AFL-BERT : Enhancing Minority Class Detection in Multi-Label Text Classification with Adaptive Focal Loss and BERT

Zakia Labd[1], Said Bahassine[2], Khalid Housni[3]
Laboratory of Research in Informatics L@RI, Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco[1, 3]
Laboratory of Artificial Intelligence, Modeling and Computational Engineering (AIMCE), Department of Computer Science, National Higher School of Arts and Crafts, Hassan II University, Casablanca, 150 Bd du Nil, Casablanca 20670, Morocco[2]

***Abstract*—Fine-tuning transformer models like Bidirectional Encoder Representations from Transformers has enhanced text classification performance. However, class imbalance remains a challenge, causing biased predictions. This study introduces an improved training strategy using a novel Adaptive Focal Loss with dynamically adjusted γ based on class frequencies. Unlike static γ values, this method emphasizes minority classes automatically. Experiments on the CMU Movie Summary dataset show Adaptive Focal Loss surpasses standard binary cross-entropy and Focal Loss, achieving an F1-score of 0.5, ROC accuracy of 0.79, and Micro Recall of 0.53. These results demonstrate the effectiveness of adaptive focusing methods in improving the detection of minority classes in imbalanced scenarios.**

***Keywords*—Adaptive focal loss; BERT; imbalanced text classification; multilabel text classification**

I. Introduction

Multi-label text classification is a crucial task in natural language processing (NLP), where each input instance may be associated with multiple labels. A major issue that faces this task is the inability of traditional methods, such as bag-of-words and Word2vec, to effectively capture the contextual relationships within longer texts, leading to classification performance. Jain U. et al. [1] address these challenges by introducing utilizing named entity recognition (NER) in conjunction with transformer models, such as BERT and DistillBERT. By extracting keywords specific to each class and employing an enhanced attention mechanism, the proposed approach improves classification accuracy.

In many real-world scenarios, the datasets are often highly imbalanced, where some minority classes can be significantly underrepresented compared to others. This class imbalance poses a hard challenge, as loss functions tend to prioritize dominant classes, resulting in poor performance on the minority classes compared to the majority one. The main issue with standard loss functions like cross-entropy [2, 3] is that they can't handle class imbalance in multi-label datasets effectively. This is because they consider all training examples equally, regardless of their relative importance or hardness level. Yasuda et al. (2024) introduced a Weighted Asymmetric Loss that combines label frequency weighting and label co-occurrence smoothing to improve performance on imbalanced multi-label text datasets [4]. Similarly, Park et al. [5] proposed a Robust Asymmetric Loss that emphasizes hard positive examples while minimizing the impact of abundant negative samples, showing strong performance on long-tailed multi-label tasks. Huang et al. [6] extended this approach with the Asymmetric Polynomial Loss, allowing for greater control over gradient contributions from positive and negative classes.

Traditional approaches have had limited success data-level methods, such as oversampling and undersampling, can balance class distributions but tend to introduce artificial patterns that are harmful to generalization. Undersampling, on the other hand, risks discarding informative data from the majority classes. In [7], the authors demonstrated that these techniques frequently failed to improve predictive performance and, in some cases, resulted in overestimated risks and miscalibration of clinical prediction models. Class weighting methods aim to address imbalance by assigning greater importance to minority classes. Finding the optimal weights often requires extensive hyperparameter tuning. Putra and Ibrohim [8] note that grid search or other optimization techniques are typically used, increasing the complexity and computational cost of model development.

Focal Loss [9] introduced an important advancement by introducing a dynamic reweighting mechanism. The most important contribution of Focal Loss is its ability to automatically down-weight readily classified instances and focus computational effort on hard cases. This is achieved by the focusing parameter γ, which balances the loss contribution in proportion to prediction confidence.

An important limitation of the Focal Loss shows in multi-label scenarios: the fixed focusing parameter cannot cope with varying class frequencies. In real-world datasets, where class distributions are highly imbalanced and interdependent, this fixed approach struggles. The same γ value will be too strong for some classes and weak for others, leading to sub-optimal model performance. Mukhoti et al. [10] observed that higher values of γ can make training unstable, especially in early stages. The study notes that while higher γ values may improve calibration, they can also result in negative outcomes if the model adjusts weights too aggressively at the beginning. Therefore, careful consideration is needed when selecting γ to avoid destabilizing the training process.

This is a consideration that requires an adaptive focusing mechanism that adjusts dynamically according to class-specific features and their interactions, and to address this problem, we propose AFL-Bert. Our contribution is three-fold, first an Adaptive Focal Loss framework that adjusts the focusing parameter (γ) dynamically according to class frequencies. In this manner, the loss function can assign more weight to rare classes without sacrificing a balance over all labels. Second, we fine-tune the whole BERT model to leverage its contextual embedding capacity and improve representation learning for multi-label classification. Third, our experiments on the CMU Movie Summary dataset show that this approach attains considerable improvements over the baseline, particularly in the detection of rare classes.

The rest of this study is structured as follows: Section II discusses related work, highlighting existing solutions for imbalanced multi-label classification. Section III outlines our proposed methodology, detailing the Adaptive Focal Loss and its integration with BERT. Section IV describes the implementation, while Section V presents experimental results. Section VI presents a discussion of findings. Section VII presents the limitations and future work. Finally, we conclude our study in Section VIII.

II. Related Work

Recent advances in multi-label text classification with imbalances have focused on employing hierarchical modeling, contrastive learning, and label co-occurrence methodologies to overcome issues like long-tail label distribution and dependency modeling. Zhang et al. put forward HCL-MTC as a novel hierarchical contrastive learning model that infers label dependence and semantic dissimilarity to achieve significantly improved classification performance [11]. Similarly, HDLTex++, presented in Expert Systems with Applications, utilizes hierarchical deep architectures that effectively manage label imbalance with a structured levels organization of text [12]. Yan et al.'s LabelCoRank model addresses long-tail using a reranking framework based on co-occurrence that enhances the accuracy of prediction of rare labels [13]. Simultaneously, HGBL combines BiLSTM with contrastive learning and Graphormer with the aim of enhancing fine-grained hierarchical label interactions, additionally underlining the importance of both local and global textual contexts [14].

Researchers have developed diverse strategies to address class imbalance in multi-label text classification. ML-OUSCA combines multilabel over-sampling and under-sampling and class alignment, which is better than K-means SMOTE and KNN-US on benchmark collections like Reuters-21578 and Enron but at the expense of losing informative information or overfitting [15]. Experiments with focal loss and mix-up techniques in extreme multi-label classification (XML) models demonstrate performance gain on datasets, particularly in transformer-based models [16]. BERT-based variants like EnvBERT [17] improve noisy environmental news classification over 80% through oversampling the data. Minority class oversampling based on imbalance ratios improves sentiment analysis classifiers [18], while Partial Label Masking (PLM) [19] enhances rare class recall through adaptive label masking during training. Generation of synthetic data through GPT-2 and LSTM models improves accuracy by 17% in instances of extreme imbalance but is prone to overfitting [20]. Advanced models like MISO [21] utilize mutual information constraints to re-embed difficult samples and achieve significant representation learning gains. Such approaches highlight the trade-off between resampling efficiency, model architecture modifications, and synthetic data quality in label imbalance resolution.

Focal Loss, introduced by Lin et al. [9], was a pivotal advancement in addressing class imbalance by down-weighting well-classified examples and focusing the model on harder cases. However, the original formulation relies on a fixed focusing parameter γ, which may not generalize well across tasks with varying class distributions. To address this, Mukhoti et al. [10] proposed calibrating deep neural networks using focal loss, highlighting that the choice of γ greatly affects model reliability and confidence estimation. Similarly, Cao et al. [24] introduced the Label-Distribution-Aware Margin Loss, which dynamically adjusts margins based on class frequency, offering an alternative way to balance class contributions. More recently, Cai et al. [22] demonstrated that combining focal loss with data augmentation and oversampling further improves performance on imbalanced multi-label tasks. These studies underscore the ongoing evolution of adaptive loss strategies, providing a strong foundation for our proposed dynamic focusing mechanism.

A. Focal Loss Overview

Focal loss, introduced by Facebook AI Research [9], is a novel method employed to mitigate the class imbalance issue for object detection. Conventional loss functions, such as cross-entropy, fail to work in cases with a huge gap between foreground and background classes. This imbalance causes the loss to predominantly arise from the readily identifiable negative instances at the expense of contributions from less frequent positive instances. The model will thus degrade during training when presented with such a gradient, where these negative classes determine the trend, hence making it less effective when trained from such rare classes. The key idea behind focal loss is to transform the standard cross-entropy loss function into a modified version that introduces a modulating factor to decrease the relative loss for highly classified samples. By doing so, the model will pay closer attention to difficult-to-classify samples, and as a result, easy negatives won't dominate the training process. Focal loss introduces a specific focusing parameter γ that can be tuned to control the level of this down-weighting effect. As γ increases, the model pays greater attention to mislabeled samples, learning from them better. Mathematically, focal loss is defined as in Eq. (1):

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} log(p_t) \quad (1)$$

In the context of focal loss, α serves as the balancing factor, while γ is referred to as the focusing parameter. The focusing parameter plays a crucial role in adjusting the rate at which easily classified examples are downweighted during training. Intuitively, this means that focal loss reduces the contribution of loss from well-classified examples, thereby allowing the model to focus more on challenging instances, such as background objects in image classification tasks.

B. Limitations of Standard Focal Loss

Recent advances in addressing class imbalance in multi-label text classification have highlighted the efficacy of focal loss. Cai et al. demonstrated that the integration of focal loss with data augmentation techniques like Easy Data Augmentation (EDA) and oversampling significantly improved model performance in the SemEval-2025 Task 9 on food hazard detection, particularly improving accuracy and F1 scores for minority classes [22]. Consequently, Sen (2021) employed focal loss within a convolutional neural network framework for genre classification of movie descriptions and noted that focal loss outperformed binary cross-entropy loss for both balanced and imbalanced datasets [23]. However, several works [10, 24] have illustrated that a fixed parameter is not ideal, as it assumes that the classes must be equally hard. This is not true in datasets where some classes appear much less frequently than others, which leads to the following issues:

- A low γ value reduces the modulation effect of focal loss, making it behave more like standard cross-entropy loss. This provides less focus on hard-to-classify examples, which is especially problematic in imbalanced datasets where minority classes often contain difficult examples that need more attention during training.
- Suboptimal Generalization: Without an adaptation process, the model cannot learn an optimal compromise between majority and minority classes and thus tends to have poor recall on a few classes and overfitting on common classes.

Despite these advancements, most prior work either employed static loss functions or required heavy manual tuning, which limited scalability. Additionally, few methods dynamically adjusted the loss function based on class frequency. This work addresses these gaps by introducing dynamically adaptive focal loss in conjunction with BERT, offering an automated solution for imbalanced multi-label text classification.

III. Methodology

This section outlines the proposed method for solving imbalanced multi-label text classification by combining Adaptive Focal Loss (AFL) with a BERT. The AFL has a dynamic focusing parameter (γ) that depends on class frequencies, enabling the model to prioritize less frequent classes more optimally. AFL-BERT: a combination of AFL with BERT leverages the transformer's contextual embeddings to create a strong and efficient classification framework, as shown in Fig. 1.

A. Introducing Adaptive Focal Loss

To overcome these limitations, we propose to dynamically adjust based on class frequencies. Prior research has pointed at the need for adaptive loss processes. Cao et al. [24] proposed Label Distribution-Aware Margin Loss, demonstrating class-dependent compensation as leading to performance improvements in imbalanced classification. Moreover, Mukhoti et al. [10] emphasized the need for calibrated deep learning models, demonstrating that static loss parameters often lead to biased training outcomes.

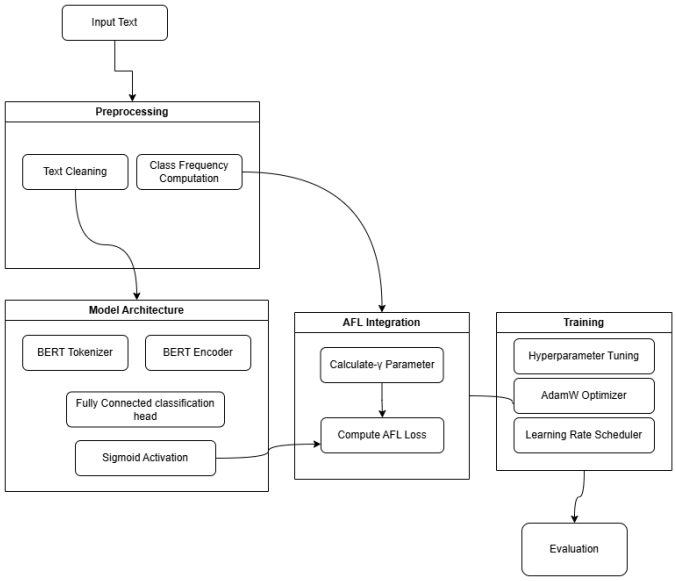

Fig. 1. Methodology of the study.

B. Mathematical Formulation

The Adaptive Focal Loss (AFL) modifies the standard Focal Loss by introducing a class-dependent focusing parameter γ_c, defined in Eq. (2):

$$\gamma_c = base_\gamma \times \frac{1}{f_c} \quad (2)$$

where, f_c represents the normalized frequency of class c, and $base_\gamma$ is a hyperparameter that sets the baseline focusing level. The AFL for a single sample in a multi-label setting is computed as shown in Eq. (3):

$$AFL(y, \hat{y}) = \sum_{c=1}^{C} -\alpha_c (1 - p_t)^{\gamma_c} \log(p_t) \quad (3)$$

where,

- p_t is the predicted probability for the true class.
- α_c is a weighting factor for class c.
- Cis the total number of classes.

This formulation aims to adapt Focal Loss to handle class imbalance by dynamically adjusting the focusing parameter γ_c based on the frequency of each class.

C. Key Features

- Dynamic Adjustment: AFL's focusing parameter adapts to class frequencies, allowing it to allocate greater emphasis to minority classes.
- Compatibility: The AFL function integrates seamlessly with existing architectures, requiring minimal changes to the training pipeline.
- Scalability: The method generalizes across datasets with varying degrees of imbalance.

IV. Experimental Setup

A. Model Architecture

The proposed model, as shown in Fig. 2, leverages BERT-based architecture as the backbone of feature extraction using the strength of pre-trained transformer layers to output high-level contextual embeddings of the input text. BERT, introduced by Devlin et al., has proven strong performance across a wide variety of natural language processing tasks due to its ability to induce deep bidirectional context, making it extremely suitable for multi-label classification [25]. BERT's tokenizer is used to tokenize the input to get input IDs, attention masks, and token type IDs, all of which are structurally sound for the transformer encoder of the model. All these tokens go into the encoder layers, resulting in contextualized word embeddings as the input passes through them for local and global linguistic dependencies.

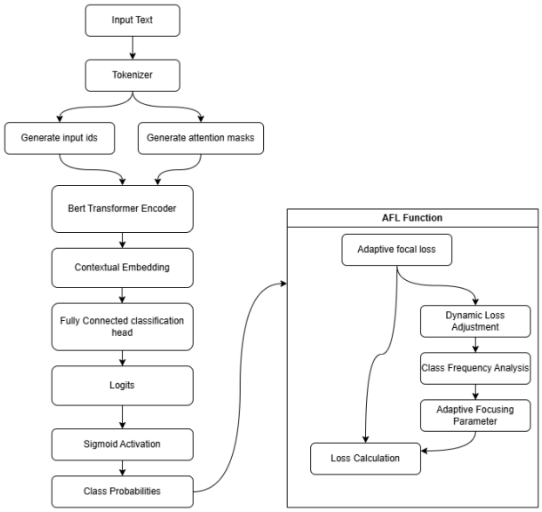

Fig. 2. AFL-Bert architecture.

A fully-connected classification head is put above the encoder for multi-label prediction. A sigmoid rather than softmax is used as the activation function, with the model then able to generate independent probability scores for each label, an important property for multi-label classification where labels can be non-exclusively relevant [26]. In a move to address class imbalance in the standard multi-label dataset, the model includes Adaptive Focal Loss (AFL) during training. AFL enhances the focal loss proposed by Lin et al. by scale-adaptively scaling the focusing factor in relation to the predicted probabilities to assign higher weights to more difficult or minority class samples without human tuning [9].

The AFL mechanism operates by taking advantage of the logits from the classification head, inserting sigmoid activation to obtain class probabilities, and dynamically calculating the loss to penalize more severely infrequent label misclassifications. This adaptive weighting helps the model focus learning on minority classes, thereby enhancing recall for minority classes without compromising overall classification performance. The combination of BERT's contextual representations and the adaptive re-weighting of AFL results in a more balanced and effective multi-label classification model [22].

B. Description of the CMU Movie Summary Corpus Dataset

The CMU Movie Summary Corpus dataset is made up of more than 200,000 movie summaries collected from differing sources, including online movie databases and user-generated content. Synopses offer a brief overview of each film's plot, characters and relevant information. While this dataset is comprised of several different files, this research will focus only on two of these components:

- movie.metadata.tsv: This file contains the metadata for 81,741 films extracted from the Freebase leak on November 4, 2012. Notably, the file contains important information such as movie genre identifiers, allowing for genre-based analyses and categorization.
- plot_summaries.txt: This file contains narrative summaries for 42,306 movies. It was extracted from the English-language Wikipedia dump on November 2, 2012. Each line in the file contains a unique identifier corresponding to the film in question, with a corresponding entry in movie.metadata.tsv.

V. Results

A. Baseline Performance

To evaluate the impact of different loss functions on fine-tuning BERT for text classification, we conducted three experiments: fine-tuning BERT with standard cross-entropy loss, fine-tuning BERT with Focal Loss with different gamma values adjusted manually, and fine-tuning BERT with Adaptive Focal Loss. Firstly, we performed a preliminary experiment using the CMU Movie LT Summary Corpus to establish baseline performance. The dataset used in the main experiments consisted of 42,303 samples, and training was performed using a pre-trained BERT model with AdamW optimizer, a batch size of 16, a learning rate of 2e-5 and incorporated a cosine learning rate scheduler with 500 warmup steps. Training spanned up to 100 epochs with early stopping applied, and evaluation metrics included accuracy, F1-score, ROC AUC, macro precision, micro precision, macro recall, and micro recall. The baseline model achieved an F1-score of 0.4707 and a ROC AUC of 0.7118.

B. Effect of Static Gamma Tuning: Focal Loss Experiments

To mitigate class imbalance, we replaced cross-entropy loss with Focal Loss and experimented with different static gamma values (γ = 1.5, 2.0, and 2.5). Among these, γ = 2.0 provided the best performance, achieving an F1-score of 0.4923 and a ROC AUC of 0.7649, while γ = 1.5 and γ = 2.5 yielded lower results. Notably, the γ = 2.5 setting resulted in marginally lower performance compared to the baseline, with an F1-score of 0.4692 and a ROC AUC of 0.7088. These findings emphasize that although tuning the focusing parameter γ improves robustness against class imbalance, static gamma values alone may not be sufficient for optimal multi-label classification. Fig. 3 illustrates the comparative effect of different gamma values across key evaluation metrics.

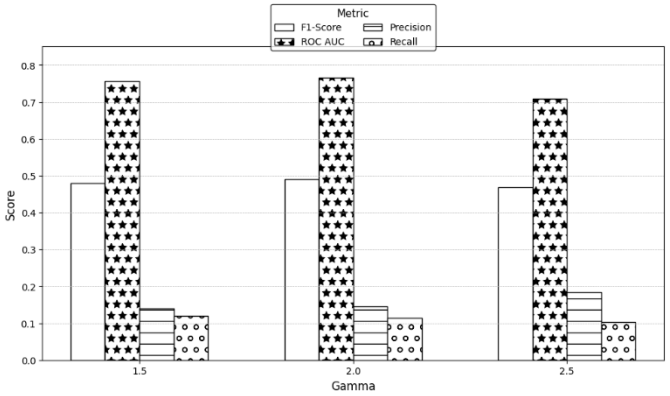

Fig. 3. The comparative impact of varying gamma values on key evaluation metrics.

C. Performance of Adaptive Focal Loss

Building on these findings, we evaluated Adaptive Focal Loss, where the focusing parameter γ is dynamically adjusted based on class frequencies. Adaptive Focal Loss outperformed all static gamma configurations, achieving the highest F1-score (0.5) and ROC AUC (0.79), while substantially improving micro recall (0.5282) without requiring manual tuning of γ. These results highlight the robustness and practical advantages of the adaptive approach, particularly in handling class imbalance in multi-label classification. Fig. 4 presents a comparative overview of the final model performances across all configurations.

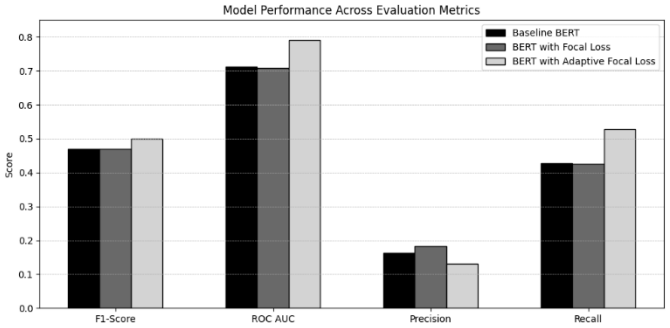

Fig. 4. Model performances across all metrics.

VI. Discussion

This study demonstrates the effectiveness of Adaptive Focal Loss in improving multi-label text classification on imbalanced data. There are three main findings from our experiments:

First, static Focal Loss can beat the conventional cross-entropy baseline on some metrics but only when it is highly tuned. Among the fixed values of γ experimented, γ = 2.0 worked best with an F1-score of 0.49 and a ROC AUC of 0.7649. However, both γ = 1.5 and γ = 2.5 were compromised in either recall or precision, demonstrating the sensitivity of Focal Loss to the γ value. This suggests the need for hyperparameter tuning when using static focusing parameters on highly imbalanced settings.

Second, Adaptive Focal Loss, where γ is adapted based on class frequencies, outperformed all static versions. It achieved the highest F1-score (0.5) and ROC AUC (0.79), and highest micro recall (0.5282), indicating it was more capable of identifying minority class instances. This was achieved without having to manually adjust γ, which is a testament to the model's robustness and its less dependence on heuristic optimization. This also makes the adaptive method more scalable and deployable in real-world settings, where optimal hyperparameters are unknown a priori. This outcome marks a progression from our earlier work, where we had studied the integration of GloVe embeddings with classical machine learning models [27] and then with deep learning models [28]. While those studies concentrated on the role of word embeddings in representation learning, the present study reverses the focus to the flexibility of the loss function and its adaptability to transformer-based models.

Third, the precision-recall trade-offs for both static and adaptive settings illustrate the merits of dynamic class weighting strategies. While the static γ = 2.5 model attained the highest precision, it did so at the cost of decreased recall and a lower F1-score compared to the adaptive approach. By contrast, Adaptive Focal Loss obtained more balanced performance in general, with significant gains in recall, an important factor in multi-label where missing relevant labels is costly.

VII. Limitations and Future Work

While Adaptive Focal Loss yielded considerable improvements, the overall F1-score (0.50) indicates that there is still room for performance gains. Potential future work may include combining adaptive loss functions with more recent backbone models like RoBERTa or DeBERTa, introducing data-level rebalancing techniques, or experimenting with more γ adaptation techniques. Also, extending this approach to other multi-label domains, such as biomedical text or legal documents, may also serve to further confirm its generalizability.

VIII. Conclusion

This work proposes a novel approach of handling class imbalance in multi-label text classification by combining Adaptive Focal Loss (AFL) and the BERT model. Through the adjustment of the focusing factor γ based on class frequencies, AFL provides a self-adjusting mechanism for assigning greater weights to minority classes and reducing the need for hyperparameter tuning. Our experimental results on the CMU Movie Summary Corpus demonstrate that this adaptive strategy consistently performs better than baseline cross-entropy and statically tuned Focal Loss model on key metrics like F1-score, ROC AUC, and micro recall.

Particularly, the adaptive mechanism achieves improved precision-recall trade-off, significantly improving minority class detection, which is a common problem with real-world multi-label problems. The BERT alignment also allows for more accurate contextual understanding and representation learning, contributing to enhanced performance overall.

These findings demonstrate the potential of adaptive loss functions to improve deep learning models for imbalanced classification tasks. Potential future directions can involve combining AFL with other transformer-based architectures like

RoBERTa or DeBERTa, applying the method to other application areas such as biomedical or legal documents, or coupling it with more advanced data augmentation strategies for further performance and generalization gains.

REFERENCES

[1] U. Jain, P. Mishra, A. Dash, and A. Pandey, "Multi-label multi-class text classification-enhanced attention in transformers with knowledge distillation," *Rev. Digit. Univ.*, vol. 23, no. 1, 2025. [Online]. Available: https://doi.org/10.22201/icat.24486736e.2025.23.1.2484.

[2] Guangxiang Zhao, Wenkai Yang, Xuancheng Ren, Lei Li, Yunfang Wu, and Xu Sun, "Well-classified examples are underestimated in classification with deep neural networks," in AAAI. 2022, pp. 9180–9189, AAAI Press.

[3] Yang Zhou and Wee Sun Lee, "None class ranking loss for document-level relation extraction," in IJCAI. 2022, pp. 4538–4544, ijcai.org.

[4] Y. Yasuda, T. Miyazaki, and J. Goto, "Weighted Asymmetric Loss for Multi-Label Text Classification on Imbalanced Data," *J. Nat. Lang. Process.*, vol. 31, no. 3, pp. 1166–1192, 2024.

[5] Park, W., Park, I., Kim, S., & Ryu, J. (2023). Robust Asymmetric Loss for Multi-Label Long-Tailed Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2711–2720.

[6] Huang, Y., Qi, J., Wang, X., & Lin, Z. (2023). Asymmetric Polynomial Loss for Multi-Label Classification. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5.

[7] J. M. Reps, P. Rijnbeek, and A. G. Sena, "The impact of resampling methods on the performance of prediction models in observational health data," *J. Big Data*, vol. 10, no. 1, pp. 1–21, 2023. [Online]. Available: https://doi.org/10.1186/s40537-023-00857-7.

[8] A. D. Putra and M. T. Ibrohim, "Class weighting technique to deal with imbalanced class problem in machine learning: methodological research," unpublished, 2023.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988. https://arxiv.org/abs/1708.02002.

[10] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems, 33, 15744-15754.

[11] W. Zhang, Y. Jiang, Y. Fang, and S. Pan, "Hierarchical Contrastive Learning for Multi-label Text Classification (HCL-MTC)," Sci. Rep., 2025. Available: https://www.nature.com/articles/s41598-025-97597-w.

[12] C. Zhang, L. Dai, C. Liu, et al., "HGBL: A Fine Granular Hierarchical Multi-Label Text Classification Model," *Neural Process. Lett.*, vol. 57, no. 1, 2025. [Online]. Available: https://doi.org/10.1007/s11063-024-11713-x.

[13] Y. Yan, J. Liu, and B.-W. Zhang, "LabelCoRank: Revolutionizing Long Tail Multi-Label Classification with Co-Occurrence Reranking," arXiv preprint, 2025. Available: https://arxiv.org/abs/2503.07968.

[14] S. Lin, F. Frasincar, and J. Klinkhamer, "Hierarchical deep learning for multi-label imbalanced text classification of economic literature," *Appl. Soft Comput.*, vol. 176, p. 113189, 2025. [Online]. Available: https://doi.org/10.1016/j.asoc.2025.113189.

[15] M. A. Islam, M. M. Rahman, and A. Basak, "Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification," Expert Systems with Applications, vol. 168, p. 114146, 2021.

[16] J. Lee, J. Kim, and S. Lee, "An Empirical Study for Class Imbalance in Extreme Multi-label Text Classification," in Proceedings of the 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), Feb. 2021, pp. 1-4, doi: 10.1109/BigComp50530.2021.00009.

[17] Jung, J., & Lee, J. (2021). EnvBERT: Multi-Label Text Classification for Imbalanced, Noisy Environmental News Data. IEEE Access, 9, 107785-107794.

[18] M. A. Alomari, M. A. Al-Taee, and A. Al-Taee, "Multilabel Sentiment Prediction by Addressing Imbalanced Class Problem Using Oversampling," in 2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2021, pp. 1-6, doi: 10.1109/JEEIT51787.2021.9429366.

[19] K. Duarte and M. Shah, "PLM: Partial Label Masking for Imbalanced Multi-label Classification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3057587.

[20] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models," Applied Sciences, vol. 11, no. 2, p. 869, Jan. 2021, doi: 10.3390/app11020869.

[21] M. Park, H. J. Song, and D.-O. Kang, "Imbalanced Classification via Feature Dictionary-Based Minority Oversampling," IEEE Access, vol. 10, pp. 34236–34245, 2022, doi: 10.1109/access.2022.3161510.

[22] X. Cai, S. Huang, Q. Yang, and H. Chen, "UJNLP at SemEval-2025 Task 9: Easy Data Augmentation and Oversampling with Focal Loss for Multi-label Text Classification," arXiv preprint arXiv:2505.00021, 2025. Available: https://arxiv.org/abs/2505.00021.

[23] S. Sen, "Focal Loss Effect on Multi-Label Text Classification Using Convolutional Neural Networks," unpublished manuscript, ResearchGate, 2021. [Online]. Available: https://www.researchgate.net/publication/372109806.

[24] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in Advances in Neural Information Processing Systems, vol. 32, pp. 1565-1576, 2019.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018. Available: https://arxiv.org/abs/1810.04805.

[26] Nam, J., Kim, J., Gurevych, I., & Mencía, E. L. (2014). Large-scale Multi-label Text Classification — Revisiting Neural Networks. In ECML PKDD. https://link.springer.com/chapter/10.1007/978-3-662-44848-9_5.

[27] Labd, Z., Bahassine, S., Housni, K., Ait Hamou Aadi, F., & Benabbes, K. (2024). Text classification supervised algorithms with term frequency-inverse document frequency and global vectors for word representation: A comparative study. IJECE, 14(1), 589–599.

[28] Labd, Z., Bahassine, S., Housni, K., & Ait Hamou Aadi, F. (2024). Reassessing GloVe Embeddings in Deep Learning: A Comparative Study with Classical ML Approaches. Procedia Computer Science, 251, 740–745.