Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution

Yi Huang
Data and Analytics Chapter
Roche (China) Holding Ltd., China
yi.huang.yh4@roche.com

Buse Gildereli
Computer Engineering
Bogazici University, Turkey
Data and Analytics Chapter
Roche Müstahzarları San A.Ş., Turkey
busegildereli@gmail.com

Abdullatif Köksal
Computer Engineering
Bogazici University, Turkey
abdullatif.koksal@boun.edu.tr

Arzucan Özgür
Computer Engineering
Bogazici University, Turkey
arzucan.ozgur@boun.edu.tr

Elif Ozkirimli
Data and Analytics Chapter
F. Hoffmann-La Roche AG
Switzerland
elif.ozkirimli@roche.com

Abstract

Multi-label text classification is a challenging task because it requires capturing label dependencies. It becomes even more challenging when class distribution is long-tailed. Resampling and re-weighting are common approaches used for addressing the class imbalance problem, however, they are not effective when there is label dependency besides class imbalance because they result in oversampling of common labels. Here, we introduce the application of balancing loss functions for multi-label text classification. We perform experiments on a general domain dataset with 90 labels (Reuters-21578) and a domain-specific dataset from PubMed with 18211 labels. We find that a distribution-balanced loss function, which inherently addresses both the class imbalance and label linkage problems, outperforms commonly used loss functions. Distribution balancing methods have been successfully used in the image recognition field. Here, we show their effectiveness in natural language processing. Source code is available at https://github.com/Roche/BalancedLossNLP.

1 Introduction

Multi-label text classification is one of the core topics in natural language processing (NLP) and is used in many applications such as search (Prabhu et al., 2018) and product categorization (Agrawal et al., 2013). It aims to find the related labels from a fixed-set of labels for a given text that may have multiple labels. Figure 1 demonstrates examples

Title 1	PENN CENTRAL <PC> SELLS U.K. UNIT
Labels 1	acq (1650), strategic-metal (16) **nickel** (8)
Title 2	U.S. MINT SEEKING OFFERS ON COPPER, NICKEL
Labels 2	copper (47), **nickel** (8)

Figure 1: Samples for multi-label text classification task from Reuters-21578 dataset. Only titles are shown for illustration. Numbers after the labels indicate total number of occurrences in the dataset.

from the Reuters-21578 multi-label text classification dataset (Hayes and Weinstein, 1990). Here, for the document with the title *PENN CENTRAL <PC> SELLS U.K. UNIT*, the aim is to find the labels *acq (acquisitions), strategic-metal*, and *nickel* from 90 labels.

Multi-label classification becomes complicated when there is a long-tailed distribution (class imbalance) and linkage (co-occurrence) of labels. Class imbalance occurs when a small subset of the labels (namely head labels) have many instances, while majority of the labels (namely tail labels) have only a few instances. For example, half of the labels in the Reuters dataset, including *copper, strategic-metal,* and *nickel*, occur in less than 5% of the training data. Label co-occurrence or label linkage is a challenge when some head labels co-occur with rare or tail labels, resulting in bias for classification to the head labels. For example, even though the la-

bel *nickel* occurs less frequently, the co-occurrence information of *nickel/copper, nickel/strategic-metal* is important for accurate modeling (Figure 1). Solutions such as resampling of the samples with less-frequent labels in classification (Estabrooks et al., 2004; Charte et al., 2015), using co-occurrence information in the model initialization (Kurata et al., 2016), or providing a hybrid solution for head and tail categories with a multi-task architecture (Yang et al., 2020) have been proposed in NLP, however they are not suitable for imbalanced datasets or they are dependent on the model architecture.

Multi-label classification has been widely studied in the computer vision (CV) domain, and recently has benefited from cost-sensitive learning through loss functions for tasks such as object recognition (Durand et al., 2019; Milletari et al., 2016), semantic segmentation (Ge et al., 2018), and medical imaging (Li et al., 2020a). Balancing loss functions such as focal loss (Lin et al., 2017), class-balanced loss (Cui et al., 2019) and distribution-balanced loss (Wu et al., 2020) provide improvements to resolve the class imbalance and co-occurrence problems in multi-label classification in CV. Loss function manipulation has also been explored (Li et al., 2020b; Cohan et al., 2020) in NLP as it works in a model architecture-agnostic fashion by explicitly embedding the solution into the objective. For example, Li et al. (2020b) has borrowed dice-based loss function from a medical image segmentation task (Milletari et al., 2016) and reported significant improvements over the standard cross-entropy loss function in several NLP tasks.

In this work, our major contribution is the introduction of the use of balancing loss functions to the NLP domain for the multi-label text classification task. We perform experiments on Reuters-21578, a general and small dataset, and PubMed, a biomedical domain-specific and large dataset. For both datasets, the distribution balancing methods not only outperform the other loss functions for the total metrics, but also lead to significant improvement for the tail labels. We suggest that the balancing loss functions provide a robust solution for addressing the challenges in multi-label text classification.

2 Loss Functions

In NLP, Binary Cross Entropy (BCE) loss is commonly used for multi-label text classification (Bengio et al., 2013). Given a dataset
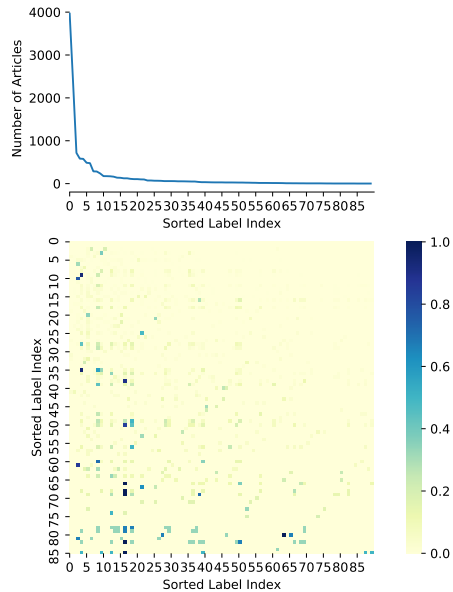

Figure 2: The long-tailed distribution and label co-occurrence for the Reuters-21578 dataset. The co-occurence matrix is color coded based on the conditional probability $p(i|j)$ of class in the ith column on class in the jth row.

$\{(x^1, y^1), ..., (x^N, y^N)\}$ with N training instances, each having a multi-label ground truth of $y^k = [y_1^k, ..., y_C^k] \in \{0, 1\}^C$ (C is the number of classes), and a classifier output $z^k = [z_1^k, ..., z_C^k] \in R$, BCE is defined as (the average reduction step is not shown for simplicity):

$$L_{BCE} = \begin{cases} -log(p_i^k) & \text{if } y_i^k = 1 \\ -log(1 - p_i^k) & \text{otherwise.} \end{cases} \quad (1)$$

The $sigmoid$ function is used for computing p_i^k, $p_i^k = \sigma(z_i^k)$. The plain BCE is vulnerable to label imbalance due to the dominance of head classes or negative instances (Durand et al., 2019). Below, we describe three alternative approaches that address the class imbalance problem in long-tailed datasets in multi-label text classification. The main idea of these balancing methods is to reweight BCE so that rare instance-label pairs intuitively get reasonable "attention".

2.1 Focal loss (FL)

By multiplying a modulating factor to BCE (with the tunable focusing parameter $\gamma \geq 0$), focal loss places a higher weight of loss on "hard-to-classify" instances predicted with low probability on ground

truth (Lin et al., 2017). For the multi-label classification task, the focal loss can be defined as:

$$L_{FL} = \begin{cases} -(1-p_i^k)^\gamma log(p_i^k) & \text{if } y_i^k = 1 \\ -(p_i^k)^\gamma log(1-p_i^k) & \text{otherwise.} \end{cases} \quad (2)$$

2.2 Class-balanced focal loss (CB)

By estimating the effective number of samples, class-balanced focal loss (Cui et al., 2019) further reweights FL to capture the diminishing marginal benefits of data, and therefore reduces redundant information of head classes. For multi-label tasks, each label with overall frequency n_i has its balancing term

$$r_{CB} = \frac{1-\beta}{1-\beta^{n_i}} \quad (3)$$

where $\beta \in [0,1)$ controls how fast the effective number grows and the loss function becomes

$$L_{CB} = \begin{cases} -r_{CB}(1-p_i^k)^\gamma log(p_i^k) & \text{if } y_i^k = 1 \\ -r_{CB}(p_i^k)^\gamma log(1-p_i^k) & \text{otherwise.} \end{cases} \quad (4)$$

2.3 Distribution-balanced loss (DB)

By integrating rebalanced weighting and negative-tolerant regularization (NTR), distribution-balanced loss first reduces redundant information of label co-occurrence, which is critical in the multi-label scenario, and then explicitly assigns lower weight on "easy-to-classify" negative instances (Wu et al., 2020).

First, to rebalance the weights, in the single-label scenario, an instance can be weighted by the resampling probability $P_i^C = \frac{1}{C}\frac{1}{n_i}$; while in the multi-label scenario, if following the same strategy, one instance with multiple labels can be over-sampled with a probability $P^I = \frac{1}{C}\sum_{y_i^k=1}\frac{1}{n_i}$. Therefore, the rebalanced weight can be normalized with $r_{DB} = P_i^C/P^I$. With a smoothing function, $\hat{r}_{DB} = \alpha + \sigma(\beta \times (r_{DB} - \mu))$, mapping r_{DB} to $[\alpha, \alpha+1]$, the rebalanced-FL (R-FL) loss function is defined as:

$$L_{R-FL} = \begin{cases} -\hat{r}_{DB}(1-p_i^k)^\gamma log(p_i^k) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}(p_i^k)^\gamma log(1-p_i^k) & \text{otherwise.} \end{cases} \quad (5)$$

Then, NTR treats the positive and negative instances of the same label differently. A scale factor λ and an intrinsic class-specific bias v_i are introduced to lower the threshold for tail classes and to avoid over-suppression.

$$L_{NTR-FL} = \begin{cases} -(1-q_i^k)^\gamma log(q_i^k) & \text{if } y_i^k = 1 \\ -\frac{1}{\lambda}(q_i^k)^\gamma log(1-q_i^k) & \text{otherwise.} \end{cases} \quad (6)$$

Table 1: Dataset Statistics

Dataset	Statistic
Reuters-21578	
Number of documents	10788
Number of labels	90
Average number of labels per instance	1.24
Average number of instances per label	148.11
PubMed	
Number of documents	224897
Number of labels	18211
Average number of labels per instance	12.30
Average number of instances per label	151.88

where $q_i^k = \sigma(z_i^k - v_i)$ for positive instances and $q_i^k = \sigma(\lambda(z_i^k - v_i))$ for negative ones. The v_i can be estimated by minimizing the loss function at the beginning of training with a scale factor κ and class prior $p_i = n_i/N$, so that

$$\hat{b}_i = -log(\frac{1}{p_i} - 1), v_i = -\kappa \times \hat{b}_i \quad (7)$$

Finally, DB integrates rebalanced weighting and NTR as

$$L_{DB} = \begin{cases} -\hat{r}_{DB}(1-q_i^k)^\gamma log(q_i^k) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^\gamma log(1-q_i^k) & \text{otherwise.} \end{cases} \quad (8)$$

3 Experiments

3.1 Datasets

Two multi-label text classification datasets of different size, property and domain are used (Table 1).

Reuters-21578 dataset (Distribution 1.0) contains documents that appeared on Reuters newswire in 1987 and that were manually annotated with 90 labels (Hayes and Weinstein, 1990). Here, we follow the train-test split used by (Yang and Liu, 1999) to obtain 7769 training (1000 among which for validation) and 3019 test documents. The labels are equally split into head (30 with $\geq$ 35 instances), medium (31 with between 8-35 instances) and tail (30 with $\leq$ 8 instances) subsets.

PubMed dataset comes from the BioASQ Challenge (License Code: 8283NLM123) providing PubMed articles with titles and abstracts, that have been manually labelled for Medical Subject Headings (MeSH) (Tsatsaronis et al., 2015; Coordinators, 2017). 224,897 articles published during 2020 and 2021 are used, among which 10,000 are used for validation and testing purpose. The 18,211 labels are split by 3-quantiles into head (6018 with

Table 2: Micro and macro F1 scores for multi-label classification of Reuters-21578 (left) and PubMed (right) using the SVM model or different loss functions. The F1 scores are reported for the total set of labels as well as for the head, medium and tail label sets, with the number of instances given in parenthesis. The experiments are performed with the SVM one-vs-rest model (SVM), the binary cross entropy (BCE), focal loss (FL), class balanced focal loss (CB), rebalanced focal loss (R-FL), negative-tolerant regularization FL (NTR-FL), distribution balance with no FL (DB-0FL), class balanced FL with negative regularization (CB-NTR) and distribution balanced loss (DB).

Model/ Loss Function	Reuters Total miF/maF	Reuters Head(≥35) miF/maF	Reuters Med(8-35) miF/maF	Reuters Tail(≤8) miF/maF	PubMed Total miF/maF	PubMed Head(≥50) miF/maF	PubMed Med(15-50) miF/maF	PubMed Tail(≤15) miF/maF
SVM	87.60/51.63	89.87/78.47	66.92/61.00	22.54/13.83	58.54/13.31	60.77/34.33	19.78/5.62	6.94/0.67
BCE	89.14/47.32	91.75/82.81	66.28/57.26	0.00/0.00	26.17/0.02	27.61/0.06	0.00/0.00	0.00/0.00
FL	89.97/56.83	91.83/82.64	76.16/70.63	27.40/15.37	58.30/13.94	60.43/33.69	26.39/8.15	8.58/0.86
CB	89.23/52.96	91.56/80.44	71.64/66.61	23.08/9.93	58.57/13.67	60.75/33.40	24.50/7.39	9.92/1.01
R-FL	89.47/54.35	91.59/80.39	72.86/66.69	25.00/14.22	57.90/14.66	59.85/34.09	30.32/9.70	11.45/1.15
NTR-FL	90.70/60.70	92.37/82.65	79.35/75.34	39.51/22.33	60.92/16.99	**63.15**/38.85	33.14/11.39	15.86/1.82
DB-0FL	89.45/57.98	91.21/82.05	77.33/71.11	31.17/19.05	58.95/15.15	60.99/34.92	31.06/10.02	14.23/1.49
CB-NTR	**90.74**/63.31	**92.46**/83.28	78.42/72.98	46.32/**32.31**	**61.07**/18.40	63.02/39.95	37.18/13.43	24.15/2.97
DB	90.62/**64.47**	92.14/**83.48**	**80.25/77.01**	**48.89**/31.39	60.63/**19.19**	62.39/**40.48**	**41.14/15.33**	**24.19/3.08**

≥ 50 instances), medium (5581 with between 15-50 instances) and tail (6612 with ≤ 15 instances) subsets.

3.2 Experimental Settings

We compare the use of different loss functions, and SVM one-vs-rest model as a classical multi-label classification baseline. For each dataset and method, we evaluate its best micro-F1 and macro-F1 scores (Wu et al., 2019; Lipton et al., 2014) for the whole label set (total) as well as different subsets of label frequency (head/medium/tail). The loss function parameters, the classification models used, and the implementation details are provided in Appendix A.

4 Results

A summary of the results of different loss functions are listed in Table 2.

There are about 10,000 documents and 90 labels in the Reuters dataset, with an average of 150 instances per label (Table 1). Figure 2 shows the long-tailed distribution where only a few labels have a high number of articles and these head labels also have high co-occurrence with other labels. The impact of the skewed distribution can also be seen from the comparison between the micro-F1 (around 90 for different loss functions) and macro-F1 (around 50-60) scores (Table 2). Furthermore, among loss functions, BCE has the lowest performance for the Reuters dataset with total macro-F1 score of 47 and tail F1 scores of 0. The PubMed dataset contains around 225,000 documents with

18,000 labels (Table 1) and the imbalance is even more pronounced for this large dataset (Figure in Appendix) and the difference between the total micro-F1 score (60) and the total macro-F1 score (around 15) is very high. Overall, SVM underperforms the proposed distribution balanced loss functions in both datasets.

Experiments with Reuters-21578 dataset. The loss functions FL, CB, R-FL and NTR-FL perform similar to BCE in head classes, yet outperform BCE in medium and tail classes, indicating the advantage of handling imbalance. DB provides the biggest improvement in tail class assignment; the tail micro-F1 score gains 21.49 from FL and 25.81 from CB. It outperforms prior works that also used this commonly used dataset, including approaches based on Binary Relevance, EncDec, CNN, CNN-RNN, Optimal Completion Distillation or attention-based GNN, that achieved micro-F1<89.9 (Nam et al., 2017; Pal et al., 2020; Tsai and Lee, 2020)

Experiments with PubMed dataset. PubMed is a biomedical domain specific, larger dataset with bigger class imbalance. For this dataset, BCE does not work efficiently, therefore we use FL as a strong baseline. With FL, the medium and tail micro-F1 scores are 26 and 9. All other loss functions outperform FL in medium and tail classes, indicating the advantage of balancing label distribution. DB again has the highest performance for all classes but the most significant improvement is achieved for the medium (micro-F1:41) and tail (micro-F1:24) classes.

Ablation Study. We further investigate the contribution of the three layers of DB by comparing DB results with R-FL, NTR-FL and DB without the focal layer (DB-0FL). As shown in Table 2, for both datasets, removing the NTR layer (R-FL) or the focal layer (DB-0FL) reduces model performance for all subsets. Removing the rebalanced weighting layer (NTR-FL) yields similar total micro-F1 (Reuters: 90, PubMed:60) but the macro-F1 as well as medium and tail F1 scores are higher with DB , showing the value of adding the rebalancing weighting layer. We also test the contribution of NTR by integrating it with CB, yielding a novel loss function CB-NTR that has not been previously explored. For both datasets, CB-NTR has better performance than CB for all class sets (Table 2). The only difference between CB-NTR and DB is the use of CB weight r_{CB} instead of the rebalancing weight $\hat{r}_{DB}$. DB has very close performance to or outperforms CB-NTR in the medium and tail classes, suggesting that the $\hat{r}_{DB}$ weight, which addresses the co-occurrence challenge, is useful.

Error Analysis. We perform an error analysis and observe that the most common errors are due to incorrect classification to similar or linked labels for all loss functions. The most common three pairs of classes confused by all loss functions for the Reuters dataset are: *platinum* and *gold*, *yen* and *money-fx*, *platinum* and *copper*. For the PubMed dataset, the most common errors are: *Pandemics* and *Betacoronavirus*, *Pandemics* and *SARS-CoV-2*, *Pneumonia, Viral* and *Betacoronavirus*, and BCE has significantly more errors for these classes compared to the other investigated loss functions.

5 Conclusion

We propose and compare the application of a series of balancing loss functions to address the class imbalance problem in multi-label text classification. We first introduce the loss function DB to NLP and design a novel loss function CB-NTR. The experiments show that the DB outperforms other approaches by considering long-tailed distribution and label co-occurrence, and its performance is robust to different datasets such as Reuters (90 labels, general domain) and PubMed (18,211 labels, biomedical domain). This study demonstrates that addressing challenges such as class imbalance and label co-occurrence through loss functions is an effective approach for multi-label text classifica-

tion. It does not require additional information and can be used with all types of neural network-based models. It may also be a powerful strategy for other NLP tasks, such as part-of-speech tagging, named entity recognition, machine reading comprehension, paraphrase identification and coreference resolution, all of which usually suffer from long-tailed distribution.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

NCBI Resource Coordinators. 2017. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 46(D1):D8–D13.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

T. Durand, N. Mehrasa, and G. Mori. 2019. Learning a deep convnet for multi-label classification with partial labels. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 647–657, Los Alamitos, CA, USA. IEEE Computer Society.

Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36.

Weifeng Ge, Sibei Yang, and Yizhou Yu. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Philip J. Hayes and Steven P. Weinstein. 1990. Construe/tis: A system for content-based indexing of a database of news stories. In *Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence*, IAAI '90, page 49–64. AAAI Press.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 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*, pages 521–526.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jianqiang Li, Guanghui Fu, Yueda Chen, Pengzhi Li, Bo Liu, Yan Pei, and Hui Feng. 2020a. A multi-label classification model for full slice brain computerised tomography image. *BMC Bioinformatics*, 21(6):200.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020b. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Los Alamitos, CA, USA. IEEE Computer Society.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize f1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. 2020. Magnet: Multi-label text classification using attention-based graph neural network. In *ICAART (2)*, pages 494–505.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Che-Ping Tsai and Hung-yi Lee. 2020. Order-free learning alleviating exposure bias in multi-label classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6038–6045. AAAI Press.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. 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 4354–4364, Hong Kong, China. Association for Computational Linguistics.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020*, pages 162–178, Cham. Springer International Publishing.

Wenshuo Yang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2020. HSCNN: A hybrid-Siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6716–6722, Online. Association for Computational Linguistics.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 42–49, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Experimental Settings

Evaluation metrics. For each dataset and method, we select the threshold with the best micro-F1 score on the validation set as our final model and evaluate its performance on the test set with micro-F1 and macro-F1 scores.

Loss function parameters. We compare the performance of DB with different loss functions, where BCE or its modifications are used. The methods include: (1) BCE with all instances and labels of the same weight. (2) FL (Lin et al., 2017): we use γ=2. (3) CB (Cui et al., 2019): we use β=0.9. (4) R-FL (Wu et al., 2020): we use α=0.1 and β=10, μ=0.9 (Reuters-21578) or 0.05 (PubMed). (5)NTR-FL (Wu et al., 2020): we use κ=0.05 and λ=2. (6) DB (Wu et al., 2020): we use same parameters with R-FL and NTR-FL when applicable.

Implementation Details. We use the *BertForSequenceClassification* backbone in *transformers* library (Wolf et al., 2020) with the bert-base-cased pretrained model (Devlin et al., 2018) for Reuters-21578 dataset and the biobert-base-cased-v1.1 pretrained model (Lee et al., 2019) for PubMed dataset. bert-base-cased and biobert-base-cased-v1.1 are base BERT models with 110 million parameters. The training data are truncated with a maximal length of 512 and grouped with a batch size of 32. We use AdamW with a weight decay of 0.01 as the optimizer, and determine the learning rate by hyperparameter search. The experiments are implemented in PyTorch. For Reuters-21578 dataset we use one-GPU (V100) experiments which takes 5 minutes for one epoch. For PubMed dataset, we use one-GPU (A100) experiments which takes 1 hour for one epoch. For the SVM one-vs-rest model, we use *scikit-learn* library (Pedregosa et al., 2011) with TF-IDF features. With hyperparameter search, we apply the linear kernel and hyper-plane shifting optimized on each validation set.

A.2 Additional Effectiveness Check

We further investigate the effectiveness of loss functions against the number of labels per instance (Table 3 in Appendix). For the Reuters dataset, we split the test instances into two groups, 2583 instances with only one label and 436 instances with multiple labels. On single-label instances, all functions from BCE to DB, have similar performance; while on multi-label instances, the performance of BCE drops more than DB. DB outperforms other
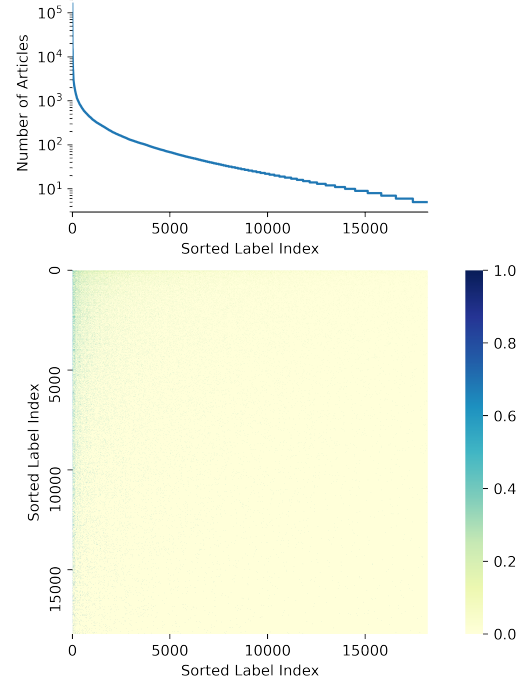

Figure 3: The long-tailed distribution and label co-occurrence for the PubMed dataset. The y-axis of distribution curve is log-scale, and the co-occurence matrix is color coded based on the quad root (for better visualization) of conditional probability $p(i|j)$ of class in the ith column on class in the jth row.

functions in micro-F1 of the multi-label instance group and macro-F1 of both groups. There are < 0.1% instances of PubMed dataset with a single label, so we divide instances into 3-quantiles by their number of labels. In each quantile, the novel NTR-FL, CB-NTR and DB outperform the rest of the models in all metrics.

Table 3: Micro and macro F1 scores for multi-label classification of Reuters-21578 (left) and PubMed (right) using different loss functions. The F1 scores are reported for the total set of labels as well as for groups split by the number of labels per instance. The experiments are performed with the binary cross entropy (BCE), focal loss (FL), class balanced focal loss (CB), rebalanced focal loss (R-FL), negative-tolerant regularization FL (NTR-FL), distribution balance with no FL (DB-0FL), class balanced FL with negative regularization (CB-NTR) and distribution balanced loss (DB).

Loss Function	Reuters Total miF/maF	Reuters Single-label miF/maF	Reuters Multi-label miF/maF	PubMed Total miF/maF	PubMed $\leq$ 9 labels miF/maF	PubMed 10-14 labels miF/maF	PubMed $\geq$ 15 labels miF/maF
BCE	89.14/47.32	94.11/41.44	76.26/33.11	26.17/0.02	16.48/0.01	27.36/0.02	30.36/0.03
FL	89.97/56.83	94.81/50.33	77.54/40.07	58.30/13.94	53.72/7.44	59.02/10.27	59.72/8.63
CB	89.23/52.96	94.10/44.72	77.27/38.80	58.57/13.67	54.41/7.40	59.21/10.11	59.82/8.51
R-FL	89.47/54.35	95.21/47.45	74.29/38.79	57.90/14.66	53.08/7.67	58.60/10.50	59.45/8.81
NTR-FL	90.70/60.70	**95.42**/51.33	78.85/44.37	60.92/16.99	**58.51**/9.07	**61.86**/12.31	61.12/10.20
DB-0FL	89.45/57.98	94.48/51.80	76.63/42.26	58.95/15.15	55.14/8.11	59.84/10.90	59.85/8.94
CB-NTR	**90.74**/63.31	95.17/51.08	79.56/49.94	**61.07**/18.40	58.29/9.67	61.72/12.97	**61.72**/10.77
DB	90.62/**64.47**	94.49/**54.31**	**81.17**/50.12	60.63/**19.19**	57.81/**9.76**	61.53/**13.49**	61.08/**11.23**