A Web-Based Intelligent System for Automated Faculty Document Evaluation

Marquez, Jethro M., Santos, Gabriel Scott., Sotomayor, Rolan C., Villarica, Mia V

College of Computer Studies Laguna State Polytechnic University Laguna, Philippines

DOI: https://doi.org/10.47772/IJRISS.2026.100600063

Received: 14 May 2026; Accepted: 19 May 2026; Published: 17 June 2026

ABSTRACT

The promotion of faculty in Philippine State Universities and Colleges is guided by National Budget Circular 461 which requires a lot of documentation within certain Key Result Areas. The process of evaluating these documents manually is tedious and error-prone, leading to long administrative delays. To automate this process, this study developed "A Web-Based Intelligent System for Automated Faculty Document Evaluation." Employing a developmental and experimental design, the researchers trained models on datasets gathered from twenty faculty members.

The system employs PyMuPDF and DocTR OCR for text extraction and a hierarchical BERT-based transformer for classification. A hybrid approach is employed for data parsing i.e. Regular Expressions for structured documents and Large Language Models for complex narratives. The results demonstrate BERT as the most suitable model with an accuracy of 98.90% for the Key Result Area and 98.69% for Criterion classification.

Granular Sub-Criterion accuracy was achieved at 92% with data augmentation and oversampling. The system is integrated with an interactive dashboard that provides real-time gap analysis and actionable recommendations. This system finally automates the assessment of academic performance while guaranteeing compliance with the institutional guidelines

Keyword: Document Classification, Transformer-based Models, Natural Language Processing (NLP), Key Result Areas (KRAs), National Budget Circular 461 (NBC 461)

INTRODUCTION

Faculty promotion systems are used not only as a mechanism for recognition but also as a motivation for educators to continually enhance their professional competencies. In the Philippines, National Budget Circular (NBC) 461 Uses the standardized framework for evaluating and ranking faculty members in State Universities and Colleges (SUCs). Faculty members must substantiate their accomplishments by submitting documentation categorized under specific Key Result Areas (KRAs). But despite the comprehensive guidelines provided in the Joint Circular, the current evaluation methodology remains manual.

To tackle this challenge, the researchers propose the development of Web-based Intelligent Faculty Document Evaluation System designed to automate the classification and scoring workflows. With the use of Transformer-based models, the system is engineered to accurately categorize documents into the correct KRAs and provides the right insights and recommendations to faculty.

Research shows the Transformer-based models in structured text analysis. Shaheen et al. (2020) Shows excellent performance in legal document classification. Studies on long-document processing, such as Park et al. (2022), proves that with the use of proper techniques, Transformers can manage classification of long and detailed texts. The combined approach of Transformer models with rule-based logic have also been proven to be successful. Alix et al. (2022) it shows how systems enforce consistency and accuracy in educational assessments.

Research Objectives

To develop a web-based system that automates the evaluation of faculty documents based on the JC guidelines by enabling efficient document uploading, classification, and organization under the appropriate Key Result Areas (KRAs).

Specifically, this research aims to:

1. To collect and analyze faculty documents into a structured and annotated dataset for model training and evaluation.
2. To design and implement a document upload and text extraction module using PyMuPDF and DocTR OCR for processing documents.
3. Experiment several transformer models including BERT, DistilBERT, and RoBERTa for classification of faculty documents to their proper KRAs.
4. To analyze the performance of the developed transformer-based classifier model with the help of performance indicators such as:

4.1. Accuracy

4.2. Precision

4.3. Recall

4.4. F1-Score

To develop a hybrid information extraction component by using Regex for extracting information from structured documents and LLMs for handling unstructured layout information.

To implement the most suitable classification algorithm and information extraction component within a web application that would log the evaluation results into Google Sheets using Google AppScript.

To test the effectiveness of the implemented model and the web-based intelligent faculty document evaluation system.

RELATED LITERATURE

Faculty evaluation is a process of measuring and assessing a university instructor's overall performance. It serves two primary functions: (1) to improve teaching quality and student learning outcomes, and (2) to ensure accountability for promotions, tenure, and institutional standards. Faculty evaluation is defined as the process of measuring, analyzing, and providing feedback on faculty members' job performance (Institute for Academic Leadership, 2022). Apart from evaluating faculty members individually, faculty evaluation systems set out a systematic framework for rewarding accomplishments and ensuring quality outcomes within institutions of higher learning. With the increasing demands for higher education institutions to produce highly qualified and internationally competitive faculty members, faculty evaluation systems must continue evolving to meet new challenges.

While such requirements to make assessment processes more stringent have been raised, another problem that involves dealing with academic documents has led to a change towards more automated document classification tools. The fact is that manual classification is not effective enough as checking papers manually is time-consuming and error-prone; moreover, it becomes harder to handle the increasing amount of paperwork due to the workload of schools and universities. That is why the use of software capable of reading academic documents and sorting them out according to their patterns became a necessity (Irin P P, 2024).

However, to successfully classify these documents, one must possess models that are able to analyze unstructured data, something that is beyond the capabilities of existing models. For example, Younus et al. (2023), in the context of Arabic handwriting recognition, were able to show that sequential models such as CNN-BiLSTM become inefficient when analyzing unstructured information. Unlike other types of models,

transformer models are well-suited to cope with such complexity, thanks to their ability to weigh context-based relations in whole sequences. The importance of this aspect becomes particularly prominent in analyzing documents containing much descriptive language.

Besides dealing with the unstructured text in documents, an efficient document classification system needs to incorporate a structural consideration in its design. Xu et al. (2025) tackled this issue by presenting a technique for joint embedding that can incorporate information about both the text and structure of a document. Using hierarchical position encoding, the model is able to consider the importance of headings, tables, and layout features that can be used effectively within faculty evaluation tools, which need to evaluate the document's content and organization.

It is equally critical to consider the efficiency of the feature extraction technique employed in transforming plain text data into features that machine learning algorithms can comprehend. Traditional approaches like TF-IDF compute the relevance of words by considering their frequencies in the entire dataset (Lesot et al., 2020). However, traditional approaches treat words as individual elements without considering the meaning of the words within the context. The current study considers dense word embeddings using the attention mechanism of BERT.

With this theoretical and technological background, we will present the development of a web-based tool that uses the transformer algorithm to automatically categorize academic papers based on their relevant Key Result Area (KRA). The reason behind developing a web-based application is to ensure that the tool is easily accessible and collaborative in nature, according to research conducted by Al-Sharhan et al. (2021). They showed that web-based applications play a critical role in ensuring increased user interactions and ease of use when working in shared organizational environments. Furthermore, for extracting the content of the various formats of the documents used by academia, text extraction and OCR tools have been incorporated.

Finally, the process of development of the system was carried out using the methods of both developmental and experimental research along with the application of a project management framework based on the concept of Kanban. This approach enabled the researchers to develop a system as well as test its effectiveness at the same time.

METHODOLOGY

Research Design

Phase 1: Model Evaluation

Both experiments and developmental research employ a research design to overcome the obstacles encountered in the evaluation of faculty documents.

The developmental research design is used in guiding the systematic development of the system. As Reeves (2006) indicates, a developmental research design is appropriate when the objective is the design and development of innovative technologies, models, or systems for problem-solving purposes. In the present case, the development of a web portal with Transformer model classification of uploaded faculty documents according to their Key Result Areas (KRAs) constitutes an innovation.

The research design that was used in this study involved experimentation whereby the performance of the three transformer-based models, namely BERT, DistilBERT, and RoBERTa, was evaluated and compared based on their accuracy and efficiency. Experimental research design as described by Creswell and Creswell (2018) refers to systematic investigations about the causality between two variables where one or more variables are manipulated to observe the effect on another variable. In this study, the intervention was carried out through the implementation of the three models, and the effects were measured using metrics like accuracy, precision, recall, and F1 score.

Phase 2: System Validation

The researchers applied the method of purposive sampling to generate 10 test batches having diverse academic content. This will ensure that the system can correctly identify, count, and classify required documents. This formula will be employed by the investigators to determine the accuracy of the system.

$$\text{Accuracy} = \frac{\text{CorrectItems}}{\text{TotalExpectedItems}} \text{ x } 100$$

Equation 1: Accuracy system validation

This evaluates each test file on whether it passes or fails based on user-defined levels of accuracy. When the accuracy is ≥80%, it indicates that the batch under test has passed, while when the accuracy is <80%, it indicates that the batch has failed.

Table 4: System Validation Threshold

Categorization	**Threshold**
Pass	≥ 80%
Fail	<80%

The output generated by this full flow testing is noted to provide functional assurance of the system prior to its implementation through documentation of the expected output against the system output to arrive at the final accuracy percentage.

Applied Concepts and Techniques

This section outlines the concepts and methods employed by the researchers to address the research problem and meet the study's objectives. The following concepts and techniques were utilized in this research:

Natural Language Processing (NLP)

It is extremely hard to connect human language with machines. That is why Natural Language Processing is so challenging. Souza et al. (2020) have gone one step further in their quest to explore the capabilities of BERT models when processing clinical data. Specifically, they wanted to find medical concepts and textual associations in clinical notes. And what they found out was truly revolutionary. No matter what generic model you use, it will not help you achieve your goal in the clinical setting unless you train it on your specific dataset. The same idea is expressed by Zhang et al. (2025), who, in their large-scale research paper, prove that all the older models like the bag of words method or TF IDF are no longer relevant. Thanks to the introduction of new pretrained architectures like RoBERTa and GPT, we can finally talk about machines understanding context.

Transformer Model

The Transformer model, first introduced by Vaswani et al. (2017) in the seminal paper “Attention Is All You Need,” revolutionized sequence-to-sequence tasks by replacing recurrence with self-attention mechanisms. This architecture enabled faster training and improved performance on tasks like machine translation. The study showed that Transformers outperformed traditional RNNs and LSTMs by allowing models to focus on different parts of a sentence simultaneously, facilitating better context understanding and scalability. Building on this, Ren et al. (2023) conducted a comprehensive survey of how Transformer models have been adopted across domains such as NLP, computer vision, and audio processing. Their findings reveal that the attention mechanism and parallelization properties of Transformers have made them the architecture of choice for a wide range of deep learning applications. The studies collectively illustrate how the Transformer has become the foundation for cutting-edge models like BERT, GPT, and Vision Transformers, setting a new paradigm in artificial intelligence research and deployment.

Evaluation Point System

Table 1. Point system

Key Result Areas	Pts	Maximum Pts
1. Instruction		100
a. Teaching Effectiveness	60	
b. Curriculum and Instructional Materials Development	30	
c. Thesis, Dissertation and Mentorship Services	10	
2. Research, Invention and Creative Work		100
a. Research Outputs	100	
b. Inventions	100	
c. Creative Works	100	
3. Extension Services		100
a. Service to the Institution	50	
b. Service to the Community	30	
c. Extension Involvement	20	
4. Professional Development		100
For All Faculty		
a. Involvement in Professional Organizations	20	
b. Continuing Development	60	
c. Awards and Recognitions	20	
***For New Entrants Only**		
a. Academic Experience	10	
b. Industry Experience	10	

The evaluation criteria are grouped into four (4) key results areas (KRAs), namely: 1) Instruction; 2) Research, Invention and Creative Work; 3) Extension; and 4) Professional Development.

The weights assigned to each KRA vary from one faculty rank to another, depending on the functions that are expected of them.

Table 2. KRA Weights per Faculty Rank

Faculty Rank	Key Result Areas (KRAs)			
	Instruction	Research, Innovation and/or Creative Work	Extension	Professional Development
Instructor (I-III)	60%	10%	20%	10%
Asst. Professor (I-IV)	50%	20%	20%	10%
Assoc. Professor (I-V)	40%	30%	20%	10%
Professor (I-VI)	30%	40%	20%	10%
Col./Univ. Professor	20%	50%	20%	10%

As the faculty rank advances, more research, invention, and creative work outputs are expected. For instance, faculty occupying Instructor positions are expected to display exemplary performance in the area of instruction. However, this should not prevent them from doing their research or extension function. On the other hand, those occupying Professor positions are expected to produce more research, invention and creative work; but these should not sacrifice the quality of instruction.

Table 3. Score Bracket and Corresponding Number of Sub-rank Increment

Score Bracket	No. of Sub-rank Increment
41-50	1 sub-rank
51-60	2 sub-ranks
61-70	3 sub-ranks
71-80	4 sub-ranks
81-90	5 sub-ranks
91-100	6 sub-ranks

Table 3 shows the number of sub-rank increments that may be awarded based on the total score received by the faculty. The faculty may be awarded a maximum of six (6) sub-ranks.

Data Collection Methods

The data collection process included the systematic extraction and organization of key attributes, document_text, KRA, Criterion, and sub_criterion which were used as the main variables for classification. These organized datasets served as the foundation for model development, testing, and training, ultimately enhancing the system's accuracy and effectiveness in evaluating faculty documents.

Documents for Classification

The dataset was obtained in raw form from the faculty position reclassification documents and processed to extract text and assign appropriate labels. The data are organized into 8 variables that will be used for KRA classification:

Table 4. Datasets for classification

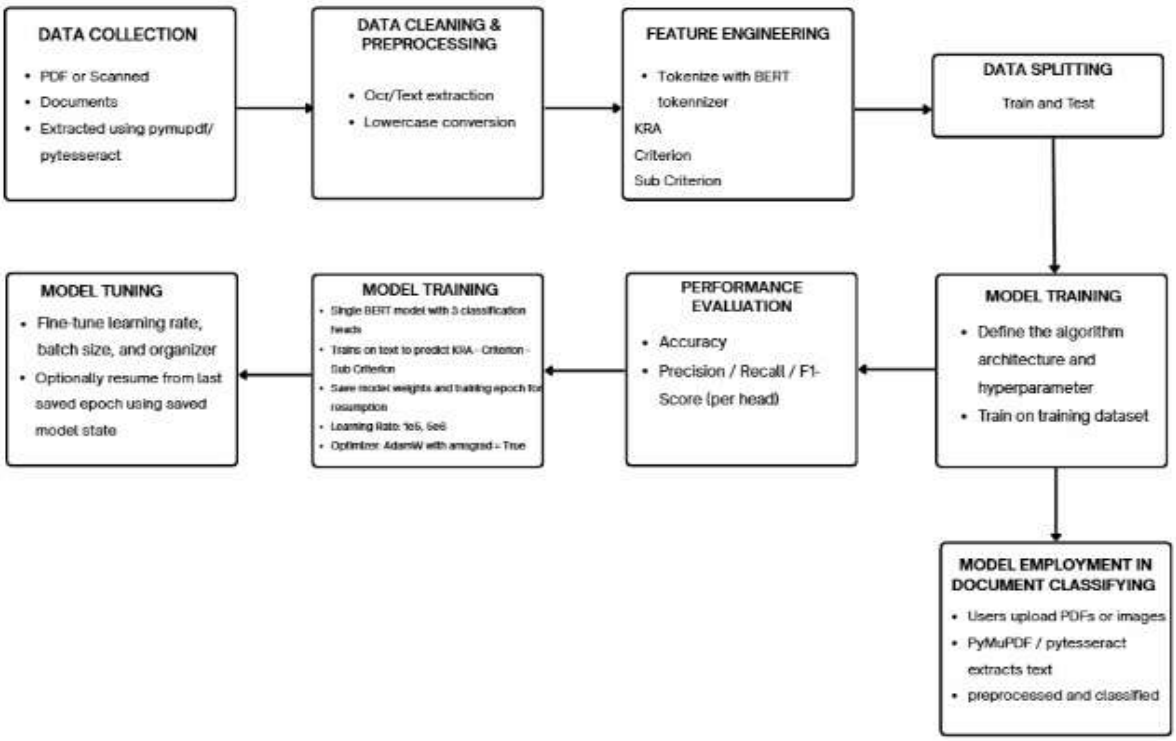
ATTRIBUTES	DESCRIPTION	VALUES
Document_Text	Textual content from documents	Covid-19 brought a significant impact on Philippine higher education institutions
KRA	Key result assigned to documents	KRA 2
Criterion	Specific criterion under the assigned KRA	Criterion A
Sub_Criterion	Specific item evaluated under the KRA.	1.1
cri_description	Description of the sub criterion.	Student's Evaluation
document_type	Category of the submitted document.	Research Output
author_name	Name of the document author.	Juan Dela Cruz
Contribution	A description of the role or effort made by the author.	50%

The researchers gathered documents from 20 faculty members, providing a high-quality dataset for training and testing the transformer model to achieve accurate results. Each document was thoroughly reviewed, organized, and encoded to ensure that all key information was captured. This well-prepared dataset helped the model recognize important patterns and improved the overall performance and reliability of the web-based intelligent evaluation system.

Data Model Generation

Figure 1 represents the entire process through which the intelligent document classification system has been developed. Data modeling process includes several critical stages like document input, data preparation, data labeling, training of the model, and testing or validation of the model among others. All these processes are very critical in order to ensure that the model can effectively classify the uploaded documents.

Figure 1: Data Model Generation Process of this study



Data Acquisition

In this part, we provide details about the methods of gathering datasets to build the model. For acquiring the dataset, we gathered information from the faculty members. The dataset consists of files in various formats, such as word, pdf, scanned pdf, and png, and it comprises a total of 2546 files. In consideration of ethical issues, the information provided by the faculty members will be stored in secure places.

Data pre-processing

This entails ensuring that faculty documentation, which comes in many different forms and varies significantly in terms of structure and language, is curated and cleaned up before applying machine learning algorithms. The raw data, collected from faculty promotion documents through PDF parsing and OCR technology, was subjected to a process of validation whereby any data record with incomplete information in vital fields like document_text, KRA, criterion, and sub_criterion was removed. This ensured that the subsequent labeling would not be affected by noise that could potentially lower the accuracy and fairness of the models. We used oversampling in conjunction with a weighted loss function due to class imbalance and high difficulty of sub_criterion classification.

Splitting the Data

We divided our dataset into three groups, each containing 70%, 15%, and 15%. The first 70% will be used by our machine learning model for training purposes. Our model needs sufficient data for its training process. Therefore, we decided to keep the remaining 30% aside for validation and testing purposes. However, this is not the only reason why we have divided our dataset in this manner. In addition, we used stratified sampling while splitting our dataset. This technique was used in order to prevent our minority classes from being marginalized.

Building and Training of Model

After splitting the data, the next step was designing the multi-head BERT structure. In particular, we sought to construct a machine that would be able to manage three different labels simultaneously. During training, batches of text were fed into the system, and then optimization was performed with backpropagation algorithms. To optimization, we employed the AdamW optimization function with a very carefully tuned learning rate of 2e 5. With such a fine setting, the machine is able to learn without ever overfitting or falling into any local minima. It is important to note that our data suffered from imbalance. To address this problem, we opted to use loss functions for weights. These weight functions would penalize the model for errors in sub-criteria prediction.

Evaluate Performance

However, accuracy alone is insufficient, particularly in cases of difficult classification. It is important to use the complete set of measures to put the model through the wringer to assess its performance. Precision, recall, and F1-score were used to determine where any potential logic could be failing. Accuracy provides us with a single data point, while the F1-score provides us with the whole picture. It serves as a means of harmonizing precision and recall, ensuring that we do not have to sacrifice sensitivity to maintain a certain level of accuracy.

Model Integration

The BERT Multihead Classification model was chosen as the model to deploy on the web-based intelligent faculty document assessment system. This included converting the saved PyTorch model's state dictionary together with its tokenizer and label encoders into a form that could be easily consumed by backend services. To ensure faster inference capabilities for real-time predictions, the conversion of model components was done efficiently.

RESULTS AND DISCUSSIONS

The findings of the current research and the aims stated at the start of the paper have been discussed in detail. The techniques used and their outcome with regard to the integration of the model into the system have also been considered.

Research Objective 1. To preprocess and to transform faculty documents into a structured dataset suitable for training and evaluating the models. The researchers organized the faculty documents into standardized categories to maintain a clean workflow and ensure consistency during the text extraction process.

The folder hierarchy adopted for document management within the repository. In this approach, the root folder, titled 'Spreadsheet,' is utilized and has individual folders that are sorted based on the faculty name. Within each unique faculty's folder, the files are classified further into individual folders depending on their Key Result Areas (KRAs).

In the extraction process, text extraction is performed using PyMuPDF (fitz), while the optical character recognition (OCR) process is performed using DocTR OCR.

The extract_text_hybrid function serves as an automated decision-making tool that determines the best way to pull information from faculty documents. It first attempts to use PyMuPDF to directly copy text from digital files, which is fast and highly accurate. If the function detects that a page is blank or contains very little text, it is detected as a scanned image or a photocopy, it automatically switches to DocTR OCR. During this fallback process, the system converts the page into a high-quality image so the OCR engine can "read" the visual

characters and convert them into editable text. This two-step approach ensures that no data is lost, regardless of whether the original document was a digital file or a scanned piece of paper.

To address class imbalance within the dataset, the researchers implemented a Random Oversampling technique to ensure that each document category is adequately represented for model training.

Figure 2. Data augmentation and Oversampling code



First, it is important to mention that the name of the class of data in our case includes all three components of the Key Result Area, criterion, and subcriterion, which constitute one common term. In cases when the number of samples was insufficient, the authors used a Surgical Preservation Augmentation method for achieving the required amount of 50 samples in each class. The described technique does not imply duplication only; rather, it creates novel samples based on existing data using a Guideline Aware Filter, which makes possible preserving technical terms and varying other elements of terminology.

Research Objective 2. To accurately classify uploaded faculty documents into their Key Result Areas (KRAs) by applying a transformer-based model. The researchers conducted experiments with three state-of-the-art transformer models: BERT, DistilBERT, and RoBERTa. These models were selected to provide a comparative analysis between established benchmarks and optimized variants. To ensure a fair and unbiased comparison, the researchers maintained consistent experimental architecture across all three models.

The architecture consists of an HMTL system, based on a shared transformer backbone that provides deep semantic representation learning from faculty data. Leveraging one encoder to generate representations from the text input ensures that the model obtains a comprehensive knowledge representation that will then be allocated among three separate classifiers. These classifiers represent the Key Result Area, the Criterion, and the Sub criterion levels. Such an HMTL solution ensures that the model optimizes all three tasks at once, meaning that the encoder representations will be impacted by both general and detailed categorical knowledge.

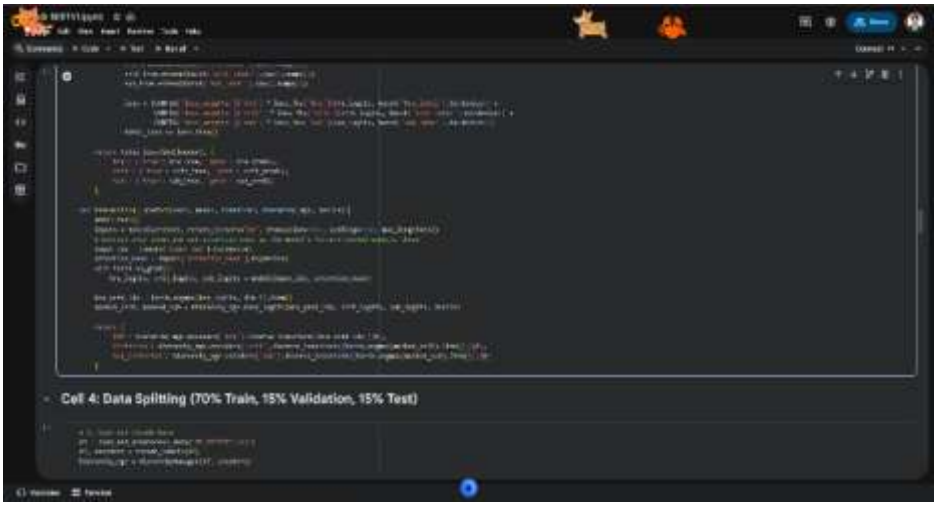
The structural consistency of the resulting output classes is ensured through the presence of a dedicated Hierarchy Manager in the proposed architecture, which imposes logical restrictions during the inference process. The Hierarchy Manager makes use of the logit masking technique, preventing the network from producing illogical paths inside the faculty hierarchy graph. This is achieved by restricting possible options for selecting the criterion and sub-criterion values based on the previously selected parent category. In effect, the inclusion of a dedicated Hierarchy Manager during inference ensures that all output categories are strictly consistent with the organizational structure.

Weighted loss functions are applied across levels to fine-tune how much we want to learn at each level. Each task (loss function) is multiplied by the weight of 0.2, 0.3, and 0.5 for Key Result Area, Criterion, and Sub criterion respectively. This amplifies the focus on learning the lowest level categories as we want to optimize for the model being able to perform at the most detailed classification level.

This prevents the model from getting a high overall score by being able to classify the high-level categories and being wrong at sub-classification. The weighted loss functions compliment the class weights as it also penalizes mistakes made on minority classes.

To make sure that the model does not provide inaccurate results, the researchers implement a hierarchical predict function.

Figure 3. hierarchical predict function



The hierarchical_predict function serves as the intelligent decision-making component that converts processed document text into a finalized set of labels. After the model analyzes the text and produces raw scores for all three levels, the function uses a specialized masking process to ensure the results are logically consistent. It first identifies the correct Key Result Area and then restricts the choices for the Criterion and Sub criterion levels so that only valid child categories can be selected. This prevents the system from making impossible combinations and ensures that the final output perfectly matches the JC Guidelines.

Research Objective 3: To the design and development considerations for an AI-based system for automating the evaluation of faculty documents, and to test the system perform in relation to DBM-CHED-JC-NO-1-S-2022 recommendations in terms of functional reliability and user acceptability. The researchers determined that BERT had the most superior performance among the three assessed architectures. A thorough evaluation of important performance metrics like accuracy, precision, recall, and the F1 score led to this decision.

Table 4. Model performances metrics

Models	KRA ACCURACY	CRITERION ACCURACY	SUB_CRITERION ACCURACY
BERT	0.9890	0.9869	0.92
DISTILBERT	0.97	0.97	0.78
ROBERTA	0.98	0.9674	0.87

The experimental results indicate that BERT achieves performance equality with RoBERTa in terms of base accuracy. This is a significant finding, as RoBERTa utilizes a substantially larger architecture and more extensive pre-training. Despite its smaller parameter size, BERT matches RoBERTa's predictive power while offering faster inference speeds and lower computational overhead. DistilBERT demonstrated the lowest performance, particularly within the sub-criterion classification.

This means that the compression methods used in DistilBERT make it harder to understand the subtle differences in language that are needed to deal with the high level of difficulty of the sub-criterion task. BERT is the best model for this application because it strikes the best balance between speed, efficiency, and accuracy in classification.

Figure 4. BERT performance metrics

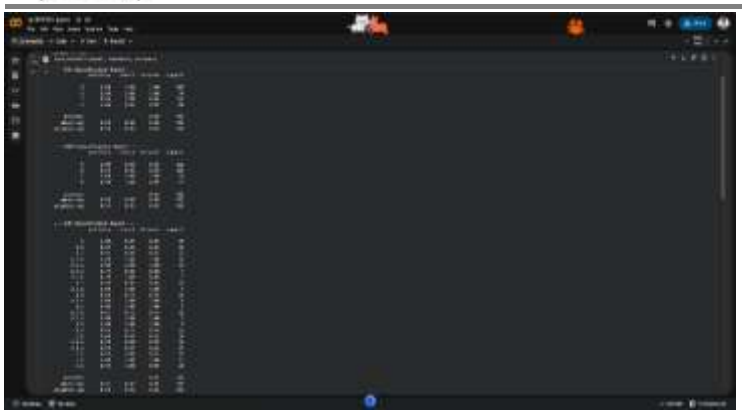


Figure 4 shows that the BERT based hierarchical classification framework achieves high levels of predictive reliability across all levels of categorization, improved by the strategic implementation of an oversampling technique. The KRA and CRIT reports demonstrate strong performance with overall accuracies of 9890% and 9869% respectively, illustrating the model's proficiency in capturing broad semantic relationships within the dataset.

For the more detailed SUB_CRI classification, oversampling was necessary to fix the class distribution. This raised the accuracy to 92% and made sure that the model could still identify items in different categories. This discovery confirms that BERT's bidirectional embeddings combined with proper data balancing work well to ensure classification accuracy regardless of the growing complexity of the hierarchy and, thus, provides an effective platform for classification automation.

Taking into account all the above-discussed data, it can be said that BERT is the most appropriate choice for the task at hand. In particular, the team developed a hybrid extraction methodology specially designed to provide metadata output into Google Sheets.

Figure 5. Information Extraction code

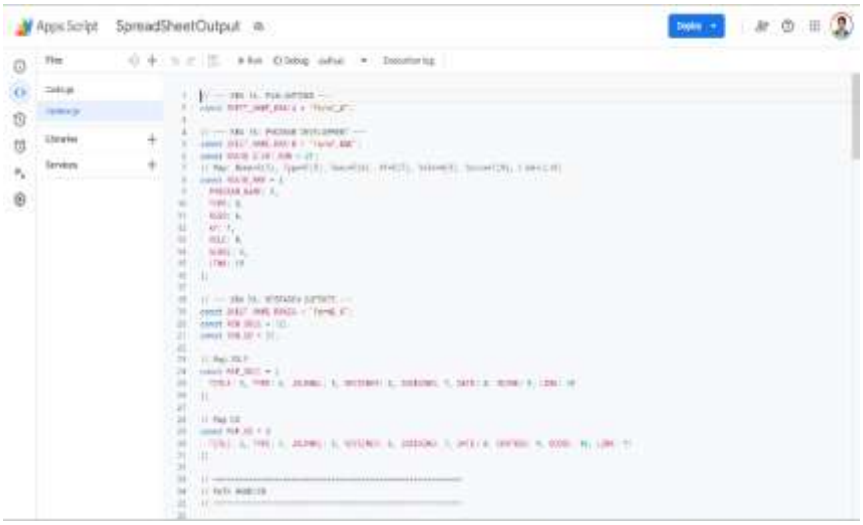


To gather the information, the authors adopted a combination technique, determining which would be most effective depending on the complexity of the document. Regex was chosen for documents of lesser complexity, where the structure and pattern of the document were relatively predictable. This allowed for quick and accurate extraction of information without extensive computations.

The extraction process transitioned into using a Large Language Model (LLM) for documents of higher complexity, containing unstructured narratives or varying formats. This was done to understand the underlying context and semantic meaning behind the text that cannot be achieved through simpler rule-based logical processes.

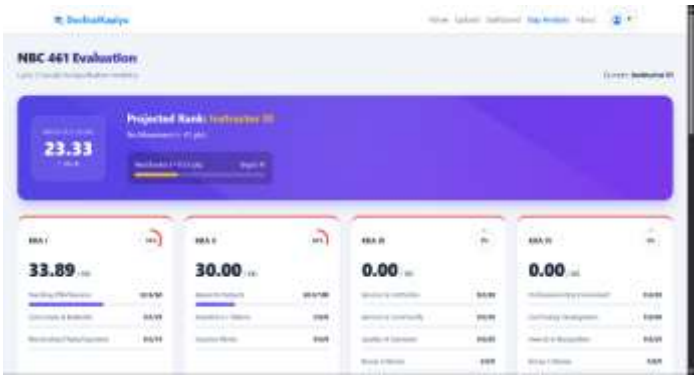
Once the metadata is extracted via the Regex or LLM modules, the system automates the data entry process into a centralized Google Spreadsheet. This integration is facilitated by Google Apps Script.

Figure 6. Google Apps Script Code



As shown in Figure 6, the Google Apps Script was used by the researchers to automate the backend process of the system and synchronize the information on it. The code connects the extraction engine to the user interface and automatically places the extracted metadata in the proper cells of the spreadsheet. In addition to that, there is a feature of dynamic provision in the code. Whenever a new user signs into the system, a customized spreadsheet will be provided to him/her automatically. Therefore, the entire process, starting from signing in to storing the data, is completely automated and scalable. Also, data isolation will be maintained for each user. The researchers implemented a system of centralized administration by ensuring that all users' spreadsheets are stored in one secured Google Drive directory. As a result of putting all the files in one drive, evaluators can access the faculty documents easily.

Figure 7. Gap Analysis UI



The software includes an instantaneous gap analysis function to evaluate how the consumer's qualifications compare to the standards set for the promotion of professors. This process identifies weaknesses in the user's performance and areas where there is a lack of compliance with requirements for the higher grade by conducting an analysis of the input data compared to the highest score possible for every Key Result Area (KRA). This immediate feedback loop ensures that consumers remain aware of their standing and any factors influencing their grades.

Figure 8. Recommendation UI



By suggesting actions based on gathered information, the algorithm helps the user maximize his or her reclassification scores. As opposed to simply rating a candidate in terms of a numerical value, the application offers suggestions for further action, such as focusing on high-yield research outputs and formalization of extension activities, which will address the identified gaps. Thanks to this function, the algorithm transforms the process of evaluation into a strategic one, where faculty members' initiatives can be directly tied to promotion and career growth.

Thanks to thoughtful choice of the architecture of the algorithm, namely, selection of the BERT model, development of hybrid extraction framework, as well as inclusion of the Google Sheets infrastructure for automated data collection, the researchers designed a coherent system with multiple strengths. The use of gap analysis and data-driven suggestions adds additional practical value to the system, making it more active rather than passive. Overall, these implementations, as well as architectural solutions, enable the accomplishment of the central aim, which is the creation and implementation of an AI-powered system for automatic evaluation and classification of faculty documents, adhering to the DBM-CHED-JC-NO-1-S-2022 guidelines.

Table 5. K-Fold Cross-Validation Results

Fold	**Number of Test Items**	**Correct Items**	**Incorrect Items**	**Mean Accuracy**
1	**200**	**172**	**28**	86%
2	**200**	**168**	**32**	84%
3	**200**	**176**	**24**	88%
4	**200**	**164**	**36**	82%
5	**200**	**180**	**20**	90%

The results show that the proposed system consistently classified faculty documents with high accuracy across all validation folds. Variations in performance were mainly caused by overlapping document categories and inconsistencies in extracted text.

To determine the overall classification performance, the mean accuracy was computed using Equation 6.

$$\text{MeanAccuracy} = \frac{86\% + 84\% + 88\% + 82\% + 90\%}{5} = 86\%$$

The mean accuracy of 86% represents the overall expected predictive performance of the system. It signifies that, on average, when the system processes an entirely new and unseen batch of faculty documents, it correctly classifies 86 out of every 100 documents into their appropriate Key Result Areas (KRAs) according to the NBC 461 guidelines.

To evaluate the consistency of the model, the standard deviation was also calculated.

$$\text{SD} = \sqrt{\frac{(86-86)^2 + (84-86)^2 + (88-86)^2 + (82-86)^2 + (90-86)^2}{5-1}} = 3.16\%$$

The standard deviation of **3.16%** indicates that the model maintained stable performance across different testing batches. These findings confirm that the proposed intelligent document evaluation system is reliable, consistent, and effective for automating faculty document classification.

SUMMARY AND CONCLUSIONS

Summary

We develop and evaluate an intelligent web-based solution that automates the faculty performance document classification and scoring task. The solution follows institutional guidelines in scoring and seeks to simplify and speed up the manual process. text extraction was done using PyMuPDF and DocTR OCR pipeline and performing text clean-up and oversampling. We fine-tuned Bert as our main model. We trained it to classify each document at three levels KRA, Criterion and Sub-criterion with weighted loss to assist the model with low frequency categories. Model scores were evaluated using DistilBERT and RoBERTa with evaluation metrics accuracy, precision, recall and F1-Score. Finally, we implemented a Hierarchy Manager within the solution that ensures all predictions fall within policy guidelines and do not generate impossible scores. Our system is deployed using Django REST framework for back-end API logic and React for the front-end. We have developed dashboards for end-users that include faculty members and administrators.

Conclusions

Based on the findings and results of the study "A Web-Based Intelligent System for Automated Faculty Document Evaluation," the researcher reached at the following conclusions:

1. The development and implementation of the web-based intelligent system successfully modernized the faculty evaluation workflow at Laguna State Polytechnic University – Santa Cruz Campus. By transitioning from a labor-intensive manual process to an automated platform, the study effectively addressed long-standing challenges related to manual processing of documents, data entry error, and manual classification of documents.

2. The experimental evaluation of transformer-based models confirmed that BERT architecture is the most optimal choice for classifying complex faculty documents. BERT's performance surpassed other models, achieving 98% accuracy on KRA level classification, 97% accuracy on Criterion level classification, and 87% accuracy on fine-grained Sub-criterion level classification. This performance allowed BERT to reach similar predictive capabilities to heavy-duty models such as RoBERTa, but with increased inference speed and decreased runtime, making BERT ideal for institutional deployment over slower-performing models such as DistilBERT.

3. The integration of a hybrid information extraction framework, combining Regular Expressions (Regex) and Large Language Models (LLMs), demonstrated robust data handling capabilities. Regex allowed accurate extraction of information from standardized documents with defined patterns. LLMs helped understand the context in unstructured documents with varied layouts. Lastly, integration with a centralized Google Sheets backend through Google Apps Script allowed us to automate generation of output into a structured, scalable, and secure dashboard for reviewers to validate professor accomplishments.

4. Beyond mere automation, the system functions as a proactive career development tool by offering real-time gap analysis and recommendations for faculty members. Through the identification of areas for improvement as well as concrete steps towards promotion to a future rank, End/promote allows the user to concentrate their efforts where they are needed most. Transitioning from passive, users are now able to take control of their professional development which allows for meaningful growth of faculty and meritocratic promotion at the university

RECOMMENDATIONS

Based on the findings of the study, the researcher proposes the following recommendations for future development and improvement:

- Future researchers should prioritize the collection of a more diverse and extensive dataset to further enhance the model's ability to classify granular document classes accurately. Continuing to grow this data set will alleviate data imbalance and allow for training on the more granular sub-criterion level, which exhibits the greatest classification difficulty. Additional categories of documents and edge cases will allow the system to function with even greater accuracy and stability.

- It is recommended that the system's infrastructure be upgraded to support institutional-scale deployment by focusing on architectural scalability and high-performance computing. Future versions must consider the issue of privacy more adequately, such as by securing faculty information and requiring passwords. Moreover, it is preferable if the transformation process is faster to minimize waiting periods when dealing with multiple documents simultaneously.

- Though the present research was able to concentrate on Laguna State Polytechnic University-Santa Cruz Campus, the scope of the project must be widened to incorporate all JC guidelines for other SUCs and HEIs in the country. Widening the future scope will allow more space for conducting faculty assessments consistently nationwide. The rationale is adaptable and will take into consideration various institutional policies without violating DBM-CHED-JC-NO-1-S-2022 guidelines.

- In order to improve the user's experience, the software must be enhanced to include an AI-powered chatbot that will serve as a guide for the staff. The bot will enable the user to pose questions on the steps to take in order to receive a promotion, document sorting techniques, and what it takes to climb up the ladder. The chatbot will function as an advisor at all times and give immediate answers, which will simplify the entire evaluation process for the academic.

LITERATURE CITED

1. Azad, M., & Kamarei, M. (2021). A study on the Effect of Telecollaboration through Social Networking on Reading Comprehension and Vocabulary Learning: The Case of Iranian Female High School EFL Learners. Journal of English Language Studies, 6(2), 171.
2. Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? Reading and Writing, 26(3), 349–379.
3. Bashar, A. (2019). Survey on evolving deep learning neural network architectures. Journal of Artificial Intelligence and Capsule Networks, 2019(2), 73–82.
4. Centre for Teaching Excellence. (n.d.). University of Waterloo.
5. Divjak, B., Grabar, D., Svetec, B., & Vondra, P. (2022). Balanced Learning Design Planning: Concept and Tool. Journal of Information and Organizational Sciences, 46(2), 361-375.
6. Flesch Reading Ease and the Flesch Kincaid Grade level. (2024, March 30). Readable.
7. Guala, B. (2024, August 30). Writing effective multiple choice questions. Knowledge Base.
8. Hashem, I., Islam, M., Haque, S. M., Jabed, Z. I., & Sakib, N. (2021). A Proposed Technique for Simultaneously Detecting DDoS and SQL Injection Attacks. International Journal of Computer Applications, 183(11), 50–57.
9. Huang, Y., & Shu, Z. (2022). Construction of Dynamic Multiparallel Foreign Language Teaching Model Based on Multicore Processor. Mathematical Problems in Engineering, 2022, 1–12.
10. Huggins, B. (2022). The College Student's Guide to Better Reading Comprehension.
11. Jiang, C. (2023). Chinese undergraduates' English reading self-efficacy, intrinsic cognitive load, boredom, and performance: A moderated mediation model. Frontiers in Psychology, 14, 1093044.
12. Kärbla, T., Uibu, K., & Männamaa, M. (2021). Teaching strategies to improve students' vocabulary and text comprehension. European Journal of Psychology of Education, 36(3), 553–572.

13. Larioque, R. R. (2019). Millennials' Reading Predicaments: How Are They Treated? International Journal of Social Science Studies, 7(4), 22.
14. Lehmann, T. H. (2022). Making Sense of Algorithms in Discrete Mathematics. International Journal of Science and Mathematics Education, 20(5), 1057–1077.
15. Minkley, N., Xu, K. M., & Krell, M. (2021). Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. Frontiers in Education, 6, 632907.
16. Obeka, Ngwoke, Okonkwo, & Ereke. (2022). Impact of Information and Computer Technology on Students' Reading Comprehension. Journal of Education and Practice.
17. Pellicer-Sánchez, A. (2022). Multimodal reading and second language learning. ITL - International Journal of Applied Linguistics, 173(1), 2–17.
18. Peng, P., Wang, W., Filderman, M. J., Zhang, W., & Lin, L. (2024). The Active Ingredient in Reading Comprehension Strategy Intervention for Struggling Readers: A Bayesian Network Meta-analysis. Review of Educational Research, 94(2), 228–267.
19. Philippine News Agency. (2022). Pandemic, school disruption stunt learners' reading skills.
20. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text (Version 3). arXiv.
21. Sareen, A. (2024, November 7). Flesch Reading Ease: Why Readability Matters in Marketing. ContentWriters Blog.
22. Sarkar, P., Posen, A., & Etemad, A. (2022). AVCAffe: A Large Scale Audio-Visual Dataset of Cognitive Load and Affect for Remote Work. arXiv.
23. Schleicher, A. (2019). PISA 2018: Insights and Interpretations.
24. Sorenson, Duncan, T., Karkada, M., Deacon, S. H., & Smith, I. M. (2021). Building Meaning: Meta-analysis of Component Skills Supporting Reading Comprehension in Children with Autism Spectrum Disorder. Autism Research, 14(5), 840–858.
25. Su, Y. (2021). Does 'WOW' translate to 'an 'A'? Exploring the effects of virtual reality assisted multimodal text on Chinese Grade 8 EFL learners' reading comprehension. Journal of Language Teaching, 1(1), 38–58.
26. UN News. (2021, March 29). 100 million more children fail basic reading skills because of COVID-19.
27. Van Der Merwe, A., Gerber, A., & Smuts, H. (2020). Guidelines for Conducting Design Science Research in Information Systems. In ICT Education (Vol. 1136, pp. 163–178). Springer.
28. Wang, N., & Issa, R. R. A. (2024). Transfer Learning-Based Question Generation for Building a Construction Safety Chatbot. Computer in Civil Engineering 2023.
29. Wijewantha, N. W. S. C. (2021a). Documentaries as Multimodal Texts to Promote Reading in the Virtual Language Classroom: An Experimental Study. Journal of English Language Teaching and Linguistics, 6(3), 633.
30. Xie, T., Cao, M., & Pan, Z. (2020). Applying Self-Assessment Manikin (SAM) to Evaluate the Affective Arousal Effects of VR Games. Proceedings of the 2020 3rd International Conference on Image and Graphics Processing, 134–138.