*Artificial Intelligence and Human-Computer Interaction*
*Y. Ye and H. Zhou (Eds.)*

*doi:10.3233/FAIA250105*

# Exploring Text Classification Methods for Bulletin Board System Posts: A Comparative Analysis of BERT, BiLSTM, and Different Loss Functions

Wenhan HU[a,b,c] and Mini Han WANG[a,b,c,d,1]
[a] *Faculty of Data Science, City University of Macao, China 999078*
[b] *Department of AI and Big Data, Zhuhai Institute of Advanced Technology Chinese Academy of Sciences, Zhuhai, China 519000*
[c] *Department of AI and Big data, Perspective Group, Zhuhai, China 519000*
[d] *Frontier Scientific Research and Innovation Development (Shenzhen) Limited Company, Shenzhen, China 518000*
ORCiD ID: Mini Han WANG https://orcid.org/0000-0002-5002-3708

**Abstract.** Multi-Label Text Classification (MLTC) is a crucial task in natural language processing (NLP), enabling the assignment of multiple labels to a single text sample, which aligns with the diverse and multifaceted nature of discussions typically found in Bulletin Board System. This study presents an investigation into text classification methodologies, leveraging a dataset comprising 388,693 entries, with 234,237 entries manually annotated for model training. The dataset encompasses diverse text data from prominent social platforms, including GitHub, H5-based forums, WeChat, QQ group chats, and more. Four distinct methods for text classification are compared: BERT and BiLSTM models with Binary Cross-Entropy (BCE) loss, BERT for feature extraction followed by BiLSTM and BCE, BERT and BiLSTM models with Focal Loss (FL), and BERT for feature extraction followed by BiLSTM and FL. The experimentation reveals insights into their performance, indicating that models utilizing pre-trained BERT for feature extraction outperform those without pre-training. Focal Loss emerges as a superior alternative to Binary Cross-Entropy, demonstrating efficacy in handling class imbalance and noisy data, thereby improving overall model accuracy and robustness. These findings underscore the importance of thoughtful model architecture and loss function selection. Future research directions include exploring ensemble methodologies, alternative pre-training techniques for BERT, and enhancing model interpretability. Keeping pace with NLP advancements and integrating cutting-edge techniques into future investigations holds promise for further advancements in model efficacy and practical utility.

**Keywords.** Multi-label text classification, natural language processing, BERT, BiLSTM, binary cross-entropy loss, focal loss, bulletin board system

[1] Corresponding Author: Mini Han WANG, 1155187855@link.cuhk.edu.hk.

## 1. Introduction

As an open-source forum for free speech, Bulletin Board System (BBS) has become vital platform for students to share information and exchange opinions. Evolving from the earliest BBS [1] to contemporary multimedia social platforms, BBS reflects the progression of internet culture. However, they face challenges such as distinguishing between true and false information and mitigating anonymous attacks, necessitating enhanced supervision and guidance from educational institutions. Moreover, the operational quality of BBS varies significantly, making standardization and monetization difficult. Despite these challenges, their substantial market potential offers considerable development opportunities. Addressing these issues effectively will ensure that BBS remains an essential communication platform for students and continue to play a crucial role in campus culture.

Incorporating Multi-Label Text Classification (MLTC) [2] into an open-source forum can significantly enhance the platform's functionality and user experience. By automatically tagging posts with multiple relevant categories, this technology improves content organization, making it easier for users to find and navigate discussions. It also aids in managing spam and offensive content, thereby supporting effective moderation and maintaining a healthy community environment. Furthermore, it enables personalized content recommendations based on users' interests, fostering greater engagement. The ability to handle diverse and complex topics through accurate categorization also ensures scalability as the forum grows. Automated insights and summaries derived from classified posts can provide valuable information on prevalent trends. Thus, Multi-Label Text Classification addresses the challenges of manual categorization and enhances the forum's efficiency, accuracy, and user satisfaction.

MLTC represents a significant task in the domain of Natural Language Processing (NLP) [3], involving the simultaneous assignment of multiple labels to a single text sample. This technique finds extensive application in various fields, including information retrieval, sentiment analysis [4], and topic detection [5]. In the context of the exponential increase in textual data on the internet, MLTC technology is crucial for enabling users to efficiently filter and identify relevant content from vast information pools. Initial research on multi-label text classification predominantly utilized traditional machine learning approaches such as Naive Bayes and Support Vector Machines (SVM). While these methods demonstrated efficacy with small-scale datasets, their limitations in feature engineering and model generalization became apparent as data volumes grew. Recent advancements in text classification tasks have been driven by deep learning techniques. Methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), attention mechanisms [6], and Transformer models [7] have shown the ability to autonomously learn complex feature representations from text data, thereby enhancing classification accuracy and efficiency. Nevertheless, despite the advancements brought by deep learning methods in MLTC, several challenges persist. These include managing label correlation, addressing class imbalance, dealing with label missingness, and implementing label compression. Current research efforts focus on modeling label correlations, analyzing specific label characteristics, handling class imbalance, mitigating label missingness, and developing effective label compression techniques.

This article delves into the application of advanced text classification methods to analyze university campus posts. Specifically, it investigates the efficacy of utilizing BERT (Bidirectional Encoder Representations from Transformers) [8] and BiLSTM

(Bidirectional Long Short-Term Memory) [4] models, alongside two distinct loss functions: Binary Cross-Entropy (BCE) [9] and Focal Loss (FL) [10]. The study leverages a sizable dataset comprising 388,693 entries extracted from general university campus posts.

This study investigates Multi-Label Text Classification (MLTC) methodologies using a dataset of 234,237 entries from various social platforms. It compares four classification methods, revealing that models with pre-trained BERT for feature extraction and Focal Loss outperform others, especially in handling class imbalance and noisy data. The results highlight the significance of model architecture and loss function selection, suggesting future research into ensemble methods, BERT pre-training techniques, and model interpretability to further enhance model efficacy and practical application in NLP.

The structure of this paper is as follows:

Introduction: It introduces the importance of BBS as an open-source forum for free speech, and the potential of integrating MLTC technology into BBS to enhance platform functionality and user experience.

Materials and Methods: It provides a detailed description of data collection, data annotation, and classification methods. The dataset includes text data from various social platforms, as well as a comparison of four different text classification methods.

Results and Discussion: It presents the performance changes of different model configurations and discusses the experimental results. It was found that models using pre-trained BERT for feature extraction performed better than those without pre-training, and Focal Loss was superior to Binary Cross-Entropy in handling class imbalance and noisy data.

Conclusion: It summarizes the experimental results and emphasizes the importance of BERT pre-training and Focal Loss in improving model performance. It also proposes future research directions, including exploring ensemble methods, alternative pre-training techniques for BERT, and enhancing model interpretability.

## 2. Materials and methods

### *2.1. Materials*

Given the specific application context of this study and the absence of publicly available datasets on open platforms, a dataset comprising approximately 388,693 entries of text data from general university campus posts was constructed. Through random sampling, 234,237 text data entries were manually identified and labeled for use in model training. The dataset utilized for model training contained a total of 9,862,950 Chinese characters. The dataset encompasses data from various social platforms, including GitHub public repository crawlers, H5-based forum websites, WeChat mini-programs, QQ group chats, QQ zones, WeChat group chats, and WeChat moments [11]. The selection of these platforms was based on their popularity and relevance to the research topic. It is worth mentioning that we have used data from all the aforementioned sources.

As of the statistical endpoint in January 2024, the data sources for this study encompass various platforms (Table 1): 11 H5-based campus wall forums and WeChat mini-program online platforms, constituting 68.09% of the dataset without temporal restrictions; 108 WeChat Moments walls and QQ Zone walls operated in matrix format, accounting for 22.19% of the dataset, collected from May 2023 to January 2024; 56

WeChat and QQ campus group chats focusing on themes like "second-hand" and "mutual aid", representing 7.27% of the dataset, gathered from May to September 2023 with reduced data during summer; and the GitHub confession wall-crawler public dataset, offering data from May 1, 2016, to May 10, 2018, comprising 2.46% of the total dataset.

**Table 1.** Detailed list of sources, quantity, and proportion of posts from regular university campus posts.

| **Data Source** | **Platform Count (units)** | **Some cases of Platform Names** | **Number of Posts (entries)** | **Percentage of Posts** |
|---|---|---|---|---|
| University H5 Forum | 1 | Zanoo Campus Market | 264668 | 68.09% |
| WeChat Mini Program | 10 | Huyou Lite, Aite Campus Circle, Duoduo Campus Circle, Sanmiao Alumni Circle, Campus Microwall, Pipi Campus, Sudden Dot Dot, Know One Campus Market, Meow Jun Alumni Circle, Wall Wall Alumni Circle, | | |
| QQ Zone Wall | 45 | Liaoning University Campus Wall (Activity Edition), Urban Life Guide, College Universal Wall, Guangke Small Assistant No.2, Below Campus Wall·Guatian Jun (Reading and Exam Edition), Shang'an Universal Wall, Jiada Universal Wall, Avionics Little Sun | 86234 | 22.19% |
| WeChat Moments Wall | 63 | City Teddy, Zhengke Institute Universal Wall (Driving School Enrollment Edition), Guangke Million (Original Science Dry Ice Cream), Hua Shang Lan Buff (Order Chat, Guangke Little Helper No.1 (Free Campus Card Edition), A Crayon Pig (Beijing Institute of Technology Pearl Campus Wall Big Account), Hua Shang Yao Yao (Order Chat, Hua Shang Bang | | |
| QQ Group Chat | 34 | Hezhou Campus Market, Zhengke Campus Circle Information Group, Zhuhai University Student Tutoring Part-time Group, Substitution Class, Restriction of Twenty Items Book Trading Group, Guangda Second-hand Book Group, Xiamen University Student | 28239 | 7.27% |
| WeChat Group Chat | 22 | University of Macau Dai Course, GGS Campus Community, Macao 2024 Idle Big Sale, Campus Wall Order, Various Substitutes, Work Part-time Group, Beijing Institute of Technology Pearl Often Wine Big Team, Urban University Flea Market, Carpooling Group, Lingang Work-study Group, Teacher's College Eating Melon Group, Teacher's College BPS Express Group | | |
| GitHub Confession Wall | 1 | Shanghai Finance Campus Confession Wall Crawler Public Datase | 9552 | 2.46% |

**Table 2.** Details about data labelling.

| Labels | Label rules | Cases |
|---|---|---|
| Drainage contact | Such text contains real contact information, such as mobile phone number, email address, social media account (such as WeChat, QQ, Xiaohongshu, Tiktok, etc.) and email format (excluding web page links), which are used to attract potential customers or users and provide the possibility of further communication. | Zhou Chuanxiong's daily settlement security contacted me 1 * * (tel. number). |
| Marketing advertising | Marketing advertising texts aim to promote the sales of products or services. They may typically include attractive titles, vivid product descriptions, prominent slogans, and calls to encourage consumers to take action. These texts may also involve user interaction, such as requesting likes, sharing, or following. | In the beginning of the school season, driving schools on campus are very popular for registration consultation. |
| Transaction interoperability | This category includes copywriting for sale, purchase, second-hand, idle, rental, and other types. These texts are used to publish buying and selling information of goods or services for communication and transactions between buyers and sellers. | 14 RMB per headset (brand new Bluetooth compatible). |
| Partner setup | This type of text involves various types of copywriting, such as matchmaking, making friends, gathering people, organizing games, shaking people, and assembling groups. These texts are used to find friends who share common interests or activities, in order to organize social activities or entertainment projects. | On Thursday, Group outings with classmates from the same school. |
| Lost items | This type of text involves lost and found items, lost and found notices, as well as requests or voluntary return to the original location (return to the owner) due to someone else or oneself taking the wrong item. These texts contain detailed descriptions of lost items for others to identify and return. | Who took my mathematic book. |
| Recruitment information | This category includes copywriting related to corporate recruitment, internships, part-time jobs, internal referrals, and part-time work. These texts are used to publish recruitment information so that job seekers can understand the job requirements and application process. | If you want to take a part-time job, take a look. Need a Chinese tutorial tutoring this afternoon. |
| Daily sharing | This category includes copywriting for sharing personal life, gossip, roast, entertainment information, exposure, chat, eating melon and other contents. These texts are used to express personal opinions, share life experiences, entertainment, and social interactions. | I went to Hainan, China to have fun. |
| Interactive Q&A | This category includes types of copywriting such as Q&A, inquiry, seeking help, consultation, and mutual assistance. These texts are used for communication and interaction between users on a certain issue or topic, to obtain information, solve problems, or share experiences. | Has the teacher established a course group yet? |
| Agency business | This type of text involves types of copywriting such as substitute classes, writing, doing, brushing, drawing, running, retrieving, and handling. These texts are used to provide agency services, such as doing homework, purchasing goods, running errands, etc., to meet the needs of others. | Agency service of student card collection: 5 dollars. |
| Love topics | This type of text involves themes such as confession, confession, separation, and love, used to express feelings for others, emotional sharing, seeking love advice, or sharing love experiences and stories. | Is there any way to stop my heart from hurting so much and break up with my boyfriend. |

As Table 2 shows, the text in the dataset encompasses various topic discussions among university students on campus forums. In this study, posts are categorized into ten major themes: traffic connection, marketing advertisements, trade exchange, partner setups, lost items, recruitment information, daily sharing, interactive Q&A, proxy services, and love topics. Given the large-scale nature of the text classification task, the dataset comprises over 388,693 text records, demanding meticulous and systematic data labeling approaches. Four distinct labeling methods were sequentially employed: regular expression matching, compound keyword filtering, ChatGLM3-6B + Prompt Engineering (open-source large model-assisted labeling), and semi-automatic labeling (semi-supervised learning). Despite initial attempts with ChatGLM3-6B, the generalized large model's performance in multi-label multi-classification of campus forum texts fell short of expectations. Thus, a semi-automatic labeling approach was adopted, leveraging semi-supervised learning techniques. This involved training a preliminary research model with labeled data, pre-labeling unlabeled data using this model, and then manually verifying and finalizing dataset labeling. Quality control measures, including labeler training, random sampling for quality assurance, and feedback loops, were implemented throughout the labeling process to ensure consistency and accuracy. By employing these labeling methods and quality control measures, we successfully constructed a dataset comprising 234,237 accurately categorized text posts, ensuring efficient model annotation, reasonable classification, and significantly enhancing labeling accuracy and consistency. This high-quality labeled dataset serves as a solid foundation for subsequent model training and analysis. In future work, we plan to further refine our labeling processes and explore automated labeling techniques to enhance efficiency and reduce costs.

### *2.2. Methods*

The study compares four different methods for text classification on the dataset. Firstly, this study utilizes BERT and Bidirectional Long Short-Term Memory (BiLSTM) models with BCE loss (Bert+BiLSTM+BCE). Secondly, this study utilizes BERT for feature extraction followed by BiLSTM and BCE(Bert-Pretraining+BiLSTM+BCE). Thirdly, this study employs BERT and BiLSTM models with FL (Bert+BiLSTM+FL). Lastly, this study employs BERT for feature extraction followed by BiLSTM and FL(Bert-Pretraining+BiLSTM+FL).

BiLSTM represents an advancement in Recurrent Neural Network (RNN) architecture, first introduced by Hochreiter and Schmidhuber in 1997. BiLSTM has garnered notable success across domains such as NLP and speech recognition due to its unique capability to capture long-term dependencies, a challenge often faced by conventional RNNs and unidirectional LSTM networks. Over time, BiLSTM has emerged as a pivotal branch within the realm of deep learning, extensively explored and applied in various related disciplines. Comprising both forward and backward LSTM layers, BiLSTM architecture efficiently captures forward and backward dependencies within sequential data, leveraging a shared output layer to integrate contextual information from both directions concurrently. This architectural design endows BiLSTM with a distinct advantage in processing sequential data. By employing LSTM units equipped with input, forget, and output gates, the model effectively learns intricate long-term dependencies, mitigating issues like gradient vanishing and explosion prevalent in traditional RNNs during prolonged sequence learning. Through the fusion of forward and backward LSTM layers, BiLSTM adeptly assimilates comprehensive sequence information. Typically, the output layer of a BiLSTM network consists of a fully connected layer, consolidating bidirectional context to yield precise predictions, often facilitated by the softmax function for classification tasks. In the domains of NLP and speech recognition, BiLSTM has demonstrated prowess in tasks such as text classification, named entity recognition, and sentiment analysis. In speech recognition, its ability to harness both forward and backward speech signal information enhances recognition accuracy. Moreover, BiLSTM's utility extends to tasks like image description generation and machine translation, where it has yielded notable achievements. As deep learning technology continues to evolve, BiLSTM's versatile applications are poised to expand, driving advancements across diverse domains. Often integrated with complementary model architectures like CNNs, BiLSTM forms hybrid CNN-BiLSTM models, capitalizing on CNNs' strengths in local feature extraction and BiLSTM's proficiency in sequential data processing.

The inception of the BERT (Bidirectional Encoder Representations from Transformers) model in 2018 by the Google AI team marks a notable advancement in pre-training language representation models rooted in the Transformer architecture. Its introduction is recognized as a significant breakthrough within the NLP domain, offering a novel approach to tackling NLP challenges. Prior to BERT, NLP predominantly relied on conventional machine learning techniques or unidirectional deep learning models, facing limitations in handling intricate linguistic phenomena. The adoption of the BERT model addressed this gap by harnessing the Transformer structure and integrating attention mechanisms, enabling the model to adeptly capture contextual nuances and substantially augment its semantic comprehension. The BERT model has revolutionized various NLP tasks, encompassing text classification, named entity recognition, and sentiment analysis. Its fundamental framework entails pre-training the model to assimilate comprehensive language knowledge, followed by task-specific fine-tuning. The model architecture primarily comprises an encoder and pre-training tasks. The encoder consists of numerous identical layers, each comprising two sub-layers: a multi-head self-attention mechanism and a feed-forward neural network. The self-attention mechanism facilitates contextual information capture within input sequences, while the feed-forward neural network further refines the extracted information. Pre-training tasks entail the Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task involves predicting masked words to grasp contextual dependencies, whereas the NSP task predicts the continuity of two sentences in a text, aiding in understanding sentence correlations. The BERT model is widely applied across NLP tasks, demonstrating prowess in text classification, named entity recognition, and sentiment analysis. In text classification, it proficiently captures intricate semantic details, achieving state-of-the-art performance across diverse datasets. In named entity recognition, BERT effectively identifies entities such as personal names, geographical locations, and organizational entities within text. In sentiment analysis, it accurately discerns sentiment expressions within text. As a pre-training language representation model grounded in Transformers, BERT holds significant academic and practical value in the realm of natural language processing. Leveraging rich language knowledge through pre-training and task-specific fine-tuning, the BERT model has attained remarkable accomplishments across diverse NLP tasks.

Binary Cross-Entropy, often abbreviated as BCE, is a widely used loss function in binary classification tasks. It measures the discrepancy between the predicted probabilities and the actual binary labels. The BCE loss is calculated as the negative log likelihood of the true labels given the predicted probabilities, which can be mathematically expressed as:

$$BCE = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})) \quad (1)$$

In binary classification, where each sample belongs to one of two classes (e.g., positive or negative), BCE loss effectively penalizes the model for predicting probabilities that diverge from the ground truth labels. It is particularly suitable when dealing with imbalanced datasets, where one class may significantly outnumber the other.

Focal Loss, denoted as FL, is a specialized loss function designed to address the issue of class imbalance and noisy data in object detection tasks, particularly in the presence of a large number of easy-to-classify examples. Proposed by Lin et al. in 2017, FL introduces a modulating factor $(1 - \hat{p}_t)^\gamma$ to down-weight the loss contribution from well-classified examples, focusing more on hard-to-classify examples. This modulating factor reduces the loss contribution from easy examples and amplifies the contribution from misclassified examples, thereby improving the model's ability to learn from challenging samples. The Focal Loss can be mathematically expressed as:

$$FL = -\sum_{t-1}^{N} ((1 - \hat{p}_t)^\gamma \cdot log(\hat{p}_t)) \quad (2)$$

where $\hat{p}_t$ is the model's estimated probability for the class with label $t, N$ is the number of classes, and $\gamma$ is a balancing factor to focus on hard-to-classify examples. FL has been found to be effective in improving the performance of models in scenarios where the class distribution is highly imbalanced or when dealing with noisy data.

Thus, during the experimental phase, a total of 388,693 data entries were collected, from which 234,237 were manually annotated for model training. In this process, the dataset is split as follows: 80% of the total number, which amounts to 187,379 data entries, is allocated to the training set; while 20% of the total number, amounting to 46,838 data entries, is designated for both validation and testing sets. Four different methods are applied to text classification on this dataset. Firstly, this study utilizes BERT and BiLSTM models with BCE loss (Bert+BiLSTM+BCE). Secondly, this study utilizes BERT for feature extraction followed by BiLSTM and BCE(Bert-Pretraining+BiLSTM+BCE). Thirdly, this study employs BERT and BiLSTM models with FL (Bert+BiLSTM+FL). Lastly, this study employs BERT for feature extraction followed by BiLSTM and FL (Bert-Pretraining+BiLSTM +FL).

## 3. Results and discussion

In our study, we focused on several key metrics to evaluate the performance of different text classification models: **Development Accuracy (dev_acc)** is crucial as it indicates how well the model performs on unseen data during the development phase. The **Macro F1 Score (macro_f1)** provides a balanced assessment of the model's precision and recall across all classes, accounting for class imbalance. The **Micro F1 Score (micro_f1)** complements the macro F1 by considering the total counts of true positives, false positives, and false negatives, offering an overall performance measure. **Recall (recall)**, or sensitivity, is important for understanding the model's ability to identify all relevant instances. **Precision (precision)** is key for evaluating the accuracy of the model's positive predictions. These metrics collectively offer a comprehensive view of model performance, enabling us to make informed comparisons between the different configurations.

The experimental results demonstrate notable variations in performance across different model configurations (Table 3 and Figure 1). Bert+BiLSTM+BCE achieves a decent dev_accuracy of 0.73 but shows relatively lower macro_f1 and micro_f1 scores compared to other models, indicating some challenges in achieving a balanced performance across all classes. Incorporating Bert pretraining significantly enhances model performance, as seen in Bert-Pretraining+BiLSTM+BCE, where both macro_f1 and micro_f1 scores notably improve to 0.59 and 0.84, respectively, along with a higher dev_accuracy of 0.77. However, the introduction of FL in both Bert+BiLSTM+FL and Bert-Pretraining+BiLSTM+FL models yields remarkable improvements across all evaluation metrics, with dev_accuracy reaching 0.83 and 0.84, and macro_f1 and micro_f1 scores substantially increasing to 0.83/0.89 and 0.84/0.89, respectively. These results underscore the effectiveness of Focal Loss in mitigating class imbalance issues and enhancing the overall performance of the models, further supported by robust recall, precision, and AUC scores.

**Table 3.** Indicator comparison of models of Bert+BiLSTM+BCE, Bert-Pretraining+BiLSTM+BCE, Bert+BiLSTM+FL, and Bert-Pretraining+FL.

| Model | dev_loss | dev_acc | macro_f1 | micro_f1 | recall | precision | auc | Sensitivity _score | specificity_ score |
|---|---|---|---|---|---|---|---|---|---|
| Bert +BiLSTM+BCE | 0.12 | 0.73 | 0.54 | 0.81 | 0.49 | 0.84 | 0.74 | 0.49 | 0.99 |
| Bert-Pretraining +BiLSTM+BCE | 0.10 | 0.77 | 0.59 | 0.84 | 0.55 | 0.85 | 0.77 | 0.55 | 0.99 |
| Bert +BiLSTM+FL | 0.01 | 0.83 | 0.83 | 0.89 | 0.77 | 0.91 | 0.88 | 0.77 | 0.99 |
| Bert-Pretraining +BiLSTM+FL | 0.01 | 0.84 | 0.84 | 0.89 | 0.78 | 0.91 | 0.89 | 0.78 | 0.99 |

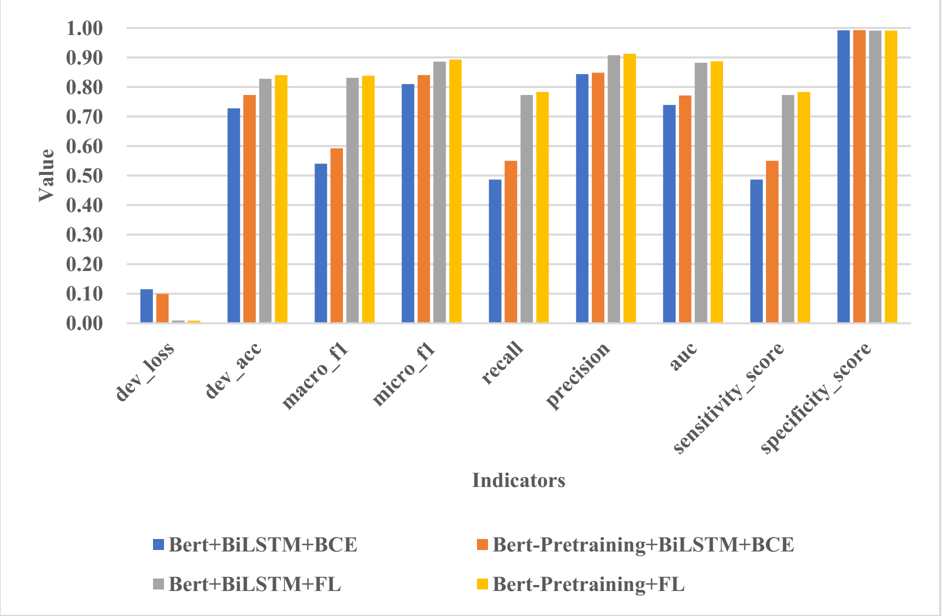


**Figure 1.** Indicator comparison of models.

In this study, BERT for the feature extraction pretraining process is advantageous compared to not pretraining for several reasons. Firstly, pretraining BERT allows the model to learn rich language representations from vast amounts of unlabeled text data, capturing complex linguistic patterns and structures. This pretraining phase enables BERT to develop a deeper understanding of language semantics, syntax, and context, which can significantly enhance its performance on downstream tasks. Additionally, pretraining helps BERT to acquire generalizable knowledge about language, making it more adaptable to diverse NLP tasks and domains. Moreover, by fine-tuning the pre-trained BERT model on specific tasks, it can leverage the learned features to achieve better performance with fewer labeled examples, thereby improving efficiency and reducing the need for extensive task-specific data. Overall, BERT's feature extraction through pretraining offers a robust foundation for subsequent task-specific learning, resulting in improved performance and generalization capabilities compared to models trained without pretraining.

Besides, Focal Loss demonstrates superiority over Binary Cross-Entropy for several reasons. Firstly, FL addresses the issue of class imbalance by down-weighting the easy examples and focusing more on hard examples during training, which is particularly beneficial for tasks with imbalanced class distributions. This characteristic helps FL to handle skewed datasets more effectively, leading to improved model performance, especially for minority classes. Additionally, FL introduces a tunable focusing parameter that allows the model to adjust the rate at which it focuses on hard examples, providing greater flexibility in optimizing the loss function to suit specific task requirements. Furthermore, FL mitigates the impact of noisy or mislabeled data points by assigning lower weights to easy examples, thereby reducing their influence on the training process and enhancing the model's robustness to noisy data. Overall, FL's ability to handle class imbalance, adapt to varying degrees of difficulty in training examples, and reduce the impact of noisy data makes it a preferable choice over BCE in scenarios where these factors are prominent.

Thus, this study offers a significant contribution by conducting an extensive examination of methodologies for text classification, utilizing a diverse dataset sourced from prominent social platforms. Through a meticulous comparison of multiple models, loss functions, and pre-training approaches, the research provides valuable insights into their performance and effectiveness within the context of the task at hand. Notably, the study highlights the superior performance of models leveraging pre-trained BERT for feature extraction, thereby emphasizing the pivotal role of pre-trained language representations in enhancing model efficacy for NLP tasks. Moreover, the identification of Focal Loss as a more effective alternative to Binary Cross-Entropy underscores its capability to address challenges such as class imbalance and noisy data, consequently bolstering overall model accuracy and resilience. These findings contribute substantially to the advancement of knowledge in the domain of natural language processing, offering valuable guidance for researchers and practitioners seeking to optimize model architectures and loss functions tailored to specific application domains. Additionally, the exploration of prospective research trajectories underscores the continual evolution of NLP techniques and the potential for further refinement and innovation in augmenting model efficacy and practical utility.

However, while this study provides valuable insights into the performance of different models and loss functions for the given task, several limitations should be acknowledged. Firstly, the evaluation metrics used may not fully capture the nuances of model performance, and additional metrics or experiments could provide a more comprehensive understanding. Secondly, the study focuses on a specific dataset and task, and the findings may not generalize to other domains or datasets. Moreover, the study does not explore the impact of hyperparameter tuning or model architecture variations, which could potentially affect the results. Additionally, the study does not investigate the computational resources required for training and inference, which is crucial for assessing the practical feasibility of deploying these models in real-world applications. Finally, while the selected models and loss functions show promising results, further research is needed to explore alternative approaches and address the limitations identified in this study.

## 4. Conclusions

In conclusion, the experimentation involving different models, loss functions, and pretraining strategies yielded valuable insights into their performance for the task at hand. The results indicate that models utilizing BERT for feature extraction, particularly when pre-trained, outperform those that do not undergo pretraining. This underscores the importance of leveraging pre-trained language representations for enhancing model performance in NLP tasks. Furthermore, Focal Loss emerges as a superior alternative to Binary Cross-Entropy, demonstrating its efficacy in handling class imbalance and noisy data, thereby improving overall model accuracy and robustness. These findings emphasize the significance of thoughtful selection and optimization of model architectures and loss functions to achieve optimal performance in specific application domains. Additionally, the study underscores the ongoing advancements in NLP techniques and the potential for further refinement and innovation in future research endeavors. Last but not least, MLTC is a crucial task in NLP, enabling the assignment of multiple labels to a single text sample, which aligns with the diverse and multifaceted nature of discussions typically found in student forums. By investigating various models, loss functions, and pre-training strategies, the study provides insights into effective techniques for text classification, which can significantly benefit the management and organization of content within open-source student forums. Furthermore, the findings offer valuable guidance for enhancing the accuracy and efficiency of MLTC systems deployed in such forums, thereby facilitating better information retrieval, sentiment analysis, and topic detection. Ultimately, the study's outcomes can contribute to optimizing the user experience, fostering constructive discussions, and maintaining a conducive environment for knowledge sharing within open-source student forums.

### *4.1. Limitation*

Despite the valuable insights offered by this study on advanced text classification methods for BBS posts, it has several limitations. The dataset, predominantly sourced from specific social platforms, may not fully represent the diversity of online communication contexts, thus affecting the generalizability of our findings. The study's scope does not extend to multimodal data, which could provide a more comprehensive understanding of posts that include images, videos, or audio. Moreover, the interpretability of the models is not deeply explored, which is crucial for establishing trust and transparency in practical applications. Additionally, the study may not sufficiently address the cross-domain applicability of the models, and the handling of class imbalance and noisy data is not thoroughly analyzed, potentially impacting the models' robustness in varied real-world scenarios.

*4.2. Potential future work*

In the trajectory of future investigations, there exist several potential research paths founded on the outcomes delineated in this inquiry. Firstly, an exploration of ensemble methodologies, amalgamating diverse models or loss functions, may yield enhancements in performance. Secondly, delving into alternative pre-training methodologies for BERT, such as domain-specific pre-training or fine-tuning, could augment its feature extraction capabilities within the specified task context. Furthermore, conducting experiments on expansive and heterogeneous datasets could furnish deeper insights into the models' generalizability and the efficacy of the loss functions scrutinized herein. Additionally, delving into the interpretability of the models and discerning the determinants steering their predictions could bolster their credibility and applicability in practical scenarios. Additionally, given the rapid strides in natural language processing, staying abreast of contemporary advancements and assimilating cutting-edge techniques into forthcoming investigations could catalyze further enhancements in model efficacy and practical utility. Lastly, integrating multi-source evidence, encompassing text, images, videos, and sounds, offers significant potential for decision-making. AI-based models leveraging multimodal data can substantially enhance human healthcare management [12], business intelligence [13], human behavior analysis [14], psychological construct description [15], and other fields.

## Funding

This research was funded by the National Natural Science Foundation of China Natural, grant number U22A2041, 82071915 and 62372047, the Shenzhen Key Laboratory of Intelligent Bioinformatics, grant number ZDSYS20220422103800001, the Shenzhen Science and Technology Program, grant number KQTD20200820113106007, the Guangdong Basic and Applied Basic Research Foundation, grant number 2022A1515220015,Youth Project of Chongqing Education Science "14th Five Year Plan", grant number K24YY3080045, Characteristic Innovation Project of Ordinary Universities in Guangdong Province, grant number 2024KTSCX226, and the Zhuhai Technology and Research Foundation, grant number ZH22036201210034PWC, 2220004000131, and 2220004002412.

## References

[1] Sun Z, Anbarasan M, Praveen Kumar D. Design of online intelligent English teaching platform based on artificial intelligence techniques. Computational Intelligence. 2021; 37: 1166-80 0824-7935.

[2] Chai Y, Li Z, Liu J, Chen L, Li F, Ji D, et al. Compositional Generalization for Multi-Label Text Classification: A Data-Augmentation Approach. 2024. p. 17727-35 2374-3468.

[3] Gin BC, ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ. 2022; 56: 303-11.

[4] Wang H, Xue L, Du W, Wang F, Li P, Chen L, et al. The effect of online investor sentiment on stock movements: an LSTM approach. Springer; 2021. p. 1-14.

[5] Hajek P, Henriques R. Predicting M&A targets using news sentiment and topic detection. Technological Forecasting and Social Change. 2024; 201: 123270 0040-1625.
[6] Xiaobing Sun and Wei Lu. "Understanding Attention for Text Classification." Annual Meeting of the Association for Computational Linguistics (2020).
[7] Friedman D, Wettig A, Chen D. Learning transformer programs. Advances in Neural Information Processing Systems. 2024;36.
[8] Lin J, Nogueira R, Yates A. Pretrained transformers for text ranking: Bert and beyond: Springer Nature; 2022.
[9] Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. Int J Adv Trends Comput Sci Eng. 2020; 9.
[10] Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr P, Dokania P. Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems. 2020; 33: 15288-99.
[11] Tai Z. Social media and contentious action: The use and users of QQ groups in China. Media and Communication. 2022; 10: 66-76 2183-439.
[12] Wang MH, Xing L, Pan Y, Gu F, Fang J, Yu X, et al. AI-Based Advanced Approaches and Dry Eye Disease Detection Based on Multi-Source Evidence: Cases, Applications, Issues, and Future Directions. Big Data Mining and Analytics. 2024; 7: 445-84.
[13] Chen Y, Li C, Wang H. Big Data and predictive analytics for business intelligence: a bibliographic study (2000–2021). Forecasting. 2022; 4: 767-86 2571-9394.
[14] Wang M, Zhao Y, Wu Q, Chen G. A YOLO-based Method for Improper Behavior Predictions. 2023 IEEE International Conference on Contemporary Computing and Communications (InC4): IEEE; 2023. p. 1-4.
[15] Kjell ONE, Kjell K, Garcia D, Sikström S. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. Psychological Methods. 2019; 24: 92 1939-463.