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 $L = \{l_1, l_2, \ldots, l_L\}$ 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

The solution to the local-versus-global problem came from combining two methods: Transformer models to understand text, and Graph Neural Networks (GNNs) to understand structure. This mix made it possible to build global models that are both efficient and aware of hierarchies. (Wang et al., 2024; Zhou et al., 2020)

GNNs are especially useful for working with label hierarchies. (Li et al., 2021) We can represent the taxonomy as a graph, where labels are nodes and parent–child links are edges. A GNN learns label representations by passing messages between connected nodes. Each node gathers information from its parents, children, and siblings. This way, the final label embeddings capture not just their own meaning, but also their place and relationships within the whole taxonomy.

A seminal work in this domain is the Hierarchy-Aware Global Model (HiAGM) by Zhou et al. (2020). HiAGM provides a blueprint for this new class of models. Its architecture consists of two main components: (Zhou et al., 2020)

- A Text Encoder (e.g., BERT, RoBERTa) that generates a powerful contextualized representation of the input document.
- A Hierarchy-Aware Structure Encoder (e.g., a Bidirectional Tree-LSTM or a specialized Hierarchy-GCN) that operates on the label graph to produce hierarchy-aware label embeddings. This encoder models dependencies in both a top-down and bottom-up fashion, allowing information to flow in both directions across the hierarchy.

HiAGM further proposes two distinct fusion variants to combine text and structure features depending on the inference regime and efficiency/inductivity trade-offs.

HiAGM-LA (Multi-Label Attention) An inductive approach that uses a multi-label attention mechanism to compute document-specific label representations by attending from the text encoder outputs to the hierarchy-aware label embeddings. Because the attention weights are computed per document, the model can generalize to unseen instances without explicitly re-running a graph propagation step for each input — label embeddings can be pre-computed and stored, and the attention operation is applied at inference time.

HiAGM-TP (Text Feature Propagation) A deductive approach that propagates text features across the label hierarchy via graph propagation: document features are attached to label nodes and a GNN is run to diffuse these features through the taxonomy. This requires executing the GNN at inference time for each instance (or batch), which can be more computationally expensive but allows richer instance-specific propagation of evidence through the hierarchy.

The principles demonstrated by HiAGM have been extended and refined in subsequent work. For instance, Xu et al. (2021) proposed a framework using a loosely coupled GCN to explicitly model not only the vertical correlations (parent-child dependencies) but also the horizontal correlations (relationships between sibling nodes at the same level). This allows the model to capture, for example, that "Computer Science" and "Electrical Engineering" are more closely related to each other than to "History," even if they share the same parent 'Science & Technology'.

4 The rise of Large Language Models

The advent of Large Language Models (LLMs) has initiated another profound shift in the NLP landscape. Their advanced reasoning and generation capabilities are redefining their role in the HMLC pipeline, moving them beyond being simple end-point classifiers to becoming integral co-pilots in various stages of the workflow.

4.1 Zero shot and few-shot classification

The most immediate impact of LLMs is their remarkable ability to perform classification tasks with little to no task-specific training data. Through in-context learning, an LLM can be prompted with a task description and a few examples (few-shot) or even no examples (zero-shot) and perform hierarchical multi-label classification. This capability represents a paradigm shift for low-resource scenarios, potentially obviating the need for extensive data collection and annotation efforts; for narrative detection, one can provide definitions of the target narratives and ask the LLM to classify a new document accordingly. (Wang et al., 2023)

For instance, Eljadiri and Nurbakova (2025) demonstrate a high-performing zero-shot agentic approach to narrative classification. Their system instantiates multiple specialized LLM agents, each responsible for a binary decision on a single narrative or subnarrative label, while a meta-agent aggregates the binary outputs into final multi-label predictions. The agents are orchestrated with AutoGen and operate without task-specific fine-tuning, which enables parallel detection across the two-level taxonomy; this design yielded competitive results on the SemEval-2025 test set and highlights the practical zero-shot LLM systems for complex hierarchical tasks. (Eljadiri and Nurbakova, 2025)

4.2 LLMs for data augmentation

One of the most promising applications of LLMs is in addressing data scarcity and class imbalance through synthetic data generation. An LLM can be prompted to act as a domain expert and generate new text samples for underrepresented (tail) classes. This can be done by paraphrasing existing examples to increase diversity or by generating entirely new, plausible examples from scratch. This approach offers a sophisticated and flexible alternative to traditional data augmentation techniques like synonym replacement or back-translation. (Cegin et al., 2025; Glazkova and Zakharova, 2024)

In more extreme cases; for example, a scenario with complete lack of labeled data, an LLM can be used to generate a large corpus of "weak" or "silver" labels. These labels, while imperfect, can serve as a starting point for training a smaller, more efficient model. Frameworks like JSDRV by Yang et al. (2024) take this a step further, using a reinforcement learning policy to select the highest-quality LLM-generated annotations for fine-tuning, creating a self-improving annotation and training pipeline (Yang et al., 2024).

4.3 LLMs as evaluation assistants

For extreme multi-label classification tasks with thousands of labels, manual evaluation is prohibitively expensive and time-consuming. Li et al. (2023) aimed to compare a label-ranking BiCross-Encoder against a SciBERT classifier in a very large-label setting (11,486 labels) while lacking a ready human-annotated test set. They demonstrated a practical approach by asking ChatGPT to rate the relevance of candidate labels: for each document they fed the top-10 model-predicted labels to ChatGPT and requested a 3-point relevance score (0 = irrelevant, 1 = somewhat relevant, 2 = highly relevant) along with brief explanations. The authors treated these LLM judgments as automatic annotations to compare models, then spot-checked a sample with subject-matter experts and reported about 60% agreement on the 3-point scale (rising to 82% when collapsing 1 and 2 into a single "relevant" class). They conclude that using ChatGPT as an evaluation assistant is a cost- and time-efficient way to bootstrap evaluation when comprehensive human annotation is infeasible. (Li et al., 2023)

5 Focus on data-specific challenges

While architectural innovations have driven significant progress, the performance of any HMLC model is ultimately constrained by the quality and characteristics of the available data. Real-world datasets for narrative detection and other complex classification tasks are rarely pristine; they are often imbalanced, small, and increasingly, multilingual.

5.1 Class imbalance and tail labels

Perhaps the most pervasive challenge in real-world classification is extreme class imbalance, often described as a long-tailed distribution. In such datasets, a small number of "head" classes are represented by a vast number of training examples, while the majority of "tail" classes are represented by very few, sometimes single-digit, examples. (Huang et al., 2021) This constitutes a severe challenge for standard training algorithms, which, when optimized for overall accuracy, tend to become biased towards the majority classes and perform poorly on the infrequent but often more interesting tail classes.

The inadequation of conventional resampling

In single-label classification, a common strategy to combat imbalance is data resampling—either oversampling the minority classes or undersampling the majority classes (Luque et al., 2019; Thabtah et al., 2020). However, this approach is fundamentally ill-suited for the multi-label context. An instance in an MLC dataset can simultaneously belong to multiple classes, for example, one highly frequent head class and one extremely rare tail class. If one were to oversample this instance to increase the representation of the tail class, one would inadvertently also increase the representation of the already-dominant head class, potentially exacerbating the overall imbalance with respect to other labels. This entanglement makes simple resampling ineffective and often counterproductive. (Yuan et al., 2024)

6 Conclusion

The evolution from traditional fine-tuning to prompt-tuning and the integration of LLMs signifies a fundamental shift in the core research objective within HMLC. The focus is moving away from pure representation learning toward task alignment and holistic system design. The research questions are becoming more systemic: "How can we use an LLM's reasoning abilities to generate the very data we need for training?" or "How can we leverage an LLM to automate the costly process of evaluation?". This represents a higher level of abstraction, focusing on aligning the downstream task with the model's pre-existing knowledge.

References

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. LLMs vs established text augmentation techniques for classification: When do the benefits outweight the costs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10476–10496, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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.

Anna V. Glazkova and Olga V. Zakharova. 2024. Evaluating llm prompts for data augmentation in multi-label classification of ecological texts. *Preprint*, arXiv:2411.14896.

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.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Li, Zi Long Zhu, Janneke van de Loo, Agnes Masip Gomez, Vikrant Yadav, Georgios Tsatsaronis, and Zubair Afzal. 2023. Enhancing extreme multi-label text classification: Addressing challenges in model, data, and evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 313–321, Singapore. Association for Computational Linguistics.

Irene Li, Tianxiao Li, Yixin Li, Ruihai Dong, and Toyotaro Suzumura. 2021. Heterogeneous graph neural networks for multi-label text classification. In *Lecture Notes in Computer Science*. ArXiv:2103.14620.

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.

Amalia Luque, A. Carrasco, Alejandro Martín, and A. D. L. Heras. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.*, 91:216–231.

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.

F. Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. 2020. Data imbalance in classification: Experimental evaluation. *Inf. Sci.*, 513:429–441.

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.

Kunze Wang, Yihao Ding, and Soyeon Caren Han. 2024. Graph neural networks for text classification: A survey. *Preprint*, arXiv:2304.11534.

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.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13423–13439, Bangkok, Thailand. Association for Computational Linguistics.

Ling Yuan, Xinyi Xu, Ping Sun, Haiping Yu, Yin Zhen Wei, and Junjie Zhou. 2024. Research of multi-label text classification based on label attention and correlation networks. *PLOS ONE*, 19(9):e0311305.

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.