LLM Pipeline for Hierarchical Narrative Classification: A Traceable Approach to Multilingual Propaganda Detection

Anonymous submission

Abstract

This paper presents Agora, a configurable and reproducible framework for robust hierarchical multi-label classification using large language models. We address critical reliability challenges in LLM-based classification through a multi-agent ensemble approach with voting consensus and an optional actor-critic self-refinement loop. Our system decomposes the hierarchical classification task into manageable steps via a state-graph pipeline, providing enhanced traceability for multilingual propaganda detection. We evaluate our approach on hierarchical narrative classification benchmarks, demonstrating improved robustness and consistency compared to baseline LLM approaches.

Keywords: hierarchical classification, propaganda detection, narrative analysis, large language models, multilingual NLP

1. Introduction

Text Classification (TC) is a foundational task in Natural Language Processing (NLP) (Zangari et al., 2024). While traditional approaches often model TC as a single-label problem, real-world texts frequently contain multiple overlapping themes, motivating the use of Multi-Label Classification (MLC) (Hu et al., 2025; Tidake and Sane, 2018). A particularly challenging yet crucial variant is Hierarchical Multi-Label Classification (HMLC), where labels are organized in a predefined hierarchy (e.g., a tree or DAG) (Liu et al., 2023). This structure is common in domains requiring nuanced analysis, such as misinformation detection, where identifying nested propaganda narratives is a key challenge.

The advent of Large Language Models (LLMs) has opened new frontiers for HMLC, enabling powerful zero-shot classification without extensive labeled data. However, this power comes with significant reliability challenges. LLMs are known to be stochastic, producing different outputs for the same input, and often exhibit low instruction fidelity, failing to consistently adhere to complex hierarchical constraints or output formats (Qin et al., 2024). These limitations hinder their deployment in high-stakes applications where robustness and verifiable reasoning are paramount.

To address these shortcomings, we introduce Agora, a multi-agent ensemble framework that significantly improves the robustness and accuracy of Large Language Models on complex Hierarchical Multi-Label Classification tasks. By aggregating parallel classifications from multiple LLM agents via voting, Agora mitigates the inherent stochasticity of LLMs and produces more reliable results than a standard single-agent approach.

2. Related Work

The application of Large Language Models (LLMs) to Hierarchical Multi-Label Classification (HMLC) has rapidly matured, moving beyond complex supervised architectures that explicitly encode label hierarchies with Graph Neural Networks (Zhou et al., 2020) towards more flexible zero-shot paradigms (Wang et al., 2023). The recent SemEval-2025 Task 10 on multilingual narrative detection serves as a clear benchmark for the current state-of-the-art. Top-performing systems demonstrate the viability of zero-shot LLM approaches, utilizing techniques such as specialized prompting strategies and retrieval-augmented generation to handle the two-level taxonomy (Singh et al., 2025; Younus and Qureshi, 2025).

In our own prior work on this task, we introduced a modular "agentic" framework that decomposed the HMLC problem into a set of parallel binary decisions, with specialized LLM agents assigned to individual labels (Eljadiri and Nurbakova, 2025). While this and other systems have proven effective, they predominantly rely on the output of a single LLM agent or a single generative pass for their final predictions. This exposes them to the inherent stochasticity of LLMs, where identical inputs can yield different classifications across runs, posing a significant challenge for reliability. The use of ensembles to improve robustness is a well-established technique in machine learning (Read et al., 2021) and has been effective for fine-tuned Transformer models in similar propaganda detection tasks (Jurkiewicz et al., 2020). However, the systematic application of ensembling to mitigate the unreliability of modern, zero-shot LLM systems in a complex HMLC context remains underexplored. Our work directly addresses this gap, proposing a multi-agent ensemble framework that aggregates votes to produce a more stable, consensus-driven classification.

Acknowledgements

Place all acknowledgments here.

References

Mohamed Nour Eljadiri and Diana Nurbakova. 2025. Team INSALyon2 at SemEval-2025 task 10: A zero-shot agentic approach to text classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 965–980, Vienna, Austria. Association for Computational Linguistics.

W. Hu, Q. Fan, H. Yan, X. Xu, S. Huang, and K. Zhang. 2025. A survey of multi-label text classification under few-shot scenarios. *Applied Sciences*, 15:8872.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Rundong Liu, Wenhan Liang, Weijun Luo, Yuxiang Song, He Zhang, Ruohua Xu, Yunfeng Li, and Ming Liu. 2023. Recent advances in hierarchical multi-label text classification: A survey.

Tianyi Qin, Yaliang Wu, Yaliang Zhang, Weiliang Liu, Jundong Li, and Philip S. Yu. 2024. Infobench: A benchmark for information disorder detection in social media. *IEEE Transactions on Knowledge and Data Engineering*, 36(5):2132–2145.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. 2021. Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research*, 70:683–718.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2025. GateNLP at SemEval-2025 task 10: Hierarchical three-step prompting for multilingual narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 148–154, Vienna, Austria. Association for Computational Linguistics.

Vaishali S. Tidake and Shirish S. Sane. 2018. Multi-label classification: a survey. *International Journal of Engineering & Technology*, 7(4.19):1045.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. 2023. Text2Topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 93–103, Singapore. Association for Computational Linguistics.

Arjumand Younus and Muhammad Atif Qureshi. 2025. nlptuducd at SemEval-2025 task 10: Narrative classification as a retrieval task through story embeddings. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1742–1746, Vienna, Austria. Association for Computational Linguistics.

A. Zangari, M. Marcuzzo, M. Rizzo, L. Giudice, A. Albarelli, and A. Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7):1199.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.