

# Classifying and Mining Disinformation Narratives with Large Language Models

mgr inż. Witold Sosnowski

This thesis advances the study of disinformation by moving from claim-level analysis to narrative-level understanding, proposing and evaluating LLM-based methods for benchmarking, mining, and analysing disinformation narratives across Europe. This dissertation investigates how Large Language Models (LLMs) can be used to identify, classify, and discover disinformation narratives. Unlike traditional approaches that focus on individual pieces of content (e.g., claims, posts, or articles), the research shifts attention to the narrative level, enabling a broader understanding of how disinformation develops, spreads, and is framed across different sources and over time.

The work introduces three key contributions:

- EUDisinfoTest – a benchmark for evaluating whether LLMs can distinguish disinformation narratives from credible narratives while assessing their robustness to different rhetorical strategies (Logos, Ethos, and Pathos).
- DiNaM (Disinformation Narrative Mining) – a large-scale pipeline that mines and maps disinformation narratives from fact-checking datasets across Europe by grouping related false claims and generating narrative representations.
- DiNO (Disinformation Narrative mining in news outlets) – an outlet-level framework that identifies and analyzes disinformation narratives within news coverage, enabling comparison of narrative prevalence and framing across media outlets and over time.

The findings demonstrate that LLM-based approaches can effectively recover coherent disinformation narratives at both global and media-outlet levels. However, the results also reveal that LLM judgments remain sensitive to rhetorical framing, suggesting that persuasive language can influence model decisions even when the underlying facts remain unchanged.

The thesis is well organized and articulated across seven chapters. The introductory chapter provides a strong motivation for the research and outlines the thesis objectives, scope, and principal contributions. Chapter 2 offers a thorough review of the relevant literature and successfully identifies significant gaps in existing research. Chapter 3 introduces and clearly defines the key concepts and theoretical foundations that guide the remainder of the study.

Chapters 4, 5, and 6 constitute the core contribution of the thesis and have resulted in publications at highly competitive Core A\* venues, including EMNLP and ACL. Chapter 4 introduces EUDisinfoTest, a large-scale benchmark grounded in expert-curated European disinformation narratives and presents a systematic empirical investigation of how contemporary Large Language Models (LLMs) respond to narrative-level disinformation expressed through different rhetorical strategies. The study identifies consistent biases and failure modes in model behaviour, demonstrating that rhetorical devices such as appeals to authority (*ethos*) and emotional framing (*pathos*) can significantly influence model judgements even when the underlying factual content remains unchanged. These findings make an important contribution to the growing literature on

the risks, vulnerabilities, and limitations of LLMs, while also providing valuable insights into the assessment of model robustness in the presence of sophisticated disinformation narratives.

Chapter 5 presents DiNaM, a novel framework for disinformation narrative mining from fact-checking articles. DiNaM aggregates instances of false information across a diverse range of fact-checking organisations and topics to construct a comprehensive view of the disinformation narrative landscape. Through this approach, the thesis identifies major narratives that recur across countries, domains, and time periods, providing valuable insights into the propagation and evolution of disinformation. The proposed method is rigorously evaluated against existing general-purpose narrative mining approaches and demonstrates consistent improvements over these baselines, recovering more coherent, meaningful, and actionable disinformation narratives. This work represents a significant methodological contribution, offering a scalable and effective approach for understanding disinformation at the narrative level.

Extending narrative mining beyond fact-checking corpora, the thesis proposes DiNO, an outlet-level instantiation of the narrative mining framework for news articles (Chapter 6). Central to DiNO is a dedicated method for extracting false claims from news content, which consists of three stages: (i) segmenting articles into candidate text spans, (ii) matching these spans against large collections of verified false claims, and (iii) applying entailment-based verification to retain only those segments that genuinely support a false claim. By combining verified false claims with large collections of news articles, DiNO produces high-precision anchors for identifying and mapping disinformation narratives within specific news outlets. The approach is specifically designed for the news media and therefore complements the fact-checking-oriented extraction framework (DiNaM) presented in previous chapter. Experimental results, including comparisons with narrative-mining baselines and expert assessments, demonstrate that DiNO reliably recovers coherent and meaningful disinformation narratives from news outlets. This work represents an important extension of the overall narrative mining framework, enabling the detection and analysis of disinformation narratives beyond curated fact-checking resources and within real-world news ecosystems.

This thesis addresses an important and timely research challenge. Given the significant societal harm caused by disinformation, it advances the state of the art by focusing on the identification and analysis of disinformation at the narrative level, rather than at the level of individual claims. This perspective provides a more comprehensive understanding of how disinformation is constructed, disseminated, and sustained across information ecosystems. In addition, I believe that this thesis will stimulate a substantial body of follow-up research, particularly once the associated code, datasets, and resources are made publicly available. The candidate's contributions are timely, methodologically rigorous, and of significant relevance to both the information retrieval and natural language processing communities. Overall, the thesis makes a strong and original contribution to knowledge and demonstrates the candidate's ability to conduct independent, high-quality research. I am therefore very pleased to support the award of the degree of Doctor of Philosophy (PhD) (subject to satisfactory performance at the viva).

Joemon M Jose

Professor, School of Computing Science, University of Glasgow.