# OT-CLASS: Optimal Transport-Enhanced Multi-label Text Classification

**Anonymous ACL submission**


## Abstract

Multi-label text classification (MLTC) aims to assign at least one label from a vast label space to a document. This task is challenging due to the large number of labels, which can range from hundreds to thousands, and the potential interdependence of labels. While previous efforts have achieved success in fully-supervised settings, they have limited performance in more practical weakly-supervised settings. Despite its potential benefits, an auxiliary task of word-to-label alignment that aligns words in the input text to the large label space has been largely overlooked in existing work. Word-to-label alignment is significant, as it provides valuable insights into how words contribute to the overall classification of a document. However, existing MLTC datasets lack ground truth labels for word-to-label alignment for supervised training. To address this limitation, we propose a novel framework called OT-CLASS, which incorporates unsupervised word-to-label alignment into MLTC using optimal transport (OT). Our framework tackles MLTC in a multi-task setting, comprising a primary task that classifies documents using a standard text classification algorithm and an auxiliary task that identifies corresponding labels for all input document words via optimal transport. Our experiments demonstrate that OT-CLASS outperforms baselines that do not utilize word-to-label alignment, highlighting its effectiveness. A detailed analysis reveals that OT-CLASS has an amplified advantage in fine-grained label spaces and appropriately influences predictions through word-to-label alignment.


## 1 Introduction

Multi-label text classification (MLTC) is a task that involves assigning at least one label from a vast label space to a document. MLTC has a wide range of downstream applications including legal judgement (Nallapati and Manning, 2008; Chalkidis et al., 2019; Aletras et al., 2016), scientific publication analysis (Mai et al., 2018; Wang et al., 2020; Lu, 2011), and e-commerce (Agrawal et al., 2013; Prabhu et al., 2018) and sentiment analysis (Cambria et al., 2014).

MLTC is challenging due to the large number of labels, which can range from hundreds to thousands, and the potential interdependence of labels. Previous research has attempted to address this challenge by modeling the label space to capture relationships between labels. The most common way is to explicitly model these relationships using Graph Neural Networks (GNNs) (Kipf and Welling, 2016; Pal et al., 2020; Vu et al., 2023) or implicitly using regularization (Zhang et al., 2021; Gopal and Yang, 2013, 2015). While these efforts have achieved success in fully-supervised settings, they have limited performance in more practical weakly-supervised.

Despite its potential benefits, an auxiliary task that aligns words in the input text to the label space, known as word-to-label alignment, has been largely overlooked in existing work. This task is significant, as it provides valuable insights into how words contribute to the overall classification of a document. However, existing MLTC datasets lack ground truth labels for word-to-label alignment for supervised training. To address this limitation, we propose a novel framework called OT-CLASS, which incorporates unsupervised word-to-label alignment into MLTC using optimal transport (OT) (Figure 1). Our framework tackles MLTC in a multi-task setting, comprising a primary task that classifies documents using a standard text classification algorithm and an auxiliary task that identifies corresponding labels for all input document words via optimal transport. Our experiments demonstrate that OT-CLASS outperforms baselines that do not utilize word-to-label alignment, highlighting its effectiveness. A detailed analysis reveals that OT-CLASS has an amplified advantage in fine-grained label spaces and appropriately influences predictions through word-to-label alignment.

## 2 Related Work

### 2.1 Multi-label Text Classification

Existing MLTC frameworks exploit the fact that there are similarities between many labels. Explicitly modeling relationships within the label space is often done by GNNs. Much existing work embeds the input documents, either using Transformer (Vaswani et al., 2017) encoders or Bidirectional LSTMs (Huang et al., 2015), and the label space separately (Pal et al., 2020; Vu et al., 2023). Some existing work creates a joint embedding space between the labels and the input documents (Chen et al., 2021; Wang et al., 2018). Other frameworks incorporate the hierarchy implicitly, primarily by regularizing the embeddings of each label in the label space by their parent label (Zhang et al., 2021; Gopal and Yang, 2013, 2015).

### 2.2 Optimal Transport in NLP

OT has been applied to many tasks within NLP. The most common task is measuring textual similarity across sentences (Wang et al., 2022; Lee et al., 2022; Arase et al., 2023a; Jiang et al., 2020). OT has also been applied to text summarization where sentences of a document are matched to potential summaries (Tang et al., 2022). Other applications of OT in NLP include natural language generation (Chen et al., 2020) and multi-lingual representation learning (Alqahtani et al., 2021). To the best of our knowledge, OTSeq2Set (Cao and Zhang, 2022) is the only work that has applied OT to MLTC. However, OTSeq2Set treats MLTC as a sequence-to-sequence task and uses the optimal transport distance as a measurement to force the model to focus on the closest labels for text classification. OT-CLASS uses optimal transport to learn an unsupervised auxiliary task of word-to-label alignment.

## 3 Methodology

We propose OT-CLASS, an optimal transport-enhanced framework to solve multi-label text classification. OT-CLASS (Figure 1) is a multi-task framework: one task attempts to learn the corresponding label a given document should be categorized under, while the other learns which tokens of the input document correspond to which labels.

### 3.1 Background

**Problem Formulation** The objective of multi-label text classification is to categorize a given document into a subset of labels in the label space. Since there are multiple labels a document can be categorized under, we operate in the multi-label classification setting. Mathematically, given an input document $\mathcal{D} = \{t_i : \forall i \in [1, |\mathcal{D}|]\}$ consisting of $|\mathcal{D}|$ tokens $t_i$, the objective is to assign labels $y \subset \mathcal{Y}$ from label space $\mathcal{Y}$ to the input document. The set of all documents $\mathcal{D}$ is denoted by $\mathcal{X} = \{\mathcal{D}_i : \forall i \in [1, |\mathcal{X}|]\}$.

**Optimal Transport** Optimal Transport is used to move mass from one distribution to another distribution as *efficient* as possible. Efficiency is measured by minimizing the total transportation cost across $m$ inputs of one distribution and $n$ inputs of another distribution. This transportation cost is denoted by $\mathbf{C} \in \mathbf{R}^{m \times n}$, which indicates the (dis)similarity of elements across both distributions. Identifying which elements should be aligned is the responsibility of the transport plan $\pi \in \mathbf{R}_+^{m \times n}$, which can be viewed as a joint probability distribution across all sets of inputs. The set of all transport plans is denoted by $\Pi = \{\pi \in \mathbf{R}_+^{m \times n} : \pi \mathbb{1}_n = a, \pi^T \mathbb{1}_m = b\}$, where $\mathbb{1}_n$ and $\mathbb{1}_m$ are the vector of ones with length $m$ and $n$, respectively. The two constraints $\pi \mathbb{1}_n = a$ and $\pi^T \mathbb{1}_m = b$ enforce that $\pi$ is a joint probability distribution. $a \in \mathbf{R}^m$ and $b \in \mathbf{R}^n$ are probability measures that assign weight to the mass of each element in their respective probability distributions. Concretely, we can define the OT problem as the following constrained minimization problem:

$$\begin{aligned} \min_{\pi \in \Pi(a,b)} \quad & \langle \mathbf{C}, \pi \rangle \\ \text{s.t.} \quad & \pi \mathbb{1}_n = a \\ & \pi^T \mathbb{1}_m = b \end{aligned} \tag{1}$$

There are multiple algorithms to solve this optimization problem; however, the most common solution is a method based on Sinkhorn's algorithm (Cuturi, 2013).

### 3.2 OT-Enhanced Multi-task Architecture

**Multi-label Text Classification** Given a set of input documents $\mathcal{X}$, we first extract its [CLS] embeddings, which represent the embeddings of the documents, using a Transformer encoder-based LLM:

$$\mathbf{E} = LLM(\mathcal{X}) \in \mathbf{R}^{|\mathcal{X}| \times d} \tag{2}$$

where $d$ is the dimensionality of each embedding. We then map the feature space to the label space by applying a linear layer on the [CLS] embeddings:

$$\hat{y} = \sigma(\mathbf{E}\mathbf{W}^T) + b \tag{3}$$

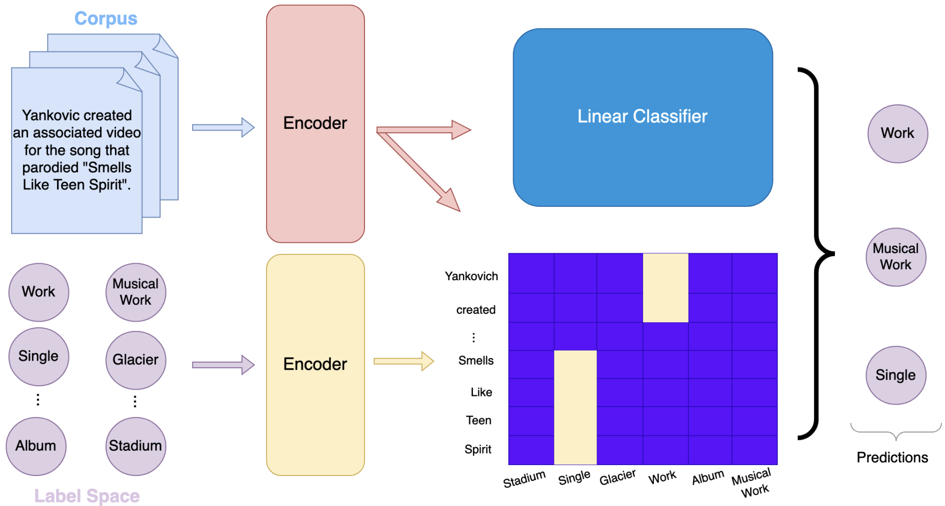


Figure 1: OT-CLASS architecture. The primary document classification task utilizes uses a given document's embedding and inputs it to a linear classifier. For the word-to-label alignment auxiliary task, both the label and document embeddings are used to form the OT plan. The yellow boxes indicate an alignment between the given document word and the label.

where $\mathbf{W} \in \mathbf{R}^{|\mathcal{Y}| \times d}$ is a learnable weight matrix, $b \in \mathbf{R}^{|\mathcal{Y}| \times 1}$ is a learnable bias vector, and $\sigma$ is the Sigmoid activation layer. For each document, we then apply binary cross-entropy loss to achieve the document classification loss:

$$l_i = [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - y_i)] \quad (4)$$

where $\hat{y}$ and $y_i$ are the predicted and ground truth labels for document $\mathcal{D}_i$, respectively. The final document classification loss across all inputs is the average loss across all instances:

$$L_{\text{cls}} = \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} l_i \quad (5)$$

**Word-to-Label Alignment** Optimal transport, as discussed in Section 3.1, aligns masses from different distributions. The distributions that OT-CLASS align are the tokens in the documents and the label space. Specifically, we align each token of the input document to any label in the label space. To do this, we first initialize the cost tensor $\mathbf{C} \in \mathbf{R}^{|\mathcal{D}| \times |\mathcal{Y}|}$ using pairwise cosine similarity:

$$\mathbf{C} = \{\cos(t_i, y_i) : \forall t_i \in \mathcal{D}, \forall y_i \in \mathcal{Y}\} \quad (6)$$

Since we wish to capture semantic similarity, we choose cosine similarity as the similarity metric, as it is one of the most common semantic similarity metrics in NLP. We then solve equation (1) using Sinkhorn's algorithm, resulting in the optimal plan $\pi^* \in \Pi$. Given the OT plan, we then compute the transportation loss as the inner product between the transport plan and the cost tensor:

| Dataset | # Train | # Test | # Labels | Depth |
|---|---|---|---|---|
| Amazon-531 | 29,487 | 19,685 | 531 | 3 |
| DBPedia-298 | 196,665 | 49,167 | 298 | 3 |

Table 1: Dataset statistics for Amazon-531 and DBPedia-298, outlining the number of data points and the depth of the hierarchy.

$$L_{\text{ot}} = \langle \mathbf{C}, \pi^* \rangle \quad (7)$$

Putting together the document classification loss from equation (5), we get the overall loss that spans both tasks:

$$L = L_{\text{cls}} + \lambda L_{\text{ot}} \quad (8)$$

where $\lambda \in [0, 1]$ is a hyper-parameter dictating the influence of the transportation alignment. It should be noted that there aren't any learnable parameters within the OT algorithm itself: The influence of our transportation alignment is seen by updating the weights of the linear layer $\mathbf{W}$ and bias vector $b$ listed in equation (3).

## 4 Experiments

### 4.1 Experimental Setup

**Dataset** We conduct experiments on two multi-label text classification datasets (Table 1): Amazon-531 (McAuley and Leskovec, 2013) and DBPedia-298 (Lehmann et al., 2015). Amazon-531 contains

Table 2: Main results of OT-CLASS on Amazon-531 and DBPedia-298 in %. The best results are **bolded**.

| Methodology | EF1 | P@1 | P@3 |
|---|---|---|---|
| **Dataset: Amazon-531** | | | |
| OT-CLASS | **85.49** | **95.76** | **84.04** |
| Fully Supervised | 83.48 | 95.20 | 82.66 |
| **Dataset: DBPedia-298** | | | |
| OT-CLASS | **97.33** | **99.49** | **97.33** |
| Fully Supervised | 97.30 | 99.40 | 97.30 |

Table 3: OT-CLASS results of using TELEClass weak labels on the Amazon-531 and DBPedia-298.

| Methodology | EF1 | P@1 | P@3 |
|---|---|---|---|
| **Dataset: Amazon-531** | | | |
| OT-CLASS | **48.78** | **66.78** | **48.34** |
| TELEClass (Zhang et al., 2024) | 46.49 | 64.30 | 46.11 |
| **Dataset: DBPedia-298** | | | |
| OT-CLASS | **50.08** | **71.78** | **50.08** |
| TELEClass (Zhang et al., 2024) | 49.59 | 69.84 | 49.59 |

reviews of various Amazon products, where the label space corresponds to various product categories. DBPedia-298 consists of Wikipedia articles and the label space represents the topics that each document could be classified as.

**Baselines** The baseline we compare against is a Transformer model fine-tuned on the groud-truth labels in the aforementioned datasets, denoted as "Fully Supervised". We also provide analysis under the weakly supervised setting, where the ground truth labels are acquired not from human annotation but from weak labels provided by TELEClass (Zhang et al., 2024).

**Evaluation Metrics** We use the **Example-F1** (Sorensen, 1948) and the **Precision at k** metrics. More details are in Appendix B.

### 4.2 Results

**Main Results** Our main results are listed in Table 2, which shows that OT-CLASS outperforms all baselines on all datasets. OT-CLASS does significantly better on Amazon-531 than DBPedia-298, which can also be seen when using OT-CLASS in the weakly-supervised scenario, using weak labels from TELEClass (Table 3). We hypothesize this is due to some words in the input document being fairly indicative of the corresponding labels.

**Multiple Granularity Analysis** We study the effectiveness of OT-CLASS across different levels of the Amazon-531 taxonomy (Table 4). OT-CLASS performs starkly better than the baseline at the most granular level. We hypothesize this is because our auxiliary task narrows the search space of labels for our primary document classification task, which is more pronounced at granular.

Table 4: OT-CLASS performance across all levels of the Amazon-531 hierarchy.

| Methodology | EF1 | P@1 | P@3 |
|---|---|---|---|
| **Level 1** | | | |
| OT-CLASS | **46.28** | **83.69** | **30.85** |
| Fully Supervised | 46.08 | 82.96 | 30.72 |
| **Level 2** | | | |
| OT-CLASS | 75.05 | **96.29** | 62.53 |
| Fully Supervised | **75.54** | 95.93 | **62.93** |
| **Level 3** | | | |
| OT-CLASS | **39.79** | **71.28** | **26.53** |
| Fully Supervised | 28.56 | 47.79 | 19.04 |

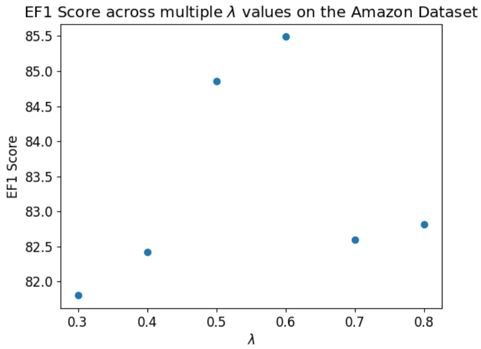


Figure 2: Analysis of multiple values of $\lambda$ over the Amazon-531 dataset.

**Word-to-label Alignment Influence** To understand the impact of OT on the final performance, we evaluate OT-CLASS over a series of $\lambda$ values. Figure 2 shows this analysis on the Amazon-531 dataset. It appears that the optimal value of $\lambda$ lies on a spectrum: too small of $\lambda$ indicates that the lack of alignment between words in the document and the label space hurt performance, whereas too large $\lambda$ indicates that OT-CLASS is paying too much attention to the word alignment.

## 5 Conclusion

We propose OT-CLASS, a novel optimal transport-enhanced framework to tackle MLTC. The OT-CLASS architecture encodes dual objectives: The primary task performs document classification while the auxiliary task performs word-to-label alignment. Experiments show that OT-CLASS achieves better performance across all baselines.

## Ethics Statement

Our work follows the ethical standards set by ACL. As we don't deal with sensitive data or domains, we do not expect and potential risks using OT-CLASS; however, we do not condone usage of our framework for any malicious motivations. We utilized all pre-trained models and datasets in a manner consistent with their existence.

## Limitations

While OT-CLASS outperforms canonical fine-tuning, we don't inject the hierarchical structure of the label space. Additionally, we speculate that finding more meaningful representations for each label, perhaps retrieving label descriptions, would let OT-CLASS better comprehend which tokens of the input document align with the label space. Furthermore, we make the assumption that each input token has a corresponding label. This maybe too harsh of an assumption, as not every token is guaranteed to have a matching label. Thus investigating variations of Optimal Transport, such as Partial Optimal Transport and Unbalanced Optimal Transport (Arase et al., 2023b), could result in better performance.

## References

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. 2013. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd international conference on World Wide Web*, pages 13–24.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuki Arase, Han Bao, and Sho Yokoi. 2023a. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986, Toronto, Canada. Association for Computational Linguistics.

Yuki Arase, Han Bao, and Sho Yokoi. 2023b. Unbalanced optimal transport for unbalanced word alignment. *arXiv preprint arXiv:2306.04116*.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Jie Cao and Yin Zhang. 2022. Otseq2set: An optimal transport enhanced sequence-to-set model for extreme multi-label text classification. *arXiv preprint arXiv:2210.14523*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

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.

Yimeng Chen, Yanyan Lan, Ruinbin Xiong, Liang Pang, Zhiming Ma, and Xueqi Cheng. 2020. Evaluating natural language generation via unbalanced optimal transport. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3730–3736. International Joint Conferences on Artificial Intelligence Organization. Main track.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

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*, pages 257–265.

Siddharth Gopal and Yiming Yang. 2015. Hierarchical bayesian inference and recursive regularization for large-scale classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(3):1–23.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5969–5979, Dublin, Ireland. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

Florian Mai, Lukas Galke, and Ansgar Scherp. 2018. Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, pages 169–178.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446, Honolulu, Hawaii. Association for Computational Linguistics.

Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. 2020. Multi-label text classification using attention-based graph neural network. *arXiv preprint arXiv:2003.11644*.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, pages 993–1002.

Thorvald Sorensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34.

Peggy Tang, Kun Hu, Rui Yan, Lei Zhang, Junbin Gao, and Zhiyong Wang. 2022. OTExtSum: Extractive Text Summarisation with Optimal Transport. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1128–1141, Seattle, United States. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Huy-The Vu, Minh-Tien Nguyen, Van-Chien Nguyen, Minh-Hieu Pham, Van-Quyet Nguyen, and Van-Hau Nguyen. 2023. Label-representative graph convolutional network for multi-label text classification. *Applied Intelligence*, 53(12):14759–14774.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Zihao Wang, Jiaheng Dou, and Yong Zhang. 2022. Unsupervised sentence textual similarity with compositional phrase semantics. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4976–4995, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *arXiv preprint arXiv:2403.00165*.

## A Implementation Details

The optimal value of $\lambda$ is 0.6 and 0.8 for the Amazon-531 and the DBPedia-298 datasets, respectively. We use `RoBERTa-base` (Liu et al., 2019) as our encoder model from the HuggingFace Transformer library[1]. All of our experiments were conducted using 2 NVIDIA A40 GPUs.

## B Evaluation Metrics

Following TELEClass (Zhang et al., 2024), we use the following two evaluation metrics below:

- **Example-F1** (Sorensen, 1948) evaluates multi-label classification without ranking:

$$\text{EF1} = \frac{2}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \frac{|y_i \cap \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{9}$$

- **Precision at k** captures the precision of predictions ranked by score:

$$P@k = \frac{1}{k} \sum_{i=1}^{|\mathcal{X}|} \frac{|y_i \cap \hat{y}_{i,:k}|}{\min(k, |y_i|)} \tag{10}$$

  where $y_{i,:k}$ represents the top-k predicted labels for document $\mathcal{D}_i$.

[1] https://github.com/huggingface/transformers