International Journal of Religion
2024
Volume: 5| Number 11 | pp. 1683 – 1694
ISSN: 2633-352X (Print) | ISSN: 2633-3538 (Online)
ijor.co.uk
DOI: https://doi.org/10.61707/9hvjfb19

Classification and Labeling Techniques of Educational Resources Based on BERT+CNN's Educational Text Classification Model

Weizhi Zhang[1], Yu Lih Jiun[2] and Jonathan Yong Chung Ee[3]

Abstract

Modern, efficient techniques of identifying and regulating digital Educational Resources (ER) are urgently needed due to the rapid development of this area of study. Due to the level of detail and emphasis on information, conventional Text Classification (TC) techniques frequently experience problems when working with ER. To overcome these challenges, this study proposes a novel approach termed BERT-CNN, which integrates the effectiveness of BERT's knowledge of context with CNN's pattern identification skills in order to attain Deep Learning (DL). The objective of the hybrid approach is to improve educational Text Labelling (TL) and TC to be more precise, successful, and robust. The BERT-CNN model is superior to conventional approaches in various measures, including accuracy, precision, recall, F1 score, and AUC-ROC, based on thorough testing and comparison. Providing essential insights into the possibilities of integrating BERT and CNN for educational purposes, this research covers detailed methods from data collection to model development, testing, and the implementation of modern TC approaches.

Keywords: *Education, Deep Learning, BERT-CNN, Text Classification, Accuracy*

INTRODUCTION

The explosion of e-books, distance education programs, and online libraries has significantly enhanced the learning environment by increasing students' exposure to plenty of knowledge. The challenge of appropriately classifying, finding, and making use of large and heterogeneous Educational Resources (ER) continues notwithstanding these improvements. Due to its heterogeneous nature and significant format diversity, ER provides a distinctive set of problems for conventional text classification (TC) techniques. Since Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machines (SVM) constitute the most commonly used basic TC scenarios, they challenge maintaining current with the complicated demands of ER. Those frameworks frequently assume that data points are autonomous and equally shared, which is rarely possible in ER, and they require significant individual feature development. In addition, they do not do an adequate job of expressing how ER's definitions shift and evolve according to the context, suggesting that the relative significance of various concepts could change dramatically.

A hybrid model termed BERT-CNN (Bidirectional Encoder Representations from Transformers and Convolutional Neural Networks) has been used in this study to deal with the shortcomings of traditional TC systems. This novel approach combines CNN's Pattern Recognition (PR) abilities with BERT's environmental DL features. The proposed approach aims to change educational TC by using BERT's transformers. BERT has the ability to recognize complex language use since its analysis takes into account the complete context of words in the ER, encompassing both the left and right sides of each word.

In education settings, where words have distinct meanings based on the background material, this is important. The CNN layer will analyze the embedded data after BERT's contextual analysis in order to detect and classify TC successfully. CNNs' stability towards positional changes within text and their abilities to find patterns across large data sets make them highly useful. The process of integration is achieved by importing the historical embedded data generated by BERT into a CNN. The CNN then uses the gained knowledge features to TC.

[1] Faculty of Engineering, UCSI University, Jalan Puncak Menara Gading, 56000 Kuala Lumpur, Malaysia E-mail: iamweizhi1025@outlook.com

[2] Faculty of Engineering, UCSI University, Jalan Puncak Menara Gading, 56000 Kuala Lumpur, Malaysia

[3] Faculty of Engineering, UCSI University, Jalan Puncak Menara Gading, 56000 Kuala Lumpur, Malaysia

Applying this order ensures accurate classification of ER while also maintaining their grammatical complexity and morphological subtleties.

The Present Investigation Has Three Primary Goals

The goal is to develop a model that can understand the context-specific nature of ER.

For better ER administration and extraction, it is essential to boost the accuracy and efficacy of TC and TL.

In order to evaluate the proposed approach with traditional TC techniques and demonstrate that it works.

The research paper has been organized in the following order: the summary of the literature is given in part 2, the approach is shown in section 3, the results and analysis will be displayed in section 4, and the work is concluded in section 5.

LITERATURE REVIEW

Several new studies indicate that integrating BERT+CNN is a successful approach to improving TC problems. Considering their techniques, results, and implications for educational TC, this part analyzes significant contributions that use this hybrid method.

In order to find expressions in text, the contributors of built a BERT-CNN model. The method they use is convolutional neural networks (CNNs) for the prediction of output following continuous training of the lexical model of words using BERT and context. Testing on the removal 2019 task 3 and ISEAR datasets demonstrated that this approach beat state-of-the-art models, with 94.7% accuracy and 94% F1-score on the first instance and 75.8% accuracy and 76% F1-score on the latter. This study emphasizes the BERT-CNN model's endurance for intricate conceptual tasks, which will be beneficial to future research with ER.

In order to classify significant Chinese news texts, introduced an innovative Local Feature Convolutional Network (LFCN) powered by BERT. As an element of the research method, they employ a Dynamic LEAD-n technique to deal with BERT's input chain length boundaries. In addition, researchers use CNN for local FE and attention mechanisms for global FE. Long TC accuracy was significantly enhanced by the framework, showing that localized FE could enhance TC outcomes.

The application of a BERT-CNN model to perform the task of economic audit word pattern classification was the main focus of [3]. Their algorithm generated superior outcomes compared to the *state-of-the-art* initial models since it utilized the self-attention mechanism to describe the text accurately and integrated BERT-CNN to extract important local textual data. In light of our work with an extensive selection of ER, this study shows how well BERT-CNN processes texts with complicated structures and correlations.

Applying a BERT-CNN model, which includes both global and local semantic parameters, product evaluation sentiment analysis has been improved. The model they developed is superior to self-sufficient BERT-CNN models about F1 scores, demonstrating the complementary features of BERT's comprehensive context information and CNN's effective local PR.

For the aim of identifying words of abuse in a context such as numerous languages, used BERT-CNN. In addition to demonstrating that CNN+BERT is superior to BERT alone, the approach they used drew attention to the importance of using pre-trained language models for specific uses. This finding is significant to this research since it suggests that pre-trained models may be effectively changed to reflect certain ERs.

METHODOLOGY

Data Collection

For this BERT-CNN model study on ER TC and TL, the comprehensive data collection process is essential. They searched for an extensive selection of text-based ER from different Chinese universities for this research. The educational variation of such institutions, which reach from primary schools to colleges and universities, was an important factor in choosing them for this ER study.

Sources of Educational Texts

Textbooks and Academic Books: To collect an array of textbooks encompassing the arts, sciences, and humanities, we collaborated with five major Chinese institutions and ten primary and secondary schools. The model we developed benefits from the comprehensive formal ER encompassed in such textbooks, the physical activity and case scenarios, and the illustrations that follow them. To further capture more specific data and terminology, books on academia were also reviewed, particularly those used for additional and continuing professional development.

Scholarly Articles: The digital libraries associated with these higher education institutions were rendered freely accessible to us; thus, we were able to collect an extensive collection of academic journals. The complex dataset is improved by the expert language, references, and summaries of these articles, which cover a wide range of participants. The dataset has been enhanced with high quality and centered around research evidence by including subjects in peer-reviewed papers.

Online ER: Online ER for emergencies is increasing in significance as a result of the digitization of educational institutions, with the present tendency toward online learning being fuelled in significant measure by concern about world health. The collection we have contains open-access ER from institution-hosted sites, multimedia instructional modules, downloadable educational PDFs, and e-texts. These ERs provide a dynamic dataset that is challenging and helpful for TC, with multimedia associations, integrated tests, and hyperlinked citations.

Government and Institutional Reports: Also featured were pedagogical studies and suggestions published by the Chinese Ministry of Education and additional HEs. In order to train the model we provide to understand more ordered and formal language, these papers serve as essential for understanding the regulatory and administration language used in ER.

The initial stage of the data collecting included obtaining electronic versions of textbooks and research papers; the second stage included collecting web resources with appropriate permissions and taking ethical factors into consideration. For publications that weren't available electronically in type, we used Optical Character Recognition (OCR) technology to make them machine-understandable. Therefore, the data we have contains all types of text-based ER, and it is finished. Data such as reference type, level of education, correction, and date of publication have been entered for every text source in the database. It is feasible to analyze the text's importance in context with the use of this data, which additionally helps with classification accuracy.

Dataset Description

A class defining what type of ER has been allocated to each object in this dataset. This makes it possible to conduct extensive evaluations and enables the BERT-CNN model to be trained quickly. The Text ID is an identification number that is associated with the document characteristics and is employed to make sure that each text is unique in the dataset. The document's content is easily determined by the Title, and its type and primary value in ER is indicated by the Source Type, which may be books, articles, internet-based materials, or reports from government agencies, among others. The document's whole content is saved in the Content resources, while the Word Count maintains track of the number of words it encompasses.

Each text in the dataset has been identified with an exact Subject Area, which specifies the field of study to which it relates, for example, Math, History, or Engineering. This enables to fill up the dataset. From preschool all through high school up to university, literature also indicates the level of education it targets. To put the text in viewpoint, readers may look at the Publication Date, which tells when it was initially released or recently changed, and the Author(s) resources, which provide the names of the individuals or institutions who authored the material, is also helpful. Several multilingual texts are additionally included in the dataset, while a great deal of texts in Chinese can be found in the Chinese Language field.

Textbooks, Research papers, Online Resources, and Government Reports are some of the classifications into which the texts in the dataset are successfully classified by the TC system, which makes use of the Source Type. The main objective of textbooks is to act as educational texts for students of every level of education. Articles in journals that have passed peer review and have made significant improvements to their subjects are known as research papers. Digital ER, such as e-books and several additional education courses, are called online ER and may be downloaded via academic portals. Government reports, the last type, are governmental publications that outline academic rules and procedures; they are the go-to places for knowledge about educational standards and behaviors. Both the training method and the accuracy of the model in TC and TL educational TC according to topic and value are enhanced by this organized technique for data classification. The dataset summary is presented in Table 1.

Table 1: Summary of Dataset

Text ID	Title	Source Type	Word Count	Subject Area	Educational Level	Publication Date	Author(s)	Language
TX837	Advanced Mathematics	Textbook	34,823	Mathematics	University	2020	Li Wei	Mandarin
TX462	Innovations in Biology	Scholarly Article	4,987	Biology	University	2022	Zhang Hua, Liu Yong	Mandarin
TX159	Learning English Online	Online Resource	15,312	English	Secondary	2021	Online English Academy	Bilingual
TX301	Education Policy 2023	Government Report	20,154	Education	General	2023	Ministry of Education	Mandarin

Preprocessing

In order to obtain the raw ER suitable for accurate modeling, the phase of preprocessing is essential. In order for the BERT-CNN model to properly analyze unorganized text, this step encodes it into a structured format. We verified that the data is perfect for DL tasks through the use of several methodical steps: tokenization, normalization, and data augmentation.

Tokenization: The first step in the text setup process is tokenization. Tokens are used to divide the text into smaller, easier-to-comprehend bits. For this model, use the BERT tokenizer, meant as $Tokenizer_{Bert}$. Not only does this tokenizer divide text into words and sub-words, but it also integrates training-friendly unique tokens.

$$T = Tokenizer_{Bert}(D)$$

where D represents the raw text document, and T denotes the sequence of tokens obtained. Tokenization helps retain the text's contextual definition, which is essential for processing the BERT model.

Normalization: Every token is normalized after the process of tokenization. At this phase, we unify the tokens' case, which is frequently lowercase, delete any punctuation, and correct common spelling or grammar errors. Enhancing the accuracy of training and minimizing model complexity relies on the normalizing procedure, which can be expressed by the function Normalize (t). This approach ensures coherence in the dataset. The normalized token set T' is obtained as follows:

$$T' = \text{Normalize}\ (T)$$

Normalization also includes the elimination of 'stop words, common occurring words in a language that suggest minimal value in understanding the text's context, such as "and", "the", "is", etc.

Data Augmentation: To enhance the robustness of our model and prevent overfitting, we applied data augmentation techniques to the preprocessed text. We employed synonym replacement and random insertion, which involve replacing a word with one of its synonyms and inserting relevant synonyms into random positions within the text. Notationally, for a token t_i in T', the augmentation function Augment (t_i) modifies T' as follows:

$$T'' = \text{Augment}\ (T')$$

where T'' represents the augmented token set. This method enriches the linguistic diversity of our training data, allowing the model to learn more generalizable patterns.

Table 2 presents the preprocessing steps for each attribute.

Table 2: Dataset description along with its Preprocessing

Attribute	Description	Preprocessing Applied
Text ID	Unique identifier for each text document in the dataset.	None
Title	Title of the document.	Tokenization, Normalization (Case Conversion, Punctuation Removal)
Source Type	The type of source (*e.g.,* Textbook, Article, Online Resource, Government Report).	None
Content	The full-text content of the document.	Tokenization, Normalization (Case Conversion, Stop Word Removal, Spelling Correction), Data Augmentation (Synonym Replacement, Random Insertion)
Word Count	The number of words in the document before preprocessing.	Update count after tokenization and augmentation
Subject Area	The academic discipline or subject area the text is associated with.	Normalization (Standardizing Terminology)
Educational Level	The educational level targeted by the text (*e.g.,* Primary, Secondary, University).	None
Publication Date	The publication or last updated date of the text.	None
Author(s)	Name(s) of the author(s) or the institution responsible for the publication.	Normalization (Standardizing Name Formats)
Language	The language of the text is primarily Mandarin, with some bilingual texts.	None

BERT-CNN Classification Model

Figure 1 presents the BERT-CNN architecture used in this work for Educational TC; this section details the architecture.

BERT Layer

The BERT layer is a pivotal component of our model, capitalizing on its capability to understand the nuanced context of language from educational texts. Leveraged from its pre-trained model, BERT is adept at generating deep contextual embeddings that significantly enrich the text representation before it undergoes classification.

At the core of BERT's functionality is its architecture based on the transformer mechanism, specifically designed to handle sequences of data. In our model, the BERT layer processes the tokenized and normalized text data, transforming each token into a vector that encapsulates both its own information and the context provided by the tokens around it. This transformation is mathematically denoted as follows:

$$E = \text{BERT}\ (T'')$$

where T'' Represents the preprocessed and tokenized text input, and E denotes the resulting set of embeddings. With *'d'* being the size of the embedded data in the BERT approach, each element of E is a d-dimensional vector. A key component of BERT's knowledge of context is its attention mechanism, which allows the model to provide different values to the impact of each word inside an identified phrase or document. Since the meaning of specific terms may change significantly based on their context, this is particularly important in ER.

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V stand for the query, key, and value matrices derived from the embeddings. d_k represents the dimensionality of the keys and queries, which is crucial for scaling the dot products in the attention mechanism. In order to get a convex tandem, the SoftMax function verifies that all of the weights add up to one. After the

BERT layer's preparation, the educational texts' embedded data E collects a profound, historical understanding of every word. Subsequent CNN layers receive these embedded data. Employing the FE from these embedded data, the CNN layers are created to do more analysis and classification of the textual material. Fundamental to the TC performance of our hybrid BERT-CNN model is the innovative sequencing of contextual embeddings from BERT to CNN's FE.

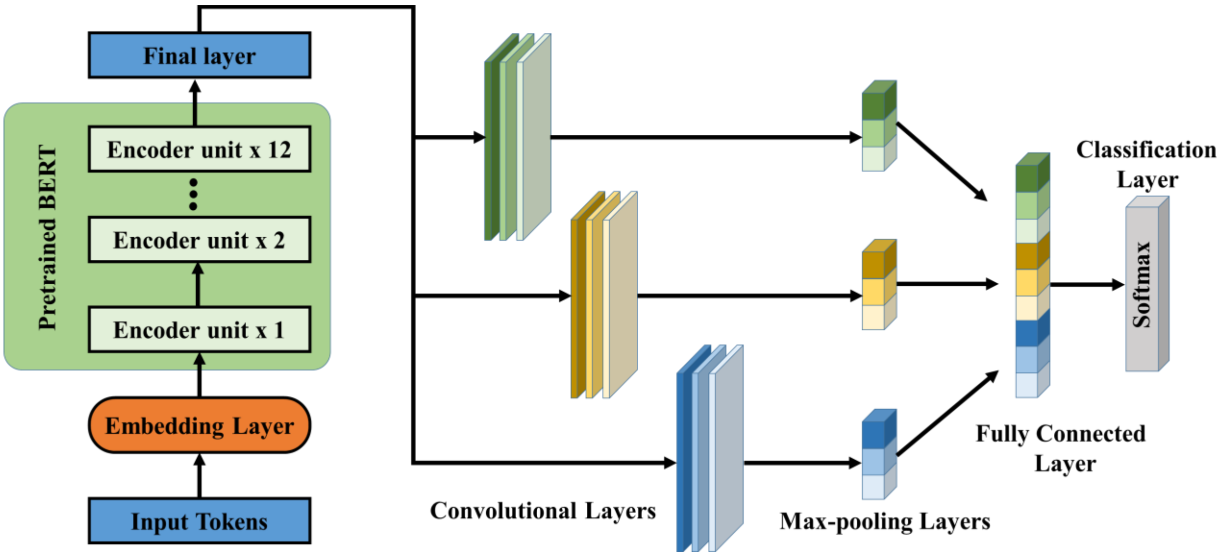

Figure 1: BERT-CNN Architecture

CNN Layer

The instructor pedagogical TC model's FE and classification algorithms depend significantly on the embedded contexts provided by the BERT+CNN layer. The primary function of the CNN layer is to apply convolution operations that detect patterns and key features across the embedded text sequences. These operations are instrumental in capturing local feature correlations in a way that is invariant to their position in the text, which is particularly valuable for recognizing linguistic patterns essential for TC.

The convolution operation in the CNN layer can be expressed mathematically as follows:

$$F = \text{ReLU}\,(E * K + b)$$

where E denotes the matrix of embeddings received from the BERT layer, K represents the kernel or filter matrix that slides over the embeddings, $*$ denotes the convolution operation, b is the bias term, and ReLU is the Rectified Linear Unit activation function that introduces non-linearity into the process. The result F comprises feature maps that encapsulate higher-level textual features derived from the embeddings.

In this context, the kernels are trained to detect specific types of features in the text, such as edges of semantic or syntactic patterns, by adjusting their weights through the learning process. Multiple kernels are typically employed, each responsible for learning to recognize different aspects of the text data, thus producing multiple feature maps. The dimensions of these kernels and the stride of the convolution determine the granularity of the analysis and the size of the output feature maps.

Once the convolutional layers have extracted these features, pooling layers are often applied to reduce the dimensionality of the feature maps, which helps reduce computational complexity and control overfitting. The most common form of pooling in text processing CNNs is max pooling, which can be described by:

$$P = \max(F)$$

Here, P represents the pooled feature map, which takes the maximum value of each region of the feature map F, effectively downsampling the information while preserving the most salient features.

The combination of multiple convolution and pooling layers forms a DL network that can extract and synthesize features at various levels of abstraction. After passing through these layers, the resultant feature maps are typically flattened into a 1-D vector and passed through one or more fully connected layers to perform the final TC. The output layer uses a SoftMax function, which calculates the probability distribution over the predefined classes:

$$\text{Output } = \text{SoftMax}\,(W \cdot P + b)$$

where W represents the weights of the fully connected layer, and b is the bias. This SoftMax output provides the probabilities for each class, facilitating the categorization of the text into one of the classes defined in our model.

Training Process

The training of the BERT-CNN model commences with the pre-processed data, which has already been tokenized, normalized, and augmented. The model is initially set with pre-trained weights from BERT, leveraging its robust foundation trained on vast amounts of text data. For the CNN layers, we initialize weights using the initialization method, which is known for its efficacy in networks with ReLu activation.

We employ the Adam optimizer, a stochastic gradient descent method that computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. The formula for the Adam optimizer can be expressed as follows:

$$\theta_{t+1} = \theta_t - \frac{\eta \cdot \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

where θ represents the parameters of the model, η is the learning rate, $\hat{m}_t$ is the bias-corrected first-moment estimate, $\hat{v}_t$ is the bias-corrected second-moment estimate, and ϵ is a small scalar added to improve numerical stability.

For the loss function, we use categorical cross-entropy, which is particularly suitable for multi-class classification tasks like ours. This loss function measures the performance of a classification model whose output is a probability value between 0 and 1 . Categorical cross-entropy is calculated as:

$$L = -\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right)$$

where M is the number of classes, y is a binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability that observation o is of class c.

To validate the model, we employ a cross-validation technique, specifically k-fold cross-validation. This method involves dividing the entire dataset into k smaller sets or folds. The training process is repeated k times, with each of the k folds used exactly once as the validation data, and the remaining $k-1$ folds are used as training data. This technique not only helps in assessing how the results of a statistical analysis will generalize to an independent data set but also mitigates the risk of model overfitting.

Labeling Technique

Our labeling system operates on the foundation of the TC results produced by the BERT-CNN model. Once a text is classified into one of the predefined categories (Textbook, Scholarly Article, Online Resource, Government Report), the system then applies a secondary layer of analysis to determine the specific labels that describe the content of the text. This secondary analysis leverages a set of predefined rules and keywords associated with each class to assign relevant labels. The TL examination is implemented in a step-by-step manner, as shown in Table 2.

Classification Output: The BERT-CNN considers first classifies all documents into one of the main classifications. In order to determine the most logical document type, this TC analyzes the context-sensitive embedded data generated by BERT using the CNN layers.

Keyword Extraction: A set of phrases and terms frequently employed for defining objects occurring within a particular class currently available. The algorithm extracts these important words from the text using algorithms for keyword extraction and Named Entity Recognition (NER), both of which are Natural Language Processing (NLP) methods. The computer programs have been optimized to find and rank words that frequently appear in educational environments.

Rule-Based Labeling: A rule-based technique is employed to create labels relying on the extracted keywords and the document's principal topic. Educational and pedagogical library classification requirements provide motivation for the standards that these professionals have developed. A "Textbook" tag could feature content-specific keywords such as "The study of math," "Grade 10," or "Higher algebra," for example.

Validation and Feedback: As a part of the process of verification, instructors manually examine a portion of the labels on the labeled articles to make sure they are accurate and appropriate. By continuously improving the keyword lists and standards based on the input of this verification step, the TL method becomes more accurate throughout the years.

Continuous Learning: The framework features an approach for continuing education to adapt to changing ER and semantics. This method maintains the labeling current and applicable when standards and ER evolve by updating the keyword extraction technologies and rule sets with new data and input from users.

Table 3: Overview of TL technique steps

Step	Description	Tools/Techniques Used	Outputs
Classification Output	Documents are classified into categories using the BERT-CNN model.	BERT-CNN	Category for each document
Keyword Extraction	Keywords are extracted based on the category.	NER, Keyword Extraction Algorithms	List of keywords
Rule-Based Labeling	Labels are assigned based on extracted keywords.	Predefined Rules	Assigned labels
Validation and Feedback	Sample labels are reviewed to ensure accuracy.	Manual Review	Feedback on label accuracy
Continuous Learning	The system updates rules and keywords based on new data and feedback.	Machine Learning Updates, User Feedback	Updated rules and keywords

MODEL EVALUATION

Performance Metrics

The metrics listed below will be used in order to assess both the reliability and accuracy of the BERT-CNN model:

Accuracy: The total accuracy of the model is determined by this statistic, which is the proportion of correctly predicted data to the actual information.

Precision: The precision of a model for forecasting is expressed as the percentage of predicted positive observations that really occurred. To the greatest extent possible, it helps when the costs related to False Positives (FP) are significant.

Recall (Sensitivity): This value evaluates the proportion of True Positives (TP) compared to the entire quantity of predicted positives. When the cost of an FP is higher than the cost of missing a TP, this is the highest priority.

F1 Score: The F1 score is the weighted average of Precision and Recall. Therefore, this score takes both False Positives (FP) and False Negatives (FN) into account. It is a good measure to use if you seek a balance between Precision and Recall.

Area Under the Curve (AUC) - ROC Curve: The AUC-ROC curve is a performance measurement for classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

To benchmark the performance of the BERT-CNN model, comparisons will be made against several well-established models in TC. These baseline models include LR, SVM, RF, NB, and Simple CNN.

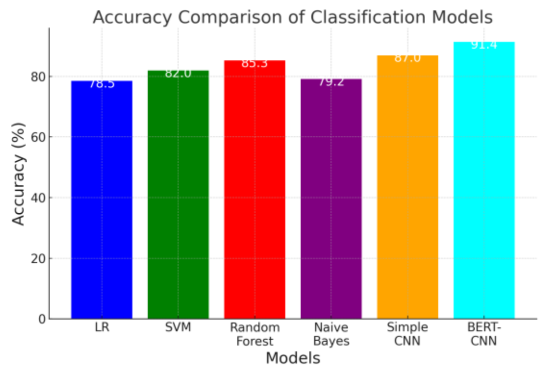

Figure 2: Accuracy Comparison of TC Models

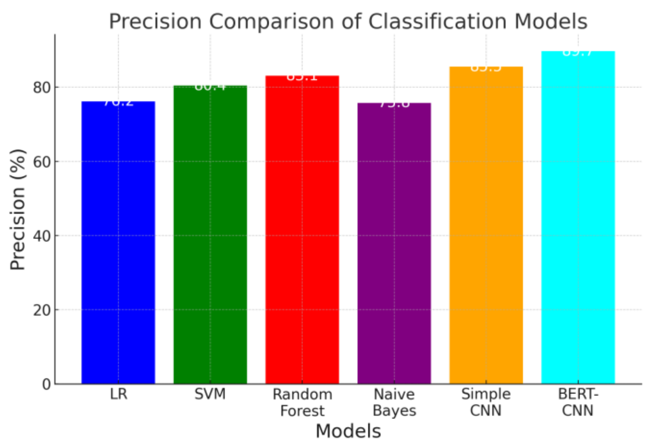

Figure 3: Precision Comparison of TC Models

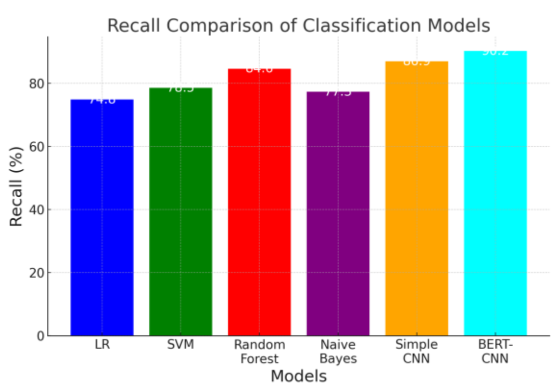

Figure 4: Recall Comparison of TC Models

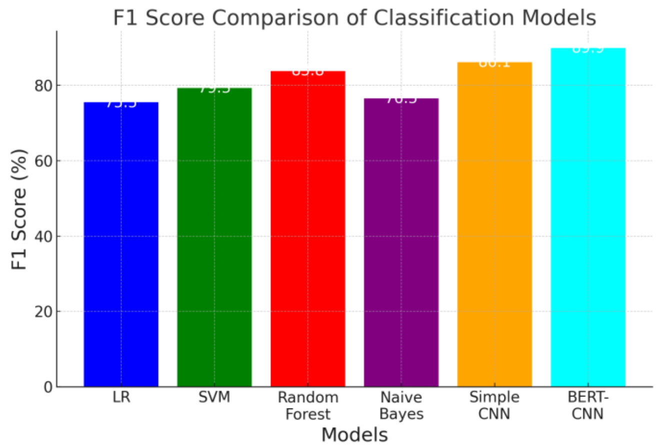

Figure 5: F1 Score Comparison of TC Models

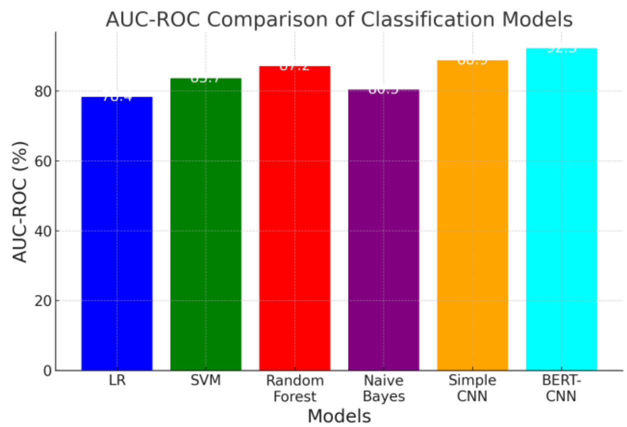

Figure 6: AUC-ROC Comparison of TC Models

Figure 2 reveals varied performance in terms of accuracy across several DL models in educational TC. LR and NB exhibit accuracies of 78.5% and 79.2%, respectively, struggling with the complexity of text data. SVM and RF show better accuracies of 82.0% and 85.3%, thanks to their capacity to manage high-dimensional data and ensemble methods that minimize overfitting. The Simple CNN model boosts performance to 87.0% by capturing local and contextual text features. At the top, the BERT-CNN model achieves the highest accuracy at 91.4%, combining BERT's contextual insights with CNN's capabilities, proving highly effective for nuanced educational TC. Precision scores, as shown in Figure 3, also vary, with LR and NB at the lower end at 76.2% and 75.8%. Improvements are noted in SVM and RF, which scored 80.4% and 83.1%, and better handling of text data complexities. The Simple CNN increases precision to 85.5%, effectively leveraging convolutional layers to capture essential textual features. The BERT-CNN model tops with a precision of 89.7%.

Recall rates displayed in Figure 4 are crucial for applications where missing a positive instance is critical—LR and NB record lower recall rates of 74.8% and 77.3%. SVM slightly improves to 78.5%, with RF at 84.6% and Simple CNN at 86.9%, enhancing their ability to capture a broader range of positive instances. The BERT-CNN model leads with a recall of 90.2%, minimizing false negatives effectively. The F1-score comparison of TC models in Figure 5 assesses the balance between precision and recall.

LR and NB show room for improvement, with F1 scores of 75.5% and 76.5%. SVM's F1 score at 79.3% and RF's at 83.8% indicate strong performance, with Simple CNN at 86.1%. The BERT-CNN model excels with an F1-score of 89.9%, optimizing precision and recall. The AUC-ROC comparison of TC models, as shown in Figure 6, measures each model's ability to differentiate between classes. LR shows moderate differentiation at 78.4%, with SVM better at 83.7% and RF strong at 87.2%. A good differentiated class is demonstrated by Basic

CNN at 88.9% and NB at 80.5%. The BERT-CNN model was superior to every other model, with an AUC-ROC of 92.3%.

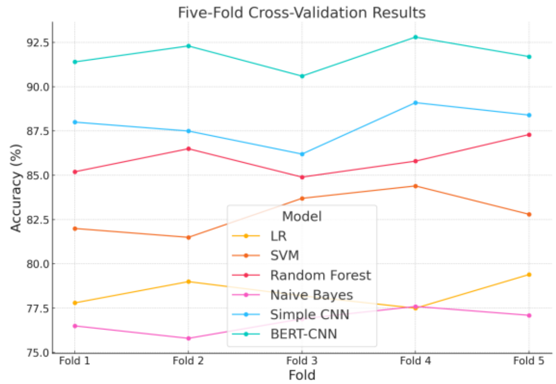

Figure 7: Five-fold cross-validation

Figure 7 presents the outcomes of five-fold cross-validation, demonstrating what distinct TC models do on distinct portions of the data. The models' reliability and stability are assessed with each fold using various sections. LR's accuracy levels are highly variable, averaging 78.4% and ranging from 77.5% to 79.4%. The model's stability and bound effectiveness are demonstrated by its maxima in Fold 5 and falls in Fold 4, which may be identified due to its ease in managing complex data exchanges. With SVM, the accuracy levels are higher, averaging 82.9% and ranging from 81.5% to 84.4%. It benefits from working with relationships that are not linear, as seen by its highest performance in Fold 4. This model's stable performance across folds suggests that it can deal with various types of data with simplicity.

RF indicates an accuracy range of 85.9% on average, with a performance range of 84.9% to 87.3%. The model's good performance in Fold 5 illustrates that it is able to successfully minimize bias and deviation via ensemble learning, providing it an excellent fit for different datasets. The average performance observed by NB is 76.8%, with a range of 75.8% to 77.6%. The model's accuracy falls short of complex models, which is indicative of its restrictions imposed by the presumptions of feature independence. Despite this, the model's accuracy usually remains constant. The accuracy of basic CNN varies between 86.2% to 89.1%, averaging 87.8%. The algorithm's superior performance in Fold 4 highlights its ability to extract contextual and local features from text, a key component toward obtaining precise TC. With a median performance of 91.8% and an interval of 90.6% to 92.8%, BERT-CNN is in first place. It is highly reliable and precise for educational TC because of its highest level in Fold 4 while maintaining high performance across all folds, which indicates its superior ability to combine deep environmental awareness with effective FE.

CONCLUSION AND FUTURE WORK

The research demonstrates that the BERT-CNN model performs well for educational Text Classification (TC) and Text Labelling (TL). The model generates far superior outcomes than traditional TC methods by integrating BERT's contextual data processing abilities with CNN's location feature recognition abilities. The key outcomes are probably enhanced abilities to cope with a broad range of educational resources (ED), such as internet-based resources and textbooks, and greater accuracy when recognizing complex language. Not only can the BERT-CNN model enhance the scientific basis for ER management, but it also provides an adaptable method that can be used in other fields that require accurate TC and TL.

Further work will be done to enhance the framework so that it may be used with greater efficiency across several languages and educational methods. A further option is to evaluate this idea in real classrooms to observe how it impacts students' ability to use digital tools and how effectively they engage with their learning. Improving ER collection, availability, and application, this investigation provides the foundation for HE systems that are more intelligent, effective, and adaptive.

REFERENCES

Wang, C., & Si, L. (2024). The Intersection of Public Policy and Public Access: Digital Inclusion, Digital Literacy Education, and Libraries. Sustainability, 16(5), 1878.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys, 55(13s), 1-40.

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer, 38(8), 2939-2970.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. Sustainable operations and computers, 3, 238-248.

Alantari, H. J., Currim, I. S., Deng, Y., & Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. International Journal of Research in Marketing, 39(1), 1-19.

Abas, A. R., Elhenawy, I., Zidan, M., & Othman, M. (2022). BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. Computers, Materials & Continua, 71(2).

Chen, X., Cong, P., & Lv, S. (2022). A long-text classification method of Chinese news based on BERT and CNN. IEEE Access, 10, 34046-34057.

Wan, C. X., & Li, B. (2022). Financial causal sentence recognition based on BERT-CNN text classification. The Journal of Supercomputing, 1-25.

Dong, J., He, F., Guo, Y., & Zhang, H. (2020, May). A commodity review sentiment analysis based on BERT-CNN model. In 2020 5th International Conference on Computer and Communication Systems (ICCCS) (pp. 143-147). IEEE.

Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. arXiv preprint arXiv:2007.13184.