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.

3. System Architecture

Our investigation into a robust zero-shot HMLC framework followed an iterative design process. We began by analyzing a promising top-down prompting approach, then developed a traceable self-refining pipeline, and finally, upon identifying its limitations, engineered a more robust multi-agent ensemble.

3.1. Baseline Inspiration: Top-Down Zero-Shot Classification

A promising recent approach to HMLC, which we refer to as H3Prompting, was introduced by Singh et al. (2025). This method decomposes the hierarchical classification task into a sequence of zero-shot LLM queries, where an initial prompt determines the top-level category, and subsequent prompts classify second-level narratives conditioned on the first-level prediction. In the spirit of reproducible research, we first aimed to replicate this method. However, the lack of publicly available implementation details prevented a direct reproduction, motivating our development of an open and configurable framework to systematically investigate zero-shot HMLC strategies.

3.2. Initial Approach: The Traceable Actor-Critic Pipeline

Our first architectural exploration, the Actor-Critic Pipeline, sought to improve the reliability and transparency of a single agent's output. While zero-shot classification is powerful, simple label outputs lack traceability, making them difficult to audit and susceptible to unverified hallucinations. To overcome this, we designed a multi-stage pipeline that reframes the task from simple classification to evidence-grounded claim generation and validation.

Our implementation uses the LangGraph framework to structure the process as a cyclical state graph (LangChain AI, 2024a). The pipeline configuration is managed by a `ConfigurableGraphBuilder`, allowing components like validation to be enabled or disabled and different LLMs to be assigned to specific nodes via a YAML file. The workflow proceeds through several stages, as illustrated in Figure 1.

3.2.1. Category Classification

An initial node determines the document's high-level topic (e.g., CC or URW) to focus subsequent stages on the relevant subset of the narrative taxonomy.

3.2.2. The Narrative Actor (Claim Generation)

The "Actor" is an LLM agent prompted to act as an expert analyst. For each narrative it identifies, it must generate a structured JSON object containing the `narrative_name`, a verbatim `evidence_quote` from the text, and reasoning that connects the two.

3.2.3. The Narrative Critic (Claim Validation)

The Actor's output is passed to the "Critic," a separate LLM agent with a distinct, skeptical persona. The Critic validates the Actor's claims against strict criteria (evidence accuracy, relevance, completeness). If the Critic finds flaws, it generates structured feedback, and the graph's conditional logic routes the state back to the Actor for a retry, incorporating the feedback into a refinement prompt.

3.2.4. Sub-narrative Actor and Critic

Once narratives are approved, the process repeats hierarchically. A Sub-narrative Actor generates claims for each parent narrative, which are then validated by a Sub-narrative Critic, also with a self-correction loop.

This evidence-grounded architecture offers significant advantages in traceability and auditability, as every prediction is grounded in a textual `evidence_quote`, a process we enhanced using LangSmith for detailed debugging and prompt refinement (LangChain AI, 2024b). However, while this pipeline showed promise, our experiments revealed that it introduced its own form of instability. The process was highly sensitive to the quality of

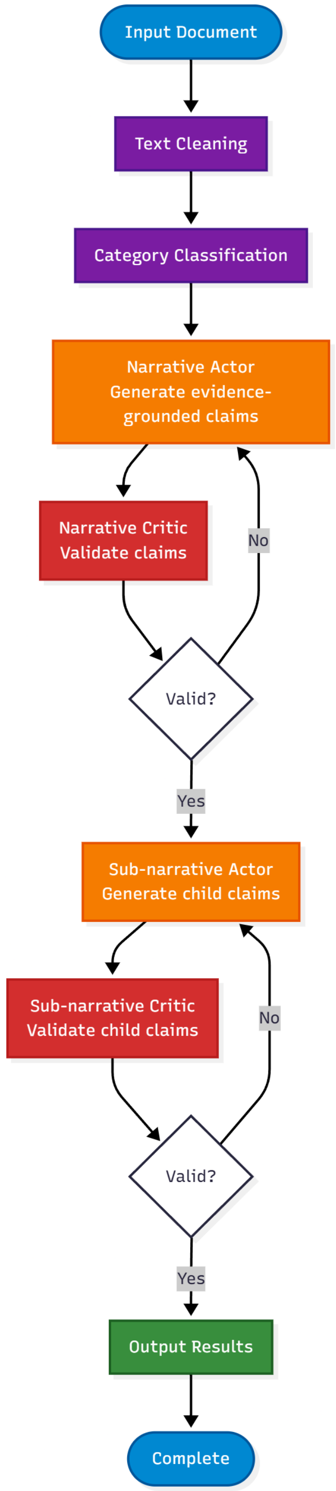

Figure 1: Actor-Critic validation pipeline architecture. The system decomposes HMLC into hierarchical stages where a Narrative Actor generates evidence-grounded claims, which are then validated by a Narrative Critic. Feedback loops enable iterative refinement before progressing to sub-narrative classification.

the Critic's feedback, which, being LLM-generated, was itself prone to stochasticity. A flawed critique could send the Actor down an unproductive refinement path, sometimes degrading performance. This reliance on a single, fallible "critic" highlighted a fundamental bottleneck and motivated our shift towards a more robust consensus mechanism.

3.3. Proposed Framework: The Agora Multi-Agent Ensemble

To overcome the limitations of a single-critic system and the broader issue of single-agent stochasticity, we developed Agora, a multi-agent ensemble framework. The core principle of Agora is to replace the judgment of a single agent (or a single critic) with the collective "wisdom of the crowd," leveraging the consensus of multiple independent agents to arrive at a more stable and accurate classification.

The framework operates via a fan-out, fan-in process for each level of the hierarchy (e.g., Narrative classification).

3.3.1. Fan-Out (Parallel Classification)

Instead of a single LLM call, Agora instantiates N independent agents (where N is a configurable parameter). The same input text and prompt are sent to all N agents, who perform the classification in parallel. This step effectively samples N independent points from the LLM's output distribution for the given task.

3.3.2. Fan-In (Aggregation via Voting)

After all N agents return their individual classifications, an Aggregation node consolidates the results using a voting scheme. We implemented and evaluated three distinct aggregation strategies:

- **Union:** The final label set includes any label proposed by at least one agent. This high-recall strategy is useful for identifying all potential labels.
- **Intersection:** The final label set includes only those labels that all agents unanimously agreed upon. This high-precision strategy filters out all but the most confident predictions.
- **Majority Vote:** The final label set includes any label proposed by more than half ($> N/2$) of the agents. This strategy provides a robust balance between precision and recall, filtering out stochastic, outlier classifications while retaining a strong consensus.

This ensemble approach directly addresses the single-point-of-failure issue observed in the Actor-Critic pipeline. Instead of relying on one critic's potentially noisy feedback, Agora relies on the statistical robustness of a majority decision, providing

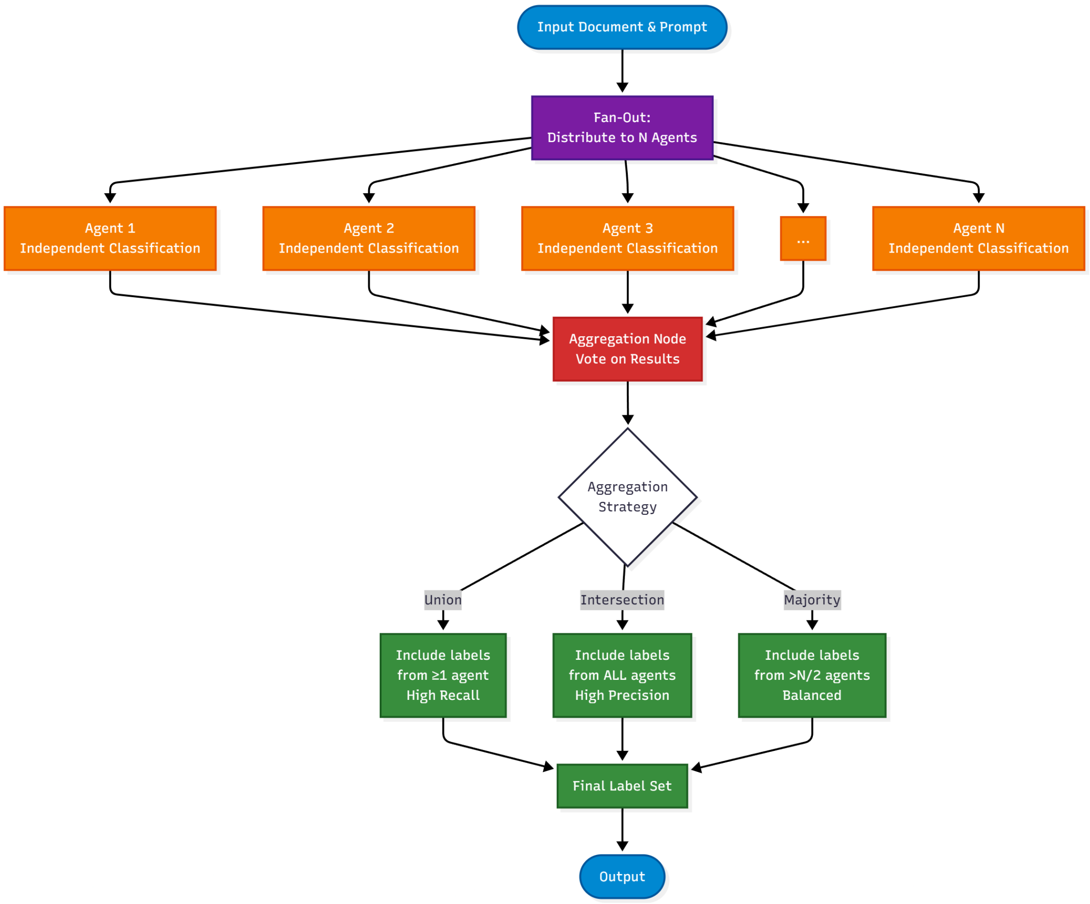

Figure 2: Agora multi-agent ensemble architecture. The framework distributes the classification task to N independent agents in parallel (Fan-Out), aggregates their results using a voting mechanism (Fan-In), and produces a final label set based on the chosen aggregation strategy (Union, Intersection, or Majority Vote).

a more reliable and conceptually simpler method for improving classification quality.

4. Experimental setup

4.1. Dataset and its Challenges

We use the dataset from the SemEval-2025 Task 10, Subtask 2, a task focused on multilingual propaganda narrative detection (Piskorski et al., 2025). While the full dataset spans five languages, our experiments focus on the English subset. A systematic analysis of the data reveals several key characteristics that motivate our architectural choices.

First, the dataset is fundamentally multi-label in nature. A majority of documents (54.0%) are assigned more than one narrative, with an average of 2.28 labels per document. This necessitates a multi-label modeling approach.

Second, and most critically, the dataset exhibits a severe class imbalance. As shown in Figure 3, the 22 narrative labels follow a classic long-tailed distribution. The most frequent narrative appears 65 times more often than the least frequent one. This sparsity poses a significant challenge for traditional supervised models, which risk overfitting on the few “head” classes and failing to generalize to the many rare but meaningful “tail” classes. This characteristic strongly motivates our adoption of a zero-shot paradigm, which does not depend on label frequencies in a training set.

4.2. System Configurations

To isolate the contributions of our different architectural choices, we define and compare three distinct zero-shot systems. All systems are implemented using the LangGraph framework (LangChain AI, 2024a) and share the same foundational prompts to ensure a fair comparison.

1. **Naive Baseline (Single-Pass Actor):** This configuration represents the most straightforward application of an LLM to the HMLC task. It consists of a single "Actor" agent that performs the classification for each hierarchical level. The output from the agent's first and

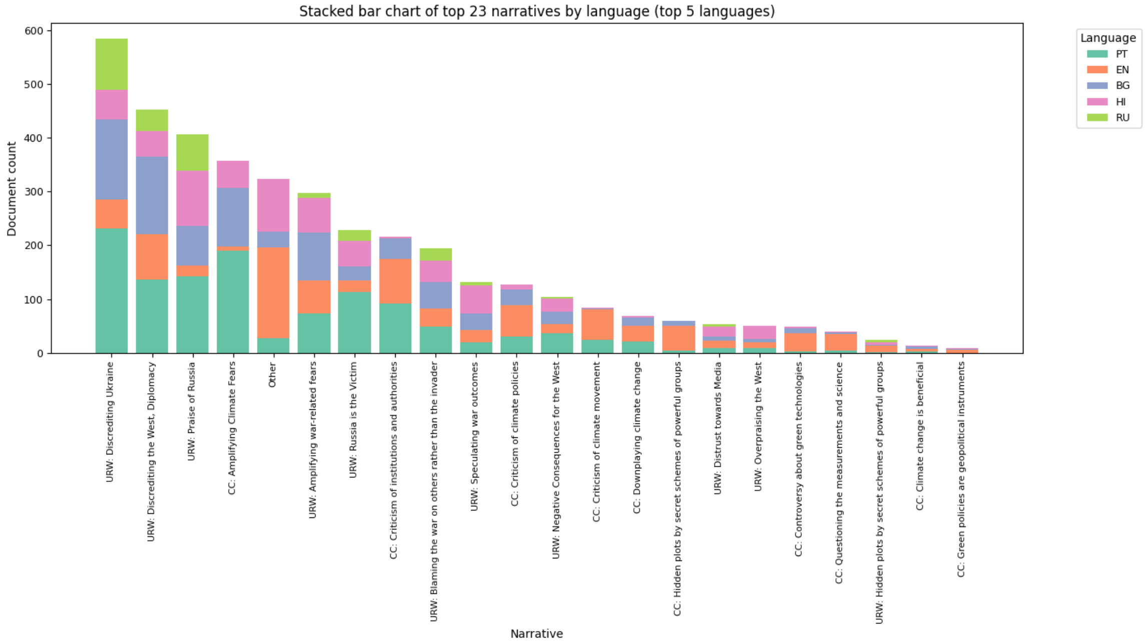

Figure 3: Long-tailed distribution of narrative labels. A small number of frequent narratives dominate the dataset, while most are rare. This severe imbalance justifies the use of zero-shot LLM approaches that do not rely on large per-class training counts.

only pass is taken as the final result. This system serves as our baseline to measure the performance of a standard, non-revising, and non-ensembled LLM classifier. For this configuration, we used gpt-5-nano as the LLM.

2. **Actor-Critic Pipeline:** This system represents our initial effort to improve reliability through self-refinement. It extends the Naive Baseline by adding a "Critic" agent that reviews and validates the Actor's output. If the Critic identifies a flaw, it provides feedback, and the Actor makes a second, revised attempt. This configuration allows us to measure the impact of a single, iterative correction loop. The Actor and Critic agents were both implemented using gpt-5-nano.

3. **Agora (Multi-Agent Ensemble):** This is our proposed framework, designed to achieve robustness through consensus. Instead of refinement, Agora employs a multi-agent ensemble where independent agents classify the text in parallel. Their outputs are then aggregated using a Majority Vote mechanism to determine the final label set. This configuration tests our primary hypothesis that consensus from multiple agents is more effective than the single-agent refinement loop. All agents in the Agora ensemble were implemented using gpt-5-nano.

4.3. Hardware Configuration

All experiments were conducted on a machine with the following specifications:

- **Processor:** Intel Core 9 Ultra
- **GPU:** NVIDIA RTX 4070 (8GB VRAM)
- **Memory:** 32GB RAM
- **Storage:** 1TB SSD

4.4. Evaluation Scope

We primarily evaluate the performance of our models on the English test set, providing detailed analysis and interpretation of the results on this language. However, to assess the multilingual capabilities of our approaches, we also conducted experiments on additional languages to validate the generalizability of our methods across different linguistic contexts.

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.

LangChain AI. 2024a. Langgraph: Build resilient language agents as graphs. Accessed: 2025-09-19.

LangChain AI. 2024b. Langsmith: Developer platform for building reliable llm applications. Accessed: 2025-09-19.

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.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval 2025)*, Vienna, Austria.

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.