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

2.1. Traditional Multi-Label Classification Approaches

The core challenge in Multi-Label Classification (MLC) has historically been the effective modeling of inter-label dependencies. Early problem-transformation methods sought to adapt single-label algorithms to this task. Approaches ranged from Binary Relevance, which simplifies the problem by training an independent classifier for each label but ignores correlations, to Classifier Chains, which attempt to capture dependencies by feeding the predictions of one classifier as features to the next in a sequence (Zhang et al., 2018; Read et al., 2011). While foundational, these methods were often superseded by deep learning models, particularly Transformers, which could learn complex label correlations implicitly from large datasets (Devlin et al., 2019).

2.2. Hierarchical Multi-Label Classification

For the more specific task of Hierarchical Multi-Label Classification (HMLC), research focused on creating architectures that explicitly leverage the label taxonomy. A dominant paradigm involved coupling a Transformer-based text encoder with a Graph Neural Network (GNN) that operates on the label graph, allowing the model to learn hierarchy-aware representations and enforce structural consistency (Zhou et al., 2020; Xu et al., 2021). However, the advent of Large Language Models (LLMs) has again shifted the landscape, enabling powerful zero-shot HMLC capabilities that bypass the need for complex, task-specific architectures and extensive supervised training (Wang et al., 2023).

2.3. Zero-Shot LLM Approaches and Current Challenges

The recent SemEval-2025 Task 10 on narrative detection serves as a clear benchmark for this modern, LLM-driven approach. State-of-the-art systems effectively use techniques like retrieval-augmented generation and specialized prompting to classify texts within the task's two-level hierarchy (Singh et al., 2025; Younus and Qureshi, 2025). In our own prior work, we introduced a modular "agentic" framework that decomposed the task into parallel binary decisions (Eljadiri and Nurbakova, 2025). Yet, these pioneering systems share a common vulnerability: their reliance on a single, non-deterministic generative pass, which exposes them to the inherent stochasticity of LLMs. While ensembling is a classic technique for improving robustness in supervised models (Jurkiewicz et al., 2020), its systematic application to mitigate the unreliability of zero-shot LLMs in HMLC remains underexplored. Our work directly addresses this gap, proposing a multi-agent ensemble framework to produce stable, consensus-driven classifications.

Acknowledgements

Place all acknowledgments here.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

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.

Linli Xu, Sijie Teng, Ruoyu Zhao, Junliang Guo, Chi Xiao, Deqiang Jiang, and Bo Ren. 2021. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2459–2468, Online and Punta Cana, Dominican Republic. 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.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

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.