Queen Mary University of London
Peter Landin, 10 Godward Square
London E1 4FZ
Tel: +44 79 3314 3400
Website: www.eecs.qmul.ac.uk

School of Electronic Engineering and Computer Science
Dr. Arkaitz Zubiaga
Associate Professor
Email: a.zubiaga@qmul.ac.uk

20th April, 2026

Review Report on the PhD thesis entitled:

'Leveraging Persuasion and Intent for Analysis and Reasoning-based Detection of Disinformation with Large Language Models'

Author: Arkadiusz Modzelewski

Supervisors: Adam Wierzbicki and Giovanni Da San Martino

Thesis motivation and contributions:

This thesis deals with the problem of online disinformation, as a timely and important issue that has severe negative impacts on the integrity of the information we find online. Research using natural language processing (NLP) methods for tackling online disinformation is still an unresolved problem and this thesis furthers research in this direction by proposing to leverage the understudied dimensions of persuasion and intent to support the dtection of disinformation.

To advance research in NLP to address disinformation focused on the above directions, the thesis makes the following four contributions:

1. It creates and publishes a new dataset which contains samples in two languages, Polish and English, and which are labelled for their intent. The aim of this dataset is to enable detection of disinformation samples where the author has a malicious intent.

2. It proposes a method where persuasion is used as a dimension to enhance the reasoning component of large language models to detect cases of disinformation.

3. It also proposes another method which in this cases uses intent as a dimension to augment the reasoning ability of large language models in the effort of detection disinformation.

4. It investigates the potential and challenges involved in detecting AI-generated persuasive text.

Thesis structure:

The thesis is organised in eight chapters, as follows:

Chapter 1 introduces the problem of online disinformation, discusses why it is important to identify and address this disinformation, and the challenges involved in doing so. It discusses the limitations of existing research which lack an ability for reasoning to reach a more informed decision about the fact that a text is deemed to convey disinformation. As an effort to further research in this direction of supporting improved reasoning in NLP models to detect disinformation, this chapter proposes exploiting persuasion and intent as supportive dimensions. It also sets forth the four objectives of thesis aligned with the four contributions stated above, i.e. creating a dataset (RO1), exploiting persuasion (RO2), exploiting intent (RO3) and studying AI-generated persuasive content (RO4). The chapter then discusses the contributions of the thesis in terms of methods and outputs, as well as in terms of publications.

Chapter 2 provides a technical background on the methods used in the thesis and particularly about large language models (LLMs). This is a well-written and technically sound chapter, which provides more textbook than scientific material, but it's a good chapter for the unfamiliar reader to understand the methods being used as a base.

Chapter 3 provides a thorough literature review on research related to this thesis. The chapter is organised in three main subsections which discuss the three core topics related to the thesis contributions: (i) disinformation detection using LLMs, (ii) computational approaches to persuasion detection, and (iii) the intent dimension of disinformation. The chapter is discussed and contextualised in relation to the thesis, and provides a summary of existing research and gaps in an additional subsection at the end.

Chapter 4 describes the creation and publication of a dataset designed to study disinformation in the Polish language, with a focus on manipulation and intent. The MIPD dataset released as part of this chapter contains over 15,000 web articles, each annotated for the following four dimensions: (I) whether it is disinformation or not (the original annotation included four categories, but the other two were subsequently removed: disinformation, misinformation, credible or hard-to-say); (ii) the intention types; (iii) the manipulation techniques used in the article (which include 11 different types listed); (iv) the theme of the article. The chapter then presents and discusses a set of benchmark experiments on this dataset, predicting both whether a web article constitutes a case of disinformation, and detecting the type of disinformation.

Chapter 5 studies how the persuasive nature of a web article can be exploited as an informative dimension that can support detection of disinformation. To address the research question as to whether disinformation detection can be supported with the input from a persuasion detection technique, this chapter proposes Persuasion-Augmented Chain of Thought (PCoT), a method based on LLMs and which incorporates persuasion for model enhancement. This chapter then presents experiments with five different datasets: two newly created datasets (MultiDis, EUDisinfo) and three existing datasets (CoAID, ISOT Fake News, ECTF). The proposed approach then follows two steps: (i) persuasion detection, and (ii) disinformation detection. The approach follows this pipeline to get a final decision. Experiment results show the effectiveness of the proposed model over an equivalent baseline model that doesn't use the persuasion information.

Chapter 6 studies how detecting the type of (malicious) intent of a web article can help with the detection of disinformation. Therefore, this is proposed in a very similar fashion to the previous chapter, but in this case using intent instead of persuasion. To do this, an intent detection process is followed by the disinformation detection model in the pipeline, making then a final decision. To support this research, this chapter introduces a new dataset called MALINT, which contains samples of disinformation annotated for the type of malicious intent. Five different types of malicious intent are identified and used for the annotation: (i) Undermining the Credibility of Public Institutions, (ii) Changing Political Views, (iii) Undermining International Organizations and Alliances, (iv) Promoting Social Stereotypes/Antagonisms, and (v) Promoting Anti-scientific Views. The annotation process is described in detail, followed by the description of the intent detection and disinformation detection approaches. Experiments demonstrate that incorporating the detected type of intent into the disinformation detection model improves over the baseline that doesn't use it.

Chapter 7 moves away from the detection of disinformation to detect persuasion in text, hence with some link and relevance to chapter 5, but slightly varying from the overall thesis topic; still, the topic is related and relevant to the problem, e.g. the methods described in this chapter could eventually be used for disinformation detection, even if they haven't been used for this purpose in the thesis. The aim of this thesis is to study whether a model can detect persuasion generated by LLMs, and this is studied by also looking at whether detecting persuasion generated by LLMs is more difficult than detecting persuasion generated by humans. A new dataset called Persuaficial is created to enable this research, in addition to using other existing datasets. Experiments conducted on persuasion detection find that detecting persuasion generated by LLMs is more difficult than detecting persuasion generated by humans. The chapter then continues and concludes with an analysis of the characteristics of LLM- vs human-generated persuasive texts.

Chapter 8 concludes the thesis by summarising the contributions and by revisiting the research questions set forth in the introduction, now giving an answer to these questions based on the experiments conducted. The chapter also discusses potential future research directions that this thesis opens up. The chapter doesn't discuss limitations of this thesis.

Quality assessment:

The thesis is overall very well written. It's very well and clearly structured throughout, provides sufficient level of detail and is easy to follow and understand. I found the technical quality of the thesis to be sound, and provides a thorough and critical analysis of a timely and important problem. The thesis produces new knowledge that is useful for the research community, especially in understanding how intent and persuasion can be important factors to support detection of disinformation.

Strengths and weaknesses:

Strengths. The thesis has created numerous new resources, especially new datasets that will be useful to the community, in addition to the knowledge produced.

The quality of the thesis has been demonstrated through publication in top-tier venues including two papers in ACL, one in EACL and one in EMNLP, which are all among the most competitive conferences in the field. This demonstrates that the core of the work has already been peer reviewed and validated by the community.

All the experimentation and analysis is very thoroughly presented and discussed, which demonstrated the high standard and quality of the research conducted.

Weaknesses. The claims that the thesis makes in the abstract that it moves away from the usual binary disinformation detection didn't seem convincing to me throughout the thesis, as most classification is defined as binary (disinformation vs credible) throughout).

The binary labelling of disinformation vs credible is confusing. Where disinformation refers to factuality (i.e. info that is factually incorrect), credible refers to a perception (i.e. perceived as being true). Disinformation can also be credible, so they are not disjoint categories. Alternatively, disinformation vs true or accurate would have been more appropriate in my view.

The thesis doesn't acknowledge its limitations, which is an essential part of any thesis and an important exercise of reflection of what hasn't been possible or couldn't be fully satisfied. This is different from what can be future research directions.

Questions for the candidate:

I found it interesting that the annotation of the MIPD dataset focuses on distinguishing misinformation and disinformation. These two differ in intent of deceiving or not, however it has been barely done previously because it's often not possible to know if the author had an intent to deceive or not, especially when we are dealing with written text and we can't speak with or scrutiinse the author's intentions. How do you know at the time of annotation if a web article intends to deceive or is not intentional, and how was this annotation conducted?

Was there a good level of agreement of annotators particularly for the categories of disinformation vs misinformation, or did you observe several cases of mistakes between these two categories?

Throughout the thesis, the two main categories studied are disinformation and credible. However, as stated above, disinformation refers to factuality and credible refers to perception. These two labels aren't mutually exclusive, as disinformation can be credible, and information that is accurate may not be necessarily credible in principle, even if it's true. How would you justify the choice of disinformation vs credible as your labels?

Queen Mary
University of London

Despite making a deep reflection on the achievements of the thesis, the contribution made, as well as the avenues for future research that the thesis has opened, it lacks a proper discussion of the limitations of the research. What would you highlight as some of the key limitations of the research in your thesis?

Final conclusion:

All in all, I consider that the thesis presents a novel and thorough investigation into an important and timely topic, in a well written and critically analysed and presented thesis. This has been validated through publication in top-tier research venues demonstrating its quality. Overall, I consider that the thesis satisfies the requirements to proceed with the oral examination.

Sincerely,

Dr. Arkaitz Zubiaga

Associate Professor and Director of Graduate Studies (Research)