

RAOCSL: A BERT-Based Strategy for Identifying Learner Confusion under Class Imbalance

Tongyu Zhao^{1,2}, Jiaying Gao³, Yu Feng^{1,2}, Yatong Zu⁴, Adriano Tavares⁵, Sandro Pinto⁵, and Hao Xu^{2,*}

¹School of Engineering, University of Minho, 4804-533 Guimarães, Portugal

²College of Computer Science and Technology, Jilin University, 130012 Changchun, China

³Artificial Intelligence School, Jilin University, 130012 Changchun, China

⁴College of Physical Education, Jilin University, 130012 Changchun, China

⁵ALGORITMI, University of Minho, 4804-533 Guimarães, Portugal

Abstract—Understanding and identifying the nature of learner confusion is important for online learning platforms. In this study, we address this problem by analyzing forum posts from large-scale online courses. However, due to the large volume of comments and frequent interactions, confusion posts are often overlooked. Existing methods and models, while capable of detecting confusion, typically rely on linguistic features of posts and community factors (e.g. votes, views) but ignore personalized contexts, such as the specific causes and types of confusion. To address this problem, we create the first deep learning dataset focused on confusion types and develop a BERT-based network to model personalized features and identify confusion types. Considering the highly imbalanced distribution of different types of confusion, we further design a novel loss function that adaptively optimizes the training weights for each type. Our method’s effectiveness is confirmed through extensive experimentation.

Index Terms—Learning confusion, Discussion forum, Text classification, Confusion characterization

I. INTRODUCTION

Discussion forums on online learning platforms are important for learner interaction, allowing them to ask questions, share opinions, and express concerns that peers or instructors can address [1]. Forum posts are valuable in capturing the emotions and confusion of learners [2]. This study aims to describe the emotions and confusion expressed by learners in forum posts and to develop automated methods for detecting these emotional states. Drawing from [3], [4], we describe “confusion” as a state in which learners encounter obstacles and are uncertain about how to proceed, often due to unclear discussion topics or technical issues within the learning interface. Although the connection between learners’ emotions, engagement, and outcomes is well-recognized, research on its impact in online environments like MOOCs is still emerging [5], [6]. Automated confusion detection serves two purposes: identifying areas in need of improvement [7], [8] and providing insights into learners’ emotional states [9], which allows for timely intervention and increased course success [10]. With the growing prevalence of large-scale learning environments, efficient and automated confusion detection has become more important than ever [11].

Despite the urgent need, there has still been limited research on confusion analysis within the course discussion forums [5], [6] [12]. Most current methods used to detect confusion rely heavily on community interactions (e.g. votes, views) and content analysis tools (e.g., LIWC, Coh-Matrix) [9], [13]–[16] [17]. Given the large volume of posts, these models may misclassify “confusion” posts if they fail to receive adequate forum responses, potentially causing some learners to fall behind despite instructor intervention. To address this problem, researchers have attempted to detect confusion through discourse analysis that is independent of community-related factors [18] [19]. [20], [21] also show that features extracted from language can effectively detect confusion, but relying only on simple lexical features may lead to misclassification. This is because confusion posts often manifest as questions, and since the information-seeking behavior in question posts and confusion posts is so similar, traditional question detection methods may not be directly applicable to confusion detection. Due to the complexity of confusion detection, this problem has become a major challenge in the realm of natural language processing [22].

To resolve this challenge, we integrate AI techniques with educational psychology to automatically identify confusion type. We treat confusion detection as a psychological characterization rather than a simple text classification task. We create the first Chinese learning confusion dataset of 10,590 posts and develop the BERT-BiLSTM-CNN model that incorporates psychological knowledge and linguistic features. Due to the highly imbalanced frequency of confusion in real-world scenarios, which may lead to model bias and poor performance, we design an effective cost-sensitive learning loss function that automatically optimizes the training weights for each type. Experimental results show significant improvements over the baseline, demonstrating the method’s effectiveness in accurately identifying confusion.

II. DATASET CONSTRUCTION

A. Dataset collection

This study aims to identify confusion types during the learning process. To achieve this, we analyze data collected from participants enrolled in the course “Applications of

*Corresponding author: xuhao@jlu.edu.cn

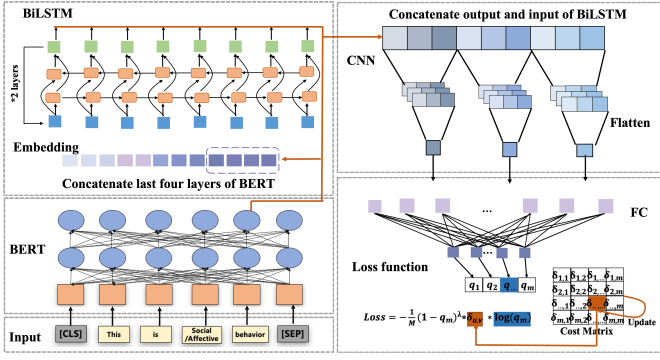


Fig. 1. The architecture of our model, includes BERT embedding, two-layer BiLSTM, three parallel CNNs, and an imbalance-aware loss function.

Mind Maps in Teaching” offered on the Chinese MOOC platform (<https://www.icourse163.org>). The course lasted for 10 weeks, from March 15, 2020, to May 24, 2020. The content, developed by a professor and two assistants based on various teaching models, primarily included video lectures, quizzes, and discussions. Participants were expected to spend 3-6 hours per week on their studies and were encouraged to engage in discussions to deepen their understanding. In total, we totally obtain 10,590 Chinese sentence-level texts about different types of confusion. As this study focuses on the identification model, the details of data preprocessing are not discussed further in this paper.

B. Coding scheme

Understanding the types of confusion in discussion forums helps instructors identify potential challenges in the learning process and promptly adjust to learners’ needs. To achieve this, we develop a coding scheme for identifying confusion type in MOOC discussions, modifying it based on Wang et al. [23] to categorize posts into six sub-types. We code a portion of the posts to create a dataset focused on confusion types. Initially, four coders with expertise in education research and deep learning were equally divided into two groups. Each group independently coded 1,000 randomly selected posts, analyzed inconsistent samples, and refined the coding scheme. Then, the two groups jointly coded an additional 2,000 posts to validate the reliability of the coding scheme. The final annotated results achieved a Cohen’s Kappa coefficient [24] of 0.84, indicating a high level of inter-coder reliability. We establish the first Chinese learning confusion dataset on this website¹. The resulting dataset contains 10,590 texts, of which 292 were labeled ‘Logistical’ type. The dataset’s imbalance factor exceeds 36, indicating a high level of imbalance. The distribution of the dataset is shown in Table I.

III. METHOD

A. Model for Identifying Learning Confusion

To identify confusion types in courses, we use the BERT model [25], pre-trained on Chinese Wikipedia, to generate

rich contextual embeddings. Instead of using the “[CLS]” token from the final layer (which may cause overfitting), we concatenate the last four layers of BERT to create a more comprehensive embedding of size $(n, 3072)$. We then integrate a Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) network for deeper feature extraction. The BERT embeddings are processed through two BiLSTM layers, each with 256 units, to capture bi-directional context, producing a vector w of size $(n, 512)$. This vector is concatenated with the original embeddings to form a vector z of size $(n, 3584)$. Next, z is passed through three parallel CNNs, each with 256 filters and kernel sizes of 2, 3, and 4, to extract fine-grained features, resulting in a final vector of size $(1, 768)$. The output is subsequently fed into a fully connected network for classification. This approach effectively leverages linguistic features, enhancing the identification of confusion types. The architecture of the model is shown in Figure 1.

B. Loss Function Design

As shown in Table I, our dataset exhibits a significant class imbalance, with the ‘Question’ class containing substantially more samples than other classes. This imbalance can lead the model to favor the majority class, thereby compromising its ability to generalize and accurately classify minority classes. To address this problem, we propose a novel cost-sensitive loss function, \mathcal{L}_ϕ , specifically designed to optimize the model parameters ϕ . We adopt a Rapid Auto-Optimization Class-Specific Loss (RAOCSL) approach that dynamically recalibrates the loss weights for various classes to reduce bias towards the majority class.

Additionally, we introduce an auxiliary loss function, \mathcal{L}_ψ , that dynamically adjusts the parameters of a matrix ψ . This matrix ψ optimizes the relative loss weights across classes to enhance the model’s ability to learn from the minority classes. The main loss function, \mathcal{L}_ϕ , is defined as follows:

$$\mathcal{L}_\phi = -\frac{1}{M} \sum_{m=1}^M [(1 - q_m^r)^\lambda \cdot \delta_{(r,s)} \cdot v_m \cdot \log(q_m^r)] \quad (1)$$

where q_m^r denotes the predicted probability of the true class r for sample m , $\delta_{(r,s)}$ represents the penalty from matrix ψ for misclassifying class r as class s , v_m is the one-hot encoded vector of the true label for sample m , M denotes the batch size, and λ is a tunable focusing parameter ($\lambda \geq 1$). To mitigate the dominance of majority classes, we introduce a modulating factor $(1 - q_m^r)^\lambda$ which reduces the loss for well-classified samples and focuses on hard-to-classify minority samples, inspired by the Focal Loss design [26]. The Adam optimizer is used to update ϕ following each batch.

After every epoch, the auxiliary loss function \mathcal{L}_ψ is applied to optimize matrix ψ . This method dynamically assigns suitable weights to each class, removing the need for complex manual settings. Research suggests that ψ is influenced by

¹<https://pan.baidu.com/s/1iKC3d4vXeqbTyG0duI5qUg?pwd=uzvq>

TABLE I
DISTRIBUTION OF CONFUSION TYPES IN THE MOOC DATASET.

Confusion type	Description	Number of the Posts
Question	These posts seek answers or help concerning the course subject matter, focusing on knowledge mastery.	5,140
Opinion	These posts expand on course content and share opinions or thoughts related to the course content.	2,227
Resource	These posts share course subject-related external resources, such as journal papers, textbooks, and newspaper articles.	533
Technical	These posts seek and provide help for technical concerns involving software, internet browsers, and course site interface and functions.	414
Social/Affective	These posts address social purposes, such as making a self-introduction and initiating study groups.	1,984
Logistical	These posts involve the availability and quality of learning materials, course policy, as well as issues related to grades and credentials.	292

factors like data distribution, inter-class separability, and classification inaccuracies [27], [28]. We will construct \mathcal{L}_ψ using these factors.

To start, we define a matrix F (dimensionally equivalent to ψ) that captures class distribution. $F_{u,v}$ represents the distribution information of class u relative to class v . Generally, classes with fewer samples are more prone to misclassification and should be assigned a higher penalty. Hence, if class u has fewer samples than class v , $F_{u,v}$ should be increased; otherwise, $F_{u,v} = 1$. The matrix F is defined as:

$$F(u, v) = \max(1, m_{\text{class}_v} / m_{\text{class}_u}) \quad (2)$$

where m_{class_u} and m_{class_v} are the numbers of samples in classes u and v , respectively. Next, to evaluate class separability, we define a matrix T that computes the mean probability difference between correct and incorrect class predictions for all samples. A smaller difference indicates lower separability. The matrix T is defined as:

$$T(u, v) = \begin{cases} -\frac{1}{M} \sum_{n=1}^M \log((q_u^n - q_v^n)^2), & u \neq v \\ 1, & u = v \end{cases} \quad (3)$$

where q_u^n is the probability of sample n being predicted as its correct class u , q_v^n is the probability when predicted as class v , and M is the number of samples in class u . The auxiliary loss \mathcal{L}_ψ for optimizing the cost matrix ψ is defined as:

$$\mathcal{L}_\psi = \|N - \delta\|_2^2 + \text{CE}_{\text{val}}(\phi) \quad (4)$$

$$N = F \cdot \exp(-T) \cdot \exp(-U) \quad (5)$$

where, CE_{val} denotes the cross-entropy loss calculated on the validation dataset, and U is a confusion matrix reflecting the current classification errors on the validation set.

IV. EXPERIMENTS AND RESULT

A. Baseline Models and Experiment Settings

To validate the effectiveness of our method, we select a range of advanced techniques as baseline models and evaluate them on our dataset. The chosen baseline models include: Machine Learning Methods: Random Forest (RF) [29] and Logistic Regression (LR) [30]. Deep Learning Methods:

TextCNN [31], TextRNN [32], FastText [33], and DPCNN [34]. Transfer Learning Methods: ERNIE [35], BERT [25], BERT-BiLSTM/CNN, and RoBERTa [36]. Except for RF and LR, other models are trained on our datasets for 20 epochs. Batch size is 8 and learning rate is 5e-5. The random seed is set to 42. For the proposed RAOCSL loss function, the diagonal of the cost matrix ψ is initialized to 1 and the other positions are initialized to 30. And a learning rate of 1e-1 is used to optimize its parameters. λ in RAOCSL is set to 2.

To assess the generalization and stability of these methods, we use five-fold cross-validation, reporting the mean and standard deviation of accuracy and macro-F1. The dataset was divided into five equal parts, with each part used in turn as the test set (20% of the data), while the remaining 80% serves as the training set. This process resulted in five distinct training-test combinations, with all experiments conducted independently. Additionally, within each training set, 5% of the data is set aside as a validation set. This validation set is used to tune model parameters during training, particularly for calculating the integral parameters of the RAOCSL loss function. This ensures that the dynamic adjustment mechanism within the loss function enhances the model's ability to identify minority classes while maintaining training stability.

B. Experimental Result

Table II compares the performance of different models on our dataset. Our method demonstrates the highest accuracy (87.9%) and macro-F1 score (29.2%), outperforming all baselines. While RF and LR show relatively high accuracy, their lower macro-F1 scores indicate limitations in handling imbalanced data. Among deep learning models, TextCNN, TextRNN, FastText, and DPCNN show some improvement, particularly in accuracy, but their macro-F1 scores remain limited. Transfer-based methods like ERNIE, BERT, and RoBERTa achieve better results, with the BERT model showing further improvement when combined with BiLSTM or CNN, indicating these additions effectively enhance feature extraction. Compared with these models, our method shows the best performance, implying our RAOCSL loss function effectively corrects the bias on the imbalanced dataset.

C. Ablation Study

To comprehensively analyze the effectiveness of the proposed components, we conduct an ablation study that evaluate

TABLE II
PERFORMANCE COMPARISON (%). \pm REPRESENTS THE STANDARD DEVIATION. ‘BERT-B/C’ DENOTES BERT-BiLSTM/CNN.

Metrics	RF	LR	TextCNN	TextRNN	FastText	DPCNN	ERNIE	BERT	BERT-B	BERT-C	RoBERTa	Ours
Accuracy	83.5 \pm 0.2	85.2 \pm 0.3	84.0 \pm 4.5	85.5 \pm 3.4	86.2 \pm 6.4	84.1 \pm 4.5	82.5 \pm 5.6	80.3 \pm 9.5	84.2 \pm 6.7	85.8 \pm 4.6	85.6 \pm 3.6	87.9\pm4.8
Macro-F1	8.13 \pm 0.5	7.91 \pm 0.6	11.2 \pm 5.3	15.4 \pm 6.2	20.5 \pm 4.4	12.0 \pm 4.5	25.1 \pm 5.6	28.1 \pm 4.4	25.8 \pm 5.7	27.9 \pm 5.7	27.7 \pm 4.6	29.2\pm5.8

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT EMBEDDING (WORD2VEC AND BERT-BASED) METHODS (%).

Metric	word2vec	Last 1	Sum 2	Sum 3	Sum 4	Concat 2	Concat 3	Concat 4
Accuracy	83.4 \pm 0.6	84.1 \pm 2.7	84.5 \pm 3.6	83.0 \pm 3.7	85.6 \pm 3.8	86.8 \pm 2.8	86.1 \pm 3.9	87.9\pm3.8
Macro-F1	12.1 \pm 1.5	26.2 \pm 5.6	27.8 \pm 6.7	27.2 \pm 6.8	28.5 \pm 5.9	28.7 \pm 6.8	29.9 \pm 5.7	30.2\pm6.8

the embedding method (“Concat 4”), adding BiLSTM and CNN on BERT respectively, and the RAOCSL loss function, which is shown in Table IV. The ‘Concat 4’ embedding method improves the performance by 1.7% in accuracy and 0.9% in macro-F1. When employing BiLSTM and CNN on BERT respectively, the accuracy is further improved from 82.0% to 85.7%, while the macro-F1 score showed a slight decrease. That implies BiLSTM and CNN can further capture discriminative information but increasing sensitivity to imbalanced data. After adding the RAOCSL loss function, the performance reaches the highest (87.9% accuracy and 29.2% macro-F1), which means RAOCSL can effectively reinforce the performance on imbalanced data. The steadily performance improvements are achieved for all components, which shows the effectiveness of our proposed method.

D. Effect of Different Embedding Methods and RAOCSL Loss Function

To explore the rationality of our embeddings, we evaluate seven methods based on the BERT-BiLSTM-CNN model: the last layer, summing, and concatenating the last 2, 3, and 4 layers, respectively. Additionally, we use word2vec [37] to generate a 3072-dimensional vector (matching the dimension of concatenating BERT’s last 4 layers) for embedding. All models are trained using the CE loss function. The results in Table III indicate that BERT-based embeddings outperform word2vec, showing that contextual representations better capture sequence-level semantics than non-contextual methods. Furthermore, concatenating the last 4 layers yields the best performance.

To evaluate the impact of λ in our loss function, we test values from 1 to 6 and compared our method with four advanced imbalanced loss functions: Cross-Entropy (CE) loss, Focal Loss (FL) [38], Cost-Weight Sensitive (CWS) loss [28], and Automatic Optimization of Class-dependent Cost (AOCC) loss [27]. Additionally, we use Random Over Sampling (ROS) [39] and Random Under Sampling (RUS) [40] to create balanced datasets with 38,464 texts (ROS) and 504 samples (RUS), respectively, and conduct experiments using CE loss. Table V shows that our RAOCSL with $\lambda = 2$ achieves the highest performance, with an accuracy of 86.9% and a macro-F1 score of 30.5%. Results for CWS suggest that relying solely

on cost-sensitive matrix parameters based on distribution may lead to suboptimal performance. The results from RUS and ROS indicate that traditional data-level methods may not be effective for highly imbalanced tasks.

TABLE IV
ABLATION STUDY ON ALL COMPONENTS. ‘BERT-B/B-C’ INDICATES BERT-BiLSTM/BiLSTM-CNN.

Concat 4	BERT-B	BERT-B-C	RAOCSL	Accuracy	macro-F1
✓				80.3 \pm 9.5	28.1 \pm 4.4
✓				82.0 \pm 5.6	29.0 \pm 4.5
✓	✓			84.2 \pm 6.5	25.8 \pm 5.7
✓	✓	✓		85.7 \pm 2.7	26.1 \pm 5.6
✓	✓	✓	✓	87.9 \pm3.8	29.2\pm5.8

TABLE V
PERFORMANCE OF DIFFERENT LOSS FUNCTIONS (%)

Loss Function	Accuracy (%)	Macro-F1 (%)
CE	85.5 \pm 3.6	29.3 \pm 6.5
FL	85.1 \pm 3.7	28.8 \pm 5.6
CWS	85.3 \pm 3.5	27.5 \pm 6.3
AOCC	85.9 \pm 3.6	29.1 \pm 6.7
ROS	85.2 \pm 3.4	22.7 \pm 5.4
RUS	28.9 \pm 3.3	18.0 \pm 2.2
RAOCSL($\lambda = 2$)	86.9\pm3.5	30.5\pm6.8

V. DISCUSSION

In this work, we explore deep neural networks for automatic identification of confusion types and create the first Chinese dataset for this task, containing 10,590 highly imbalanced text samples labeled into six classes. Our model improves text understanding by using BERT’s last four layers as embeddings for a BiLSTM-CNN network. We also propose a novel loss function, RAOCSL, which effectively addresses class imbalance issues. Experimental results demonstrate that RAOCSL significantly enhances model performance in accuracy and macro-F1 score.

VI. ACKNOWLEDGMENTS

This research is supported by the the Ministry of Education of the People’s Republic of China (No. 2021180010), Department of Science and Technology of Jilin Province, China (No. 20230201086GX).

REFERENCES

- [1] Michael Yee, Anindya Roy, Meghan Perdue, Consuelo Cuevas, Keegan Quigley, Ana Bell, Ahaan Rungta, and Shigeru Miyagawa, "Ai-assisted analysis of content, structure, and sentiment in mooc discussion forums," in *Frontiers in Education*. Frontiers Media SA, 2023, vol. 8, p. 1250846.
- [2] Hai Min Dai, Timothy Teo, Natasha Anne Rappa, and Fang Huang, "Explaining chinese university students' continuance learning intention in the mooc setting: A modified expectation confirmation model perspective," *Computers & Education*, vol. 150, pp. 103850, 2020.
- [3] Reinhard Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," *Educational psychology review*, vol. 18, pp. 315–341, 2006.
- [4] Krista R Muis, Marianne Chevrier, Courtney A Denton, and Kelsey M Losenno, "Epistemic emotions and epistemic cognition predict critical thinking about socio-scientific issues," in *Frontiers in Education*. Frontiers Media SA, 2021, vol. 6, p. 669908.
- [5] Jason M Lodge, Gregor Kennedy, Lori Lockyer, Amael Arguel, and Mariya Pachman, "Understanding difficulties and resulting confusion in learning: An integrative review," in *Frontiers in Education*. Frontiers Media SA, 2018, vol. 3, p. 49.
- [6] Mehdi Badali, Javad Hatami, Seyyed Kazem Banihashem, Ebrahim Rahimi, Omid Noroozi, and Zahra Eslami, "The role of motivation in moocs' retention rates: a systematic literature review," *Research and Practice in Technology Enhanced Learning*, vol. 17, no. 1, pp. 1–20, 2022.
- [7] Bowen Liu, Wanli Xing, Yifang Zeng, and Yonghe Wu, "Quantifying the influence of achievement emotions for student learning in moocs," *Journal of Educational Computing Research*, vol. 59, no. 3, pp. 429–452, 2021.
- [8] Sannyuya Liu, Shiqi Liu, Zhi Liu, Xian Peng, and Zongkai Yang, "Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement," *Computers & Education*, vol. 181, pp. 104461, 2022.
- [9] Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke, "Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips," *International Educational Data Mining Society*, 2015.
- [10] Zhi Liu, Qianhui Tang, Fan Ouyang, Taotao Long, and Sannyuya Liu, "Profiling students' learning engagement in mooc discussions to identify learning achievement: An automated configurational approach," *Computers & Education*, vol. 219, pp. 105109, 2024.
- [11] Wei Wang, Yongyong Zhao, Yenchun Jim Wu, and Mark Goh, "Factors of dropout from moocs: a bibliometric review," *Library Hi Tech*, vol. 41, no. 2, pp. 432–453, 2023.
- [12] Yinjuan Shao, Jingjing Zhang, Eamon Costello, and Mark Brown, "Public perceptions towards moocs on social media: an alternative perspective to understand personal learning experiences of moocs," *Interactive Learning Environments*, vol. 31, no. 2, pp. 670–682, 2023.
- [13] Omaira Almatrafi, Aditya Johri, and Huzefa Rangwala, "Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums," *Computers & Education*, vol. 118, pp. 1–9, 2018.
- [14] Diyi Yang, Robert E Kraut, and Carolyn P Rose, "Exploring the effect of student confusion in massive open online courses," *Journal of Educational Data Mining*, vol. 8, no. 1, pp. 52–83, 2016.
- [15] Ziheng Zeng, Snigdha Chaturvedi, and Suma Bhat, "Learner affect through the looking glass: Characterization and detection of confusion in online courses," *International Educational Data Mining Society*, 2017.
- [16] Tongyu Zhao, Jiaying Gao, Yu Feng, Jian Li, Yatong Zu, Sandro Pinto, Adriano Tavares, and Hao Xu, "Feature visualization and attribution analysis of confusion for massive open online course," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2023, vol. 45.
- [17] Gaganpreet Bhajji, M Ali Akber Dewan, and Fuhua Lin, "Unveiling uncertainty: Supporting learners through nlp-driven confusion identification," in *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, 2023, pp. 0137–0142.
- [18] Thushari Atapattu, Katrina Falkner, Menasha Thilakarathne, Lavendini Sivaneasharajah, and Rangana Jayashanka, "What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 878–888, 2020.
- [19] Xiaohui Tao, Aaron Shannon-Honson, Patrick Delaney, Christopher Dann, Haoran Xie, Yan Li, and Shirley O'Neill, "Towards an understanding of the engagement and emotional behaviour of mooc students using sentiment and semantic features," *Computers and Education: Artificial Intelligence*, vol. 4, pp. 100116, 2023.
- [20] Abdessamad Chanaa and Nour-Eddine El Faddouli, "Bert and prerequisite based ontology for predicting learner's confusion in moocs discussion forums," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*. Springer, 2020, pp. 54–58.
- [21] Hanxiang Du and Wanli Xing, "Leveraging explainability for discussion forum classification: Using confusion detection as an example," *Distance Education*, vol. 44, no. 1, pp. 190–205, 2023.
- [22] Qiaorong Zhang and Lin Sun, "A cnn-bi-lstm model for mooc forum post classification," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 18, no. 21, pp. 89–101, 2023.
- [23] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé, "Investigating how student's cognitive behavior in mooc discussion forums affect learning gains," *International Educational Data Mining Society*, 2015.
- [24] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [25] Jacob Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [26] Sangwook Park and Mounya Elhilali, "Time-balanced focal loss for audio event detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 311–315.
- [27] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3573–3587, 2017.
- [28] Imranul Ashrafi, Muntasir Mohammad, Arani Shawkat Mauree, Galib Md Azraf Nijhum, Redwanul Karim, Nabeel Mohammed, and Sifat Momen, "Banner: A cost-sensitive contextualized model for bangla named entity recognition," *IEEE Access*, vol. 8, pp. 58206–58226, 2020.
- [29] Leo Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
- [30] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009.
- [31] Y Chen, "Convolutional neural network for sentence classification (unpublished master's thesis). university of waterloo, canada," 2015.
- [32] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101*, 2016.
- [33] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.
- [34] Rie Johnson and Tong Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 562–570.
- [35] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu, "Ernie: Enhanced representation through knowledge integration," *arXiv preprint arXiv:1904.09223*, 2019.
- [36] Y Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [37] Tomas Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [39] David Masko and Paulina Hensman, "The impact of imbalanced training data for convolutional neural networks," 2015.
- [40] Hansang Lee, Minseok Park, and Junmo Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3713–3717.