Revisiting Hierarchical Text Classification: Inference and Metrics

Roman Plaud[1,2], Matthieu Labeau[1], Antoine Saillenfest[2], Thomas Bonald[1]
[1] Institut Polytechnique de Paris
[2] Onepoint, 29 rue des Sablons, 75016, Paris, France
{roman.plaud, matthieu.labeau, thomas.bonald}@telecom-paris.fr
a.saillenfest@groupeonepoint.com

Abstract

Hierarchical text classification (HTC) is the task of assigning labels to a text within a structured space organized as a hierarchy. Recent works treat HTC as a conventional multilabel classification problem, therefore evaluating it as such. We instead propose to evaluate models based on specifically designed hierarchical metrics and we demonstrate the intricacy of metric choice and prediction inference method. We introduce a new challenging dataset and we evaluate recent sophisticated models against a range of simple but strong baselines, including a new theoretically motivated loss. Finally, we show that those baselines are very often competitive with the latest models. This highlights the importance of carefully considering the evaluation methodology when proposing new methods for HTC. Code implementation and dataset are available at https://github.com/RomanPlaud/revisitingHTC.

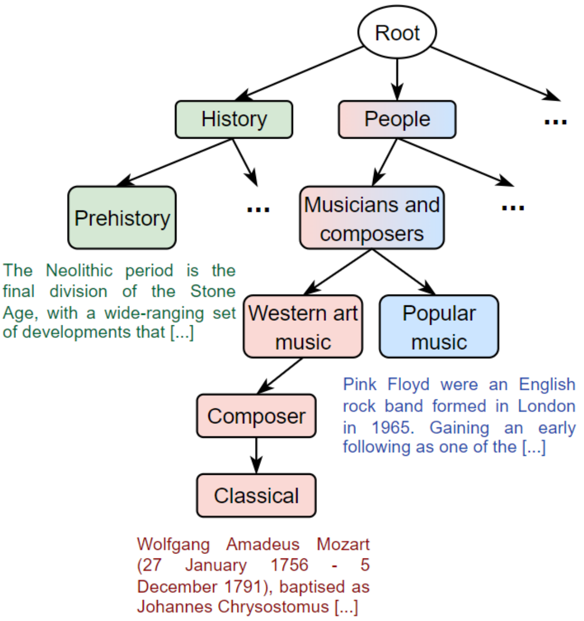
Root
History
People
...
Prehistory
...
Musicians and composers
...
The Neolithic period is the final division of the Stone Age, with a wide-ranging set of developments that [...]
Western art music
Popular music
Composer
Pink Floyd were an English rock band formed in London in 1965. Gaining an early following as one of the [...]
Classical
Wolfgang Amadeus Mozart (27 January 1756 - 5 December 1791), baptised as Johannes Chrysostomus [...]

Figure 1: Extract of the taxonomy of our new dataset Hierarchical WikiVitals. Each colored path is the set of labels of the same color.

1 Introduction

Text classification is a long-studied problem that may involve various types of label sets. In particular, Hierarchical Text Classification (HTC) includes labels that exhibit a hierarchical structure with parent-child relationships. The structure that emerges from these relationships is either a tree (Kowsari et al., 2018; Lewis et al., 2004; Lyubinets et al., 2018; Aly et al., 2019; Sandhaus, 2008) or a Directed Acyclic Graph (DAG) (Bertinetto et al., 2020). Each input text then comes with a set of labels that form one or more paths in the hierarchy. A first crucial challenge in HTC lies in accurately evaluating model performance. This requires metrics that are sensitive to the severity of prediction errors, penalizing mistakes with larger distances within the hierarchy tree. While pioneering efforts have been made by Kiritchenko et al. (2006), Silla and Freitas (2011), Kosmopoulos et al. (2014) and Amigo and Delgado (2022), evaluation in the context of hierarchical classification remains an ongoing research area.

There is a substantial body of literature addressing HTC. The most recent methods produce text representations which are *hierarchy-aware*, as they integrate information about the label hierarchy (Song et al., 2023; Zhou et al., 2020; Deng et al., 2021; Wang et al., 2022b,a; Jiang et al., 2022; Chen et al., 2021; Zhu et al., 2023, 2024; Yu et al., 2023). However, we believe that the evaluation of these models has been insufficiently investigated: in those works, the task is evaluated as standard multi-label classification. Here, we plan to explore what this implies; especially, looking at how predictions are inferred from an estimated probability distribution – which we consider an under-addressed challenge. We provide new insights, emphasizing the intricacy of inference and evaluation, which cannot be considered separately.

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 231–242
November 15-16, 2024 ©2024 Association for Computational Linguistics

To complete this investigation, we introduce a new English benchmark dataset, Hierarchical WikiVitals (HWV), which we intend to be significantly more challenging than the usual HTC benchmarks in English (see Figure 1 for an extract of the taxonomy). We experiment within our proposed framework, verifying the performance of recent models against simpler methods, among which loss functions (Bertinetto et al., 2020; Vaswani et al., 2022; Zhang et al., 2021) we design to be able to integrate hierarchical information, based on the conditional softmax. Overall, our contributions are:

1. We propose to quantitatively evaluate HTC methods based on specifically designed hierarchical metrics and with a rigorous methodology.
2. We present Hierarchical WikiVitals, a novel high-quality HTC dataset, extracted from Wikipedia. Equipped with a deep and complex hierarchy, it provides a harder challenge.
3. We conduct extensive experiments on three popular HTC datasets and HWV, introducing a novel loss function. When combined with a BERT model, this approach achieves competitive results against recent advanced models.

Our results show that state-of-the-art models do not necessarily encode hierarchical information well, and are surpassed by our simpler loss on HWV.

Problem definition

Hierarchical Text classification (HTC) is a subtask of text classification which consists of assigning to an input text $x \in \mathcal{X}$ a set of labels $Y \subset \mathcal{Y}$, where the label space $\mathcal{Y}$ exhibits parent-child relationships. We call hierarchy the directed graph $\mathcal{H} = (\mathcal{Y}, \mathcal{E})$, where $\mathcal{E} \subset \mathcal{Y}^2$ is the set of edges, which goes from a parent to its children. We restrain our study to the case where $\mathcal{H}$ is a tree. We follow the notations of Valmadre (2022) and call $\mathbf{r} \in \mathcal{Y}$ the unique root node and $\mathcal{L}$ the set of leaf nodes. For a node $y \in \mathcal{Y}\backslash\{\mathbf{r}\}$ we denote $\pi(y)$ its unique parent, $\mathcal{C}(y) \subset \mathcal{Y}$ the set of its children and $\mathcal{A}(y)$ the set of its ancestors (defined inclusively).

A label set Y of an input x cannot be arbitrary: if $y \in Y$ then, due to the parent relations, we necessarily observe that $\mathcal{A}(y) \subset Y$. An even more restrictive framework is the *single-path leaf labels* setting, where $Y = \mathcal{A}(l)$ for a given $l \in \mathcal{L}$ (Y is a single path and reaches a leaf).

We study methods mapping an input text x to a conditional distribution $\mathbb{P}(\cdot|x)$ over $\mathcal{Y}$, whose estimation is denoted $\hat{\mathrm{P}}(\cdot|x)$. Lastly, what we call *inference rule* is the way of producing a set of binary predictions from a probability distribution. For example predictions can be obtained by thresholding $\hat{\mathrm{P}}(\cdot|x)$ to τ as follows : $\hat{Y}_\tau = \{y \in \mathcal{Y}, \hat{\mathrm{P}}(y|x) > \tau\}$.

2 Related Work

2.1 Hierarchical Text Classification

Hierarchical classification problems, including the particular case of HTC, are typically dealt with through either a *local* approach or a *global* one. We refer to the original definition made by Silla and Freitas (2011) according to which the difference between the two categories lies in the training phase. Indeed, local methods imply training a collection of specialized classifiers, *e.g.* one for each node, for each parent node or even one for each level; and during its training each classifier is unaware of the holistic structure of the hierarchy (Zangari et al., 2024). While often computationally costly, it has proven to be effective to capture crucial local information. Along those lines, Banerjee et al. (2019) propose to link the parameters of a parent classifier and those of its children, following the idea of transferring knowledge from parent nodes to their descendants (Shimura et al., 2018; Huang et al., 2019; Wehrmann et al., 2018). Conversely, global methods involve a unique model that directly incorporates the whole hierarchical information in their predictions. There exist very different types of global approaches, from which we can draw two broad categories: losses incorporating hierarchical penalties and hierarchy-aware models.

Hierarchical penalties. The idea of these methods is generally to use a standard binary cross-entropy (BCE), and add penalization terms that incorporate hierarchical information. Gopal and Yang (2013) and Zhang et al. (2021) propose regularization based on hypernymy, either acting on the parameter space or the outputted probability space, while Vaswani et al. (2022) introduce an enhanced BCE loss, named CHAMP, which penalizes false positives based on their distance to the ground truth in the hierarchy tree.

Hierarchy-aware models. To incorporate the structural constraints of the hierarchy into prediction, Mao et al. (2019) propose a reinforcement learning approach, while Aly et al. (2019) introduce an architecture based on capsule networks. However, recent works have achieved state-of-the-art results by combining a text encoder with a structure encoder applied to the label hierarchy. This concept was first introduced by Zhou et al. (2020), who utilized graph convolution networks as the hierarchy encoder. Building on this foundational work, Jiang et al. (2022) and Wang et al. (2024) developed methods to better incorporate local hierarchy information. Wang et al. (2022a) proposed a contrastive learning approach, while Zhu et al. (2023) designed a method to encode hierarchy with the guidance of structural entropy. Zhu et al. (2024) combined both of these ideas. These developments follow earlier works on the same concept (Chen et al., 2020; Zhang et al., 2022; Deng et al., 2021; Chen et al., 2021; Wang et al., 2021). It is important to note that these models are typically trained with a BCE loss or one of its penalized versions (Zhang et al., 2021).

2.2 Hierarchical prediction

Making a prediction in HTC involves two seemingly irreconcilable difficulties: one has to decide between making independent predictions, which may lead to *coherence* issues (e.g., predicting a child without predicting its parent), or employing a top-down inference approach, which may cause *error propagation* issues (Yang and Cardie, 2013; Song et al., 2012). Recent hierarchy-aware models predominantly operate within the former framework, training and evaluating the model as a simple multi-label classifier, at the price of ignoring potentially badly structured predictions. In this work, we will experiment with both approaches.

2.3 Hierarchical classification evaluation

Evaluation in the context of hierarchical classification is a long-studied problem (Kosmopoulos et al., 2014; Amigo and Delgado, 2022; Costa et al., 2007) from which arise multiple questions. First, diverse setups exist, implying different assumptions on the labeling structure: while we previously introduced the *single-path leaf label* framework, multi-path hierarchies exist, or even inputs with only non-leaf labels. It is therefore important to design metrics that are **agnostic to the hierarchical classification framework**. Then, a hierarchical metric must indeed be hierarchical. This means it should take into account the severity of an error based on the known hierarchy: intuitively, predicting a *Bulldog* instead of a *Terrier* should be less penalized than predicting a *Unicorn* instead of a *Terrier*. Amigo and Delgado (2022) identify a set of properties an evaluation metric should possess for hierarchical classification, and classifies them in a taxonomy of metrics differentiating between **multi-label metrics** (label-based, example-based, ranking-metrics) and **hierarchical metrics** (pair-based, set-based). We heavily rely on this seminal work when it comes to choose which metric to use to evaluate different methods. Finally, the inference rule should be chosen in accordance with the metric. The bayesian decision theory literature (Berger, 1985) aims at finding an optimal rule given the metric of interest. However, little consideration was given to this issue in the context of hierarchical classification and ad hoc and non-statistically grounded inference methodology are often chosen: for example, recent HTC literature mostly performs inference through thresholding the estimated probability distribution with $\tau = 0.5$. We can think of other inference methodology, based on top-down or bottom-up inference rules. It is then crucial to find metrics that either come with a properly grounded prediction rule, or **do not depend on an inference methodology** but rather account for the whole probability distribution, which implies evaluating at different operating points. In the next part, we will re-introduce metrics in the light of the three listed requirements.

3 Evaluation metrics

The aforementioned inference rule used in recent HTC literature corresponds to a classical multi-label evaluation methodology: computing a F1-score (*micro* and *macro*) with $\tau = 0.5$. In what follows, we show that this thresholding scheme is suboptimal and we introduce the metrics we use in our experiments. We will then motivate the use of an inference-free evaluation methodology.

3.1 Multi-label metrics

There is a large array of methods for multi-label evaluation; Wu and Zhou (2016), through unifying notations, proposed a set of 11 different metrics. Among them, we keep the *micro* and *macro* F1-score computed upon scores obtained through a 0.5 threshold, as it is generally done in HTC literature. We add a simple metric corresponding to the fraction of misclassified labels: the Hamming Loss, which we also couple to a 0.5 thresholding inference rule.[1]

[1] This optimal inference holds in case of label independence (Dembczyński et al., 2012) which is not the case here.

3.2 Hierarchical metrics

We introduce hF1-score which we identify to be relevant to our evaluation framework. We note that a prediction is *coherent* if $z \in \hat{Y} \Rightarrow \mathcal{A}(z) \subset \hat{Y}$.

Hierarchical F1-score. Introduced by Kiritchenko et al. (2006), this **set-based** measure consists in augmenting $\hat{Y}$ with all its ancestors as follows :

$$\hat{Y}^{\text{aug}} = \bigcup_{\hat{y} \in \hat{Y}} \mathcal{A}(\hat{y}) \tag{1}$$

And to compute the hierarchical precision, recall and F1-score are as follows :

$$\mathrm{hP}(\mathrm{Y}, \hat{\mathrm{Y}}) = \frac{\left|\hat{Y}^{\text{aug}} \cap Y\right|}{\left|\hat{Y}^{\text{aug}}\right|} \quad \mathrm{hR}(\mathrm{Y}, \hat{\mathrm{Y}}) = \frac{\left|\hat{Y}^{\text{aug}} \cap Y\right|}{|Y|}$$

$$\mathrm{hF1}(\mathrm{Y}, \hat{\mathrm{Y}}) = \frac{2 \cdot \mathrm{hP}(\mathrm{Y}, \hat{\mathrm{Y}}) \cdot \mathrm{hR}(\mathrm{Y}, \hat{\mathrm{Y}})}{\mathrm{hP}(\mathrm{Y}, \hat{\mathrm{Y}}) + \mathrm{hR}(\mathrm{Y}, \hat{\mathrm{Y}})}$$

It is a simple extension of the F1-score to hierarchical classification. In the multi-label setting, there are several methods of aggregation to compute a global F1-score[2]. We define here a per-instance hF1-score as per Kosmopoulos et al. (2014) which is then averaged over all inputs (referred as *samples* setting). In its very first introduction, it was defined in a *micro* fashion by Kiritchenko et al. (2006) (see Appendix C.2 Plaud et al. (2024) for full definitions).

Proposition 1 *In micro and samples settings, if every prediction* $\hat{Y}$ *is coherent, then hF1 and F1 are strictly equal.*

Motivations. Hierarchical F1-score considers an ancestor overlap between ground truth and predicted labels therefore accounting for **mistake severity** and is also **agnostic to the hierarchical classification framework**. Moreover, Proposition 1 (whose proof is detailed in Appendix C.2 Plaud et al. (2024)) draws a link between example-based multi-label metrics and set-based hierarchical metrics proving that it was therefore relevant to employ the *micro* F1-score as it is done in recent literature, as long as predictions are coherent. Finally, hF1-score incorporates **all desirable hierarchical properties** as listed by Amigo and Delgado (2022), except that it does not completely capture the *specificity* (*i.e* the level of uncertainty left by predicting a given node).

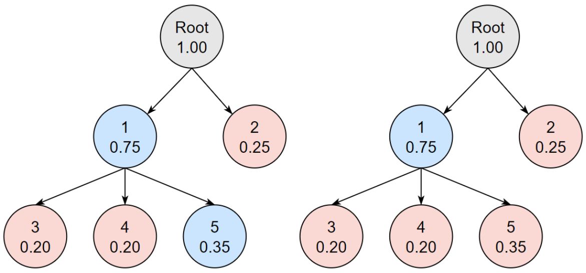

Figure 2: Example of a conditional distribution estimation over a simple hierarchy and corresponding predicted nodes (in blue) for different thresholds (0.3 on the left, 0.5 on the right).

Other hierarchical metrics. As explained in previous section, hF1-score is imperfect as it assumes an equivalence between depth and specificity. To solve this issue, Valmadre (2022) has proposed an information-based hierarchical F1-score, introduced in Appendix A (Plaud et al., 2024). There also exist constrained versions of **multi-label F1-scores** (Yu et al., 2022; Ji et al., 2023) which account for coherence issues: a correct prediction for a label node is valid only if all its ancestor nodes are correct predictions.
Although these metrics might seem pertinent, we have chosen not to utilize them, as they do not globally influence the ranking of methods when compared to their standard metric counterparts. We thoroughly detail our reasons in Appendix A (Plaud et al., 2024). An important number of context-dependent hierarchical metrics were also introduced (Sun and Lim, 2001; Bi and Kwok, 2015), which we will not discuss here as we aim for agnosticism to the hierarchical classification context.

3.3 Inference methodology

In this section, we begin by motivating our argument against the practice of using a BCE-based loss and $\tau = 0.5$ to produce predictions. While this corresponds to minimizing the multilabel Hamming loss in case of label independence (Dembczyński et al., 2012), there is to the best of our knowledge no evidence of the optimality of such a predictor in a hierarchical setting. Rather, tools such as *risk minimization* can provide a way to obtain a statistically grounded inference methodology optimizing the chosen metric, from an estimation of $\mathbb{P}(\cdot|x)$, obtained by a model for a given x. In particular, it is possible to show that the optimal threshold for the F1-scores depends on $\mathbb{P}(\cdot|x)$; we detail the proof in Appendix C.1.2 (Plaud et al., 2024). Though, a

[2] See for example the Scikit-learn documentation.

simple counter-example is enough to invalidate the choice of 0.5: such an example is depicted in Figure 2. It shows a coherent and exhaustive probability distribution $\mathbb{P}(\cdot|x)$, for a given x. Thresholding to 0.5 would lead to predict $\{1\}$, while a simple computation, detailed in Appendix C.1 (Plaud et al., 2024), gives:

$$\mathbb{E}[\text{hF1}(\text{Y}, \{1\})|X = x] = 0.5$$
$$\mathbb{E}[\text{hF1}(\text{Y}, \{1, 5\})|X = x] = 0.55$$

which shows that in a *single path leaf label* setting it is strictly better to predict $\{1, 5\}$ when aiming at maximizing the hF1-score. With Proposition 1 in mind, this simple example shows theoretically **the sub-optimality of the current state-of-the-art models inference methodology**. As the optimal threshold is unknown, we need to design an evaluation framework which does not depend on an ad-hoc inference rule to avoid introducing non statistically grounded methods. Following recommendations given by Valmadre (2022), we hence do away with inference rules and we construct precision-recall curves for hF1 by browsing all possible thresholds. From these curves, we compute the Area Under Curve (AUC).

4 Simple conditional loss-based methods

As a counterpart to the existing state-of-the-art consisting mainly of BCE-based approaches, we introduce several loss-based methods that incorporates local information, all relying on estimating **conditional probabilities**.

4.1 Conditional softmax cross-entropy

As outlined in Problem Definition, we focus on methods that, given an input text x, produce an estimated distribution $\hat{\text{P}}(\cdot|x)$ over $\mathcal{Y}$. We propose here to associate a modern text encoder to the conditional softmax (Redmon and Farhadi, 2017), which inherently incorporates the hierarchy structure by producing a hierarchy-coherent probability distribution, and coupling it with a cross-entropy loss. We detail in this section the modeling and training associated with it. Let us consider an input text x with its corresponding label set Y; a text encoder is first used to produce an embedded representation $h_x \in \mathbb{R}^d$ of x.

Conditional softmax. The conditional softmax first maps h_x to $s_x \in \mathbb{R}^{|\mathcal{Y}|}$ through a standard linear mapping:

$$s_x = W h_x + b \tag{2}$$

where $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $b \in \mathbb{R}^{|\mathcal{Y}|}$. Then, a softmax is applied to each brotherhood as follows:

$$\hat{\text{P}}(y|x, \pi(y)) = \frac{\exp s_x^{[y]}}{\sum\limits_{z \in \mathcal{C}(\pi(y))} \exp s_x^{[z]}} \tag{3}$$

We recall that $\pi(y)$ denotes the parent node of node y, and $\mathcal{C}(\pi(y))$ represents the set of children of $\pi(y)$, which includes y. The term $s_x^{[y]}$ refers to the entry of s_x associated with node y.
Hence, the logits s_x are used to model the conditional probability of a node **given** its parent. For example, this could represent the probability of an instance x to belong to the class *Bulldog*, conditioned on it being a *Dog*.

Cross-entropy. The contribution to the loss of the pair (x, Y) is given by a standard leaf nodes cross-entropy (as if we were in a standard monolabel multiclass classification problem over leaf nodes). With our modelisation it can further be decomposed as:

$$\begin{aligned} l_{\text{CSoft}}(x, Y) &= -\log \hat{\text{P}}(y^{\text{leaf}}|x) \\ &= -\sum_{y \in Y} \log \hat{\text{P}}(y|x, \pi(y)) \end{aligned} \tag{4}$$

where we denote y^{leaf} the unique leaf node of Y.

Outputted conditional distribution. The probability of $y \in \mathcal{Y}$ is computed by a standard conditionality factorization :

$$\hat{\text{P}}(y|x) = \prod_{z \in \mathcal{A}(y)} \hat{\text{P}}(z|x, \pi(z))$$

Motivations. Contrary to BCE-based methods, this modelisation directly incorporates the hierarchy structure prior of labels. Besides, the outputted probability distribution is coherent and exhaustive. It is more powerful than a leaf nodes softmax, as it decomposes the leaf probability estimation into several sub-problems. It is also computationally cheap, with a $\mathcal{O}(|\mathcal{Y}|)$ time complexity.

4.2 Logit-adjusted conditional softmax

We then propose an enhanced version of the conditional softmax, in order to improve its robustness to data imbalance. This is particularly important for our newly introduced HWV dataset, which has around half of labels having less than 10 instances in total. Our proposal is motivated by Zhou et al. (2020), who suggest that integrating the prior probability distribution in the model is relevant to the

HTC task, which is confirmed by their experimental results. Their approach involves initializing (or fixing) the weights of the structure encoder using this pre-computed prior distribution. Hence, we draw inspiration from Menon et al. (2021) and introduce the logit-adjusted conditional softmax cross-entropy. Equation (3) becomes:

$$\hat{\mathrm{P}}(y|x, \pi(y)) = \frac{e^{s_x^{[y]} + \tau \log \nu(y|\pi(y))}}{\sum\limits_{z \in \mathcal{C}(\pi(y))} e^{s_x^{[z]} + \tau \log \nu(z|\pi(z))}}$$

where $\nu(y|\pi(y))$ is an estimation of $\mathbb{P}(y|\pi(y))$[3] and τ a hyperparameter. Equation (4) remains unchanged. Comprehensive details on the adaptation of the logit-adjusted softmax to our case, along with the theoretical justifications, are provided in Appendix C.3 (Plaud et al., 2024). We expect this loss to enhance performances on the under-represented classes.

4.3 Conditional sigmoid binary cross-entropy

In practice, several real-world datasets consistently used in recent literature to evaluate HTC models (Lewis et al., 2004; Aly et al., 2019) are multi-path. As the conditional softmax is not designed for multi-path labels, we propose to use a conditional sigmoid loss, introduced by Brust and Denzler (2020). It follows a similar intuition to the conditional softmax: sigmoids are applied to each entry of s_x, modeling the conditional probability of the node given its parent. Hence, the contribution to the loss of a pair (x, Y) is given by a **masked** cross-entropy[4]:

$$\begin{aligned} l_{\mathrm{CSig}}(x, Y) = &-\sum_{z \in Y} \log(\hat{\mathrm{P}}(z|x, \pi(z))) \\ &+ \sum_{u \in \mathcal{C}(\pi(z)) \setminus \{z\}} \log(1 - \hat{\mathrm{P}}(u|x, \pi(z))) \end{aligned}$$

Proposition 2 *Let $x \in \mathcal{X}$, $Y \subset \mathcal{Y}$ and W defined as per Equation 2 then*

$$\frac{\partial l_{\mathrm{CSoft}}(x, Y)}{\partial W} = \frac{\partial l_{\mathrm{CSig}}(x, Y)}{\partial W}$$

Proof can be found in Appendix C.4 (Plaud et al., 2024). While the conditional sigmoid was not motivated by theoretical arguments in Brust and Denzler (2020), Proposition 2 proves that gradients computed for this loss and the conditional softmax cross-entropy loss are equivalent. This loss then allows to deal with both multi-path and non-exhaustive datasets while having similar properties to conditional softmax.[5]

Dataset	Train/Val/Test	#nodes (#leaves)	#nodes per level	Avg. #labels per sample
HWV (SPL)	6,408/1,602 2,003	1186 (953)	11-109-381- 437-244-4	3.7
WOS (SPL)	30,070/7,518 9,397	141 (134)	7-134	2.0
RCV1 (MP)	23,149/ - 781,265	103 (82)	4-55-43-1	3.2
BGC (MP)	58,715/14,785 18,394	146 (120)	7-46-77-16	3.0

Table 1: Key statistics of the selected datasets. **SPL** indicates that the dataset enters the *single path leaf labels* setting, and **MP** that it is multi-path; d represents the maximum depth of the label hierarchy.

5 Experimental settings

In this section, we introduce the existing datasets and models we experiment with; we also present our new dataset, Hierarchical WikiVitals (HWV).

5.1 Datasets

We will verify the performance of our proposed approaches versus baselines and recent state-of-the-art models on hierarchical metrics on three widely used datasets in the HTC literature, which is mainly applied to English data: Web-of-Science (WOS) (Kowsari et al., 2018), RCV1-V2 (Lewis et al., 2004) and BGC (Aly et al., 2019). Data statistics are displayed in Table 1: those datasets have in common a relatively large number of training samples, a sizable number of nodes, and a low depth of the label structure. We contribute to HTC benchmarking by releasing Hierarchical WikiVitals, which we aim to present a more difficult challenge.

HWV Dataset Texts are extracted from the abstracts of the *vital* articles of Wikipedia, level 4 [6] as of June 2021. This project involves a handmade hierarchical categorization of the selected articles, which are themselves put through high scrutiny with respect to their quality. The resulting dataset is a *single path leaf label* dataset, a constraint only fulfilled by WOS. As the number of nodes and the

[3] In practice, we estimate it by computing an empirical probability on train set for each label. It is not trainable.

[4] See Fig. 2b of Brust and Denzler (2020) for visual understanding of the mask

[5] However, no logit-adjusted version of it can be properly derived.

[6] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles/Level/4

	HWV				WOS			
Method	Hamming L. (in ‰) ↓	F1-score (in %) ↑		hF1 AUC (in %) ↑	Hamming L. (in ‰) ↓	F1-score (in %) ↑		hF1 AUC (in %) ↑
		micro	macro			micro	macro	
BCE	0.854$_{\pm0.010}$	85.86$_{\pm0.15}$	45.56$_{\pm0.58}$	89.23$_{\pm0.13}$	3.627$_{\pm0.015}$	87.03$_{\pm0.05}$	81.19$_{\pm0.12}$	**89.18**$_{\pm0.10}$
CHAMP	0.786$_{\pm0.009}$	87.14$_{\pm0.15}$	50.90$_{\pm0.24}$	89.87$_{\pm0.19}$	3.637$_{\pm0.037}$	87.01$_{\pm0.13}$	81.23$_{\pm0.18}$	88.74$_{\pm0.08}$
HBGL	-	-	-	-	**3.584**$_{\pm0.027}$	**87.22**$_{\pm0.10}$	**81.86**$_{\pm0.19}$	89.00$_{\pm0.10}$
HGCLR	0.922$_{\pm0.020}$	84.92$_{\pm0.37}$	44.89$_{\pm1.38}$	88.35$_{\pm0.35}$	3.727$_{\pm0.077}$	86.63$_{\pm0.27}$	80.04$_{\pm0.45}$	**89.23**$_{\pm0.22}$
HITIN	**0.776**$_{\pm0.006}$	**87.49**$_{\pm0.08}$	51.73$_{\pm0.42}$	90.72$_{\pm0.16}$	3.655$_{\pm0.028}$	87.05$_{\pm0.10}$	81.49$_{\pm0.06}$	88.92$_{\pm0.04}$
Leaf Softmax	0.950$_{\pm0.036}$	84.79$_{\pm0.57}$	51.49$_{\pm0.52}$	88.55$_{\pm0.47}$	3.987$_{\pm0.059}$	85.91$_{\pm0.25}$	80.02$_{\pm0.29}$	88.62$_{\pm0.08}$
Conditional Sigmoid	0.801$_{\pm0.011}$	87.01$_{\pm0.19}$	52.27$_{\pm0.82}$	90.40$_{\pm0.17}$	3.692$_{\pm0.067}$	86.86$_{\pm0.23}$	81.07$_{\pm0.30}$	88.78$_{\pm0.17}$
Conditional Softmax	0.788$_{\pm0.015}$	**87.49**$_{\pm0.10}$	53.79$_{\pm0.65}$	**90.94**$_{\pm0.09}$	3.869$_{\pm0.086}$	86.27$_{\pm0.17}$	80.25$_{\pm0.33}$	88.77$_{\pm0.07}$
Cond. Softmax + LA (ours)	**0.782**$_{\pm0.004}$	**87.51**$_{\pm0.07}$	**54.39**$_{\pm0.58}$	**90.97**$_{\pm0.05}$	3.837$_{\pm0.038}$	86.35$_{\pm0.12}$	80.11$_{\pm0.26}$	88.90$_{\pm0.10}$

Table 2: Performance evaluation metrics (and 95% confidence interval) on the test sets of the WOS and HWV datasets for the implemented models. Best results for each metric are highlighted in bold. The HBGL model was too large to fit in the memory of a 32GB GPU on the HWV dataset.

	RCV1				BGC			
Method	Hamming L. (in ‰) ↓	F1-score (in %) ↑		hF1 AUC (in %) ↑	Hamming L. (in ‰) ↓	F1-score (in %) ↑		hF1 AUC (in %) ↑
		micro	macro			micro	macro	
BCE	8.225$_{\pm0.148}$	86.65$_{\pm0.30}$	66.47$_{\pm1.49}$	**93.66**$_{\pm0.19}$	**7.788**$_{\pm0.071}$	**80.51**$_{\pm0.21}$	62.33$_{\pm1.36}$	**90.26**$_{\pm0.29}$
CHAMP	8.565$_{\pm0.234}$	85.93$_{\pm0.66}$	62.86$_{\pm3.64}$	93.12$_{\pm0.33}$	**7.775**$_{\pm0.081}$	**80.54**$_{\pm0.20}$	63.58$_{\pm0.49}$	**90.19**$_{\pm0.22}$
HBGL	**8.122**$_{\pm0.071}$	**87.11**$_{\pm0.12}$	**70.20**$_{\pm0.33}$	93.35$_{\pm0.14}$	8.092$_{\pm0.045}$	80.19$_{\pm0.11}$	**65.94**$_{\pm0.18}$	88.08$_{\pm0.10}$
HGCLR	8.761$_{\pm0.276}$	86.11$_{\pm0.26}$	67.49$_{\pm0.61}$	93.27$_{\pm0.14}$	8.054$_{\pm0.171}$	80.16$_{\pm0.29}$	63.58$_{\pm0.40}$	89.81$_{\pm0.17}$
HITIN	8.583$_{\pm0.188}$	85.72$_{\pm0.60}$	60.00$_{\pm5.15}$	93.04$_{\pm0.24}$	7.981$_{\pm0.096}$	**80.36**$_{\pm0.21}$	61.62$_{\pm1.47}$	**90.08**$_{\pm0.16}$
Conditional Sigmoid	8.652$_{\pm0.316}$	85.77$_{\pm0.71}$	63.90$_{\pm2.45}$	93.23$_{\pm0.36}$	7.954$_{\pm0.202}$	80.24$_{\pm0.46}$	62.65$_{\pm0.64}$	**90.07**$_{\pm0.40}$

Table 3: Performance evaluation metrics (and 95% confidence interval) on the test sets of the RCV1 and BGC datasets for the implemented models. Best results for each metric are highlighted in bold.

depth of the hierarchy are higher than for the previously cited datasets, HWV is much more challenging. It is also characterized by a very imbalanced label distribution with $\sim 50\%$ of labels having less than 10 examples in the whole dataset. We show in Figure 1 three observations from our new dataset, illustrating how much leaf nodes depth can vary (ranging from 2 to 6). Comprehensive details regarding the building process of the quality of data of HWV are provided in Appendix B (Plaud et al., 2024).

5.2 Models

We propose to compare very different HTC models, ranging from simple baselines to the most recent state-of-the-art approaches. For fair comparison between them, we use a pre-trained BERT[7] model (Devlin et al., 2019) as text encoder, adopting the standard [CLS] representation as h_x for every model. We list below all the different models evaluated. **BERT + BCE** is the simplest baseline, treating the problem as a multi-label task, without using any information from the hierarchical structure of labels. **BERT + Leaf Softmax** outputs a distribution over leaves, and hence is only fitted for single-path leaf label settings. **BERT + CHAMP** implements the penalization of false positives based on their shortest-path distance to the ground label set in the tree (Vaswani et al., 2022). **BERT + Conditional {Softmax, logit-adjusted Softmax, Sigmoid}** are our proposed methods, detailed in Section 4.1. **Hitin** (Zhu et al., 2023), **HBGL** (Jiang et al., 2022), **HGCLR** (Wang et al., 2022a) are among the most recent models, proposing respectively to separately encode the label hierarchy in an efficient manner, to incorporate both global and local information when encoding the label hierarchy, by considering subgraphs, and to use contrastive learning and exploiting the label hierarchy to create plausible corrupted examples.

5.3 Training details

We use `bert-base-uncased` model from the transformers library (Wolf et al., 2020) as text encoder (110M parameters). Our implementation is based on Hitin.[8] Each of our baselines is trained for 20 epochs on a V100 GPU of 32GB with a batch size of 16. We used an AdamW optimizer with initial learning rate of $2 \cdot 10^{-5}$ and with a warmup period of 10% of the training steps. For HBGL[9], Hitin and HGCLR[10], we rely on implementation guidelines

[7] `https://huggingface.co/bert-base-uncased`

[8] `https://github.com/Rooooyy/HiTIN`

[9] `https://github.com/kongds/HBGL`

[10] `https://github.com/wzh9969/contrastive-htc`

to conduct experiments. For datasets not used in the original papers, we performed a grid-search hyperparameter optimization. Our results are derived from averaging over four separate training runs, each initialized with distinct random seeds, ensuring the robustness and fairness of our evaluation methodology.

6 Results and Analysis

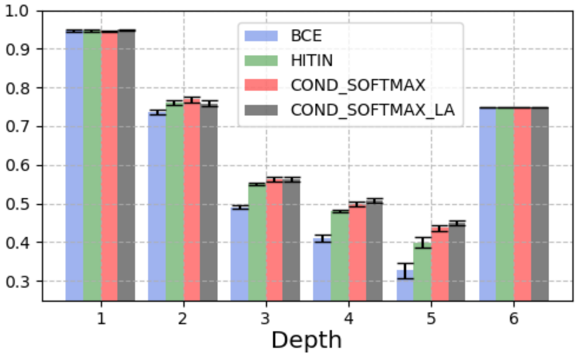

Figure 3: Averaged Macro F1-Scores on the test set per depth for different models and for the HWV dataset. The error bars represent a 95% confidence interval.

We start our investigation by evaluating models on our newly proposed dataset, HWV. Results are shown in Table 2. Unfortunately, the HBGL architecture could not run for HWV, requiring memory above the capacity of our GPUs. **On this dataset, we note the overall superiority of our newly introduced logit-adjusted conditional softmax loss and its vanilla version.** The latest models fail to obtain the best results, which is surprising given the complex hierarchy and label imbalance. We hence emit the hypothesis that while *hierarchy-aware* models were proven useful on simpler datasets, they fail to capture that complexity on HWV. To investigate why it performs better, we display in Figures 3 & 4 averaged macro F1-scores over classes. Figure 3 corresponds to averages of scores based on label depth: we observe that the higher the depth the higher the improvement brought by conditional softmax and its logit-adjusted version is (except for depth 6 which has only 4 classes inside). Figure 4 seems to hint that the improvement of the logit-adjusted conditional softmax vs. a vanilla conditional softmax lies in its ability to correctly classify *under-represented* classes. Until the third decile of the label count distribution, our newly introduced method is statistically better. We could have expected such a result, as this loss was specifically designed to deal with label imbalance (see Appendix C.3 (Plaud et al., 2024)). Obviously, depth is strongly correlated with *under-representation* of labels. We then conduct an ablation study with respect to the label hierarchy, by cutting the HWV hierarchy at depth 2. By doing so, the hierarchy becomes shallow and the label *imbalance* remains. Table 4 presents the results obtained from this modified dataset. In this scenario, state-of-the-art models catch up with our conditional softmax losses and Hitin reclaim a marginal lead across all metrics. Furthermore, we observe that our logit-adjusted conditional softmax remains better than the vanilla conditional softmax, especially on *macro* F1-score. These two observations allow us to refine our conclusions. First, the superiority of the vanilla conditional softmax on HWV vs. recent state-of-the-art methods seems to stem from the hierarchy complexity: **a conditional modelisation allows to better classify deep classes**. Second, the logit-adjusted version proves to be useful in presence of label imbalance as we can see with *macro* F1-score metrics, which are statistically better than the vanilla version in both versions of HWV dataset.

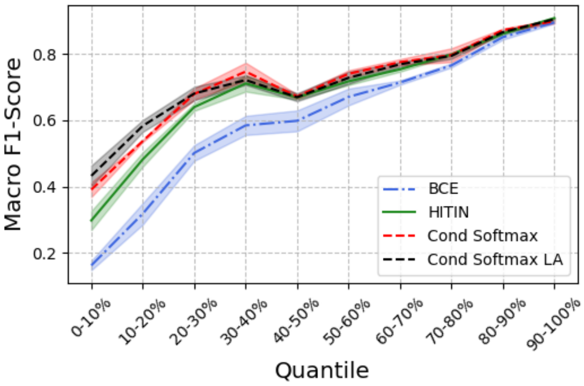

Figure 4: Averaged Macro F1-Scores on the test set by quantiles of label counts distribution in the training set for different models and for the HWV dataset. The shaded regions represent a 95% confidence interval.

On WOS, simpler baselines reach remarkable results. Despite the marginal superiority of HBGL, it is noteworthy that the **BERT+BCE model is in the top performances across all metrics**, while not using label hierarchy information. On this dataset, our new method, while competitive, lags behind.

These results are coherent with conclusions drawn with HWV dataset : the WOS dataset has

	HWV (depth 2)			
Method	Hamming L. (in ‰) ↓	F1-score (in %) ↑		hF1 AUC (in %) ↑
		micro	macro	
BCE	2.367$_{\pm 0.030}$	**92.89**$_{\pm 0.06}$	78.42$_{\pm 0.31}$	**94.67**$_{\pm 0.15}$
HITIN	**2.316**$_{\pm 0.068}$	**93.05**$_{\pm 0.20}$	**79.59**$_{\pm 0.43}$	**94.79**$_{\pm 0.18}$
Cond. Soft.	2.450$_{\pm 0.087}$	92.65$_{\pm 0.26}$	78.40$_{\pm 1.06}$	**94.73**$_{\pm 0.18}$
Cond. S. L.A.	2.432$_{\pm 0.072}$	**92.89**$_{\pm 0.22}$	**79.38**$_{\pm 0.26}$	**94.77**$_{\pm 0.18}$

Table 4: Performance evaluation metrics (and 95% confidence interval) on the test sets of the cutted HWV dataset for the implemented models. Best results for each metric are highlighted in bold.

a low complexity, both in terms of depth (maximum depth of 2) and distribution of labels (only one class has less than 40 examples in the dataset). On multi-path datasets, our observations align closely with what we noticed on WOS: we observe in Table 3 that a straightforward BCE loss consistently yields great results across datasets and metrics. Hierarchical metrics clearly highlight this phenomenon. In fact, model rankings in multi-label F1 scores and hierarchical F1 scores only keep consistent for HWV: for the three other datasets, the **structure-aware threshold-independent metrics put the BCE baseline to the top**.

We believe those results allow us to draw two main lessons: first, that hierarchical metrics bring useful insights on HTC evaluation, and are necessary to properly evaluate models on their capacity to encode label structure, which our results show to be lacking. Second, that when used on a more challenging dataset, state-of-the-art hierarchy-aware HTC models are less able to integrate that complex hierarchical information into their prediction than a simple model trained with conditional softmax cross-entropy.

7 Conclusion

In this paper, we come back upon recent progress in HTC, and propose to investigate its evaluation. To do so, we begin by showing the limitations of the inference and metrics that are commonly used in the recent literature. We instead propose to use existing hierarchical metrics, and an associated inference method. Then, we introduce a new and challenging dataset, Hierarchical WikiVitals; our experiments show that recent sophisticated hierarchy-aware models have trouble integrating hierarchy information in any better way than simple baselines. We finally propose simple hierarchical losses, able to better integrate hierarchy information on our dataset. In the future, we plan to investigate the inference mechanism for hierarchical metrics, through which we will aim to make a direct contribution to improving models on HTC tasks.

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Enrique Amigo and Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300, Florence, Italy. Association for Computational Linguistics.

James O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer Series in Statistics. Springer, New York.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. 2020. Making better mistakes: Leveraging class hierarchies with deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei Bi and Jame T. Kwok. 2015. Bayes-optimal hierarchical multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):2907–2918.

Clemens-Alexander Brust and Joachim Denzler. 2020. Integrating domain knowledge: Using hierarchies to improve deep classifiers. In *Pattern Recognition*, pages 3–16, Cham. Springer International Publishing.

Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. 2020. Hyperbolic interaction model for hierarchical multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7496–7503.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4370–4379, Online. Association for Computational Linguistics.

Eduardo Costa, Ana Lorena, Andre Carvalho, and Alex Freitas. 2007. A review of performance evaluation measures for hierarchical classifiers. *AAAI Workshop - Technical Report*.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. 2012. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88:5–45.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip Yu. 2021. HTCInfoMax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3259–3265, Online. Association for Computational Linguistics.

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.

Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 257–265, New York, NY, USA. Association for Computing Machinery.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1051–1060.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2918–2933, Toronto, Canada. Association for Computational Linguistics.

Ting Jiang, Deqing Wang, Leilei Sun, Zhongzhi Chen, Fuzhen Zhuang, and Qinghong Yang. 2022. Exploiting global and local hierarchies for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4030–4039, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Advances in Artificial Intelligence*, pages 395–406, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2014. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865.

Kamran Kowsari, Donald Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew Gerber, and Laura Barnes. 2018. Web of science dataset.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr):361–397.

Volodymyr Lyubinets, Taras Boiko, and Deon Nicholas. 2018. Automated labeling of bugs and tickets using attention-based mechanisms in recurrent neural networks. In *2018 IEEE Second International Conference on Data Stream Mining and Processing (DSMP)*, pages 271–275.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 445–455, Hong Kong, China. Association for Computational Linguistics.

Aditya Krishna Menon, Andreas Veit, Ankit Singh Rawat, Himanshu Jain, Sadeep Jayasumana, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR) 2021*.

Roman Plaud, Matthieu Labeau, Antoine Saillenfest, and Thomas Bonald. 2024. Revisiting hierarchical text classification: Inference and metrics. ArXiv preprint.

Joseph Redmon and Ali Farhadi. 2017. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium*, 6(12):e26752.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Brussels, Belgium. Association for Computational Linguistics.

Carlos Silla and Alex Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

Hyun-Je Song, Jeong-Woo Son, Tae-Gil Noh, Seong-Bae Park, and Sang-Jo Lee. 2012. A cost sensitive part-of-speech tagging: Differentiating serious errors from minor errors. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1025–1034, Jeju Island, Korea. Association for Computational Linguistics.

Junru Song, Feifei Wang, and Yang Yang. 2023. Peer-label assisted hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3758, Toronto, Canada. Association for Computational Linguistics.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchial text classification and evaluation. pages 521–528.

Jack Valmadre. 2022. Hierarchical classification at multiple operating points. In *Advances in Neural Information Processing Systems*, volume 35, pages 18034–18045. Curran Associates, Inc.

Ashwin Vaswani, Gaurav Aggarwal, Praneeth Netrapalli, and Narayan G Hegde. 2022. All mistakes are not equal: Comprehensive hierarchy aware multi-label predictions (champ).

Boyan Wang, Xuegang Hu, Peipei Li, and Philip S. Yu. 2021. Cognitive structure learning model for hierarchical multi-label text classification. *Knowledge-Based Systems*, 218:106876.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022a. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3740–3751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, and Houfeng Wang. 2024. Utilizing local hierarchy with adversarial training for hierarchical text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17326–17336, Torino, Italia. ELRA and ICCL.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xi-Zhu Wu and Zhi-Hua Zhou. 2016. A unified view of multi-label performance measures. In *International Conference on Machine Learning*.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1865–1869, New York, NY, USA. Association for Computing Machinery.

Simon Chi Lok Yu, Jie He, Victor Basulto, and Jeff Pan. 2023. Instances and labels: Hierarchy-aware joint supervised contrastive learning for hierarchical multi-label text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8858–8875, Singapore. Association for Computational Linguistics.

Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7).

Xinyi Zhang, Jiahao Xu, Charlie Soh, and Lihui Chen. 2022. La-hcn: Label-based attention for hierarchical multi-label text classification neural network. *Expert Systems with Applications*, 187:115922.

Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. Match: Metadata-aware text classification in a large hierarchy. In *Proceedings of the Web Conference 2021*, pages 3246–3257.

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.

He Zhu, Junran Wu, Ruomei Liu, Yue Hou, Ze Yuan, Shangzhe Li, Yicheng Pan, and Ke Xu. 2024. HILL:

Hierarchy-aware information lossless contrastive learning for hierarchical text classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4731–4745, Mexico City, Mexico. Association for Computational Linguistics.

He Zhu, Chong Zhang, Junjie Huang, Junran Wu, and Ke Xu. 2023. HiTIN: Hierarchy-aware tree isomorphism network for hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7809–7821, Toronto, Canada. Association for Computational Linguistics.