Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 20, 2025

ENHANCING NON-FORMAL LEARNING CERTIFICATE CLASSIFICATION WITH TEXT AUGMENTATION: A COMPARISON OF CHARACTER, TOKEN, AND SEMANTIC APPROACHES

I Gede Susrama Mas Diyasa*	School of Magister Information Technology, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	igsusrama.if@upnjatim.ac.id
Eva Yulia Puspaningrum	School of Informatics, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	evapuspaningrum.if@upnjatim.ac.id
Dimas Saputra	School of Informatics, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	21081010151@student.upnjatim.ac.id
Wan Suryani Wan Awang	School of Informatics and Computing, Universiti Sultan Zainal Abidin Besut Campus, Besut, Malaysia	suryani@unisza.edu.my

* Corresponding author

ABSTRACT

Aim/Purpose — The purpose of this paper is to address the gap in the recognition of prior learning (RPL) by automating the classification of non-formal learning certificates using deep learning techniques. This study aims to evaluate the effectiveness of different text augmentation strategies—character-level, token-level, and semantic-level—in improving the classification accuracy of these certificates, which are crucial for bridging the skills gap in the digital economy.

Accepting Editor Dirk Frosch-Wilke | Received: March 21, 2025 | Revised: May 6, May 17, 2025 | Accepted: May 19, 2025.
Cite as: Mas Diyasa, I. G. S., Puspaningrum, E. Y., Saputra, D., & Wan Awang, W. S. (2025). Enhancing non-formal learning certificate classification with text augmentation: A comparison of character, token, and semantic approaches. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 18.
https://doi.org/10.28945/5525

Background	Traditional education systems often overlook skills gained through non-formal learning, creating a gap between industry needs and academic qualifications. This paper addresses this by using BERT-based deep learning models to classify non-formal learning certificates, enhanced by text augmentation techniques to improve accuracy in mapping them to formal academic standards.
Methodology	This study employs a deep learning approach using Bidirectional Encoder Representations from Transformers (BERT) to classify non-formal learning certificates into seven core computer science courses. The research utilizes text augmentation techniques at character, token, and semantic levels to improve classification accuracy. A dataset of 525 certificates, collected through data gathering, was preprocessed using Optical Character Recognition (OCR) to extract text from PDF documents, followed by cleaning and augmentation before training the BERT model.
Contribution	This paper addresses the growing need for efficient Recognition of Prior Learning (RPL) in the context of rapidly advancing knowledge, particularly in the AI era, where non-formal learning is becoming increasingly important. We present a novel approach to automating the classification and validation of non-formal learning certificates using deep learning techniques. The study evaluates and compares character-level, token-level, and semantic-level text augmentation methods to improve the accuracy of certificate classification. What sets this research apart is the systematic assessment of which augmentation method best enhances model performance for RPL tasks, providing new insights into optimizing deep learning models for this purpose. The findings aim to reduce human error and improve the efficiency of RPL implementation, offering a scalable solution for better integrating or converting non-formal learning into formal educational systems.
Findings	The study found that token-level augmentations, particularly word insertion and word deletion, significantly improved classification accuracy, with validation accuracies exceeding 88%. Character-level augmentations also contributed to model performance, but with slightly lower accuracy. Semantic-level augmentation via back translation showed the least impact. These results demonstrate that token-level text augmentations offer the most effective strategy for enhancing the classification of non-formal learning certificates in the context of Recognition of Prior Learning (RPL).
Recommendations for Practitioners	Practitioners should focus on token-level text augmentation techniques, like word insertion and deletion, to improve the accuracy of machine learning models for classifying non-formal learning certificates, enabling better integration into formal education and employment pathways.
Recommendations for Researchers	Researchers should explore combining multiple augmentation techniques (e.g., token-level and semantic-level) and investigate advanced models like BERT-large or multilingual variants for improved classification accuracy. Additionally, examining the impact of different OCR tools and preprocessing strategies could further enhance non-formal learning certificate recognition.
Impact on Society	The findings of this study have significant implications for improving access to education and employment opportunities. By enhancing the recognition of prior learning through automated classification of non-formal learning certificates, this research supports a more inclusive and equitable education system. It

can help individuals, particularly those with non-traditional educational backgrounds, gain recognition for their skills, ultimately bridging the skills gap in the workforce and promoting lifelong learning in the digital economy.

Future Research — Future research should focus on expanding the dataset to include multilingual certificates, which would enhance the model's ability to generalize across different languages and cultural contexts. Additionally, researchers could investigate the use of hybrid models that combine BERT with other machine learning techniques to further improve classification accuracy. Exploring the integration of real-world data sources, such as employer-verified work experience and additional non-formal learning formats, could also provide a more comprehensive approach to recognizing prior learning.

Keywords — document classification, text augmentation, recognition of past learning (RPL), BERT, non-formal learning

INTRODUCTION

In today's rapidly evolving digital landscape, the demand for skills in emerging fields, especially in AI, has outpaced traditional educational systems. Recognition of Prior Learning (RPL) offers a solution to bridge the gap between formal education and industry needs, particularly in recognizing non-formal learning experiences. However, accurately classifying non-formal learning certificates remains a significant challenge (Domvo, 2022), especially as the volume of these certificates escalates exponentially.

This study aims to investigate how different text augmentation techniques can improve the accuracy and efficiency of classifying non-formal learning certificates using deep learning models. The research questions guiding this study is: Which text augmentation method—character-level, token-level, or semantic-level—most effectively improves document classification accuracy? The main objective is to identify the most effective augmentation approach to automate certificate classification, facilitating more efficient and scalable implementation of RPL in educational institutions.

Recognition of Prior Learning (RPL) involves evaluating and acknowledging skills and knowledge gained outside formal educational settings (Winstanley & Cunningham, 2023), such as through online courses, certifications, or on-the-job training. Traditional education systems tend to overlook Recognition of Prior Learning (RPL) because they prioritize formal qualifications over skills and knowledge gained through non-formal learning experiences. However, traditional educational systems often fail to keep pace with the rapid advancements in knowledge, especially in fields like AI, where industry needs evolve faster than traditional curricula can adapt.

This research focuses on classifying certificates from non-formal learning experiences into seven core courses within the field of computer science, namely Machine Learning, Web Programming, Interface Design, Computer Networks, Game Applications, Mobile Programming, and Project Management. By conducting this classification, we can determine the equivalencies between industry-acquired skills and formal academic curricula, facilitating a more effective recognition process.

To address the complexities of classifying unstructured certificate data, this study employs Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), which overcomes critical limitations of traditional machine learning (ML) models like Support Vector Machines (SVM) (Zamsuri et al., 2023) and K-Nearest Neighbors (KNN) (Sharma et al., 2020). While SVM/KNN rely on static features (e.g., TF-IDF) that fail to capture context or semantic nuances, BERT leverages contextual embeddings to dynamically understand the meaning and relationships within the data (Z. Guo & Nguyen, 2020). BERT's bidirectional architecture dynamically models word dependencies,

enabling reliable classification of ambiguous or noisy text (e.g., OCR errors that provide noise), Traditional ML models, however, struggle with such noise due to their reliance on rigid, context-agnostic features like TF-IDF (Chen et al., 2023).

Moreover, traditional ML models require large-labeled datasets to generalize effectively. In contrast, BERT leverages pre-trained knowledge from vast corpora and enabling strong performance with minimal labeled data examples (Lu et al., 2021). This efficiency, combined with synthetic data augmentation, enables BERT to achieve powerful performance even when training data is limited, a critical advantage for real-world certificate classification tasks.

Moreover, a critical challenge in document-based classification tasks is the availability of sufficient labeled data for training models. In response, this study incorporates synthetic data augmentation techniques to artificially expand the training dataset, enhancing the model's ability to generalize and handle diverse inputs. These techniques include character-level, token-level, and semantic-level augmentations. Previous research, such as that by Şahin (2022), has shown that character-level augmentation outperforms other techniques in certain tasks, particularly when linguistic nuances need to be preserved. However, their research focused on tasks like Part-of-Speech (POS) tagging and Dependency Parsing (DEP), while document classification tasks based on NLP especially those involving certificates remain largely unexplored.

This study aims to fill this gap by comparing the effectiveness of character-level, token-level, and semantic-level augmentations specifically for document NLP based classification tasks. By evaluating these techniques, we seek to determine which approach best enhances the accuracy of models trained on datasets derived from non-formal learning certificates. The findings will contribute to a deeper understanding of how text augmentation can optimize document-level classification, with significant implications for recognizing prior learning and improving educational outcomes in the digital age.

RELATED WORKS

Previous studies, like a Naive Bayes-based system (Zhang et al., 2024), automate the classification of educational reform documents but are limited to form-based documents and Chinese language processing. While such studies provide valuable insights into the classification of educational reform documents, they focus on structured formats and specific languages, unlike our research which targets the classification of unstructured educational certificates in English. Moreover, unlike these studies using traditional ML models, our research leverages BERT, a large language model (LLM), to classify unstructured English educational certificates.

Other previous studies, such as the comparison of text-image fusion models for high school diploma certificate classification (Atmaja Perdana et al., 2020), focus on image-based classification using CNNs and OCR, mainly addressing structured certificates. In contrast, our research advances the field by using BERT, a text-based model, to classify unstructured non-formal educational certificates, overcoming the limitations of traditional image-based approaches and addressing a broader range of document types.

DOCUMENT CLASSIFICATION

Document classification, as described by Gopi et al. (2022), involves organizing text documents into predefined categories based on their content. This process typically uses supervised learning, where a model is trained on a labeled dataset consisting of documents and their corresponding category labels (Ibrahem Al-Obaydy et al., 2022). The goal is to develop a model capable of automatically categorizing new, unseen documents into the appropriate categories. According to Labd et al. (2024), document classification has broad applications, including spam filtering, sentiment analysis, topic categorization, and information retrieval.

Various machine learning algorithms, such as K-nearest neighbors (Zamsuri et al., 2023), Support Vector Machines (SVM) (Sharma et al., 2020), and deep learning models, are commonly used for these tasks, as discussed by Jiang et al. (2024). The choice of algorithm depends on factors like data type and task complexity, with deep learning models, particularly Bidirectional Encoder Representations from Transformers (BERT), becoming more popular due to their ability to capture contextual relationships within text (Röttger & Pierrehumbert, 2021).

Image-based document classification often uses multimodal approaches, combining visual features and text embeddings from noisy OCR outputs using tools like FastText to improve accuracy, as demonstrated by (Audebert et al., 2020), but these methods face challenges with dataset quality and fail to capture deep contextual relationships from the machine learning model used, making them less effective for text-heavy tasks compared to BERT-based models. While Röttger and Pierrehumbert (2021) demonstrate the effectiveness of BERT in temporal adaptation for text-based document classification using the Reddit Time Corpus, their work does not address image-based document classification, which introduces additional challenges such as OCR error noise that are central to this study.

In our research, we intentionally did not remove OCR-generated noise, instead focusing on observing the impact of naturally occurring noise in the OCR output. This approach allows us to explore how machine learning models, particularly BERT, handle noisy, unstructured datasets without preprocessing or denoising steps.

Text Augmentation

Text augmentation, as described by Shorten et al. (2021), is the process of modifying textual data to create new, synthetic examples while preserving the original meaning. This technique is widely employed to expand both the diversity and quantity of data available for machine learning models, thereby enhancing their performance. By incorporating subtle alterations into the text, augmentation simulates real-world variations, ultimately bolstering model accuracy and enabling them to handle a broader range of inputs (Shorten et al., 2021).

In text augmentation, there are several levels at which modifications can occur, ranging from character-level to word/token-level, and even sentence-level. Each level involves different techniques, with character-level augmentations typically focusing on changes like character insertion or deletion, while word/token-level augmentations modify the text by adding or removing words. Sentence-level augmentation, as discussed by Şahin (2022), may involve rephrasing or altering the structure of entire sentences, providing a more comprehensive approach to enhancing textual data. These varying levels provide a wide array of methods to enhance training datasets, making them more varied and complex. Text augmentation has been shown to significantly enhance BERT's performance, particularly in small and highly imbalanced datasets. For instance, previous research by Hu et al. (2022) demonstrated a 15.6–40.4% F1 score increase with BERT augmentation compared to the base model, especially when dealing with limited data (e.g., 500 training documents) and high imbalance ratios (e.g., 9:1). This approach, combined with BERT fine-tuning, offers a durable solution for improving classification accuracy in challenging text classification tasks.

As shown in Figure 1, text data augmentation techniques can be categorized into two broad types: the modification of raw data, often referred to as data space, and the alteration of processed data representations, which is known as feature space (Bayer et al., 2023). At the character level, methods such as character insertion, substitution, and swapping can be employed to introduce variations in the data. Koopmans et al. (2023) highlight that character-level augmentation can improve model training by generating realistic noise, particularly in datasets that include common errors or misspellings found in everyday writing. Moving to the token level, Wang (2024) explains that this form of augmentation focuses on manipulating words or tokens within the text. Techniques like word insertion and deletion are commonly used to introduce additional noise and complexity while preserving the overall coherence of the text. In contrast, semantic-level augmentation aims to retain the underly-

ing meaning of the text while altering its structure. One popular technique at this level is back translation, which involves translating the original text into a different language and then translating it back, thereby generating paraphrases that maintain the semantic content but offer a different structure (Chi et al., 2024).

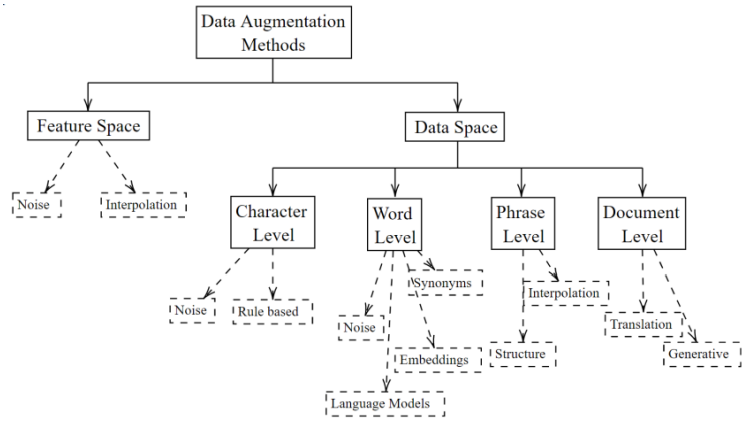

Figure 1. Taxonomy for different text data augmentation methods (Bayer et al., 2023)

Character insertion, as described by Ke et al. (2023), is a text augmentation technique that involves randomly inserting characters at various positions within a given text. By introducing small perturbations to the original text, this method creates synthetically augmented examples that can enhance the stability of machine learning models. Another technique at the character level is character deletion, where characters are randomly removed from the text (Zaiton & Al-Ansary, 2024). This technique introduces noise by omitting specific characters, which simulates common text issues such as typos, misprints, or incomplete word entries, ultimately helping to improve the model's ability to handle such variations.

Word insertion, as explained by B. Guo et al. (2022), is a token-level text augmentation technique that involves randomly inserting new words into the original text. These words are typically selected from a predefined vocabulary or thesaurus, and their inclusion may either alter the meaning or increase the complexity of the text. On the other hand, word deletion, as discussed by Wei and Zou (2019), involves removing words from the text. This process introduces variations by omitting certain parts of a sentence while preserving its overall structure. This technique simulates situations where information is missing or where a writer might omit specific words, thus challenging the model to handle incomplete input.

Back translation, according to Mohiuddin et al. (2021), is a semantic-level text augmentation technique that involves translating a text into a different language and then translating it back to its original form. The idea behind back translation is to introduce variations in phrasing and structure while maintaining the original meaning of the text. This method proves especially effective for enhancing the diversity of training data and improving the generalization of machine learning models by introducing linguistic variations (Edunov et al., 2018).

BERT Classification Task

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model with a bidirectional design that is particularly effective for recognizing and understanding text (Devlin et al., 2018; Vu et al., 2021). Unlike previous unidirectional models, BERT processes text in both directions, which allows it to capture a richer context and better understand the meaning of words in relation to their surrounding context (Areshey & Mathkour, 2023; Zheng et al., 2022). This architecture has led to significant advancements in various natural language processing (NLP) tasks, including text classification, question answering, and named entity recognition (Alshalan & Al-Khalifa, 2020; Sayeed et al., 2023).

BERT also outperforms traditional ML models like SVM/KNN by leveraging bidirectional contextual understanding and large pre-trained vocabularies, addressing their reliance on static features (e.g., TF-IDF) and limited generalization. As shown in previous research (Jin et al., 2019), BERT's architecture demonstrates durability against adversarial attacks by maintaining semantic integrity, grammaticality, and computational efficiency, even with perturbed inputs.

Related to traditional ML models, recent studies (Cunha et al., 2025) have highlighted the effectiveness of large language models (LLMs) like LLaMA, BERT, DeepSeek, and Mistral in text classification tasks, outperforming traditional models. However, these models come with significantly higher computational costs. For example, fine-tuning LLaMA3-70B for classification tasks has shown superior results, though at a greater computational expense.

BERT tokenization and embedding

Tokenization in BERT is a crucial preprocessing step that converts input text into a format suitable for processing (Schick & Schütze, 2020). Using WordPiece tokenization, BERT efficiently handles a wide variety of words and subwords, enabling it to represent rare or out-of-vocabulary words by breaking them into smaller units (Rogers et al., 2020). In addition, BERT employs special tokens like "CLS" for classification tasks and "SEP" to separate input segments, which is particularly useful for tasks involving multiple sentences (Ettinger, 2020). This tokenization process significantly enhances BERT's ability to generalize across languages and in tasks such as sentiment analysis and question answering (Özçift et al., 2021).

BERT architecture and training

BERT's architecture consists of stacked layers of transformer encoder blocks, each containing multi-head attention layers and position-wise feed-forward layers as shown in Figure 2 (Tran & Le, 2023). The multi-head attention mechanism allows the model to assign different weights to each word or token based on its relevance within the sentence, thereby providing a dynamic understanding of context (Ma et al., 2022).

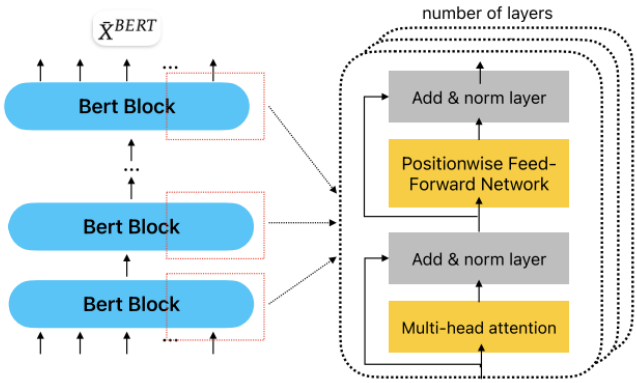

Figure 2. BERT encoder architecture (Tran & Le, 2023)

Another key feature of BERT is its positionwise feedforward network (Y. Li & Rockinger, 2024), which consists of fully connected layers that process each token in the input sequence independently. This design allows the model to learn complex representations while maintaining the order of tokens through positional encodings.

BERT Pre-training and fine-tuning process

The training process of the model, as outlined by Yang et al. (2024), is divided into two stages: pre-training and fine-tuning. During pre-training, the model is trained on large, unlabeled datasets, allowing BERT to learn general language patterns and contextual relationships (D. Li et al., 2021). Fine-tuning, on the other hand, is a more specialized process in which the pre-trained model is further

trained on specific labeled datasets (Da Rocha Junqueira et al., 2023). These datasets, such as Multi-Genre Natural Language Inference (MNLI), Named Entity Recognition (NER), or sentiment analysis, help optimize the model's performance for particular tasks (Su, 2024).

BERT's impact and configurations

BERT has had a significant impact on the performance of many NLP tasks, largely due to its pre-training process. The model has been developed in several variants, such as the BERT-base-uncased and BERT-base-cased versions. BERT's bidirectional architecture allows it to capture the full context of words more effectively than previous models, which typically processed text in a unidirectional manner (Gupta, 2024). This bidirectional contextual understanding enables BERT to dynamically resolve ambiguities and outperform traditional ML models like SVM or KNN, which rely on static features (e.g., TF-IDF) and lack the ability to interpret word relationships in context.

The impact of BERT on NLP has been transformative, driving the creation of various adaptations and variations, including RoBERTa, ALBERT, and domain-specific models like BioBERT and FinBERT (Dhole et al., 2024).

Computational Approach

Traditional image-based document classification, such as the multimodal research approach by Audebert et al. (2020), struggles with limited real-world dataset diversity and models (e.g., CNN/MLP) that lack contextual understanding. To overcome these limitations, this study employs text augmentation techniques to synthetically expand a small dataset of 75 certificates on each label into 700 certificate-text data total in this research. By diversifying text data, these methods simulate real-world variations (e.g., OCR errors, formatting inconsistencies), enabling resilient model training without relying on large-scale image datasets.

This study systematically applies character-level, token-level, and semantic-level augmentations to enhance data variability. While character-level methods introduce noise akin to OCR errors, token-level augmentations mimic missing or redundant text in certificates. Semantic-level techniques preserve meaning while altering phrasing, ensuring the model learns contextual relationships critical for classifying unstructured certificate data.

BERT model is chosen for its ability to capture bidirectional context and leverage large pre-trained vocabularies, addressing the contextual gaps of traditional models. While studies by Röttger and Pierrehumbert (2021) demonstrate BERT's success in text-based document classification, their work relies on pre-digitized text data. In contrast, this study addresses image-based document classification, where text extraction from scanned certificates introduces OCR errors and formatting inconsistencies. By combining BERT with text augmentation techniques, this study approach ensuring sturdiness in scenarios where data is limited and not pre-cleaned/pre-structured.

METHODOLOGY

The methodology of this study involves a systematic approach to document classification, starting with data gathering and preprocessing. This process followed by the application of various text augmentation techniques, model training, and evaluation to optimize classification performance as shown in the Figure 3.

The Figure 3 illustrates the workflow of a machine learning-based document classification system that used in this research study. With a focus on the preprocessing steps and the integration of various text augmentation techniques. The process begins with data gathering, where relevant text documents are collected. This data then subjected to data conversion and data cleaning to transform the raw input into a usable format.

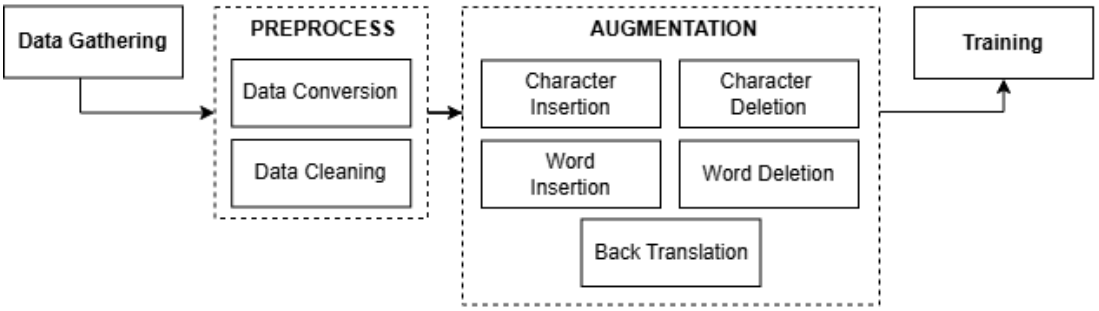

Figure 3. Research workflow

Once the data was preprocessed shown in the Figure 3, augmentation techniques are applied to enhance the dataset, including methods such as character insertion, character deletion, word insertion, word deletion, and back translation. These augmentations introduce variations into the text, enriching the training data. After augmentation, the model undergoes training, where the augmented data was used to refine the model's ability to classify documents. Finally, the results are evaluated through test evaluation, where the model's performance is assessed based on accuracy and other metrics, helping determine the most effective augmentation strategy.

STEP 1: DATA GATHERING

The data gathering process in this research utilized web scraping, an automated technique to extract relevant information from websites, focusing specifically on retrieving certificates related to various course categories from Google Search results. Python-based code implemented using Playwright (Almabruk et al., 2025), is a browser automation library, with Chromium in non-headless mode to interact directly with Google Search pages. Advanced search operators such as filetype:pdf and intitle:certificate were used to narrow down results, targeting certificates related to specific topics like machine learning or programming. Throughout this process, careful attention was given to the ethical considerations, including verifying that the data being collected did not violate any copyright restrictions. The research ensured that the data gathered was publicly available.

Once the queries were executed, the scraping tool extracted key details, including titles, URLs, and snippets from the search results. Duplicate titles were automatically numbered to avoid redundancy, and the program navigated through multiple result pages, clicking "next" until all relevant data was gathered. The extracted information was stored in a structured format, such as a list of dictionaries, for further processing. The URLs of the collected certificates were then used to automate the downloading of the PDF files, which were saved with filenames based on the document titles to ensure proper organization and easy identification. This systematic approach streamlined the data management process, making the collected data ready for further analysis. Sample data of certificate as shown in Figure 4, used in this research.

(a) (b)

Figure 4. Sample of certificates data

In this study, only English-language data was utilized to ensure consistency and alignment with the model's configuration. A total of 525 PDF files were collected through web scraping. As illustrated in Figure 4, the data certificate samples are presented, with (a) showcasing an example of a machine learning certificate, while (b) displays an example of a web programming certificate. The data was evenly distributed across seven course-related classes: Machine Learning, Web Programming, User Interface Design, Computer Networking, Game Applications, Mobile Programming, and Project Management.

STEP 2: DATA PREPROCESSING

In this study, data processing will focus on transforming raw certificate data into a structured format suitable for classification. The process includes converting PDF files into images, extracting text using OCR, and cleaning the data by standardizing text and removing unnecessary elements to improve model accuracy and efficiency.

Data conversion

Data conversion plays a pivotal role in transforming raw certificate data into a structured format that is suitable for analysis and classification. As illustrated in Figure 5, this study involves converting PDF data into tabular text data in several steps.

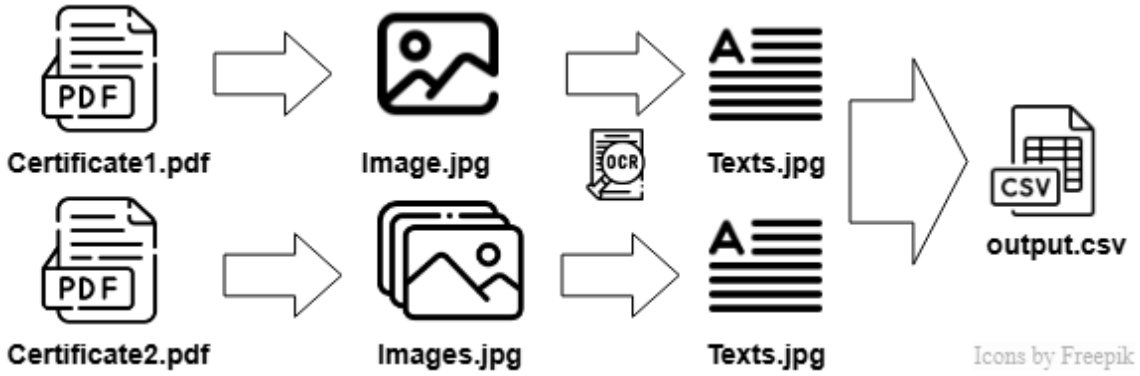

Figure 5. Data conversion steps

Data processing in this study as shown in Figure 5 begins by converting PDF certificates into high-quality grayscale images using tools like Poppler, which helps preserve the visual layout and reduces complexities in text extraction (Mursari & Wibowo, 2021). These images (e.g., Image.jpg) represent the first page of the certificate. If the certificate has multiple pages, multiple images (e.g., Images.jpg, Images2.jpg) are generated, with each image corresponding to a single page of the certificate. Optical Character Recognition (OCR) technology, such as PyTesseract, was then used to extract machine-readable text from these images, recognizing characters, words, and sentences.

Finally, the extracted text was organized into a structured CSV file, where each row corresponds to one certificate. Each row also includes the certificate's category label (to identify the certificate type) and the file path to the original PDF. This structure ensures that each row of data contains the details of certificate information.

Data cleaning

In this research, data cleaning involves three key processes: case folding, filter stopword, and stemming. Case folding standardizes the text by converting all characters to lowercase, which helps reduce variability caused by differences in letter casing. This ensures that words like "Machine" and "machine" are treated as the same entity during model training. Additionally, filter stopword removes common words such as "the" and "is" which do not contribute significantly to the meaning of the text. These stopwords often appear frequently but add little value to the classification task, making their removal essential for improving the model's focus on meaningful terms.

The final step in data cleaning was stemming, which reduces words to their base or root form to minimize redundancy and improve generalization. For example, words like "running" and "runner" are both reduced to "run," ensuring that variations of the same word are treated uniformly. This process not only simplifies the dataset but also enhances the model's ability to recognize patterns across similar words.

STEP 3: DATA AUGMENTATION

In this study, five distinct datasets will be generated, each utilizing a different text augmentation technique. These techniques—character-level, token-level, and semantic-level augmentations— were selected to handle OCR errors and formatting inconsistencies in non-formal learning certificates. Character-level techniques like insertion and deletion simulate OCR noise, while token-level augmentations help the model classify certificates despite missing or redundant text. Whereas semantic-level augmentations can add vocabulary variety, they focus on altering the phrasing while preserving the meaning, which can introduce more linguistic diversity without changing the core content.

Character level augmentation

Character-level augmentation is a technique used to generate synthetic variations of text data by making modifications at the character level. This approach is particularly useful for increasing dataset diversity and improving model performance. In this study, two primary methods of character-level augmentation were implemented: character insertion and character deletion.

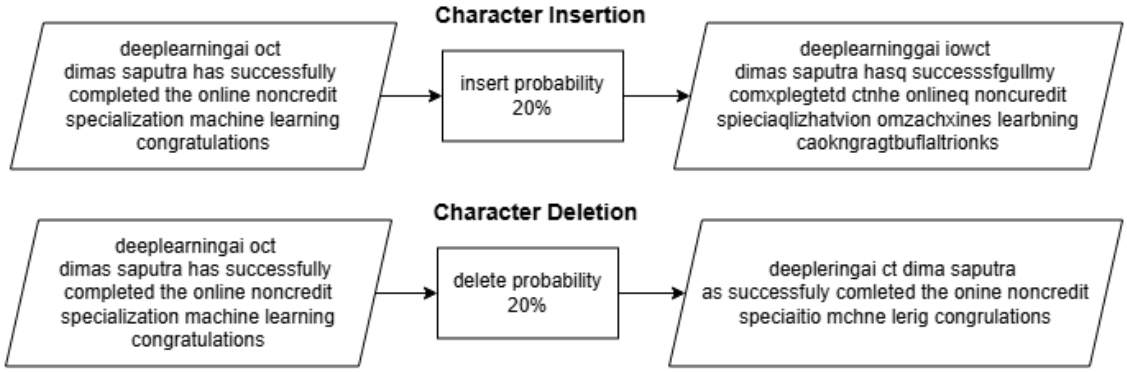

Figure 6. Character-level augmentation

The first method, character insertion like shown in Figure 6. This process introduces noise into the text while preserving its overall structure and meaning. In the implementation, each character in the text has a specified probability (e.g., 20%) of being followed by a randomly chosen character from a predefined alphabet.

The second method, character deletion like shown in Figure 6. In the implementation, each character in the text has a specified probability (e.g., 20%) of being deleted, which introduces variations while preserving the overall context of the text.

Token level augmentation

Token-level augmentation is a technique used to generate synthetic variations of text data by making modifications at the word level. In this study, two primary methods of token-level augmentation were implemented: word insertion and word deletion.

The first method, word insertion like shown in Figure 7. This process introduces noise into the text while preserving its overall structure and meaning. In the implementation, each word in the text has a specified probability (e.g., 20%) of being followed by a randomly chosen word from a predefined vocabulary.

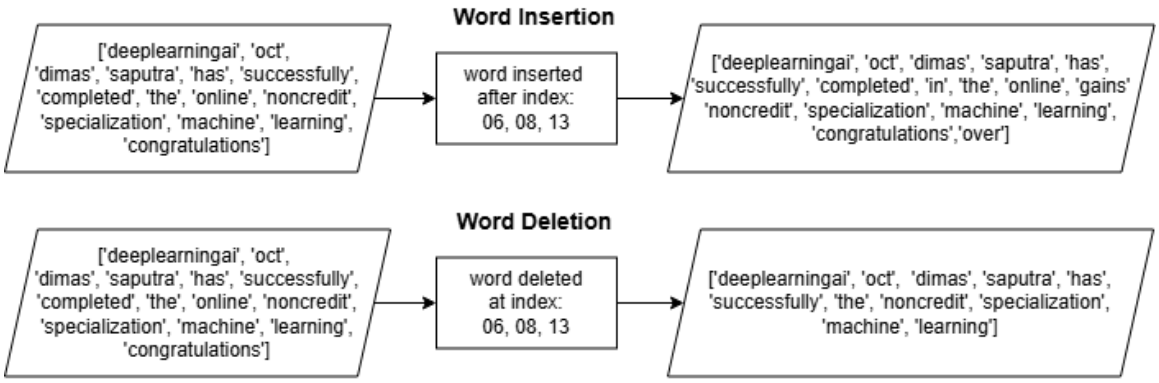

Figure 7. Token-level augmentation

The second method shown in Figure 7, word deletion, focuses on removing words from the text at random positions like shown in Figure 7. This process introduces noise into the text while preserving its overall structure and meaning. In the implementation, each word in the text has a specified probability (e.g., 20%) of being removed.

Semantic augmentation

Semantic-level augmentation focuses on preserving the meaning of the text while introducing variations in its structure or wording. One of the most effective techniques for achieving this is back translation, which involves translating a text from its original language into another language and then translating it back into the original language.

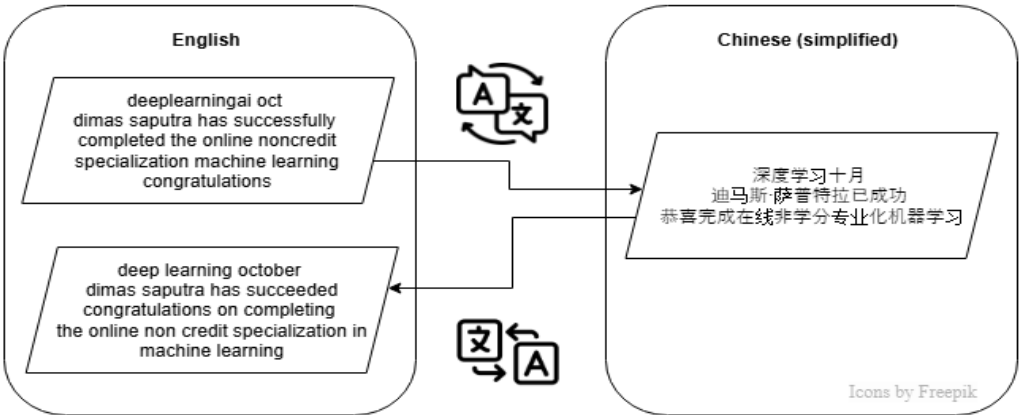

Figure 8. Semantic-level augmentation: back translation

In this study, the back translation method was implemented to augment the dataset by translating English text into Chinese (Simplified) and then translating it back into English, as shown in Figure 8. This process introduces linguistic diversity into the training set, enabling the model to learn from varied sentence structures while preserving the original meaning of the text.

Step 4: Classification Using BERT

Involves training a BERT-based model for text classification using preprocessed data. The dataset is first split into training, validation, and test sets to evaluate the model's performance effectively. 80% of the data was used for training, while the remaining 20% is further divided equally for validation and testing. The model utilizes a pretrained BERT-base-uncased model, configured with a tokenizer and model settings that accommodate 7 label categories for classification.

The BERT-based model used in this research was fine-tuned to classify certificates into predefined categories. Fine-tuning leverages a pretrained BERT-base-uncased model, adjusting it to the specific task by training on the labeled dataset (Arslan & Cruz, 2024). This allows the model to learn context-specific patterns that improve classification accuracy.

During training, the model was compiled using the AdamW optimizer to ensure stable and efficient training. For loss computation, the hf_compute_loss function from Hugging Face was used, which defaults to SparseCategoricalCrossentropy. This loss function is particularly suitable for multi-class classification tasks, where the target labels are integers representing class indices rather than one-hot encoded vectors. The model is optimized for accuracy, and during training, the loss is minimized to improve classification performance on the dataset.

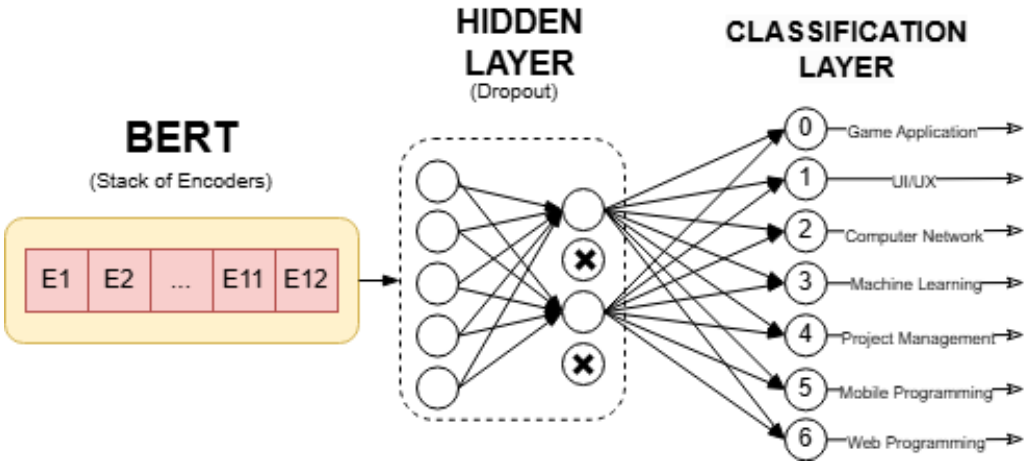

Figure 9. Model architecture

As shown in Figure 9, the model architecture consists of the BERT main layer (a stack of encoders), followed by a dropout layer in the hidden layer to prevent overfitting. The final layer, the classifier, uses a dense layer with a softmax activation function. The classifier outputs 7 logits corresponding to the 7 categories, and softmax was applied to these logits to produce probabilities. The category with the highest probability is selected as the predicted label.

$$f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \quad (1)$$

The formula for softmax is shown in Equation 1. In the softmax formula (Mehra et al., 2023), x represents the output values from the previous layer, e refers to the logits for each predicted class, N is the total number of output values, and i is the element whose probability is being calculated. The sum is computed over all logits j, from 1 to N, to normalize the output into a probability distribution. This ensures that the sum of all label probabilities equals 1.

EXPERIMENT SETUP

HYPERPARAMETER TRAINING

Hyperparameter tuning is a critical step in optimizing the performance of machine learning models, particularly when working with complex architectures like BERT. In this study, several key hyperparameters were carefully selected and fine-tuned to achieve optimal classification results. These hyperparameters include the optimizer, learning rate, number of epochs, and batch size. By systematically adjusting these parameters, the goal was to identify the most effective configuration for both character-level and token-level augmentations. The AdamW optimizer was chosen for its efficiency

in handling large-scale models like BERT, combining adaptive learning rates with weight decay regularization to prevent overfitting during training. For this research, the learning rate was set to 5e-5, and epsilon was configured at 1e-08, ensuring stable convergence while maintaining computational efficiency.

The number of epochs was fixed at 5 to balance between model convergence and computational cost. Each epoch represents a complete pass through the entire training dataset, allowing the model sufficient exposure to the data without excessive resource consumption. The batch size was set to 8, striking an appropriate balance between frequent updates to the model weights and efficient utilization of GPU memory. This configuration ensures that the models can adapt effectively to the diverse range of textual inputs generated by the augmentation techniques, ultimately enhancing their ability to classify certificates accurately.

TRAINING SCENARIO

In this research, the BERT-base-uncased model was utilized to evaluate the effectiveness of various text augmentation techniques in improving classification performance. The primary objective was to determine which augmentation method yields the most efficient results and to compare the relative performance of character-level augmentations (Character Insertion [CI] and Character Deletion [CD]) versus token-level augmentations (Word Insertion [WI] and Word Deletion [WD]). Four distinct datasets are created using these augmentation techniques, and each dataset is used to train the model independently.

Each of the four datasets was trained independently using the BERT-base-uncased model, ensuring a controlled comparison of the augmentation techniques. The training process involves feeding the augmented data into the model and allowing it to learn the patterns and features associated with each class. This approach enables a direct assessment of how each augmentation technique impacts the model's ability to classify certificates accurately. The results from these experiments will provide insights into which augmentation strategy—character-level or token-level—is better suited for improving model performance in this specific task.

Table 1. Training scenario

MODEL	AUGMENTATION
BERT-base-uncased	Character Insertion
	Character Deletion
	Word Insertion
	Word Deletion
	Back Translation

As shown in Table 1, the training scenarios used in this study focused on applying various data augmentation techniques to the BERT-base-uncased model. The study aims to provide valuable insights into optimizing text classification models using BERT, particularly for scenarios where data variability and balance play crucial roles. By systematically assessing the effect of different augmentation strategies, the research seeks to identify best practices for improving model generalization across diverse datasets.

RESULTS

ACCURACY RESULTS

The model's performance across different data augmentation techniques shows significant improvement in both training and validation accuracy. Word-based augmentations, particularly word insertion and deletion, resulted in the highest accuracy levels, demonstrating strong generalization capabilities.

In addition to word-based augmentations, character-level techniques, such as character insertion and deletion, also contributed to improvements in model performance, though to a lesser extent. These augmentations added noise to the text, helping the model generalize better to minor variations and typos in the input data.

However, semantic-level augmentation, specifically back translation, showed more modest improvements. While it introduced more linguistic diversity into the training set, its impact on accuracy was not as significant as word-level augmentations, suggesting that word-based variations are more effective in enhancing the model's ability to classify non-formal learning certificates.

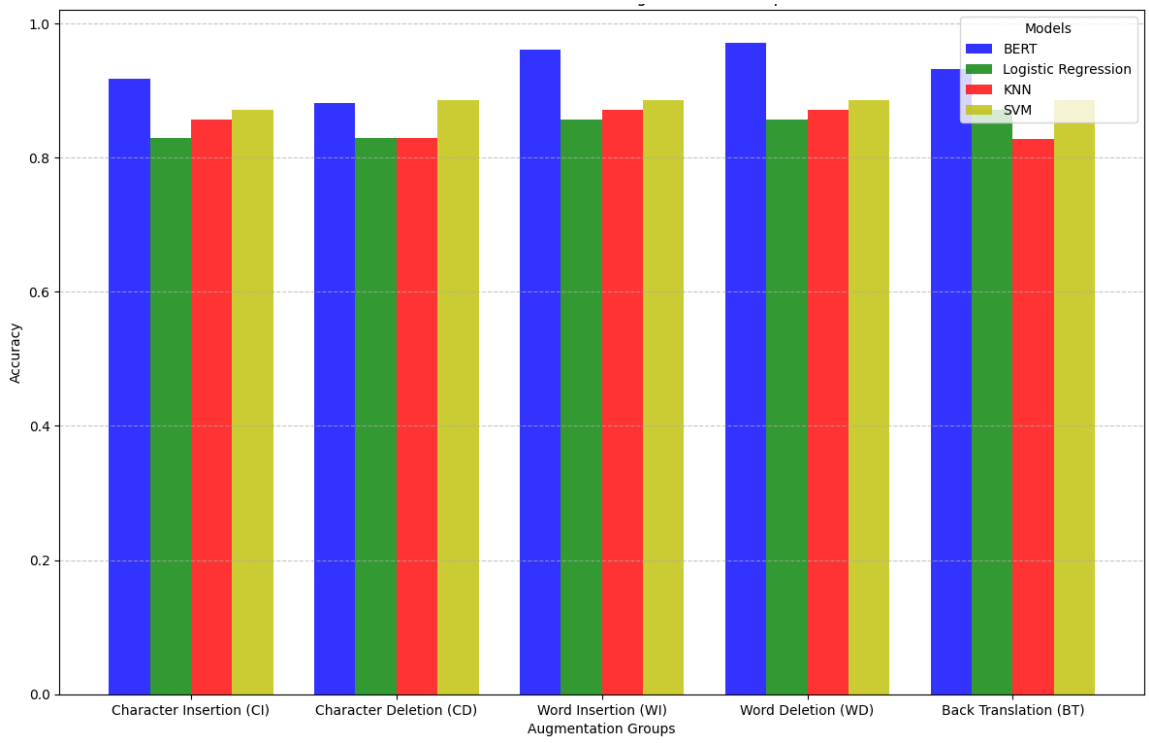

Figure 10. Model accuracies across augmentation method

On comparison BERT with traditional ML models as shown in the figure 10, BERT significantly outperforms Logistic Regression, KNN, and SVM across all five text augmentation techniques, achieving the highest validation accuracies. Word-based augmentations (word insertion: 96.1%, word deletion: 97.1%) highlight BERT's strongest performance, while character-based methods (character insertion: 91.8%, character deletion: 88.2%) and back translation (93.2%) follow.

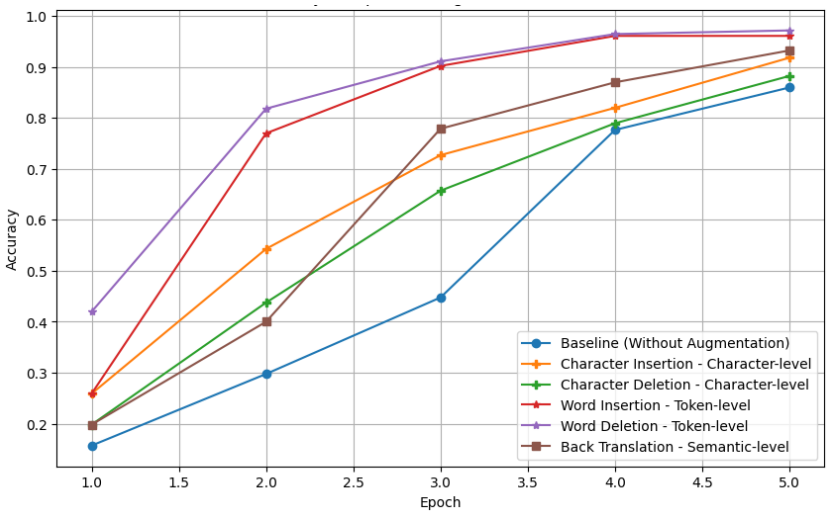

Figure 11. Accuracy comparison on various scenarios

The Figure 11 illustrates the training accuracy results across different data augmentation scenarios over five epochs. The accuracy for each scenario improves progressively with each epoch. Among the augmentation techniques, word deletion yields the highest accuracy, starting at 0.4196 in the first epoch and reaching 0.9714 by the fifth epoch. Word insertion also shows strong performance, with accuracy increasing from 0.2589 in the first epoch to 0.9607 in the fifth epoch, stabilizing at the higher value in later epochs. Back translation follows, showing steady improvements from 0.1982 to 0.9321. Character insertion and character deletion show similar trends, but their accuracy levels are slightly lower than the word-based augmentations, with character insertion achieving 0.9179 and character deletion reaching 0.8821 by the fifth epoch. All augmentation techniques, except for character deletion and back translation, surpass the baseline accuracy of 0.8462, demonstrating their effectiveness in improving the classification performance over the course of the training epochs. Among these augmentation techniques, all surpass the baseline accuracy of 0.8595 by the fifth epoch, with word deletion achieving the highest accuracy.

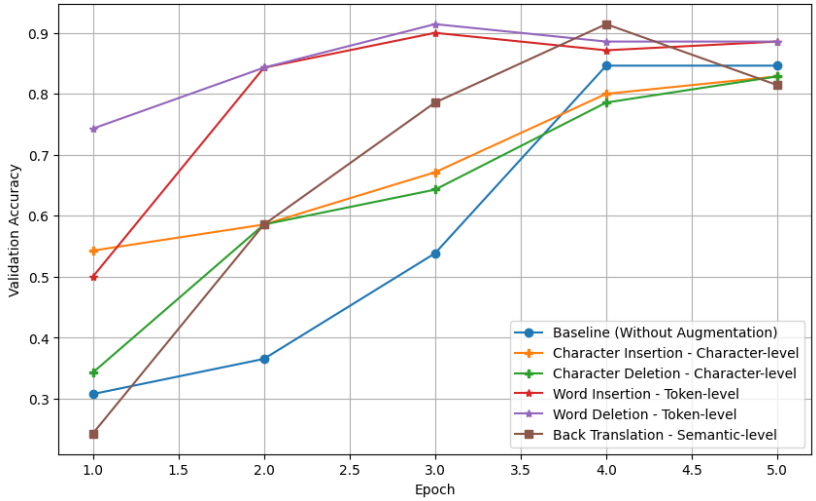

Figure 12. Validation accuracy comparison augmentation scenarios

The Figure 12 presents the validation accuracy results across different data augmentation scenarios over five epochs. Word deletion consistently shows the best validation performance, starting at 0.7429 in the first epoch and maintaining high accuracy throughout, with a slight drop after the

fourth epoch, stabilizing at 0.8857 in the fifth epoch. Word insertion also performs well, with a significant improvement from 0.5000 in the first epoch to 0.8857 in the fifth epoch. However, it shows a slight decline in performance during the fourth epoch, before recovering in the fifth. Character insertion and character deletion both exhibit moderate improvements across the epochs, with character insertion reaching 0.8286 by the fifth epoch and character deletion achieving the same value. Back translation, while showing improvement, lags behind the other techniques, starting at 0.2429 in the first epoch and gradually improving to 0.8143 by the fifth epoch. Among these augmentation techniques, word insertion and word insertion surpass the baseline validation accuracy of 0.8462 by the fifth epoch, with both word insertion and word deletion achieving the highest validation performance.

Confusion Matrix

The confusion matrix for the BERT model on the five augmented datasets shows the model's ability to classify multiple categories accurately. It highlights strong performance, with most instances correctly classified, though a few misclassifications occur between similar categories. These misclassifications suggest that the model may struggle with subtle differences between closely related classes, but overall, the results demonstrate the model's durable in handling diverse input variations. Further tuning or data adjustments could potentially reduce these errors.

The results derived from the confusion matrices, displayed in Figure 13, across the five perturbation scenarios character insertion (a), character deletion (b), word insertion (c), word deletion (d), and back translation (e) demonstrate varying classification performances for different types of dataset alterations.

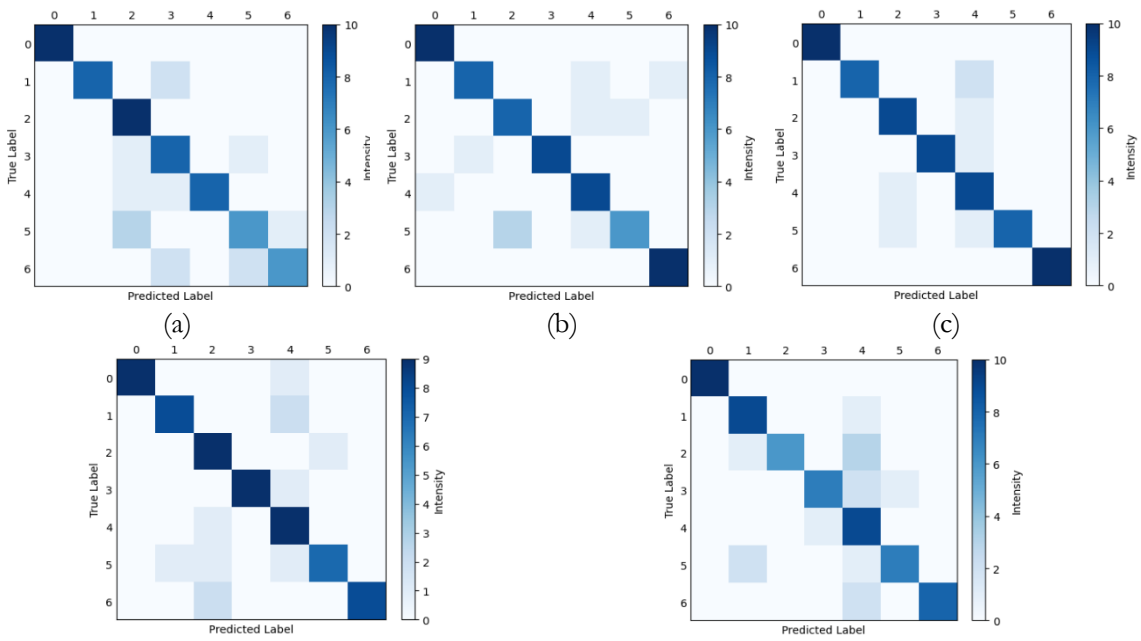

Figure 13. Confusion matrix results

Shown in Figure 13, scenario (a), character insertion, the "game application" and "computer networks" labels showed no prediction errors, while other labels experienced some misclassifications. A total of 56 data points were correctly predicted, with 14 misclassifications. In scenario (b), character deletion, the "game application" and "web programming" labels exhibited no prediction errors, while other labels had some misclassifications. This scenario had 60 correctly predicted instances and 10 misclassifications. In scenario (c), word insertion, the "game application" and "web programming"

labels also showed no prediction errors, while other labels had some misclassifications. Here, 63 data points were accurately predicted, with 7 misclassifications. Scenario (d), word deletion, resulted in misclassifications across all labels, yielding 59 correct predictions and 11 misclassifications. Finally, in scenario (e), back translation, the "game application" label showed no prediction errors, while the other labels had some misclassifications, resulting in 56 correctly predicted data points and 14 misclassifications.

CLASSIFICATION REPORTS

The BERT model's classification report on the Word Deletion dataset highlights its performance across precision, recall, and F1-scores, showing strength and area for improvement shown in Table 2.

Table 2. Classification report result

AUGMENTATION	PRECISION		RECALL		F1-SCORE	
	Macro	Weighted	Macro	Weighted	Macro	Weighted
- (Baseline)	0.83	0.83	0.78	0.77	0.77	0.77
Character Insertion	0.83	0.83	0.80	0.80	0.80	0.80
Character Deletion	0.86	0.86	0.86	0.86	0.85	0.85
Word Insertion	**0.92**	**0.92**	**0.90**	**0.90**	**0.90**	**0.90**
Word Deletion	0.87	0.87	0.84	0.84	0.85	0.85
Back Translation	0.86	0.86	0.80	0.80	0.81	0.81

As shown in Table 2, the classification report details the BERT model's performance across various text augmentation techniques, with precision, recall, and F1-scores broken down by macro and weighted values. The first row (baseline) shows the model's performance without any text augmentation, with precision, recall, and F1-scores of 0.83 (macro) and 0.77 (weighted), 0.78 (macro) and 0.77 (weighted), and 0.77 (macro) and 0.77 (weighted) respectively, providing a reference point for comparison.

Character Insertion (row 2) and Character Deletion (row 3) both exhibit moderate improvements, with precision scores of 0.83 (macro) and 0.86 (weighted), as seen in column 2. Their recall scores are 0.80 (macro) and 0.86 (weighted), found in column 3, and their F1-scores are 0.80 (macro) and 0.85 (weighted), in column 4. This indicates that while these methods add variability to the data, they have less impact on accuracy compared to other techniques.

Word Insertion (row 4) shows the highest performance, with precision scores of 0.92 (macro and weighted) in column 2, recall scores of 0.90 (macro and weighted) in column 3, and F1-scores of 0.90 (macro and weighted) in column 4, highlighting its effectiveness in diversifying the dataset and improving accuracy. Word Deletion (row 5) also shows strong results, with precision of 0.87 (macro and weighted) in column 2, recall of 0.84 (macro and weighted) in column 3, and F1-scores of 0.85 (macro and weighted) in column 4, demonstrating its sturdy in handling incomplete text. Lastly, Back Translation (row 6) shows the least improvement, with precision and recall of 0.86 and 0.80 (macro and weighted) in columns 2 and 3, respectively, and an F1-score of 0.81 (macro and weighted) in column 4.

STATISTICAL ANALYSIS

Additionally descriptive analysis as explained before, we also need to conduct inferential statistical analysis to determine the statistical significance of observed differences in performance metrics between various approaches and models. To evaluate the performance of the augmentation techniques, we first conduct a statistical analysis of accuracy results using mean, median, standard deviation (STD), and the Shapiro-Wilk test to assess data normality, with all metrics calculated from 5 training

iterations for each scenario. Subsequently, we perform pairwise comparisons of F1-score results across augmentation techniques using paired t-tests and calculate p-values to determine statistical significance, ensuring reliable conclusions based on consistent experimental runs.

Table 3. Accuracy results statistical analysis

Model	Augmentation	Mean	Median	STD	Shapiro-Wilk
BERT	CI	0.827	0.939	0.261	0.001
	CD	0.924	0.929	0.034	0.747
	WI	0.970	0.968	0.008	0.803
	WD	0.944	0.961	0.056	0.035
	BT	0.937	0.961	0.049	0.104

As shown in Table 3, the table summarizes the statistical analysis of accuracy results for different augmentation methods under the BERT model, revealing that Word Insertion (WI) achieved the highest mean and median accuracy scores, indicating its superior performance. Character Insertion (CI) showed the lowest mean accuracy and high variability, as evidenced by its large standard deviation and non-normal distribution. The other methods, including Character Deletion (CD), Word Deletion (WD), and Back Translation (BT), demonstrated moderate performance with varying levels of consistency and normality in their distributions.

Table 4. Paired augmentation techniques F1-score results statistical analysis

Augmentation (1)	Augmentation (2)	t-test	p-value	Significance (<0.05)
CI	CD	0.066	0.9504	No significant
CI	WI	-5.926	0.0041	Significant
CI	WD	-3.467	0.0256	Significant
CI	BT	-1.779	0.1499	No significant
CD	WI	-3.392	0.0275	Significant
CD	WD	-1.906	0.1293	No significant
CD	BT	-1.960	0.1216	No significant
WI	WD	2.867	0.0456	Significant
WI	BT	2.746	0.0516	No significant
WD	BT	-0.136	0.8984	No significant

Shown above in Table 4, presents the results of paired t-tests comparing F1-scores across different augmentation techniques, revealing that Word Insertion (WI) significantly outperforms Character Insertion (CI), Character Deletion (CD), and Word Deletion (WD), as indicated by p-values below 0.05. Conversely, no significant differences were found between Character Insertion and Back Translation (BT), or between various other pairs such as CD and WD. These findings highlight the superior effectiveness of token-level augmentations like Word Insertion in enhancing classification performance compared to character-level methods.

DISCUSSION AND FINDINGS

This study builds upon existing research, highlighting the effectiveness of token-level text augmentation techniques—specifically word insertion and word deletion—in improving document classification accuracy. While prior studies, such as Şahin (2022), showed that character-level augmentations

were useful in linguistic tasks, our findings demonstrate that token-level techniques significantly outperform them, especially when dealing with noisy OCR data in non-formal learning certificates. Additionally, our results further validate the effectiveness of BERT in document classification but suggest that augmenting it with these techniques can yield even better performance than using BERT alone, especially in noisy, unstructured data scenarios.

However, some anomalies were observed in our findings, particularly with the back translation augmentation. While it introduced linguistic diversity, its impact on model accuracy was notably smaller compared to token-level techniques, suggesting that semantic-level changes may not always be as effective for improving classification performance. Additionally, while character-level augmentations did contribute to model resilient, their improvements were less significant, especially when handling noisy OCR data, indicating a potential limitation in their applicability for this specific task.

For practitioners, the key takeaway is to prioritize token-level augmentations, such as word insertion and deletion, when working with OCR data from non-formal learning certificates, as these techniques significantly improve classification accuracy. While character-level augmentations offer some benefits in enhancing model performance, their impact is less pronounced, particularly in the presence of noisy OCR data. Additionally, back translation, although it introduces linguistic variety, has a smaller effect on accuracy compared to token-level techniques. Therefore, focusing on token-level augmentations and fine-tuning BERT with augmented datasets will optimize performance. Using high-quality OCR tools can also enhance accuracy, and expanding this methodology to multilingual datasets and continuously updating models as new data becomes available will ensure the long-term effectiveness of the model in real-world applications.

CONCLUSION, FUTURE RESEARCH AND LIMITATIONS

This study aimed to investigate how different text augmentation methods—character-level, token-level, and semantic-level—affect the performance of BERT in classifying non-formal learning certificates into computer science course categories. The research objectives were to determine which augmentation strategy improves classification accuracy and to assess the impact of these techniques on model generalization, particularly with noisy OCR data.

The key findings from this study demonstrate that word-level text augmentation techniques, such as word insertion and word deletion, were the most effective in enhancing BERT's performance for classifying non-formal learning certificates. This study further validated that BERT, enhanced by token-level text augmentation strategies, can accurately classify non-formal learning certificates into computer science course categories, achieving over 88% validation accuracy. Moreover descriptive analysis (e.g., accuracy, loss, f1-score, etc.) this finding is statistically validated through paired t-tests and p-value, which confirmed significant differences in accuracy and F1-scores between word-level and other augmentation methods ($p < 0.05$). These methods contributed significantly to improving the model's accuracy and generalization, enabling it to handle diverse certificate data more stable. Word insertion and deletion introduced controlled variability into the text, allowing the model to learn from a broader range of inputs. As a result, these augmentation techniques led to the highest improvements in both training and validation accuracy, with word deletion achieving the best validation results, demonstrating the model's adaptability to variations in certificate content.

While other methods, such as character-level augmentation and back translation, also contributed to model performance, they were less impactful than word-level modifications. Inferential analysis, including p-value comparisons and normality checks via the Shapiro-Wilk test, further confirmed the superiority of token-level approaches over character-level and semantic-level techniques. Character-level augmentation, though useful for adding noise at the character level, did not produce the same level of improvement as word-level techniques. Similarly, back translation, which translated text between languages and returned it to its original form, provided some benefits but was less effective due to challenges in maintaining semantic consistency. These findings emphasize the importance of

selecting the most suitable augmentation strategy, with word-level techniques proving to be the most effective for document classification tasks in the context of Recognition of Prior Learning (RPL).

However, there are some limitations to this research. The dataset, while diverse in course categories, remains limited in size, which may impact the model's generalizability. Additionally, the study focused solely on English-language certificates, which could restrict the model's performance in multilingual contexts. The reliance on synthetic data augmentation is also a limitation, as it may not fully replicate real-world text variations found in certificates across different domains. These factors could influence the accuracy of the model when applied to larger-scale or more varied datasets.

For future research, expanding the dataset to include a wider range of languages and labels is essential for improving the model's applicability. Additionally, experimenting with different OCR tools could enhance the text extraction process and improve accuracy in real-world scenarios. Exploring various BERT configurations, such as BERT-large, BERT-base, or BERT-multilingual, might provide insights into the optimal model settings for different tasks. Finally, investigating other NLP models or hybrid approaches could further enhance document classification performance, offering more flexibility for real-world applications in RPL programs. These efforts would help create more scalable solutions to automate the recognition of prior learning, contributing to broader access to education and employment opportunities for diverse populations.

REFERENCES

Almabruk, S., Abdalhamid, S., & Almabruk, T. (2025). Comparative reliability analysis of selenium and playwright: evaluating automated software testing tools. *Asian Journal of Research in Computer Science*, *18*(1), 34–44. https://doi.org/10.9734/ajrcos/2025/v18i1546

Alshalan, R., & Al-Khalifa, H. (2020). A deep learning approach for automatic hate speech detection in the Saudi twittersphere. *Applied Sciences*, *10*(23), 8614. https://doi.org/10.3390/app10238614

Areshey, A., & Mathkour, H. (2023). Transfer learning for sentiment classification using Bidirectional Encoder Representations from Transformers (BERT) model. *Sensors*, *23*(11). https://doi.org/10.3390/s23115232

Arslan, M., & Cruz, C. (2024). Business text classification with imbalanced data and moderately large label spaces for digital transformation. *Applied Network Science*, *9*(1). https://doi.org/10.1007/s41109-024-00623-5

Atmaja Perdana, C. R., Adi Nugroho, H., & Ardiyanto, I. (2020). Comparison of text-image fusion models for high school diploma certificate classification. *Communications in Science and Technology*, *5*(1), 5–9. https://doi.org/10.21924/cst.5.1.2020.172

Audebert, N., Herold, C., Slimani, K., & Vidal, C. (2020). Multimodal deep networks for text and image-based document classification. In *Communications in Computer and Information Science: Vol. 1167 CCIS* (pp. 427–443). Springer. https://doi.org/10.1007/978-3-030-43823-4_35

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2023). A survey on data augmentation for text classification. *ACM Computing Surveys*, *55*(7), 1–39. https://doi.org/10.1145/3544558

Chen, X., He, B., Hui, K., Sun, L., & Sun, Y. (2023). Dealing with textual noise for robust and effective BERT re-ranking. *Information Processing & Management*, *60*(1), 103135. https://doi.org/10.1016/j.ipm.2022.103135

Chi, W. W., Tang, T. Y., Salleh, N. M., Mukred, M., AlSalman, H., & Zohaib, M. (2024). Data augmentation with semantic enrichment for deep learning invoice text classification. *IEEE Access*, *12*, 57326–57344. https://doi.org/10.1109/ACCESS.2024.3387860

Cunha, W., Rocha, L., & Gonçalves, M. A. (2025). *A thorough benchmark of automatic text classification: From traditional approaches to large language models*. arXiv preprint. https://arxiv.org/abs/2504.01930

Da Rocha Junqueira, J., Da Silva, F., Costa, W., Carvalho, R., Bender, A., Correa, U., & Freitas, L. (2023). BERTimbau in action: An investigation of its abilities in sentiment analysis, aspect extraction, hate speech detection, and irony detection. In *The International FLAIRS Conference Proceedings*, *36*. https://doi.org/10.32473/flairs.36.133186

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint. http://arxiv.org/abs/1810.04805

Dhole, Y. B., Sharma, D. R., Patil, S., & Chavan, D. (2024). Unveiling market sentiments: Finbert-powered analysis of stock news headlines. *International Journal of Scientific Research in Engineering and Management*, *08*(10), 1–6. https://doi.org/10.55041/IJSREM38254

Domvo, W. (2022). The challenges surrounding recognition of prior learning for refugees in European universities. *Australian and New Zealand Journal of European Studies*, *14*(1). https://doi.org/10.30722/anzjes.vol14.iss1.15855

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500. https://doi.org/10.18653/v1/D18-1045

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34–48. https://doi.org/10.1162/tacl_a_00298

Gopi, K., Sushma, G., & Vallepu, S. R. (2022). Classification of large documents using machine learning techniques. *International Journal of Engineering Technology and Management Sciences*, *6*(5), 898–904. https://doi.org/10.46647/ijetms.2022.v06i05.137

Guo, B., Han, S., & Huang, H. (2022). *Selective text augmentation with word roles for low-resource text classification*. arXiv preprint. http://arxiv.org/abs/2209.01560

Guo, Z., & Nguyen, M. L. (2020). Document-level neural machine translation using BERT as context encoder. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, 101–107. https://doi.org/10.18653/v1/2020.aacl-srw.15

Gupta, R. (2024). Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing. *Информатика. Экономика. Управление - Informatics. Economics. Management*, *3*(1), 0311–0320. https://doi.org/10.47813/2782-5280-2024-3-1-0311-0320

Hu, L., Li, C., Wang, W., Pang, B., & Shang, Y. (2022). Performance evaluation of text augmentation methods with BERT on small-sized, imbalanced datasets. *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*, 125–133. https://doi.org/10.1109/CogMI56440.2022.00027

Ibrahem Al-Obaydy, W. N., Hashim, H. A., AbdulKhaleq Najm, Y., & Jalal, A. A. (2022). Document classification using term frequency-inverse document frequency and K-means clustering. *Indonesian Journal of Electrical Engineering and Computer Science*, *27*(3), 1517–1524. https://doi.org/10.11591/ijeecs.v27.i3.pp1517-1524

Jiang, S., Hu, J., Magee, C. L., & Luo, J. (2024). Deep learning for technical document classification. *IEEE Transactions on Engineering Management*, *71*, 1163–1179. https://doi.org/10.1109/TEM.2022.3152216

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2019). *Is BERT really robust? A strong baseline for natural language attack on text classification and entailment*. arXiv preprint. http://arxiv.org/abs/1907.11932

Ke, W., Tian, Z., Liu, Q., Wang, P., Gao, J., & Qi, R. (2023). Towards incremental NER data augmentation via syntactic-aware insertion transformer. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 5104–5112. https://doi.org/10.24963/ijcai.2023/567v

Koopmans, L., Dhali, M., & Schomaker, L. (2023). The effects of character-level data augmentation on style-based dating of historical manuscripts. *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods*, pp. 124–135. https://doi.org/10.5220/0011699500003411

Labd, Z., Bahassine, S., Housni, K., Aadi, F. Z. A. H., & Benabbes, K. (2024). Text classification supervised algorithms with term frequency inverse document frequency and global vectors for word representation: A comparative study. *International Journal of Electrical and Computer Engineering*, *14*(1), 589–599. https://doi.org/10.11591/ijece.v14i1.pp589-599

Li, D., Xiong, Y., Hu, B., Tang, B., Peng, W., & Chen, Q. (2021). Drug knowledge discovery via multi-task learning and pre-trained models. *BMC Medical Informatics and Decision Making*, *21*. https://doi.org/10.1186/s12911-021-01614-7

Li, Y., & Rockinger, M. (2024). Unfolding the transitions in sustainability reporting. *Sustainability (Switzerland)*, *16*(2). https://doi.org/10.3390/su16020809

Lu, J., Henchion, M., Bacher, I., & Mac Namee, B. (2021). *A sentence-level hierarchical BERT model for document classification with limited labelled data*. arXiv preprint. http://arxiv.org/abs/2106.06738

Ma, K., Tan, Y., Xie, Z., Qiu, Q., & Chen, S. (2022). Chinese toponym recognition with variant neural structures from social media messages based on BERT methods. *Journal of Geographical Systems*, *24*(2), 143–169. https://doi.org/10.1007/s10109-022-00375-9

Mehra, S., Raut, G., Purkayastha, R. D., Vishvakarma, S. K., & Biasizzo, A. (2023). An empirical evaluation of enhanced performance softmax function in deep learning. *IEEE Access*, *11*, 34912–34924. https://doi.org/10.1109/ACCESS.2023.3265327

Mohiuddin, T., Bari, M. S., & Joty, S. (2021). AugVic: Exploiting BiText vicinity for low-resource NMT. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3034–3045. https://doi.org/10.18653/v1/2021.findings-acl.267

Mursari, L. R., & Wibowo, A. (2021). The effectiveness of image preprocessing on digital handwritten scripts recognition with the implementation of OCR Tesseract. *Computer Engineering and Applications Journal*, *10*(3), 177–186. https://doi.org/10.18495/comengapp.v10i3.386

Özçift, A., Akarsu, K., Yumuk, F., & Söylemez, C. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): An empirical case study for Turkish. *Automatika*, *62*(2), 226–238. https://doi.org/10.1080/00051144.2021.1922150

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Röttger, P., & Pierrehumbert, J. (2021). Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2400–2412. https://doi.org/10.18653/v1/2021.findings-emnlp.206

Şahin, G. G. (2022). To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP. *Computational Linguistics*, *48*(1), 5–42. https://doi.org/10.1162/coli_a_00425

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A review of applications in sentiment analysis. *HighTech and Innovation Journal*, *4*(2), 453–462. https://doi.org/10.28991/HIJ-2023-04-02-015

Schick, T., & Schütze, H. (2020). Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8766–8774. https://doi.org/10.1609/aaai.v34i05.6403

Sharma, S. K., Sharma, N. K., & Potter, P. P. (2020). Fusion approach for document classification using random forest and SVM. *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, 231–234. https://doi.org/10.1109/SMART50582.2020.9337131

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00492-0

Su, Z. (2024). Applications of BERT in sentimental analysis. *Applied and Computational Engineering*, *92*(1), 147–152. https://doi.org/10.54254/2755-2721/92/20241711

Tran, Q. D. L., & Le, A. C. (2023). Exploring bi-directional context for improved chatbot response generation using deep reinforcement learning. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13085041

Vu, V.-H., Nguyen, Q.-P., Tunyan, E. V., & Ock, C.-Y. (2021). Improving the performance of Vietnamese–Korean neural machine translation with contextual embedding. *Applied Sciences*, *11*(23), 11119. https://doi.org/10.3390/app112311119

Wang, Z. (2024). Probabilistic linguistic knowledge and token-level text augmentation.. In M. Abbas (Ed.), *Practical solutions for diverse real-world NLP applications. Signals and communication technology*. Springer. https://doi.org/10.1007/978-3-031-44260-5_1

Wei, J., & Zou, K. (2019). *EDA: easy data augmentation techniques for boosting performance on text classification tasks.* arXiv preprint. http://arxiv.org/abs/1901.11196

Winstanley, D., & Cunningham, C. (2023). A descriptive literature review of recognition of prior learning for vocational learners in emergency medical care in South Africa. *South African Journal of Higher Education*. https://doi.org/10.20853/37-4-5313

Yang, T., Sucholutsky, I., Jen, K.-Y., & Schonlau, M. (2024). exKidneyBERT: A language model for kidney transplant pathology reports and the crucial role of extended vocabularies. *PeerJ Computer Science*, *10*, e1888. https://doi.org/10.7717/peerj-cs.1888

Zaiton, H., & Al-Ansary, S. (2024). Natural language processing approaches to text data augmentation: A computational linguistic analysis. *International Journal of Arabic-English Studies*. https://doi.org/10.33806/ijaes.v25i1.682

Zamsuri, A., Defit, S., & Nurcahyo, G. W. (2023). Classification of multiple emotions in Indonesian text using the K-Nearest neighbor method. *Journal of Applied Engineering and Technological Science (JAETS)*, *4*(2), 1012–1021. https://doi.org/10.37385/jaets.v4i2.1964

Zhang, P., Ma, Z., Ren, Z., Wang, H., Zhang, C., Wan, Q., & Sun, D. (2024). Design of an automatic classification system for educational reform documents based on naive bayes algorithm. *Mathematics*, *12*(8), 1127. https://doi.org/10.3390/math12081127

Zheng, C., Wang, Z., & He, J. (2022). BERT-based mixed question answering matching model. *2022 11th International Conference of Information and Communication Technology (ICTech))*, pp. 355–358. https://doi.org/10.1109/ICTech55460.2022.00077

AUTHORS

I Gede Susrama Mas Diyasa received a doctoral degree in electrical engineering at the Sepuluh Nopember Institute of Technology Surabaya, specializing in intelligent systems. He is attached to the School of Magister Information Technology, Faculty of Computer Science, University of Pembangunan Nasional Veteran Jawa Timur. His research fields include intelligent systems, especially in the biomedical field. Several studies focus on developing technology and algorithms that can be used to improve diagnosis and treatment in the medical field, as well as several studies related to facial detection in smart cities based on intelligent systems.

Eva Yulia Puspaningrum received her Bachelor's and Master's degrees in Information Technology from National Development University "Veteran" East Java and Sepuluh Nopember Institute of Technology, respectively. She is a lecturer at the School of Informatics, Faculty of Computer Science, University of Pembangunan Nasional Veteran Jawa Timur. Her research interests include intelligent computing, visual computing, information retrieval, and data mining.

Dimas Saputra is a final-year student in Informatics at the Faculty of Computer Science, University of Pembangunan Nasional Veteran Jawa Timur. Currently in his final semester, he is dedicating his efforts to his final project. Dimas research interests are centered on machine learning, particularly in text-related applications such as text classification and Named Entity Recognition (NER).

Wan Suryani Wan Awang received her PhD from Cardiff University, UK, MSc from Universiti Malaysia Terengganu (UMT), Postgraduate Diploma in Advanced Computing from Bristol University, and BSc from Sheffield Hallam University. She is a Senior Lecturer at the School of of Informatics and Computing, Universiti Sultan Zainal Abidin Besut Campus. Her research interests include database systems, cloud computing, and grid computing.