Wiley
Journal of Sensors
Volume 2024, Article ID 3857343, 12 pages
https://doi.org/10.1155/2024/3857343

WILEY

Research Article

Enhancing College Student Education and Management through Semisupervised Learning

Jiqing Huang,[1,2] **Hua Zhao**,[3] **and Xiaojun Jin**[4]

[1]*College of Educational Sciences, Yangzhou University, Yangzhou, Jiangsu 225009, China*
[2]*The Communist Youth League Committee, Jiangsu Ocean University, Lianyungang, Jiangsu 222005, China*
[3]*School of Mechanical Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu 222005, China*
[4]*Shandong Laboratory of Advanced Agricultural Sciences at Weifang, Peking University Institute of Advanced Agricultural Sciences, Weifang, Shandong 261325, China*

Correspondence should be addressed to Xiaojun Jin; xiaojun.jin@outlook.com

Received 13 December 2023; Revised 8 July 2024; Accepted 23 July 2024

Academic Editor: Sangsoon Lim

College student education and management can be enhanced through a data-driven approach involving student surveys, academic records, and text analysis to understand student interests and concerns. Effective categorization of relevant topics enables universities to provide tailored support and educational content, thus improving the quality of education and fostering student success and well-being by adapting to evolving student needs and aspirations. The primary contribution of this work is demonstrating the effectiveness of semisupervised learning methods in educational content classification, providing a robust solution for enhancing college student education and management with limited labeled data. The objective of this study was to evaluate the feasibility of using semisupervised learning methods in educational content classification using the Yahoo Answers dataset. For the Yahoo_500 dataset, the supervised neural network achieved a best evaluation accuracy of 0.6565, an average precision of 0.6539, an average recall of 0.6565, and an average F_1 score of 0.6547. In contrast, semisupervised approaches, Dash, FixMatch, and FreeMatch, consistently demonstrated superior performance. Among the evaluated semisupervised architectures, FreeMatch achieved the highest best evaluation accuracy (0.6759), average precision (0.6739), average recall (0.6759), and average F_1 score (0.6744). The Yahoo_2000 dataset, which benefited from an increased labeled data pool, exhibited a similar trend with semisupervised approaches consistently surpassing the supervised approach. FreeMatch maintained its top performing position in several categories, including computers and Internet, consumer electronics, and business and finance, with impressive F_1 score of ≥ 0.753. Overall, the semisupervised approaches prove highly effective in improving model performance, highlighting its practical advantages. These results underscore the robustness of semisupervised approaches and their capability to improve classification performance even with limited labeled data. Employing semisupervised learning on the Yahoo Answers dataset and additional data sources for college student management and education can be a powerful tool for gaining insights into students' interests and concerns.

1. Introduction

Enhancing college student education and management is of paramount importance in today's educational landscape [1]. Through a multifaceted approach, including the analysis of various data sources such as student surveys, online content, and additional data like academic records, extracurricular activities, and social interactions, universities can gain profound insights into the ever-evolving interests and concerns of their student body [2, 3]. This holistic understanding empowers educators and administrators to tailor their strategies and support services to the specific needs and aspirations of students [4]. By providing personalized guidance and relevant resources, universities can facilitate a more focused, effective, and timely response to students' evolving requirements [5, 6]. Moreover, proactive management that anticipates and adapts to students' needs can greatly enhance their overall educational experience [7]. Through this proactive approach, universities can ensure that students receive not only a high-quality education but also the support and resources they need to thrive and succeed [8]. This approach fosters a dynamic and responsive educational environment, ultimately benefiting both students and institutions in the pursuit of educational excellence [9, 10].

In college student education and management, emotions and psychological analysis also play a crucial role. Analyzing the emotions and psychology expressed in the texts and communications of students (e.g., posts, academic papers, social media content, or surveys) is a fundamental tool for understanding their psychological well-being [11, 12]. It offers insights into their prevailing emotional states, concerns, and responses to academic and personal challenges. A facet of analysis provides universities with a unique window into the emotional lives of students, enabling them to offer targeted emotional support and mental health resources [13, 14]. It is imperative to acknowledge that the emotional well-being of students is intricately linked to their academic success and overall satisfaction with their college experience [15]. Institutions that can effectively gauge and address these emotional needs are better equipped to provide a holistic education that nurtures not only students' intellectual growth but also their mental and emotional resilience. Nevertheless, the analysis of student texts and communications necessitates effective categorization of the topics that are most relevant to students. By gaining a comprehensive understanding of what students care about, universities can not only provide emotional support but also deliver educational content and resources that resonate with their interests, challenges, and aspirations [16, 17].

Recently, an increasing number of studies have leveraged artificial intelligence techniques, notably machine learning, to explore the use of text analysis for sentiment perception in the context of education or management [18, 19, 20]. For instance, [21] employed machine learning techniques to tackle sentiment analysis challenges in blog, review, and forum texts written in English, Dutch, and French. Their work focused on addressing issues such as noisy text, entity sentiment attribution, and the limited size of the training dataset. Using a combination of unigram and linguistic features, their approach achieved an accuracy rate of ~83% for English texts. In contrast, the accuracy rates for Dutch and French texts were slightly lower due to the greater linguistic diversity in those languages. These experiments provided valuable insights into the transferability of the learned models across different domains and languages, highlighting the crucial role of linguistic nuances in sentiment analysis. Salinca utilized multiple feature extraction methods and machine learning models to investigate the Yelp Challenge dataset for automatic sentiment classification, achieving an impressive accuracy of 94.4% with Linear SVC and SGD classifiers [22]. The study also underscores the potential for further accuracy enhancements by incorporating advanced linguistic features such as bigrams, trigrams, and part-of-speech tagging. Halde et al. [23] used machine learning to investigate the impact of students' psychology and study skills on academic performance. Their study, which utilized real-time data from final-year students, revealed that motivation and information processing play vital roles in performance prediction, leading to a 4–6% improvement in accuracy [23]. Zhang [24] employed machine learning-based psychology evaluation of college students to identify critical factors affecting practical capabilities and leadership awareness. The research provided a thorough assessment by utilizing the EN-TOPSIS framework for building an innovative health service system.

By utilizing artificial intelligence to analyze the emotions, psychology, interests, or concerns of college students, the essential task involves accurately categorizing their relevant textual materials. This categorization allows college educators or administrators to engage in targeted communication and interventions based on the emotions, psychology, interests, or concerns encapsulated within the materials [25]. In the long run, this approach is poised to ultimately result in more effective education and management of college students. Despite the growing recognition of the importance of educational content analysis in student management, a significant challenge exists in efficiently collecting and analyzing relevant data [16, 26]. Currently, the assessment of students' interests and concerns often relies on labor-intensive manual judgments or self-report surveys [27]. These methods, while widely used, have inherent limitations in terms of accuracy and scalability. They not only consume significant time but are also susceptible to biases and errors due to the complexity and dynamic nature of students' interests and concerns. Furthermore, conducting analysis of students' interests and concerns demands substantial human resources and time investments, making the scaling to large-scale applications a formidable task [28].

Semisupervised learning is a machine learning paradigm that combines elements of both supervised and unsupervised learning. It leverages a small amount of labeled data along with a larger pool of unlabeled data to train models, offering several advantages over traditional supervised learning methods [29, 30]. Firstly, it is often more cost-effective and less labor-intensive, as obtaining labeled data can be expensive and time-consuming. Secondly, semisupervised learning can lead to more robust models by harnessing additional information from the unlabeled data, resulting in better generalization to new, unseen examples [31]. Semisupervised learning finds widespread applications across various domains, including natural language processing, computer vision, and speech recognition [32]. In natural language processing, it has been employed for tasks such as part-of-speech tagging, named entity recognition, and machine translation [33, 34]. In the area of computer vision, semisupervised learning has been applied to image classification [35, 36], object detection [37, 38], and segmentation [39, 40].

In the context of enhancing college student education and management, semisupervised learning holds great potential. With the availability of vast amounts of textual data on social media platforms, blogs, and other online sources, efficient organization and access to educational content can significantly benefit from semisupervised approaches. By leveraging both labeled data, such as manually annotated text in a small dataset, and unlabeled data, like unannotated social media posts, models can better capture the nuances of human emotions and opinions. Moreover, this approach can provide richer data, enabling educational institutions to gain a deeper understanding of their students and offer personalized support, ultimately enhancing student management practices.

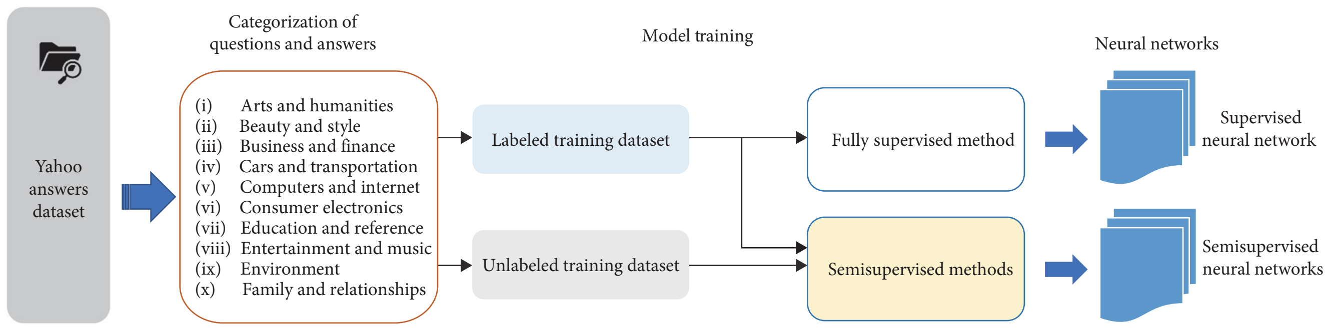

Figure 1: The flowchart of the overall process.

The ultimate goal is to use semisupervised learning techniques to enhance higher education management practices, where effective organization and retrieval of educational resources and information play a pivotal role. Through the utilization of deep learning models to categorize textural materials, we aim to provide educators and students with a more streamlined and efficient means of accessing relevant educational content. This study explored semisupervised learning methods in the educational domain, addressing the limitations of existing studies by leveraging both labeled and unlabeled data for more effective content classification. The specific objectives of this study were to (1) investigate the feasibility of using semisupervised learning methods, including Dash, FixMatch, and FreeMatch, in educational content classification, with a focus on the Yahoo Answers dataset, and (2) compare and analyze the results of both supervised method and various semisupervised methods in educational content classification using the Yahoo Answers dataset.

2. Materials and Methods

2.1. Overview. In the present research work, four different deep learning architectures, encompassing both supervised and semisupervised methods, were trained to effectively sort and categorize the questions and answers within the Yahoo Answers dataset (Figure 1). By classifying this user-generated content into the appropriate categories, we can further enhance its applicability to improve university-level education and management. The baseline architectures used in this study were as follows:

(i) The first baseline architecture was a supervised method with BERT as the backbone network. BERT, pretrained on a large corpus of text data, provided contextual embeddings which were fed into additional layers for the classification task. After preparing the text data and transforming it into a suitable format for BERT, fully connected layers were added on top. The combined model, consisting of the pretrained BERT and the added classification layers, was then fine-tuned using labeled data. The supervised method in this study was referred to as the BERT-based text classifier, trained solely on labeled data. This method was utilized as a benchmark for supervised learning to assess the performance of subsequent semisupervised methods.

(ii) Dash was employed as the second baseline to evaluate the effectiveness of semisupervised learning. Dash is a powerful and versatile framework widely used in semisupervised learning, providing a comprehensive platform for creating interactive web-based applications for data visualization and analysis.

(iii) The third baseline was FixMatch. FixMatch is a semisupervised learning method that focuses on consistency regularization. It leverages a small amount of labeled data and a large pool of unlabeled data. In FixMatch, a model is trained on both labeled and unlabeled data, and it enforces consistency between the predictions on pseudolabeled unlabeled data and the actual labels on labeled data. This technique helps the model generalize better and improves its performance on the labeled data. FixMatch has been widely adopted in various applications, especially in scenarios where acquiring labeled data is expensive or time-consuming.

(iv) FreeMatch is another semisupervised learning framework that aims to enhance model robustness and accuracy. It introduces a concept called "Free Examples," which are selected from unlabeled data based on their confidence scores. Free Examples are the most confidently predicted unlabeled samples that the model uses as additional training data. This approach distinguishes FreeMatch from other semisupervised methods by dynamically selecting high-confidence samples. By incorporating free examples into the training process, FreeMatch achieves improved performance and robustness against noise in the unlabeled data. It is particularly beneficial in situations where the quality of unlabeled data can vary.

To ensure a fair comparison, all these models utilize the same backbone network (BERT) and are evaluated on the same task. By comparing the performance of these models, we aim to gain deeper insights into the relative advantages and disadvantages of supervised and semisupervised learning methods for this specific task.

2.2. Dataset. The public Yahoo Answers dataset was utilized to train the neural networks, leveraging the extensive textual data to explore novel approaches in the field of college student education and management. The Yahoo Answers dataset is a vast and diverse collection of user-generated content

Table 1: The number of data samples used to establish the training, validation, and testing datasets of the supervised and semisupervised methods.

Dataset	Labeled	Unlabeled	Validation	Testing
Yahoo_500	500	500,000	50,000	60,000
Yahoo_2000	2000			

Table 2: Hyperparameters used for training the supervised and semisupervised neural networks.

Deep learning architecture	Backbone	Iteration	Optimizer	Base learning rate	Weight decay	Batch size	Eval batch size	Momentum
Supervised	BERT	102,400	AdamW	0.0001	0.0005	4	4	0.9
Dash	BERT	102,400	AdamW	0.0001	0.0005	4	4	0.9
FixMatch	BERT	102,400	AdamW	0.0001	0.0005	4	4	0.9
FreeMatch	BERT	102,400	AdamW	0.0001	0.0005	4	4	0.9

from Yahoo Answers, a popular online Q&A platform. It comprises a wide range of questions, answers, and user interactions spanning various topics and categories. Yahoo Answers allows users to ask questions and receive answers from the community, and the dataset reflects this collaborative and diverse nature. The dataset contains text data that includes questions posed by users, answers provided by the community, and additional metadata such as user IDs, timestamps, and category labels. To provide a more comprehensive understanding, the dataset includes ~4.4 million questions and over 14 million answers, categorized into 10 main categories and over 1,000 subcategories, ranging from "education and reference" to "computers and Internet." Moreover, it extends beyond the mere Q&A format, incorporating additional layers of interaction such as comments, votes, and user profiles. This intricate network of interactions enables researchers to delve into the dynamics of user engagement, community building, and information sharing within online platforms. These interactions are represented through metadata, which includes over 2 million comments and voting data that helps in understanding the popularity and reliability of responses. The Yahoo Answers dataset is valuable for natural language processing (NLP) and machine learning research due to its real-world data, user behavior insights, and various research applications. Researchers have used it for tasks like question answering, sentiment analysis, and studying information diffusion and misinformation in online communities. The dataset's diverse and extensive content makes it an excellent resource for training robust models that require understanding of nuanced human interactions and language use in varied contexts.

2.3. Training and Testing. To assess the performance of semisupervised methods, we conducted training and validation using the datasets detailed in Table 1. The data employed for training the semisupervised neural networks were categorized into two groups. One group comprises 500 labeled data samples (Yahoo_500), while the other includes 2000 labeled data samples (Yahoo_2000). Both Yahoo_500 and Yahoo_2000 utilize the same pool of 500,000 unlabeled data samples, maintaining consistency across different training scenarios. Each validation set comprises 50,000 data samples, and each testing set contains 60,000 data samples, drawn from separate partitions of the Yahoo Answers dataset to avoid overlap with the training data. This separation guarantees that our evaluation metrics accurately reflect the models' ability to generalize to unseen data. All datasets, including labeled, unlabeled, validation, and testing sets, were randomly selected from the broader Yahoo Answers dataset to ensure they are representative and diverse.

The supervised and semisupervised neural networks, including Dash, FixMatch, and FreeMatch, were trained and tested on the PyTorch deep learning environment (https://pytorch.org/; version 1.8.1; Facebook, San Jose, California, United States) using an NVIDIA RTX 3080 graphics processing unit (GPU) with 128 GB of memory. The same hyperparameter values for each neural network, as presented in Table 2, were adopted to ensure a fair comparison between the evaluated neural networks. Using the same hyperparameters ensures that any performance differences are due to the learning methods themselves rather than variations in the training settings.

For both supervised and semisupervised neural networks, validation and testing results were presented in a binary classification confusion matrix which included four conditions, including true positive (*tp*), false positive (*fp*), true negative (*tn*), and false negative (*fn*). *tp* represents the instances where the neural networks correctly identified positive cases. In our context, it signifies the questions or answers that were accurately classified as relevant or belonging to a specific category. *fp* indicates the instances where the neural networks incorrectly identified cases as positive when they were actually negative. This would represent questions or answers mistakenly categorized as relevant or belonging to a category when they should not have been. *tn* reflects the instances where the neural networks correctly identified negative cases. In the context of our training and testing, this would represent questions or answers accurately classified as irrelevant or not belonging to a particular category. *fn* highlights the instances where the neural networks incorrectly identified cases as negative when they were genuinely positive. In our scenario, this would signify questions or answers that should have been categorized as relevant or belonging to a category but were missed or misclassified.

Precision, recall, and F_1 score were used as classification metrics to evaluate the performance of the supervised and

Table 3: The performances of the supervised and semisupervised neural networks in validation datasets.

Deep learning architecture	Dataset	Best eval acc	Avg_pre	Avg_rec	Avg_F_1_score
Supervised	Yahoo_500	0.6565	0.6539	0.6565	0.6547
Dash	Yahoo_500	0.6683	0.6677	0.6683	0.6602
FixMatch	Yahoo_500	0.6636	0.6624	0.6636	0.6557
FreeMatch	Yahoo_500	0.6759	0.6739	0.6759	0.6744
Supervised	Yahoo_2000	0.6723	0.6774	0.6723	0.6703
Dash	Yahoo_2000	0.6928	0.6923	0.6928	0.6878
FixMatch	Yahoo_2000	0.6952	0.6903	0.6952	0.6879
FreeMatch	Yahoo_2000	0.6971	0.6948	0.6971	0.6950

semisupervised neural networks. Precision is a metric that measures the accuracy of the neural networks' positive predictions. It calculates the proportion of true positive predictions (correctly identified positive cases) over the total positive predictions (true positives plus false positives). Precision enables us to evaluate the accurate classification of questions or answers that are relevant or category-specific by the neural networks. It measures the proportion of true positive predictions among the identified instances:

$$\text{Precision} = \frac{tp}{tp + fp}. \tag{1}$$

Recall, also known as sensitivity or true positive rate, evaluates the neural networks' ability to identify all the relevant or category-specific instances. It calculates the ratio of true positive predictions over the total actual positive instances. In the context of our dataset, recall measures the neural networks' effectiveness in capturing all the relevant questions or answers:

$$\text{Recall} = \frac{tp}{tp + fn}. \tag{2}$$

The F_1 score is the harmonic mean of precision and recall and provides a balanced measure of the neural networks' performance. It combines both precision and recall into a single metric by taking their harmonic mean. A higher F_1 score suggests that the neural networks have a better balance between precision and recall:

$$F_1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

These metrics collectively offer a comprehensive understanding of the neural networks' strengths and weaknesses when classifying data on the Yahoo Answers dataset. Precision emphasizes the neural networks' precision in making positive predictions, recall highlights its ability to find all relevant instances, and the F_1 score provides a holistic evaluation of its overall classification performance, considering both precision and recall. These metrics are essential for fine-tuning and optimizing the neural networks for better results in real-world applications.

For the validation data, the performance of supervised and semisupervised neural networks was evaluated using the average precision, average recall, and average F_1 score. These average metrics are indispensable for evaluating the neural networks' performance in the context of multiclass or multicategory classification tasks. They provide a comprehensive perspective on how the neural networks manage the trade-off between precision and recall across diverse classes, thereby facilitating a comprehensive assessment of its overall efficacy in data classification:

$$\text{avg\_pre} = \frac{1}{C_N} \sum_{i=1}^{C_N} \text{Precision}, \tag{4}$$

$$\text{avg\_rec} = \frac{1}{C_N} \sum_{i=1}^{C_N} \text{Recall}, \tag{5}$$

$$\text{avg\_}f_1 = \frac{1}{C_N} \sum_{i=1}^{C_N} F_1, \tag{6}$$

where C_N was the number of classes. In our case, the value of C_N is 10.

3. Results

Table 3 presents the performances of supervised and semisupervised neural networks, including Dash, FixMatch, and FreeMatch, on classifying data on the Yahoo Answers dataset in the validation dataset. For the Yahoo_500 dataset, the supervised neural network achieved a best evaluation accuracy of 0.6565, an average precision of 0.6539, an average recall of 0.6565, and an average F_1 score of 0.6547. In contrast, semisupervised approaches, Dash, FixMatch, and FreeMatch, consistently demonstrated superior performance. FreeMatch, for instance, achieved the highest best evaluation accuracy (0.6759), average precision (0.6739), average recall (0.6759), and average F_1 score (0.6744). This underscores the advantage of incorporating unlabeled data in enhancing model performance, even when labeled data is scarce.

Compared to the Yahoo_500 dataset, the findings from the Yahoo_2000 dataset demonstrated a consistent pattern. The expansion of the labeled data pool in Yahoo_2000 resulted in a slight enhancement in performance across all deep learning architectures. The semisupervised methods

Table 4: The performances of the supervised and semisupervised neural networks in Yahoo_500 testing datasets.

Deep learning architecture	Classes	Precision	Recall	F_1 score
Supervised	Arts and humanities	0.547	0.497	0.521
	Beauty and style	0.628	0.659	0.643
	Business and finance	0.655	0.717	0.685
	Cars and transportation	0.261	0.604	0.365
	Computers and Internet	0.874	0.682	0.766
	Consumer electronics	0.809	0.788	0.798
	Education and reference	0.486	0.444	0.464
	Entertainment and music	0.647	0.611	0.628
	Environment	0.712	0.657	0.684
	Family and relationships	0.716	0.677	0.696
Dash	Arts and humanities	0.374	0.687	0.484
	Beauty and style	0.628	0.723	0.672
	Business and finance	0.786	0.72	0.751
	Cars and transportation	0.524	0.487	0.505
	Computers and Internet	0.88	0.78	0.827
	Consumer electronics	0.86	0.853	0.857
	Education and reference	0.46	0.421	0.439
	Entertainment and music	0.664	0.737	0.699
	Environment	0.839	0.562	0.673
	Family and relationships	0.611	0.768	0.681
FixMatch	Arts and humanities	0.582	0.483	0.528
	Beauty and style	0.584	0.66	0.62
	Business and finance	0.84	0.556	0.669
	Cars and transportation	0.406	0.504	0.449
	Computers and Internet	0.819	0.806	0.813
	Consumer electronics	0.826	0.798	0.811
	Education and reference	0.388	0.642	0.483
	Entertainment and music	0.675	0.647	0.66
	Environment	0.832	0.606	0.701
	Family and relationships	0.445	0.851	0.584
FreeMatch	Arts and humanities	0.522	0.543	0.532
	Beauty and style	0.653	0.646	0.65
	Business and finance	0.76	0.747	0.754
	Cars and transportation	0.477	0.508	0.492
	Computers and Internet	0.803	0.823	0.813
	Consumer electronics	0.853	0.821	0.836
	Education and reference	0.51	0.543	0.526
	Entertainment and music	0.697	0.652	0.674
	Environment	0.773	0.696	0.733
	Family and relationships	0.702	0.747	0.724

continued to outperform the supervised approach, with FreeMatch retaining its position as the leading framework, attaining an impressive best evaluation accuracy of 0.6971, an average precision of 0.6948, an average recall of 0.6971, and an average F_1 score of 0.6950. The results for both Yahoo_500 and Yahoo_2000 datasets affirm the conclusion that semisupervised methods, Dash, FixMatch, and FreeMatch, surpass the supervised approach. Notably, this trend highlights the potential of semisupervised learning in efficiently utilizing unlabeled data, making it less sensitive to the quantity of labeled data. This interpretation underscores the robustness of semisupervised methods and their capability to improve classification performance even with limited labeled data.

Tables 4 and 5 provide a detailed evaluation of supervised and semisupervised methods' performance on the testing dataset in terms of precision, recall, and F_1 score. Each deep learning model was evaluated in 10 categories, including arts and humanities, beauty and style, business and finance, cars and transportation, computers and Internet, consumer electronics, education and reference, entertainment and music, environment, and family and relationships.

Table 5: The performances of the supervised and semisupervised neural networks in Yahoo_2000 testing datasets.

Deep learning architecture	Classes	Precision	Recall	F_1 score
Supervised	Arts and humanities	0.454	0.656	0.537
	Beauty and style	0.626	0.687	0.655
	Business and finance	0.791	0.644	0.71
	Cars and transportation	0.463	0.5	0.48
	Computers and Internet	0.824	0.799	0.811
	Consumer electronics	0.834	0.842	0.838
	Education and reference	0.538	0.46	0.496
	Entertainment and music	0.652	0.68	0.665
	Environment	0.742	0.686	0.713
	Family and relationships	0.72	0.715	0.718
Dash	Arts and humanities	0.512	0.657	0.575
	Beauty and style	0.692	0.689	0.691
	Business and finance	0.813	0.701	0.753
	Cars and transportation	0.543	0.479	0.509
	Computers and Internet	0.879	0.783	0.829
	Consumer electronics	0.881	0.836	0.858
	Education and reference	0.418	0.63	0.502
	Entertainment and music	0.66	0.758	0.706
	Environment	0.771	0.665	0.714
	Family and relationships	0.764	0.724	0.744
FixMatch	Arts and humanities	0.503	0.676	0.577
	Beauty and style	0.726	0.652	0.687
	Business and finance	0.753	0.756	0.754
	Cars and transportation	0.467	0.564	0.511
	Computers and Internet	0.889	0.761	0.82
	Consumer electronics	0.897	0.806	0.849
	Education and reference	0.422	0.619	0.502
	Entertainment and music	0.689	0.723	0.705
	Environment	0.82	0.63	0.713
	Family and relationships	0.801	0.728	0.762
FreeMatch	Arts and humanities	0.499	0.656	0.567
	Beauty and style	0.677	0.663	0.67
	Business and finance	0.784	0.725	0.753
	Cars and transportation	0.587	0.471	0.523
	Computers and Internet	0.821	0.845	0.833
	Consumer electronics	0.851	0.869	0.86
	Education and reference	0.524	0.502	0.513
	Entertainment and music	0.691	0.73	0.71
	Environment	0.762	0.729	0.745
	Family and relationships	0.715	0.79	0.751

In the Yahoo_500 dataset, semisupervised models consistently outperformed the purely supervised approach across various categories. Notably, FreeMatch demonstrated robust performance, achieving impressive F_1 scores. For example, in the consumer electronics category, FreeMatch achieved an outstanding F_1 score of 0.836, indicating its prowess in harnessing unlabeled data. However, the cars and transportation category posed a challenge for all models, with F_1 scores hovering around 0.45. Education and reference also exhibited similar challenges, which could be attributed to the fewer samples in these categories, highlighting the need for further research on addressing class imbalance or fine-tuning. Family and relationships exhibited varied results across models, while computers and Internet showed strong performance for all models, with Dash achieving an F_1 score of 0.827.

The Yahoo_2000 dataset, which benefited from an increased labeled data pool, exhibited a similar trend with semisupervised models consistently surpassing the supervised approach. FreeMatch maintained its top-performing position in several categories, including computers and Internet, consumer electronics, and business and finance, with impressive F_1 scores of ≥ 0.753. The family and relationships category showcased differences in

Table 6: Cross-validation results for supervised and semisupervised methods.

Deep learning architecture	Dataset	Cross-validation dataset	Avg_pre	Avg_rec	Avg_F_1_score
Supervised	Yahoo_500	Dataset 1	0.6345	0.6332	0.6338
	Yahoo_500	Dataset 2	0.6349	0.6295	0.6322
	Yahoo_500	Dataset 3	0.6347	0.6342	0.6344
Dash	Yahoo_500	Dataset 1	0.6670	0.6566	0.6618
	Yahoo_500	Dataset 2	0.6700	0.6563	0.6631
	Yahoo_500	Dataset 3	0.6694	0.6571	0.6632
FixMatch	Yahoo_500	Dataset 1	0.6577	0.6406	0.6490
	Yahoo_500	Dataset 2	0.6549	0.6364	0.6455
	Yahoo_500	Dataset 3	0.6563	0.6350	0.6455
FreeMatch	Yahoo_500	Dataset 1	0.6739	0.6759	0.6749
	Yahoo_500	Dataset 2	0.6689	0.6700	0.6694
	Yahoo_500	Dataset 3	0.6698	0.6712	0.6705
Supervised	Yahoo_2000	Dataset 1	0.6631	0.6603	0.6617
	Yahoo_2000	Dataset 2	0.6618	0.6596	0.6607
	Yahoo_2000	Dataset 3	0.6625	0.6608	0.6616
Dash	Yahoo_2000	Dataset 1	0.6923	0.6928	0.6925
	Yahoo_2000	Dataset 2	0.6932	0.6903	0.6917
	Yahoo_2000	Dataset 3	0.6924	0.6908	0.6916
FixMatch	Yahoo_2000	Dataset 1	0.6876	0.6925	0.6900
	Yahoo_2000	Dataset 2	0.6870	0.6904	0.6887
	Yahoo_2000	Dataset 3	0.6885	0.6911	0.6898
FreeMatch	Yahoo_2000	Dataset 1	0.6968	0.6896	0.6932
	Yahoo_2000	Dataset 2	0.6948	0.6971	0.6959
	Yahoo_2000	Dataset 3	0.6951	0.6959	0.6955

Table 7: Descriptive statistics for cross-validation F_1 scores.

Group	N	Mean	Std. deviation	Std. error
Supervised/500	3	0.633467	0.0011372	0.0006566
Dash/500	3	0.662700	0.0007810	0.0004509
FixMatch/500	3	0.646667	0.0020207	0.0011667
FreeMatch/500	3	0.671600	0.0029103	0.0016803
Supervised/2000	3	0.661333	0.0005508	0.0003180
Dash/2000	3	0.691933	0.0004933	0.0002848
FixMatch/2000	3	0.689500	0.0007000	0.0004041
FreeMatch/2000	3	0.694867	0.0014572	0.0008413
Total	24	0.669008	0.0213469	0.0043574

model performance, while education and reference remained challenging for all models, with F_1 scores around 0.5. Overall, the semisupervised approaches prove highly effective in improving model performance, highlighting its practical advantages when labeled data is limited.

To further validate our findings, we conducted threefold cross-validation experiments on the Yahoo Answers dataset. The results, presented in Table 6, demonstrated the performance consistency of both supervised and semisupervised methods across different subsets of the data.

The cross-validation results confirmed the robustness of the semisupervised methods. For both 500 and 2000 labeled data points, the semisupervised approaches (Dash, FixMatch, and FreeMatch) consistently outperformed the fully supervised approach. Notably, FreeMatch demonstrated the highest average F_1 scores across all datasets, further validating its effectiveness in leveraging unlabeled data.

These results aligned with the previous findings and underscored the advantage of incorporating unlabeled data to enhance model performance. The consistent performance improvements observed in the cross-validation experiments highlighted the reliability and generalizability of the semisupervised methods used in this study.

The descriptive statistics for the cross-validation F_1 scores are presented in Table 7. FreeMatch demonstrated the highest average F_1 scores for both 500 and 2000 labeled data samples,

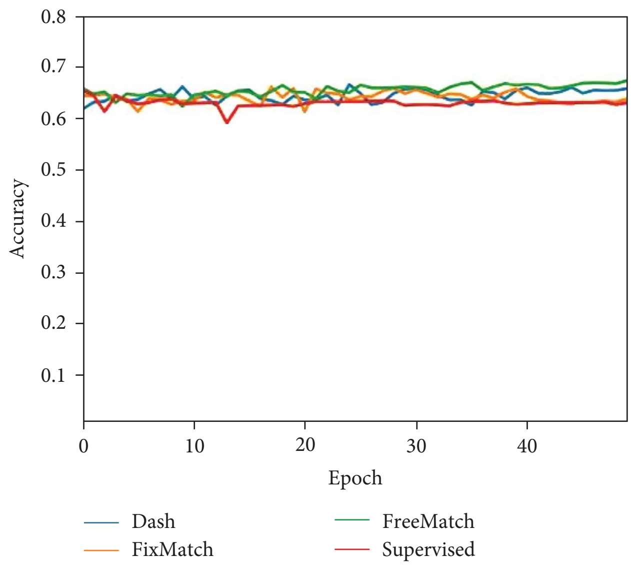

FIGURE 2: Comparison of accuracy evaluation of the supervised and semisupervised neural networks in Yahoo_500 testing datasets.

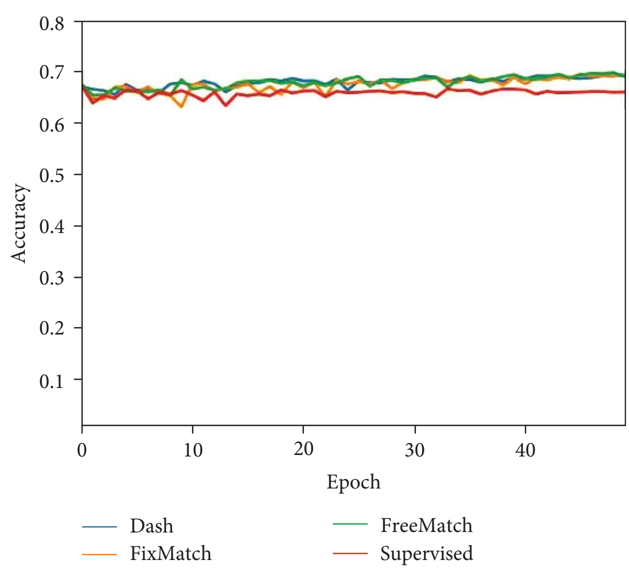

FIGURE 3: Comparison of accuracy evaluation of the supervised and semisupervised neural networks in Yahoo_2000 testing datasets.

with mean F_1 scores of 0.6716 and 0.6949, respectively. The standard deviations for the semisupervised methods were relatively low, indicating stable performance across different folds of the cross-validation. For instance, Dash/500 had a standard deviation of 0.0007810, and FreeMatch/2000 had a standard deviation of 0.0014572, further reinforcing the robustness of these methods. Moreover, the standard error values in Table 7 were relatively small, suggesting that the sample mean F_1 scores were reliable estimates of the true population mean. This highlighted the stable and consistent performance of the models.

Figures 2 and 3 illustrate the evolution of accuracy throughout the training process for both supervised and semisupervised methods. Within the Yahoo_500 dataset, FreeMatch demonstrated a consistently smooth training accuracy curve from epochs 0–50. It commenced with relatively high accuracy, indicating its ability to effectively leverage labeled and unlabeled data. The curve exhibited minimal fluctuations, highlighting its robust performance. However, the purely supervised approach started with the highest initial accuracy compared to semisupervised methods but exhibited limited improvement during training. In contrast, semisupervised methods exhibited a more consistent and upward trajectory. Among the various semisupervised architectures, FreeMatch consistently outperformed the other methods in this task. It demonstrated superior accuracy levels that were sustained throughout the entire training process. Dash and FixMatch displayed similar trajectories but at slightly lower overall accuracy. This illustrated the potential of different semisupervised techniques to boost performance, with FreeMatch being the most effective in this case.

The transition from Yahoo_500 to Yahoo_2000, which involved a larger pool of labeled data, showed notable changes in the accuracy curves. In both supervised and semisupervised cases, the increase in labeled data led to improved performance. However, even with this expansion in labeled data, semisupervised methods still maintained their superiority over purely supervised methods in Yahoo_2000. The rise in accuracy was more gradual in Yahoo_2000 due to the diminishing returns of additional labeled samples. These observations emphasized the advantages of semisupervised learning methods, especially FreeMatch, in effectively utilizing both labeled and unlabeled data for text classification tasks. The inclusion of more labeled data in Yahoo_2000 further corroborates the benefits of semisupervised learning, showcasing its resilience and superiority over supervised approaches, which experience diminishing returns as more labeled data are added.

In summary, the training accuracy data suggest that the semisupervised models, including FreeMatch, Dash, and FixMatch gradually adapt to the data and eventually reach a stable level of accuracy, outperforming the purely supervised approach. This observation aligns with the final performance results on the testing dataset, as the semisupervised models exhibit higher F_1 scores and better generalization. The use of unlabeled data is a significant factor in the enhanced performance of these models.

4. Discussion

Semisupervised learning has shown promise in the context of class-imbalanced data within the Yahoo Answers dataset. However, the effectiveness of these methods is not uniform across all categories, given the severe class imbalance in some. Future research should delve deeper into this issue to ensure a comprehensive understanding. Investigating the impact of class imbalance on model performance is essential. The underrepresentation of certain classes in specific categories might lead to reduced model effectiveness, which necessitates tailored solutions. One such approach is the development of adaptive semisupervised strategies that can account for the imbalance and provide more nuanced training. This could include techniques to redistribute underrepresented classes or methods for oversampling them during the training process. Additionally, techniques like Synthetic Minority Over-sampling Technique (SMOTE) can be explored to address class imbalance by generating synthetic samples for the minority classes. Furthermore, researchers should explore the integration of external knowledge sources, such as pretrained language models, to mitigate the impact of class imbalance. These resources may help the models generalize more effectively in categories that are particularly challenging due to class imbalance. Additionally, introducing more fine-grained evaluation metrics like precision–recall curves or class-specific F_1 scores can offer a deeper understanding of model performance under class-imbalanced conditions. Lastly, active learning strategies should be considered as a means to address imbalance by selectively acquiring labels for underrepresented classes.

While semisupervised models have demonstrated effectiveness, there is still room for improvement. To enhance accuracy, it is essential to consider various model-related factors. One avenue of future research involves refining the consistency mechanisms within semisupervised models. These mechanisms include data augmentation techniques and consistency loss functions. Experimenting with different forms of data augmentation or consistency loss functions that adapt to specific tasks or data conditions can yield significant performance gains. Moreover, the quality of pseudolabels and the level of confidence used for selecting unlabeled samples in self-training strategies should be carefully tuned. This fine-tuning can significantly improve model accuracy. Exploring different model architectures, especially those tailored for semisupervised learning, may also lead to enhanced performance. Another avenue involves the integration of transfer learning paradigms, allowing semisupervised models to leverage pretrained models or domain-specific embeddings, potentially enhancing accuracy across various categories.

It should be noted that we measured the training time required for both fully supervised and semisupervised methods. The fully supervised method required ~5.5 hr to complete the training. In contrast, the semisupervised methods took about 7.5 hr. The increased training time for the semisupervised methods can be attributed to the additional computational steps needed to effectively utilize the unlabeled data, which adds complexity and computational load to the training process.

The Yahoo Answers dataset, with its diverse range of questions and answers, allows us to explore new avenues for improving the educational experience. This study embarks on a pioneering journey that establishes the viability of integrating semisupervised methods into the broader landscape of educational content analysis and management. Our research aims to bridge the gap between a large volume of user-generated content and its practical application in university education management, ultimately contributing to a more effective and organized educational environment.

College student education and management involve the processing of a substantial amount of diverse data, including academic performance, behavioral data, and other pertinent information. The successful application of semisupervised methods highlights a pivotal insight—through the effective utilization of unlabeled data, institutions can enhance model performance and accuracy. This implies that universities can gain deeper insights into students' needs, trends, and behaviors, ultimately enabling them to better cater to these requirements. For instance, by collecting data from various sources such as student surveys, online content, and possibly additional data like academic records, extracurricular activities, and social interactions, universities can classify the subjects or topics that students are most interested in. This holistic approach can provide valuable insights into students' current priorities and concerns. Understanding the specific areas of interest can enable educational institutions and administrators to tailor their approaches, resources, and support systems accordingly. Moreover, by identifying common themes like academic challenges, mental health concerns, career prospects, or personal development, educators and administrators can direct their efforts more effectively. They can develop proactive strategies to address the prevalent issues or provide targeted support where it is needed the most.

Numerous universities face the challenge of limited labeled data, making the adoption of purely supervised methods impractical. Semisupervised methods shine in such situations, not only for their ability to enhance model performance but also for their proficiency in leveraging unlabeled data. This is of paramount significance for universities striving to improve student management in resource-constrained environments. Employing semisupervised learning on the Yahoo Answers dataset and additional data sources for college student education and management is a powerful tool for gaining insights into students' interests and concerns. By understanding their priorities, educators and administrators can optimize their efforts, ensuring a more focused and impactful approach to education and support. In addition, the superior performance of semisupervised methods has significant practical implications for student management systems. The scalability and cost-effectiveness derived from semisupervised methods allow educational institutions to implement more efficient and responsive student management systems, capable of quickly adapting to changing student needs and providing personalized support at a lower cost. Building upon our research, future investigations can delve further into the application of semisupervised methods in university student management. This could encompass optimizing models to adapt to various types of student data and exploring advanced techniques for maximizing the potential of unlabeled data.

5. Conclusions

This study demonstrated the feasibility of using semisupervised learning methods in educational content classification using the Yahoo Answers dataset. The proposal of integrating semisupervised learning methods promises to revolutionize the way we approach the categorization of the topics that are most relevant to students, offering a more efficient and scalable solution for improved student education and management. For both Yahoo_500 and Yahoo_2000 datasets, the semisupervised methods consistently outperformed the supervised approach.

FreeMatch was the leading framework for Yahoo Answers classification, attaining an impressive best evaluation accuracy of 0.6971, an average precision of 0.6948, an average recall of 0.6971, and an average F_1 score of 0.6950 in validation datasets. The transition from Yahoo_500 to Yahoo_2000, which involved a larger pool of labeled data, showed notable changes in the accuracy evaluation. FreeMatch maintained its top-performing position in several categories, including computers and Internet, consumer electronics, and business and finance, with impressive F_1 scores of ≥ 0.753. The family and relationships category showcased differences in model performance, while education and reference remained challenging for all models. The inclusion of more labeled data in Yahoo_2000 further corroborates the benefits of semisupervised learning, showcasing its resilience and superiority over supervised approaches, which experience diminishing returns as more labeled data are added. We anticipate that by categorizing the topics that are most relevant to students and thus gaining a comprehensive understanding of what students care about, universities can improve the overall quality of education and management, fostering student success and well-being through the creation of an environment that adapts to the evolving needs of its student body. However, there are potential limitations, such as the need to try more diverse datasets and the challenges posed by class imbalances. Future work could explore optimizing semisupervised models to address these limitations and further enhance their applicability.

Data Availability

The data presented in this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Major Philosophy and Social Sciences Project for Higher Education Institutions of Jiangsu Province, grant number 2020SJZDA005, and the Philosophy and Social Sciences Special Project for Higher Education Institutions of Jiangsu Province, grant number 2021SJB0901.

References

[1] D. R. Garcia, R. Barber, and A. Molnar, "Profiting from public education: education management organizations and student achievement," *Teachers College Record: The Voice of Scholarship in Education*, vol. 111, no. 5, pp. 1352–1379, 2009.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[3] P. Kotamraju and O. Blackman, "Meeting the 2020 American Graduation Initiative (AGI) goal of increasing postsecondary graduation rates and completions: a macro perspective of community college student educational attainment," *Community College Journal of Research and Practice*, vol. 35, no. 3, pp. 202–219, 2011.

[4] C. MacCann, G. J. Fogarty, and R. D. Roberts, "Strategies for success in education: time management is more important for part-time than full-time community college students," *Learning and Individual Differences*, vol. 22, no. 5, pp. 618–623, 2012.

[5] H. C. Riker and D. A. DeCoster, "The educational role in college student housing," *Journal of College & University Student Housing*, vol. 35, no. 2, Article ID 80, 2008.

[6] L. K. Trevino, "Moral reasoning and business ethics: implications for research, education, and management," *Journal of Business Ethics*, vol. 11, pp. 445–459, 1992.

[7] C. C. Ho, H. L. Lee, W. K. Lo, and K. F. A. Lui, "Developing a chatbot for college student programme advisement," in *2018 International Symposium on Educational Technology (ISET)*, pp. 52–56, IEEE, Osaka, Japan, August 2018.

[8] R. A. Henry, J. G. Weber, and D. Yarbrough, "Money management practices of college students," *College Student Journal*, vol. 35, no. 2, pp. 244–244, 2001.

[9] S. Peterson, "Community college student-parents: priorities for persistence," *Community College Journal of Research and Practice*, vol. 40, no. 5, pp. 370–384, 2016.

[10] E. J. Krumrei-Mancuso, F. B. Newton, E. Kim, and D. Wilcox, "Psychosocial factors predicting first-year college student success," *Journal of College Student Development*, vol. 54, no. 3, pp. 247–266, 2013.

[11] J. P. Bean and S. B. Eaton, "A psychological model of college student retention," *Reworking the Student Departure Puzzle*, vol. 1, no. 48–61, Article ID 12, 2000.

[12] S. A. Hardy, S. W. Francis, B. L. Zamboanga, S. Y. Kim, S. G. Anderson, and L. F. Forthun, "The roles of identity formation and moral identity in college student mental health, health-risk behaviors, and psychological well-being," *Journal of Clinical Psychology*, vol. 69, no. 4, pp. 364–382, 2013.

[13] V. F. Filak and K. M. Sheldon, "Student psychological need satisfaction and college teacher-course evaluations," *Educational Psychology*, vol. 23, no. 3, pp. 235–247, 2003.

[14] J. W. You, "The relationship among college students' psychological capital, learning empowerment, and engagement," *Learning and Individual Differences*, vol. 49, pp. 17–24, 2016.

[15] C.-C. D. C. Wang and C. Castañeda-Sound, "The role of generational status, self-esteem, academic self-efficacy, and perceived social support in college students' psychological well-being," *Journal of College Counseling*, vol. 11, no. 2, pp. 101–118, 2008.

[16] J. Ives and M. Castillo-Montoya, "First-generation college students as academic learners: a systematic review," *Review of Educational Research*, vol. 90, no. 2, pp. 139–178, 2020.

[17] W. Sun, L. Wu, Y. Lu, X. Zha, and M. Cui, "Research on the application of big data in the management and service of college students," *IOP Conference Series: Materials Science and Engineering*, vol. 806, Article ID 012056, 2020.

[18] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: automated machine learning approach," *Computers & Electrical Engineering*, vol. 89, Article ID 106903, 2021.

[19] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, 2019.

[20] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning

session data," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 381–407, 2019.
[21] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information Retrieval*, vol. 12, no. 5, pp. 526–558, 2009.
[22] A. Salinca, "Business reviews classification using sentiment analysis," in *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pp. 247–250, IEEE, Timisoara, Romania, September 2015.
[23] R. R. Halde, A. Deshpande, and A. Mahajan, "Psychology assisted prediction of academic performance using machine learning," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 431–435, IEEE, Bangalore, India, May 2016.
[24] X. Zhang, "Machine learning-based psychology evaluation of college students for building innovative health service system," *Journal of Environmental and Public Health*, vol. 2022, Article ID 6302756, 7 pages, 2022.
[25] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Computers in Human Behavior*, vol. 98, pp. 166–173, 2019.
[26] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: a machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, 2020.
[27] H. Luan and C.-C. Tsai, "A review of using machine learning approaches for precision education," *Educational Technology & Society*, vol. 24, no. 1, pp. 250–266, 2021.
[28] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, Article ID 106189, 2020.
[29] Y.-F. Li and D.-M. Liang, "Safe semi-supervised learning: a brief introduction," *Frontiers of Computer Science*, vol. 13, no. 4, pp. 669–676, 2019.
[30] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.
[31] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 22243–22255, Curran Associates Inc., Vancouver, BC, Canada, December 2020.
[32] Y. Chong, Y. Ding, Q. Yan, and S. Pan, "Graph-based semi-supervised learning: a review," *Neurocomputing*, vol. 408, pp. 216–230, 2020.
[33] A. Ahadh, G. V. Binish, and R. Srinivasan, "Text mining of accident reports using semi-supervised keyword extraction and topic modeling," *Process Safety and Environmental Protection*, vol. 155, pp. 455–465, 2021.
[34] E. Hagberg, D. Hagerman, R. Johansson et al., "Semi-supervised learning with natural language processing for right ventricle classification in echocardiography—a scalable approach," *Computers in Biology and Medicine*, vol. 143, Article ID 105282, 2022.
[35] F. Nie, L. Tian, R. Wang, and X. Li, "Multiview semi-supervised learning model for image classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 12, pp. 2389–2400, 2019.
[36] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3429–3440, 2020.
[37] B. Chen, W. Chen, S. Yang et al., "Label matching semi-supervised object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14381–14390, IEEE, New Orleans, LA, USA, June 2022.
[38] G. Li, X. Li, Y. Wang, Y. Wu, D. Liang, and S. Zhang, "Pseco: pseudo labeling and consistency training for semi-supervised object detection," in *European Conference on Computer Vision*, pp. 457–472, Springer, Cham, November 2022.
[39] R. Mendel, L. A. De Souza, D. Rauber, J. P. Papa, and C. Palm, "Semi-supervised segmentation based on error-correcting supervision," in *Computer Vision-ECCV 2020*, vol. 12374, pp. 141–157, Springer, Cham, Glasgow, UK, 2020.
[40] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622, IEEE, Nashville, TN, USA, June 2021.