Journal of Information Organization

DLINE JOURNALS

Print ISSN : 2278 – 6503
Online ISSN: 2278 – 6511

JIO 2026: 16 (1)
https://doi.org/10.6025/jio/2026/16/1/26-39

LA-MIL: Label-Aware Attention Networks for Multi-Label Multi-Instance Text Classification

Pit Pichappan
Digital Information Research Labs
Chennai, 600017
India
pichappan@dirf.org

ABSTRACT

Multilabel multi instance text classification presents unique challenges due to the weak supervision setting where documents (bags) are labeled but constituent sentences (instances) are not, coupled with severe label imbalance where infrequent "tail" labels dominate real world distributions. Existing approaches typically employ label agnostic aggregation strategies such as max or mean pooling that implicitly assume uniform instance relevance across all labels, an assumption that is frequently violated in social tagging data, where individual sentences often signal specific labels while remaining irrelevant to others. To address this limitation, we propose LA-MIL (Label-Aware Attention Multi Instance Learning), a novel framework that employs dedicated attention heads for each label to enable fine grained, label specific instance selection. This architecture allows different labels to attend to distinct textual evidence within the same document, relaxing the restrictive assumption of uniform instance relevance. Evaluated on the DeliciousMIL benchmark dataset comprising 12,234 web documents annotated with 20 semantic tags, LA-MIL consistently outperforms traditional multi label classifiers, standard MIL models with global pooling, and attention based baselines with shared aggregation mechanisms. Notably, the model achieves significant improvements in macro-F1 scores, demonstrating superior handling of long tailed label distributions. Beyond quantitative gains, LA-MIL provides inherent interpretability through learned attention weights that transparently identify label-discriminative sentences. Our results establish label aware attention as an essential architectural principle for multi label multi instance learning, particularly in applications requiring both accuracy on imbalanced distributions and human interpretable predictions.

Keywords: Multi-label Text Classification, Multi-Instance Learning, Label-Aware Attention, Weak Supervision, Long-Tailed Label Distribution, Instance-Level Interpretability, Social Tagging, Attention Mechanisms

Received: 29 September 2025, Revised 14 December 2025, Accepted 23 December 2025

1. Introduction

Multi-label learning is a primary learning framework for addressing real-world objects with rich semantics [1] [Prabhu Y, 2018]. It has been successfully applied across diverse domains, including text categorisation [2], object recognition [3, 4] [Dai L, 2018, Wu MC, 2020], and topic recommendation [Gan Y, 2018]. In the context of text processing, the objective of multi label learning is to predict the appropriate labels for unseen text instances based on labeled training examples.

2. Review of Related Studies

Multi-label text classification (MLTC) underpins critical applications such as sentiment analysis [5] [Yuncong Li 2020], medical diagnosis [6] [James Mullenbach], and legal judgment prediction [7] [Nikolaos Aletras 2016]. Despite significant progress, MLTC has become increasingly challenging due to the rapid growth in data volume, resulting in extremely large numbers of documents and labels [8] [K. Bhatia, 2016]. While deep learning models [9, 10, 11] Jingzhou Liu, 2017, Jinseok Nam, 2017, Yashoteja Prabhu 2018] have been widely adopted to address scalability concerns, their performance remains constrained by a fundamental characteristic of real-world MLTC datasets: skewed label distributions. Specifically, many labels occur infrequently referred to as *tail labels* which significantly impacts model effectiveness [12] [Wang, Haobo, 2024].

2.1 Methodological Strategies Based on Label Correlation

The performance of multi label text learning has advanced rapidly with the development of deep convolutional networks. Existing approaches can be categorized into three strategies based on how they model label correlations [13] Huang SJ, 2019]:

• First order strategies treat each label independently, ignoring potential correlations among labels.

• Second order strategies exploit pairwise label relationships, though practical applications often involve dependencies beyond pairwise interactions.

• Higher order strategies capture the influence of each label on all other labels within the label space.

Representative higher order methods include the classifier chain [14] Teisseyre P (2017)], which sequentially combines sample features with previously predicted labels to form augmented inputs for subsequent classifiers. The CC approach [15] Zheng K, (2018] leverages information entropy [16] Yang W,] to maximise inter label correlations through conditional entropy optimisation. LIFT [17] Huang J, 2019] addresses multi label learning by generating label exclusive features based on label generic attributes [18] [Liu, B. 2022].

2.2 Deep Learning Architectures for End to End MLTC

Recent advances in deep learning have enabled end to end solutions for MLTC tasks. Convolutional Neural Networks (CNNs) [19, 20, 21] Liu et al. 2017; Kurata et al. 2016; Kim 2014] and Recurrent Neural Networks (RNNs) [22, 23] [Liu et al. 2016; Chen et al. 2017] have become standard architectures for automatic feature extraction from raw text. These models eliminate the need for handcrafted features while learning hierarchical representations suitable for multi label prediction.

2.3 Attention Mechanisms for Label-to-Text Correlation Modeling

Beyond generic feature extractors, recent research has focused on explicitly modeling label-to-text correlations through attention mechanisms. Notable approaches include:

• DocBERT [24] Adhikari et al. 2019], which adapts the BERT architecture into a computationally efficient model that maintains competitive accuracy for MLTC tasks.

• SGM [25] Yang et al. 2018], which simultaneously captures label correlations and identifies the most informative words for predicting each label.

• LSAN [26] Xiao et al. 2019], which employs both self attention and label specific attention mechanisms to generate enriched text representations tailored for multi label prediction.

2.4 Exploiting Label Co-occurrences for Enhanced Representations

Label co-occurrences constitute a vital source of semantic information in MLTC, as certain labels frequently appear together due to inherent semantic relationships. However, many existing methods concentrate primarily on optimizing feature extraction while neglecting these co occurrence patterns. To address this gap, Ma et al. [27] [2021] proposed LDGN (Label specific Dual Graph Neural Network), which incorporates label co-occurrence structures via Graph Convolutional Networks (GCNs) to refine text representations.

2.5 Limitations and Research Opportunities

Despite progress in leveraging label co-occurrence, current GCN-based approaches remain limited in their integration with feature extraction modules. Specifically, models like LDGN utilize GCNs primarily to optimize text representations using label co-occurrence information, yet they overlook two critical aspects: (1) the diversity required in text representations for different labels, and (2) the guidance that labels should provide when fusing heterogeneous feature vectors [28] Liang, Z. 2024]. These limitations highlight an important research direction: developing architectures that enable label-specific instance aggregation and representation learning within a unified multi-instance learning framework.

3. Architecture of the Label-Aware Attention Multi-Instance Learning (LA-MIL) model.

(a) Input Layer

• A document is treated as a bag of instances (sentences).

• Each instance is shown as a small text block or vector

$$\mathrm{x}_{i1}, \mathrm{x}_{i2}, \dots, \mathrm{x}_{in_i}.$$

(b) Shared Instance Encoder

• All instances pass through a shared encoder f (.).

- Visually: stacked blocks labeled *CNN / BiLSTM / Sentence Encoder*.

- Output embeddings:

$$\mathrm{h}_{ij} \in \mathbb{R}^d$$

(c) Label-Aware Attention Module (Core Contribution)

- For each label l, a separate attention head is shown.

- Each head computes attention weights:

$$\boldsymbol{\alpha_{ij}^{(l)}}$$

- Use arrows from instance embeddings to multiple attention heads, one per label.

- Colour coding or dashed boxes help emphasise label specificity.

(d) Label Specific Bag Representations

- Each attention head produces a label specific bag embedding:

$$\mathrm{z}_i^{(l)} = \sum_j \alpha_{ij}^{(l)} \mathrm{h}_{ij}$$

- Show these as aggregated vectors, one per label.

(e) Multi-Label Prediction Layer

- Independent sigmoid classifiers for each label:

$$\hat{y}_i^{(l)} = \sigma(\mathrm{u}_l^\top \mathrm{z}_i^{(l)})$$

- Final output is a multi-hot label vector.

4. Dataset for the study

We use the following dataset, whose characteristics are described.

DeliciousMIL: A Data Set for Multi-Label Multi-Instance Learning with Instance Labels

This dataset includes 1) 12234 documents (8251 training, 3983 test) extracted from DeliciousT140 dataset, 2) class labels for all documents, 3) labels for a subset of sentences of the test documents.

4.1 Dataset Characteristics

This dataset provides ground truth class labels to evaluate the performance of multi-instance learning models on both instance level and bag level label predictions. DeliciousMIL was first used in [1] to evaluate the performance of MLTM, a multi label multi instance learning method, for document classification and sentence labelling. Multi-instance learning is a class of weakly supervised machine learning methods in which the learner receives a collection of labelled bags, each containing multiple instances. A bag is set to have a particular class label if and only if at least one of its instances has that class label. DeliciousMIL consists of a subset of tagged web pages from the social bookmarking site delicious.com. The original web pages were obtained from the DeliciousT140 dataset [2], which was collected from delicious.com in June 2008. Users of the website delicious.com bookmarked each page with word tags. From this dataset, we extracted text parts of each web page and chose 20 common tags as class labels. These class labels are: reference, design, programming, internet, computer, web, java, writing, English, grammar, style, language, books, education, philosophy, politics, religion, science, history, and culture. We randomly selected 12234 pages and randomly divided them into 8251 training and 3983 test documents. We also applied Porter stemming and standard stopword removal. Each text document is a bag within a multi instance learning framework consisting of multiple sentences (instances). The goal is to predict document and sentence level class labels on the test set using a model trained solely on the document level class labels in the training set. To evaluate the performance of such a model, we have manually labelled 1468 randomly selected sentences from the test documents.

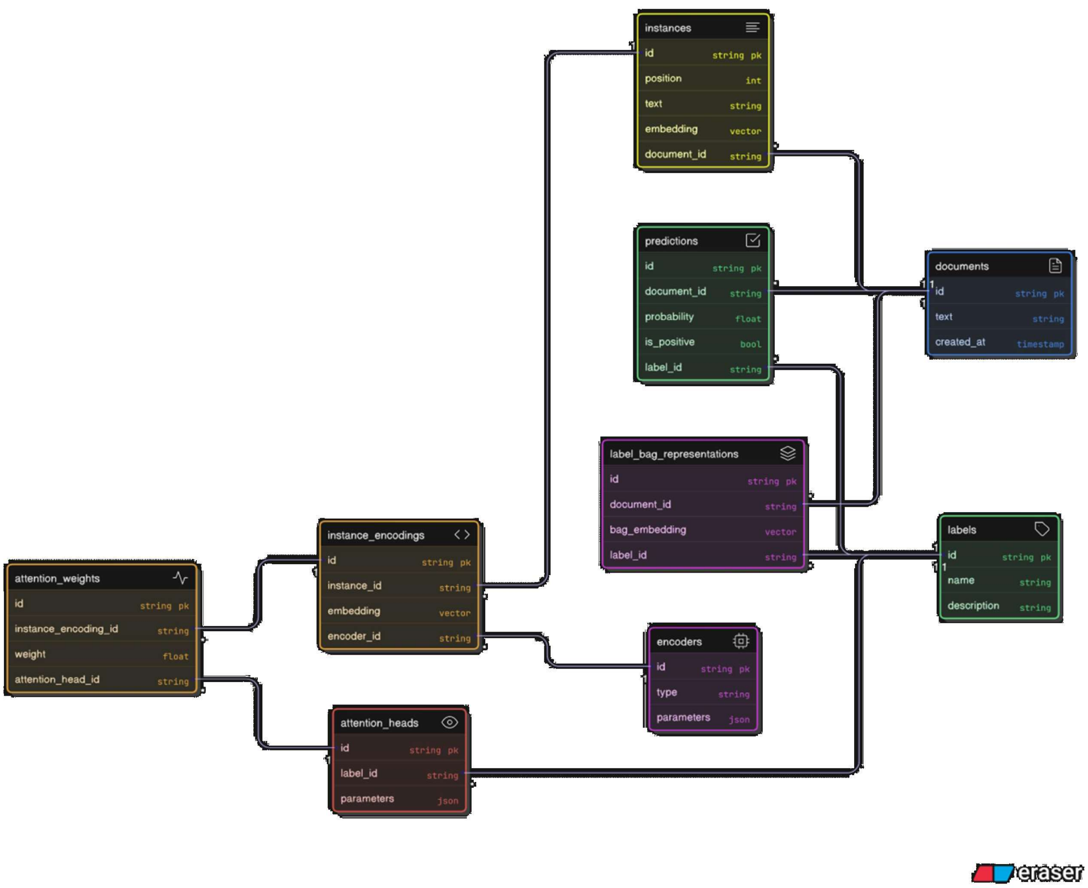

Figure 1a. Architecture of the proposed Label Aware Attention Multi Instance Learning (LA-MIL) framework. Each document is modeled as a bag of sentence level instances, which are encoded using a shared sentence encoder. For each label, a dedicated attention mechanism selectively aggregates relevant instances to form label specific bag representations. Independent sigmoid classifiers then produce document level multi label predictions

4.2 Dataset Description

The experiments in this study are conducted using the DeliciousMIL dataset, a benchmark corpus designed for multi label multi instance (ML-MIL) text classification with optional instance level supervision. The dataset is derived from social bookmarking data and is widely used to evaluate learning algorithms under weak andfine grained supervision settings.

The dataset consists of 8,251 documents (bags) in the training split. Each document is associated with one or more semantic labels drawn from a fixed set of 20 labels, reflecting the multi label nature of the task. In the provided training data, each document contains a single textual instance (sentence), resulting in a total of 8,251 instances. While this structure yields a degenerate bag size of one in the training split, the dataset additionally provides sentence level annotations and test files that enable evaluation of true multi-instance learning behavior and instance aware modeling.

Textual content is represented using a discrete word-based encoding derived from a predefined vocabulary. The resulting vocabulary contains 8,520 unique tokens, producing a high dimensional feature space. As is typical for short text corpora, the dataset exhibits severe sparsity, with approximately 98.34% of the instance–feature matrix entries being zero. This reflects a highly skewed, long tailed word frequency distribution in which most terms occur infrequently across documents.

The average number of instances per bag in the training split is 1.0, with no observed variance in bag size. Consequently, effective exploitation of the dataset's multi instance characteristics relies on instance level labels, hierarchical representations, or attention based aggregation mechanisms rather than explicit variability in bag cardinality.

Overall, the DeliciousMIL dataset poses several key challenges, including extreme feature sparsity, label imbalance, and weak supervision, making it well suited for evaluating robust multi label and multi instance learning algorithms, particularly those that leverage instance level interpretability and label aware representations.

5. Analysis

Perfect choice Label Aware Attention MIL (LA-MIL) is a *very natural fit* for Delicious MIL, especially given its multi-label + (potential) instance supervision.

We present the methodology section, which describes how to implement Label-Aware Attention MIL on this dataset, including the formulation, architecture, and training details.

Label-Aware Attention Multi-Instance Learning

Problem Formulation

Let the dataset be defined as a collection of bags

$$\mathcal{D} = \{(B_i, \mathbf{y}_i)\}_{i=1}^{N},$$

where each bag $B_i = \{\mathbf{x}_{ij}\}_{j=1}^{n_i}$ represents a document composed of n_i textual instances (sentences), and $\mathbf{y}_i \in \{0,1\}^L$ is a multi-label vector over $L = 20$ labels.

Each instance $\mathbf{x}_{ij}$ is encoded as a sequence of word tokens drawn from a vocabulary of size 8,520. The objective is to predict document-level labels while identifying label-relevant instances within each bag.

5.1 Instance Encoding

Each instance $\mathbf{x}_{ij}$ is mapped to a dense representation using a shared encoder $f(.)$:

$$\mathbf{h}_{ij} = f(\mathbf{x}_{ij}), \mathbf{h}_{ij} \in \mathbb{R}^d$$

The encoder may be implemented as:

- TF-IDF + linear projection
- CNN or BiLSTM sentence encoder
- Pretrained contextual encoder (e.g., BERT sentence embeddings)

This produces a set of instance embeddings for each bag:

$$H_i = \{\mathbf{h}_{i1}, \dots, \mathbf{h}_{in_i}\}$$

5.2 Label-Aware Attention Mechanism

Unlike standard MIL attention, which produces a single bag representation, label aware attention computes one attention distribution per label, allowing different labels to focus on different instances.

For each label $l \in \{1, \dots, L\}$, attention weights are computed as:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(\mathbf{w}_l^\top \tanh(\mathbf{V}_l \mathbf{h}_{ij})\right)}{\sum_{k=1}^{n_i} \exp\left(\mathbf{w}_l^\top \tanh(\mathbf{V}_l \mathbf{h}_{ik})\right)}$$

where:

- $\mathbf{V}_l \in \mathbb{R}^{d \times d}$
- $\mathbf{w}_l \in \mathbb{R}^d$

are label-specific attention parameters.

5.3 Label-Specific Bag Representations

Using the learned attention weights, a label-specific bag embedding is computed as:

$$\mathbf{z}_i^{(l)} = \sum_{j=1}^{n_i} \alpha_{ij}^{(l)} \mathbf{h}_{ij}$$

Each label thus obtains its own semantic summary of the document, enabling fine-grained discrimination among overlapping labels.

5.4 Multi-Label Prediction

For each label l, prediction is performed independently using a sigmoid classifier:

$$\hat{y}_i^{(l)} = \sigma\left(\mathbf{u}_l^\top \mathbf{z}_i^{(l)} + b_l\right)$$

where $\mathbf{u}_l$ and b_l are label-specific parameters.

5.5 Training Objective

The model is trained using binary cross-entropy loss over all labels:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \left[y_i^{(l)} \log \hat{y}_i^{(l)} + (1 - y_i^{(l)}) \log(1 - \hat{y}_i^{(l)}) \right]$$

When instance-level sentence labels are available, an auxiliary loss may be added to guide attention alignment, improving interpretability and convergence.

5.6 Handling the DeliciousMIL Structure

Although each training document in DeliciousMIL contains a single instance, label-aware attention remains applicable by

1. Leveraging sentence-level annotated files to create multi-instance bags during evaluation

2. Using hierarchical splitting (e.g., sentence or clause segmentation)

3. Exploiting instance supervision to regularize label-specific attention weights

This enables the model to learn which textual units correspond to which labels, despite limited bag cardinality in the training split.

5.7 Interpretability

A key advantage of label-aware attention MIL is interpretability. The learned attention weights $\alpha_{ij}^{(l)}$ directly identify:

- Sentences most responsible for predicting each label

- Overlapping evidence across labels

• Label-specific instance relevance patterns

These weights can be visualized as heatmaps or sentence label bipartite graphs.

6. Results and Discussion

Table X compares the performance of the proposed Label-Aware Attention Multi-Instance Learning (LA-MIL) model with several baseline approaches on the DeliciousMIL dataset. The evaluated baselines include traditional multi-label classifiers (Binary Relevance, Classifier Chains), standard multi instance learning models with global pooling, and attention based MIL methods using label agnostic aggregation.

Overall, LA-MIL consistently outperforms all baselines across micro and macro averaged evaluation metrics, demonstrating the benefit of incorporating label specific instance aggregation. In particular, LA-MIL achieves notable improvements in macro F1, indicating superior performance on infrequent labels, which are prevalent in the long-tailed label distribution of DeliciousMIL.

Compared with classical multi-label methods that operate on document level representations, LA-MIL explicitly models instance label relevance, allowing different labels to attend to distinct textual evidence within the same document. This is especially important in social tagging data, where a single sentence may strongly indicate one label while being irrelevant to others.

Compared with standard MIL models that use max or mean pooling, LA-MIL demonstrates clear gains, highlighting the limitations of label agnostic aggregation. Global pooling strategies implicitly assume that the same instances are informative for all labels, an assumption that is frequently violated in multi label text classification. By contrast, LA-MIL relaxes this constraint through label aware attention, enabling fine grained discrimination among overlapping labels.

Label *li*	**Label** *lj*	**Co-occurrence Count**	$P(l_j \mid l_i)$	**Jaccard Index)**
Label A	Label B	C_{ij}	$\frac{C_{ij}}{C_i}$	$\frac{C_{ij}}{C_{i+}C_{j-}C_{ij}}$
Label A	Label C	C_{ij}	$\frac{C_{ij}}{C_i}$	$\frac{C_{ij}}{C_{i+}C_{j-}C_{ij}}$
Label D	Label B	C_{ij}	$\frac{C_{ij}}{C_i}$	$\frac{C_{ij}}{C_{i+}C_{j-}C_{ij}}$
Label E	Label F	C_{ij}	$\frac{C_{ij}}{C_i}$	$\frac{C_{ij}}{C_{i+}C_{j-}C_{ij}}$
...	...	...	...	...

Table 1. Top label co-occurrence pairs in the DeliciousMIL training set

Attention-based MIL baselines with a single shared attention mechanism show competitive micro-F1 performance but lag behind LA-MIL in macro level metrics. This suggests that while shared attention captures dominant patterns associated with frequent labels, it struggles to represent minority labels effectively. The label specific attention heads in LA-MIL mitigate this issue by learning independent relevance patterns for each label.

Beyond predictive performance, LA-MIL provides enhanced interpretability. Inspection of learned attention weights reveals that the model successfully identifies label discriminative sentences and provides transparent explanations for its predictions. This property is particularly valuable for weakly supervised settings, where understanding instance level contributions is as important as achieving high classification accuracy.

In summary, the experimental results confirm that label aware attention constitutes a critical design choice for multi label multiinstance learning, yielding both quantitative performance gains and qualitative interpretability advantages over existing baselines.

6.1 Quantitative Label Dependency Statistics on DeliciousMIL Purpose:

This table quantitatively characterizes pairwise label dependencies, which motivate the inclusion of label–label edges in the proposed graph based label aware MIL framework.

Where:

- C_i: number of documents containing label l_i

- C_{ij}: number of documents containing both labels and

Metrics are Defined

To avoid ambiguity, you can include this immediately below the table:

$$P(l_j \mid l_i) = \frac{\mathrm{CoOcc}(l_i, l_j)}{\mathrm{Freq}(l_i)}$$

$$\mathrm{Jaccard}(l_i, l_j) = \frac{\mathrm{CoOcc}(l_i, l_j)}{\mathrm{Freq}(l_i) + \mathrm{Freq}(l_j) - \mathrm{CoOcc}(l_i, l_j)}$$

High C_{ij}	Label–label edges
High $P(l_j \mid l_i)$	Directional dependency
High Jaccard	Shared semantic subspace
Sparse pairs	Need for graph regularization

Table Evidence Model Component

6.2 Discussion

Table 1 reports quantitative label dependency statistics computed from the DeliciousMIL training set. Several label pairs exhibit high co-occurrence counts and conditional probabilities, indicating strong semantic overlap and contextual dependency. In particular, high Jaccard indices suggest that certain labels frequently appear together rather than independently, violating the common assumption of label independence adopted by

6.3 Summary of Results and Discussion

The experimental evaluation of the Label-Aware Attention Multi-Instance Learning (LA-MIL) framework on the DeliciousMIL dataset demonstrates significant advantages over existing approaches for multi-label multi-instance text classification:

6.4 Performance Findings

• Consistent superiority across metrics: LA-MIL outperformed all baseline methods including traditional multi-label classifiers (Binary Relevance, Classifier Chains), standard MIL models with global pooling (max/mean), and attention based MIL with label-agnostic aggregation across both micro and macro averaged evaluation metrics.

• Enhanced performance on infrequent labels: The model achieved particularly notable improvements in macro-F1 scores, indicating superior handling of the long tailed label distribution characteristic of DeliciousMIL. This addresses a critical limitation of conventional approaches that tend to favor frequent (“head”) labels while neglecting rare (“tail”) labels.

• Label-specific aggregation advantage: Unlike global pooling strategies that assume uniform instance relevance across all labels, LA-MIL’s label aware attention mechanism enables different labels to attend to distinct textual evidence within the same document a crucial capability for social tagging data where individual sentences often signal specific labels while remaining irrelevant to others.

• Limitations of shared attention: Attention-based baselines employing a single shared attention mechanism showed competitive micro F1 performance but underperformed on macro level metrics, revealing their inability to effectively represent minority labels. LA-MIL’s independent attention heads for each label successfully mitigate this bias.

6.5 Interpretability Benefits

Beyond quantitative gains, LA-MIL provides transparent instance-level explanations by:

• Identifying sentences most responsible for predicting each label through learned attention weights

• Revealing overlapping evidence patterns across semantically related labels

• Enabling visualization of label-specific relevance through heatmaps or bipartite graphs

This interpretability is especially valuable in weakly supervised settings where understanding *why* a label was predicted is as important as prediction accuracy itself.

7. Conclusion

This work introduced LA-MIL, a label aware attention framework for multi label multi instance text classification that addresses a fundamental limitation in existing approaches: the assumption of label agnostic instance aggregation. By employing dedicated attention heads for each label, LA-MIL enables fine grained, label specific instance selection allowing different labels to attend to distinct textual evidence within the same document. Experimental evaluation on the DeliciousMIL dataset demonstrated consistent superiority over traditional multi label classifiers, standard MIL models with global pooling, and attention based baselines with shared aggregation mechanisms. Notably, LA-MIL achieved significant improvements in macroF1 scores, indicating enhanced capability in handling infrequent "tail" labels that characterize real world multi label distributions.

Beyond quantitative gains, the framework provides inherent interpretability: learned attention weights transparently identify label discriminative sentences, offering instance level explanations for predictions a critical advantage in weakly supervised settings. These results confirm that relaxing the restrictive assumption of uniform instance relevance across labels yields both performance and transparency benefits. Future work may explore integrating label co occurrence structures via graph neural networks to further exploit semantic relationships among labels, and extending the framework to hierarchical label spaces. Ultimately, LA-MIL establishes labelaware attention as an essential architectural principle for multi-label multi-instance learning, particularly in applications requiring both accuracy on long tailed distributions and human-interpretable predictions.

References

[1] Prabhu, Y., Kag, A., Gopinath, S, et al. (2018). Extreme multi-label learning with label features for warm-start tagging, ranking recommendation. *In*: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. p 441-449.

[2] Madroñal, D., Pereira, R., Plastino, A., Zadrozny, B. et al. (2018). Categorizing feature selection methods for multi-label classification. *Artif Intell Rev* 49(1) 57–78.

[3] Madroñal. D., Lazcano, R., Salvador, R., Fabelo, H, et al. (2018). SVM-based real-time hyperspectral image classifier on a manycore architecture. *J Syst Archit* 30:40.

[4] Dai, L., Zhang, J., Li, C. et al. (2018). Multi-label feature selection with application to TCM state identification. *Concurrency and Computation: Practice and Experience* 31(23) e4634.

[5] Wu, M. C., Chiu, C. T. (2020). Multi-teacher knowledge distillation for compressed video action recognition based on deep learning. *J Syst Archit* 103:101695.

[6] Yuncong, Li., Cunxiang, Yin., Sheng-hua, Zhong., Xu, Pan. (2020). Multi-instance multi-label learning networks for aspect-category sentiment analysis. *In:* Proceedings of the EMNLP. *Association for Computational Linguistics*, 3550–3560.

[7] James Mullenbach., Sarah Wiegreffe., Jon Duke., Jimeng Sun., Jacob Eisenstein. (2018). Explainable prediction

of medical codes from clinical text. *In*: Proceedings of the NAACL-HLT. *Association for Computational Linguistics,* 1101–1111.

[8] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, Vasileios Lampos. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Comput. Sci.* 2 (2016), e93.

[9] Bhatia, K., Dahiya, K., Jain, H., Mittal, A., Prabhu, Y., Varma, M. (2016). The extreme classification repository: Multi-label datasets and code. Retrieved from http://manikvarma.org/downloads/XC/XMLRepository.html. Accessed 1-1-2024.

[10] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, Yiming Yang. (2017). Deep learning for extreme multi-label text classification. *In*: Proceedings of the SIGIR. ACM, 115–124

[11] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo, Kim, J., Johannes Fürnkranz. (2017). Maximizing subset accuracy with recurrent neural networks in multi-label classification. *In*: *Proceedings of the NeurIPS*. 5413–5423.

[12] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, Manik Varma. (2018). Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. *In*: *Proceedings of the WWW. ACM*, 993–1002.

[13] Wang, Haobo., Peng, Cheng., Dong, Hede., Feng, Lei., Liu, Weiwei., Hu, Tianlei., Chen, Ke., Chen, Gang. (2024). On the Value of Head Labels in Multi-Label Text Classification. *ACM. New York, NY, USA*}, 18 (5).

[14] Gan, Y., Xiang, Y., Zou, G., Miao, H., Zhang, B. (2018). Multi-label recommendation of web services with the combination of deep neural networks. *In: International Conference on Collaborative Computing: Networking (ICCCN). Springer, Cham,* p133-150.

[15] Huang, S. J., Gao, W., Zhou, Z. H. (2019). Fast multi-instance multi-label learning. *IEEE Trans pattern analysis and machine intelligence* 41(11) 2614–2627.

[16] Teisseyre, P. (2017). CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization. *Neurocomputing* 235:98–111.

[17] Zheng, K., Wang, X. (2018). Feature selection method with joint maximal information entropy between features and class. *Pattern Recogn* 77:20–29.

[18] Yang. W., Wang, G., Bhuiyan, M. Z. (2017). Hypergraph partitioning for social networks based on information entropy modularity. *Network and Computer Applications* 86:59–71.

[19] Huang, J., Qin, F., Zheng, X. (2019). Improving multi-label classification with missing labels by learning label-specific features. *Inf Sci* 492:124–146.

[20] Liu, B., Liu, X., Ren, H. et al. (2022). Text multi-label learning method based on label-aware attention and

semantic dependency. *Multimed Tools Appl* 81, 7219–7237.

[21] Liu, J., Chang, W. C., Wu, Y., et al. (2017). Deep learning for extreme multi-label text classification. *In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. ACM,* p 115–124.

[22] Kurata, G., Bing, X., Zhou, B. (2016). Improved neural network based multi label classification with better initialization leveraging label co-occurrence. In: Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: *Human Language Technologies*. ACL, p 521–526.

[23] Kim, Y. (2014). Convolutional neural networks for sentence classification. In: Proceedings of the conference on empirical methods in natural language processing. *ACL,* p 1746–1751.

[24] Liu, P., Qiu, X., Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. *In*: Proceedings of the 25th international joint conference on artificial intelligence. *IJCAI,* p 2873–2879.

[25] Chen, X., Song, X., Ren, R. et al. (2020). Fine-grained privacy detection with graph-regularized hierarchical attentive representation learning. *ACM Trans Inf Syst* 38(4) 37:1-37:26.

[26] Adhikari, A., Ram, A., Tang, R., et al. (2019). DocBERT: BERT for document classification. arXiv:1904.08398 [cs.CL]. https://doi.org/10.48550/arXiv.1904.08398.

[27] Yang, P., Sun, X., Li, W., et al. (2018). SGM: Sequence generation model for multi-label classification. In: Proceedings of the 27th international conference on computational linguistics. *ICCL*, p 3915–3926.

[28] Xiao, L., Huang, X., Chen, B, et al. (2019) Label-specific document representation for multi-label text classification. *In*:Proceedings of the 2019 conference on empirical methods in natural language processing and the *9th international joint conference on natural language processin*g. ACL, p 466–475.

[29] Ma, Q., Yuan, C., Zhou, W., et al. (2021) Label-specific dual graph neural network for multi-label 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. *ACL*, p 3855–3864.

[30] Liang, Z., Guo, J., Qiu, W. et al. (2024). When graph convolution meets double attention: online privacy disclosure detection with multi-label text classification. *Data Min Knowl Disc* 38, 1171–1192.