Instructions for *ACL Proceedings

Nour Eljadiri
University of Passau
INSA Lyon
mohamed.eljadiri@insa-lyon.fr

Abstract

The task of assigning multiple, hierarchically-structured labels to text documents, known as Hierarchical Multi-Label Classification (HMLC), is critical in domains from scientific archiving to legal analysis. This review traces the methodological evolution of HMLC, beginning with foundational problem transformation methods like Binary Relevance and Classifier Chains, which primarily address the challenge of label correlation. We then examine the paradigm shift introduced by pre-trained Transformers, dissecting the dichotomy between local, top-down approaches prone to error propagation and global, hierarchy-aware models that integrate structural constraints via specialized loss functions or Graph Neural Networks. Finally, we explore the current frontier, where Large Language Models (LLMs) are reframing the task through generative paradigms, enabled by techniques such as LLM-powered data augmentation, instruction fine-tuning, and Parameter-Efficient Fine-Tuning (PEFT). This narrative highlights a progression towards increasingly sophisticated methods for embedding hierarchical prior knowledge into statistical models.

1 Introduction

Text Classification (TC) stands as one of the most foundational and widely researched tasks within the domain of Natural Language Processing (NLP). (Zangari et al., 2024) In its most common formulation, TC involves supervised learning algorithms designed to map a given piece of text, or document, to a predefined set of labels or categories. (Zangari et al., 2024) Historically, this has often been a multiclass problem, where each document is assigned to exactly one class from a set of mutually exclusive options. However, the increasing complexity and richness of information in modern digital text have rendered this single-label paradigm insufficient.(Hu et al., 2025) Many documents, from news articles and scientific papers to legal filings and product descriptions, simultaneously encompass multiple topics or themes. (Tidake and Sane, 2018)

This reality gave rise to Multi-Label classification (MLC), a more challenging variant of text classification where each data sample can be associated with one or multiple labels simultaneously.(Tidake and Sane, 2018) The core challenge in MLC, which distinguishes it from simply running multiple independent binary classifiers, is the presence of correlations between labels (Tidake and Sane, 2018), that is, the assignment of one label often provides strong statistical evidence for or against the assignment of another, and effectively modeling these inter-label dependencies has become a central focus of research in the field. (Huang et al., 2024; Tidake and Sane, 2018)

Formally, in MLC, the goal is to learn a function $f : X \to 2^L$ that maps an instance $x \in X$ to a subset of labels $Y \subseteq L$, where $L = \{l_1, l_2, \dots, l_L\}$ is the finite set of all possible labels. The number of labels associated with an instance is not fixed and can vary. (Tidake and Sane, 2018)

Hierarchical Multi-Label Text Classification (HMLC), the primary subject of this review, introduces a further layer of complexity and structure to the MLC problem. HMLC is defined as a classification task where instances may not only belong to multiple classes simultaneously, but where these classes are themselves organized within a predefined hierarchy

(Liu et al., 2023). This hierarchical structure, typically represented as a tree or a Directed Acyclic Graph (DAG), formalizes the relationships among the labels, arranging them from broader, coarse-grained categories at higher levels to more specific, fine-grained ones. (Liu et al., 2023)

This structured approach is particularly relevant for analyzing the sophisticated communication strategies found in online media. The rapid spread of online news has increased exposure to deceptive narratives and manipulation attempts, especially during major crisis events like geopolitical conflicts. To support research in this area, tasks such as SemEval-2025 Task 10 have been established, focusing on automated narrative classification (Piskorski et al., 2025). The goal is to categorize news articles by assigning them multiple labels from a two-level taxonomy of predefined narratives and subnarratives. Addressing this HMLC problem requires models that can understand nuanced content while respecting the explicit hierarchical dependencies between labels.

The task of narrative detection in text is an ideal application for HMLC. Narratives are structured frameworks of meaning that shape the interpretation of events and issues. A single text can invoke multiple, often nested, narratives. For example, a news report on an international incident might simultaneously employ a broad "National Security" narrative, a more specific "Foreign Aggression" sub-narrative, and a granular "Economic Sanctions" micro-narrative. An HMLC framework can model this structure, capturing both the multiple narrative elements present and their hierarchical relationships.

This review will trace the methodological evolution of HMLC. We begin by examining foundational problem transformation techniques such as Binary Relevance (Zhang et al., 2018b), Classifier Chains (Li et al., 2024; Weng et al., 2020), and the Label Powerset method (Shan et al., 2018). We then transition to the current state-of-the-art, dominated by deep learning models leveraging Transformer architectures like BERT (Devlin et al., 2019) and its multilingual variants such as XLM-RoBERTa (Conneau et al., 2020). Finally, we will cover specialized strategies like hierarchical classification models (Sadat and Caragea, 2022) and graph-based methods (Peng et al., 2021; Gong et al., 2020) designed to explicitly model the structured taxonomies inherent to HMLC.

2 Foundational paradigms of MLC

The main challenge in multi-label classification (MLC) was how to adapt algorithms designed for single-label (binary or multiclass) problems to handle multiple labels per instance. The core issue these methods faced was the presence of label correlations: in real-world data, labels are often not independent. For example, a news article tagged "Politics" is much more likely to also be tagged "Elections" than "Sports." Treating each label as a separate binary problem ignores these dependencies, potentially leading to suboptimal predictions.

Classical MLC methods can be seen as different strategies for balancing computational simplicity with the need to model label correlations. Some methods, like Binary Relevance, treat each label independently for simplicity, but this can miss important relationships between labels. Others, such as Classifier Chains or Label Powerset, explicitly model these correlations, but at the cost of increased computational complexity. The choice of method reflects a trade-off between efficiency and the ability to capture the true structure of the data.

To address this, early research focused on a family of techniques known as problem transformation methods, which decompose the multi-label task into one or more single-label problems. These methods are algorithm-independent, allowing any standard classifier to be applied. The three canonical approaches represent distinct strategies for managing label dependencies.

2.1 Binary Relevance

Binary Relevance (BR) is the most intuitive approach. It decomposes the MLC problem with a label set of size $|\mathcal{L}|$ into independent binary classification problems. For each label, a separate classifier is trained to predict its presence or absence, effectively ignoring all other labels. (Zhang et al., 2018b) The primary advantage of BR is its simplicity and efficiency, as the classifiers can be trained in parallel. Its main drawback, however, is the foundational label independence assumption, which completely disregards label correlations and can lead to lower predictive accuracy and logically incoherent label set predictions. (Sucar et al., 2014)

2.2 Label Powerset

In contrast to decompositional methods, the Label Powerset (LP) method reframes the entire problem at once. It converts the multi-label task into a standard multi-class problem by mapping each distinct set of co-occurring labels to a single, unique class. For instance, the label sets 'Politics, Elections' and 'Sports, Weather' would become two separate classes for a multi-class classifier to learn. (Read et al., 2011)

The main advantage of this approach is its ability to perfectly model the dependencies between labels for all combinations it has seen, as these correlations are baked into the class definitions (Sucar et al., 2014). However, this strategy is often impractical. The number of potential classes can become unmanageably large as the label set grows, a problem known as combinatorial explosion. This leads to a highly sparse class distribution where many label sets appear only a few times, making it difficult to train a robust model (Cherman et al., 2011). Critically, the LP method cannot generalize to predict any combination of labels that did not appear in the training data.

2.3 Classifier Chains

The Classifier Chains (CC) method was proposed as a novel approach to overcome the stark trade-off between the label-independent BR and the computationally explosive LP. It seeks to model label dependencies while maintaining the efficiency of a binary relevance framework (Read et al., 2011).

Like BR, CC trains $|\mathcal{L}|$ binary classifiers. However, instead of being independent, these classifiers are linked in a chain. The first classifier in the chain, $\mathcal{C}_1$, predicts the presence or absence of the first label. The predictions of this classifier are then used as additional features for the second classifier, $\mathcal{C}_2$, which predicts the second label. This process continues down the chain, with each classifier potentially benefiting from the predictions of all previous classifiers. This allows CC to capture label dependencies while still being relatively efficient to train. (Read et al., 2011) The order of the chain can significantly impact performance, as earlier classifiers influence later ones (Read et al., 2021). To mitigate this, ensemble methods that average predictions over multiple random chain orders are often employed (Sucar et al., 2014; Zhang et al., 2018a).

3 Hierearchy aware architectures

The limitations inherent in classical problem transformation methods, particularly their struggles with large label spaces and their inability to deeply integrate structural information, precipitated a paradigm shift toward deep learning. We will focus on how neural architectures evolved from treating the label hierarchy as a post-hoc constraint to using it as a central component of the learning process. This evolution is characterized by a fundamental architectural divergence between local and global approaches, a conflict that was ultimately resolved through the powerful synthesis of Transformer-based text encoders and Graph Neural Networks (GNNs) for structure encoding.

3.1 From flat to hierarchical models

Early deep-learning methods for hierarchical multi-label classification often treated the task as a standard (flat) multi-label problem and ignored the label taxonomy. While these models learned strong text representations, they did not use the hierarchy's relationships. That matters because mistakes at higher levels of the taxonomy are semantically worse than small, nearby errors: for example, assigning a paper on "Quantum Mechanics" to "Arts" is much more serious than confusing "Physics" with "Chemistry." Flat models cannot distinguish these degrees of error. (Xu et al., 2021)

The paradigm shift occurred with the recognition that the hierarchy could be treated as a feature to guide learning, rather than just an output format. By making models "hierarchy-aware," it becomes possible to share statistical strength between parent and child nodes. For example, the few training instances available for a rare, specific label like "Superstring Theory" can be supplemented by the more abundant data from its parent labels, "String Theory" and "Theoretical Physics." This is especially crucial for improving performance on the long tail of infrequent labels that characterizes most real-world HMLC datasets. (Zangari et al., 2024)

3.2 Local vs global approaches

The first generation of truly hierarchical models diverged into two main architectural philosophies: local and global. This division reflects a fundamental trade-off between capturing fine-grained, localized class relationships and maintaining a holistic, computationally tractable view of the entire label space.

3.2.1 Local approaches (Top down)

The local approach decomposes the hierarchical classification problem into a set of smaller, more manageable classification tasks distributed across the taxonomy. This is typically implemented as a top-down or "divide and conquer" strategy.

During inference, an instance is typically classified in a top-down manner. It is first evaluated by the classifier at the root; if a positive prediction is made for a node, the instance is then passed down to the classifiers of its children, and this process continues until a leaf node is reached or no further positive predictions are made.(Romero et al., 2022). While this approach excels at capturing the specific features that distinguish between closely related sibling classes, it suffers from a critical problem: **error propagation.** A single misclassification at a higher level of the hierarchy can irreversibly steer the prediction down an incorrect path, making it impossible to classify the instance into its correct, more specific sub-categories. (Wehrmann et al., 2018)

3.2.2 Global approaches (single classifier)

In contrast, the global approach uses a single, unified model to predict all labels in the hierarchy simultaneously. This is typically framed as a large multi-label classification problem where the output layer corresponds to the entire set of labels in the taxonomy. (Wehrmann et al., 2018)

The primary advantage of the global approach is that it inherently avoids the error propagation problem of local models, as all decisions are made in parallel by a single classifier. This makes the model more robust to errors at higher levels. Furthermore, global models are often more computationally efficient, as they require training only one model instead of a potentially large cascade of local classifiers. However, early global models faced a significant challenge: they struggled to effectively incorporate the complex structural information of the entire hierarchy into a single model. By treating the problem as a flat multi-label task, they often failed to capture the nuanced, local distinctions between sibling classes and could underfit the hierarchical relationships, thereby losing the very information that hierarchical classification aims to exploit. (Wehrmann et al., 2018; Zhou et al., 2020)

3.3 Encoding the Hierarchy with Transformers and Graph Neural Networks

In recent years a common, simple pattern has emerged and is widely used in hierarchy-aware text models: use a Transformer (or other strong text encoder) to build contextual representations for the instance, and use a Graph Neural Network to encode the hierarchy or other structural information (labels, label co-occurrence, or corpus/document graphs). The two modules play complementary roles. The Transformer captures rich, local and contextual features from the text, while the GNN injects global relational signals about the label taxonomy (or the corpus) so that information can flow between related labels or documents. At training time these components are typically joined so that the text encoder and the graph encoder are learned end-to-end: the Transformer produces node features, the GNN propagates structural context, and the final classifier combines both signals to make multi-label predictions.

The practical upshot for hierarchical multi-label classification is simple: keep the powerful Transformer encoder for the instance-level signal, model the labels and their relations with a compact graph, and let a GNN mediate the exchange. This combination is robust to error propagation (compared to pure top-down local models) while still preserving the ability to make fine, label-specific distinctions.

References

E. A. Cherman, M. C. Monard, and J. Metz. 2011. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

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.

Jibing Gong, Hongyuan Ma, Zhiyong Teng, Qi Teng, Hekai Zhang, Linfeng Du, Shuai Chen, Md Zakirul Alam Bhuiyan, Jianhua Li, and Mingsheng Liu. 2020. Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification. *IEEE Access*, 8:30885–30896.

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.

S. Huang, W. Hu, B. Lu, Q. Fan, X. Xu, X. Zhou, and H. Yan. 2024. Application of label correlation in multi-label classification: A survey. *Applied Sciences*, 14:9034.

Xinyu Li, Jiaman Ding, and Shuang Hu. 2024. Relative entropy and pagerank-based classifier chains for multi-label classification. *IEEE Access*, 12:87665–87674.

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. *Preprint*, arXiv:2307.16265.

Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip S. Yu, and Lifang He. 2021. Hierarchical taxonomy-aware and attentional graph capsule rcnns for large-scale multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2505–2519.

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.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

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

Miguel Romero, Jorge Finke, and Camilo Rocha. 2022. A top-down supervised learning approach to hierarchical multi-label classification in networks. *Applied Network Science*, 7(1):8.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jincheng Shan, Chenping Hou, Wenzhang Zhuge, and Dongyun Yi. 2018. Co-learning binary classifiers for lp-based multi-label classification. In Yuxin Peng, Kai Yu, Jiwen Lu, and Xingpeng Jiang, editors, *Intelligence Science and Big Data Engineering*, volume 11266, pages 443–453. Springer International Publishing, Cham.

L. Sucar, C. Bielza, E. Morales, Pablo Hernandez-Leal, Julio H. Zaragoza, and P. Larrañaga. 2014. Multi-label classification with bayesian network-based chain classifiers. *Pattern Recognit. Lett.*, 41:14–22.

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

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. 2018. Hierarchical multi-label classification networks. In *Proceedings of the 35th International*

Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084. PMLR.

Wei Weng, Da-Han Wang, Chin-Ling Chen, Juan Wen, and Shun-Xiang Wu. 2020. Label specific features-based classifier chains for multi-label classification. *IEEE Access*, 8:51265–51275.

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.

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.

M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng. 2018a. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018b. 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.