz Modzelewski is the first author in all four papers, but proper explanations of his role, paper by paper, are clearly missing in the thesis.

3. Thesis structure

The thesis has a very appropriate structure in terms of its general plan. Most chapters present rich content, but with a notable exception of Chapter 3 "Literature Review", which is short and very brief. It rather signals different problems and solutions than presents. The main body of the thesis is quite lengthy: around 120 pages.

However, a closer look reveals that the chapters are in fact adapted conference papers, so the whole thesis more resembles a description of the body of works, including four conference papers, not a thesis planned and written from scratch. It is worth noting that the papers have been adapted quite well to be

interlinked chapters. However, some inconsistencies and gaps in the fluency of the narration can be noticed. For instance, in Section 7.1 "Human Persuasion Datasets", the question about the cut-off and materials known to LLMs is out of sadness forgotten, while it was carefully taken into account in all other chapters. Also on page 70, BERT-based methods from the previous chapter seem to be forgotten, and a new set of BERT-based methods is developed in this chapter.

There is also a very visible lack of a combination of both methods for disinformation detection proposed in the thesis, namely: supported by intent recognition and persuasion signal recognition into one scheme based on the Chain of Thought approach to disinformation detection.

4. Correctness

Can we trust what is claimed in the dissertation?

All works included in the dissertation have been carefully planned, conducted in a very systematic and thorough way, and carefully evaluated. For instance, it is highly appreciated that in most evaluation experiments in the thesis, best efforts are made to ensure that test data sets do not overlap with the texts potentially used for pre-training the tested LLMs.

All the decisions and steps undertaken are very well described.

Nevertheless, some points for the discussion or minor drawbacks can be noticed.

Are the argument s correct? Indicate the flaws you have noticed, if any.

Concerning the proposed text classification or detection methods based on LLMs, it is surprising that in the case of most of them, if not all, there is no clear reference to a few-shot prompting scheme. Moreover, there are no clear attempts in the thesis to use a few-shot prompting scheme as a baseline.

In some prompts in the appendix, examples appear as part of the prompt, but there is no discussion on the selection of examples and their use in the method description, as well as evaluation. This negligence of a few-shot prompting technique, a rather standard one recently, is a strange drawback of the thesis. It is not clear to what extent the performance increase observed in the experiments in relation to the methods compared with is due to the inclusion in the prompts text material resembling examples, and to what extent it is due to a kind of prior solving of a supporting task. For example (Page 96), "improvements can still occur even when individual intents are misclassified. This indicates that intent signals may partially compensate for one another and that the benefits of intent-based reasoning are not strictly dependent on perfect intent recognition." – but, maybe, this effect is due to the information about intent fulfilling the role of an example that directs the internal state of LLM towards areas closer to disinformation detection skills? Without a more thorough comparison to a few-shot prompting, such questions are left unanswered.

Mostly, smaller open LLMs are used in the experiments, which is a very good tendency. However, large commercial LLMs capable of reasoning are also used for comparison. Their reasoning skills are mentioned, but the solutions proposed in the thesis are not appropriately compared to baseline solutions based on reasoning LLMs. This is especially missing and surprising, as both core methods of the thesis are based on a semi-reasoning scheme, i.e. a combination of two steps (or iterations), where the first one leads to enrichment of the context used in the next phase of the final disinformation detection. This

scheme resembles a kind of pre-planned, simplified reasoning, so it would be good to compare it with LLM-based reasoning with appropriate system prompts.

Moreover, after reading the dissertation, it is hard to learn how both the steps of supporting pre-analysis and the final disinformation detection are combined. We can guess that it is done by a deterministic algorithm controlling two calls to LLMs, but this guess is hard to clarify on the basis of both the main content and the appendices.

In case of the author's own approaches of the author, all prompts are more or less presented in the appendices (and also the whole repository is available on open licence, which is highly appreciated), but in the case of the baseline methods, the prompts used are often not explained, e.g. in Section 5.5.2 "Prompting Methods Comparison".

5. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of ***Information and Communication Technology****. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections?*

Chapters 2 and 3 demonstrate a good understanding of the relevant developments in the area of disinformation and persuasion detection. In addition, Chapter 2 and the information spread in Chapters 4-6 show a good understanding of the basic NLP techniques, language models, Large Language Models and their adaptation to different tasks. It is especially worth noticing that the candidate presented expertise in the development of annotated language resources and their use in evaluation.

What is your opinion on the list of references? What is the degree of its completeness?

The list of references is very long, and it is very comprehensive from the point of view of the main goals of the thesis. However, different specific solutions that are included in the proposed methods touch upon different problems studied in broader NLP, and with respect to them, the bibliography is very limited. For instance, there is a lack of references to the recognition of human-origin vs LLM-generated texts. In the same way, the area of multitask learning in NLP, at least in relation to fine-tuning of language models, is not represented in the bibliography. However, these aspects have also been omitted in the description of the proposed methods, so their absence in the bibliography is the consequence.

6. Other remarks

Additional detailed comments:

p 19: "to fine-tune the underlying language model" — in what sense does the model not change due to the input prompt?

p 21: " Rather than using surface-level credibility judgments, this work focuses on obtaining earned credibility evaluations for novel datasets" — but such an approach has limited applicability as such techniques are very laborious.

p 38: „ In this study, we wanted to focus on a binary classification: disinformation versus credible articles." — but such an approach results in simplification of the task, e.g. misinformation can be harder to distinguish from disinformation, than from credible articles

p 42: „ Experts placed 105 articles in that category. We have removed these articles from the dataset, considering them as articles that did not reach a consensus." — once again, this seems to be a simplification of the dataset and the represented task, as ambiguous classes often pose the most difficult challenges.

p 43: „ neither manipulations nor intents are specific to the topic of the articles." — but it was not checked if any correlation is present.

p 46: „ PolBERT when the base model was HerBERT, and PolBERTa when the base model was Polish RoBERTa." — the distintion of the names is very hard to be spotted

p 47: „4.4.1 Polish Disinformation Detection" — the results for the BERT-like models may suggest that the task represented by the dataset is relatively simple, too simple? This may go in line with the simplifications introduced and noted in the comments above.

p 48: Sec. 4.4.2 — What was a statistical baseline for the manipulation techniques discovery? Was the IAA equal for the different techniques?

p 49: " annotation quality and methodological rigor." — but IAA before consensus is not reported, so we do not know what is the real quality of the dataset, the changes introduced by the consensus are useful, but blurs the consistency of annotations across the dataset.

p 58: " IP establishes the context and overrides alignment tuning," — this is only an assumption or intention, but not verified in the experiments.

p 61: " and prompts identification of those present in the text" — something unclear

p 67: " Tables 5.14 and 5.15 provide further key insights into the relationship between persuasion strategies and disinformation across different models and prompting methods." — Was there any attempt to compare this with co-occurrences observed in the manually annotated data?

p 70: " We also compare the BERT performance on unseen data to LLMs with baseline methods and with" — but why have the methods and results from the previous chapter been forgotten?

p 80: "They reached 65.19% agreement on the more complex multilabel intent task." — what does this mean? What kind of the IAA measure was used?

p 80: "This improved the reliability and quality of the dataset. " — but this isn't a trustworthy indicator of the difficulty of the task

p 82: "Table 6.2 details the distribution of the five malicious intent categories in the dataset." — The distribution is very balanced; is this due to the nature of the data or the result of manual preselection?

p 82: " The most frequent intent pair is UIOA and UCPI," — Is not UIOA intrinsically correlated with UCPI (by definition)?

p 86: "Multilabel Detection. " — it was not discussed that the multilabel results are higher than the binary ones, that not often happens.

p 91: Table 6.9 – a typo, repeated value

P 97: " leaving the linguistic differences between human-written and LLM-generated persuasive texts unexplored." — Is not this question of the training process? LLM very much depend on the selection of text and post-training etc. This difference is interesting from the point of the LLM-based data augmentation, but it is risky to claim that it is an intrinsic feature of LLM-generated persuasive texts in general, i.e. of any LLM used, as it seems to be argued for in the thesis.

P 100: "Controlling the generation process through explicit instructions is crucial for our study, as it ensures that the resulting LLM-generated persuasive texts remain semantically comparable to human-written texts." — In what sense of [illegible]

P 123: " summa4 rizes" — a typo

Bibliography: there are quite frequent errors in the use of capital letters, i.e., their lack in abbreviations or proper names

7. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the A*ct of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

- *Does the dissertation present an original solution to a scientific problem?*
- *Does the candidate possess general theoretical knowledge and understanding of the discipline of* ***Information and Communication Technology****?*
- *Does the dissertation support the claim that the candidate is able to conduct scientific work?*

are positive and recommend proceeding with further steps of the PhD procedure and public defence of the thesis of Mr Arkadiusz Modzelewski.

Maciej Piasecki

Signature

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276