Beyond BERT: Exploring the Efficacy of RoBERTa and ALBERT in Supervised Multiclass Text Classification

Christian Y. Sy[1], Lany L. Maceda[2], Mary Joy P. Canon[3], Nancy M. Flores[4]
Department of Computer Science and Information Technology-Bicol University,
College of Science, Legazpi City, Philippines[1, 2, 3]
College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, Philippines[4]

***Abstract*—This study investigates the performance of transformer-based machine learning models, specifically BERT, RoBERTa, and ALBERT, in multiclass text classification within the context of the Universal Access to Quality Tertiary Education (UAQTE) program. The aim is to systematically categorize and analyze qualitative responses to uncover domain-specific patterns in students' experiences. Through rigorous evaluation of various hyperparameter configurations, consistent enhancements in model performance are observed with smaller batch sizes and increased epochs, while optimal learning rates further boost accuracy. However, achieving an optimal balance between sequence length and model efficacy presents nuanced challenges, with instances of overfitting emerging after a certain number of epochs. Notably, the findings underscore the effectiveness of the UAQTE program in addressing student needs, particularly evident in categories such as "Family Support" and "Financial Support," with RoBERTa emerging as a standout choice due to its stable performance during training. Future research should focus on fine-tuning hyperparameter values and adopting continuous monitoring mechanisms to reduce overfitting. Furthermore, ongoing review and modification of educational efforts, informed by evidence-based decision-making and stakeholder feedback, is critical to fulfill students' changing needs effectively.**

Keywords*—Multi-class text classification; Bidirectional Encoder Representations from Transformers (BERT); RoBERTa; ALBERT; Universal Access to Quality Tertiary Education (UAQTE) program; educational policy reforms

I. Introduction

The free tertiary education program, referred to as the Universal Access to Quality Tertiary Education (UAQTE) program, was initiated in the Philippines as a significant development in educational policy. The objective of this initiative is to increase access to tertiary education for all eligible Filipino students [1]. As the program progresses, understanding students' diverse experiences becomes crucial for evaluating its impact [2]. While qualitative responses offer rich narratives [3], manual categorization of these diverse accounts can be overwhelming. Therefore, this study employs transformer-based machine learning models like BERT, RoBERTa, and ALBERT to automate text classification [4].

By systematically analyzing student responses, the research aims to uncover nuanced perceptions of the UAQTE program's impact. The objective is to gain valuable insights into each student's distinctive experiences within the UAQTE framework. Leveraging prominent models for automated multiclass text classification [5], the study seeks to categorize student responses and unveil domain-specific insights systematically. This involves aligning experiences with predefined classes identified through collaboration with domain experts, ensuring a nuanced and contextualized understanding [6] of the diverse impacts of the UAQTE program on students.

Furthermore, the research evaluates the performance of these models in accurately categorizing qualitative responses, contributing to the advancement of machine learning techniques in educational research. This assessment ensures the reliability and effectiveness of the machine learning models [7] employed in the study. The primary research contribution lies in developing and applying advanced machine learning methods to analyze qualitative data in educational contexts [8], providing a novel approach to understanding the effects of educational policies like the UAQTE program. Ultimately, this research aims to contribute to informed policymaking and enhance educational initiatives for the benefit of Filipino students, thereby advancing the objectives of the UAQTE program.

II. Related Works

This section delves into the literature surrounding machine learning applications, particularly those relevant to implementing transformer-based models.

A. Machine Learning in Educational Policy

The integration of machine learning (ML) into educational policy signifies an innovative strategy for shaping the course of education [9], [10], [11], specifically in the context of revolutionary endeavors such as the Philippines' Universal Access to Quality Tertiary Education (UAQTE) program. Machine learning algorithms are highly effective tools for managing large data sets, presenting an opportunity to reform how policymakers understand, assess, and improve educational initiatives [12], [13].

Conventional assessment techniques might find capturing the intricacies and varied nature of student experiences challenging, underscoring the importance of adopting advanced data-driven methodologies. ML algorithms excel in uncovering patterns and trends within extensive datasets, offering a depth

of analysis that traditional methods might overlook [14], [15]. This capability becomes invaluable when assessing the effectiveness of educational programs, including initiatives like UAQTE.

The integration of ML into educational policy represents a significant leap forward. It facilitates a more thorough, dynamic, and nuanced assessment, providing policymakers with actionable insights into what aspects of the program are succeeding and where improvements are needed [16], [17]. As education systems worldwide navigate the complexities of providing equitable and quality education, the synergy between ML and educational policy becomes a pivotal force in shaping a more responsive and effective future for education.

B. Transformer-based Models

Transformer-based models, harnessing contextual relationships and language patterns through an architecture emphasizing parallel processing and self-attention mechanisms [18], [19], [20] represent a groundbreaking advancement in natural language processing (NLP). Unlike conventional recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformers operate simultaneously, enabling a holistic assessment of the entire context. This parallelized architecture enhances the model's ability to effectively capture long-range dependencies and subtle contextual nuances within the input sequence [21], [22].

Transformer-based models undergo initial pre-training on extensive corpora, enabling them to comprehensively understand language structures and patterns [23], [23], [24]. This foundational pre-training phase equips the models with a wealth of linguistic understanding. Following this, the models demonstrate adaptability by undergoing fine-tuning on domain-specific datasets, allowing them to tailor their knowledge to specific applications [25], [26]. This inherent adaptability renders them versatile across various tasks and domains, showcasing their efficacy in diverse applications [27], [28].

One notable application of transformer-based models is in qualitative data analysis, especially in domains like education policy. The models can perform multiclass text classification, categorizing and extracting insights from qualitative responses systematically. This capability becomes particularly valuable when evaluating the impact of programs like the UAQTE initiative, providing a data-driven lens to understand the diverse experiences of student beneficiaries.

C. Multiclass Text Classification

Leveraging advanced capabilities in natural language processing, transformer-based models exhibit remarkable proficiency in multiclass text classification, surpassing traditional text classification tasks [29], [30], [31]. This competence is deeply rooted in the objective of categorizing textual data into more than two predefined classes or categories. In multiclass text classification, each document or piece of text is precisely assigned to one specific class from a set of multiple classes, a task crucial for accurately discerning the most appropriate category or label based on the input text's content, themes, or characteristics [32], [33].

The adeptness of transformer-based models in this research facilitates a systematic understanding and categorization of qualitative responses [34], [35] from student beneficiaries within the UAQTE initiative. This structured approach guarantees precision in assigning text to relevant categories and serves as an effective tool for extracting insights into the diverse nature of students' experiences. Fortified with attention mechanisms and contextual understanding, the models capture subtle distinctions in qualitative data, providing a comprehensive and nuanced understanding of its impact [36], [37].

The proficiency in multiclass text classification offered by transformer-based models elevates their role as invaluable assets in navigating the complexities of education policy analysis, especially when seeking to comprehend the multifaceted dimensions of student experiences within specific programs like UAQTE. Ultimately, this capability empowers policymakers with nuanced, data-driven perspectives, facilitating well-informed choices in the dynamic landscape of education policy.

D. Role of Domain Experts

The involvement of domain experts in crafting and refining predefined categories is a well-established practice [38], [39], [40]. Their expertise ensures that the categories encapsulate the diverse dimensions of qualitative responses [41], [42]. This proactive role positions domain experts as key architects in aligning the model with the intricacies of the specific research domain, contributing significantly to ensuring that class labels are accurate and contextually relevant [43], [44].

Moreover, domain experts continue to play a critical role in the ongoing validation of machine learning models [45], [46]. As these models generate predictions on new or unseen data, domain experts validate the accuracy of these predictions against the true labels they provide. This validation process is a robust quality control mechanism, ensuring the alignment of model predictions with the ground truth. Establishing a feedback loop between domain experts and machine learning models contributes to the continual enhancement of the classification system [47], [48].

This iterative collaboration bolsters the accuracy of machine learning models and cultivates a dynamic understanding of qualitative data within the specific research domain. By actively participating in providing true labels and validating model predictions, domain experts ensure that multiclass text classification models are not only accurate but also ethically sound [49], [50]. Their dual role as architects of the categorization framework and validators of model predictions positions them as indispensable contributors to the success of the entire machine learning process within the field of education policy analysis.

E. BERT, RoBERTa, and ALBERT

In the broader context of natural language processing and machine learning, the integration of advanced models such as Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), and A Lite BERT (ALBERT) has become a focal point of research, particularly in the domain of multiclass text classification. These models demonstrate exceptional proficiency in achieving fine-grained categorization objectives, allowing for

systematically classifying qualitative responses into multiple predefined classes [51]. The emphasis is on their ability to capture nuanced distinctions, going beyond traditional categorization methods.

BERT pioneered bidirectional training, considering both left and right contexts in all layers [52], [53]. In refining this approach, RoBERTa removed the next sentence prediction objective and integrated dynamic masking during training [54], [55]. ALBERT, addressing computational challenges, implemented cross-layer parameter sharing and a factorized embedding parameterization, enhancing efficiency [56], [57].

Recognized for its general applicability, BERT's larger model size may pose computational intensity [58]. RoBERTa, optimized for larger mini-batches, showcases improved efficiency [59]. ALBERT, designed for parameter efficiency, strikes a balance between reduced parameters and competitive performance [60]. These distinctions significantly impact their versatility in categorizing information, spanning various themes or topics with potential applications across diverse domains.

The bidirectional attention mechanisms and extensive pre-training of BERT, RoBERTa, and ALBERT equip them with a profound understanding of context and relationships within the text [61]. This comprehensive comprehension positions them as valuable tools for many multiclass text classification applications. They provide insights crucial for refining models, enhancing accuracy, and facilitating informed decision-making. Furthermore, the iterative nature of machine learning emphasizes ongoing feedback loops, allowing continuous adjustments for enhanced model performance. This iterative refinement ensures that multiclass text classification models, whether BERT, RoBERTa, or ALBERT, progressively excel in handling diverse textual data [62].

F. Evaluation Metrics for Text Classification

Evaluation metrics are essential benchmarks for assessing the effectiveness of text classification models, providing valuable insights into their performance and generalization capabilities [63]. In the domain of text classification, various metrics are utilized to measure accuracy and reliability. Training accuracy assesses the model's proficiency in classifying instances within the training dataset, while validation accuracy evaluates its ability to generalize to new data without overfitting. Test accuracy offers a final evaluation of the model's performance on unseen data [64], [65]. Precision, recall, and F1-score provide nuanced assessments of its ability to correctly classify positive instances and balance false positives and false negatives [66], [67].

The confusion matrix visually represents the model's predictions, facilitating a detailed analysis of its performance across different classes. Additionally, the involvement of domain experts is crucial in providing true labels and validating the model's predictions against ground truth, thereby enhancing the reliability and credibility of the text classification process [68], [69].

These models, through the utilization of data-driven techniques, provide a comprehensive understanding of initiatives like the Universal Access to Quality Tertiary Education (UAQTE) program, revealing valuable insights derived from student experiences. The collaboration between machine learning algorithms and domain experts not only verifies model predictions but also ensures the precision of classifications, thereby reinforcing the credibility of the analysis. As these models evolve, they offer the potential to influence the development of more adaptive and efficient education policies in the future.

III. METHODOLOGY

The methodology employed is outlined in this section. Fig. 1 presents the information processing phases and delineates the steps, encompassing data preparation, tokenization and formatting, model training, model evaluation, hyperparameter tuning, and inference.

Fig. 1. Information processing phases.

A. Data Preparation

The "Boses Ko" or "My Voice" is a toolkit developed, with its grassroots approach that is instrumental in gathering data directly from student beneficiaries of the UAQTE program. This prioritization of perspectives from those directly involved ensures the authenticity and relevance of the collected data. The qualitative question guiding the study, "Write your experiences as one of the beneficiaries of the UAQTE program," further focuses the data collection process on soliciting responses specifically tailored to understanding student experiences within the program.

The sample size of 3,325 student responses, selected from State Universities and Colleges (SUCs), provides a diverse representation necessary for comprehensive examination across various institutional contexts. Data cleaning involves removing non-English, duplicate, non-grantee, and blank responses, which helps ensure the dataset's quality and consistency. Text standardization techniques, such as converting the cleaned dataset to lowercase and eliminating special characters, punctuation marks, and digits, further streamline the text representation, reducing noise and interference with the modeling process. These steps enhance the dataset's suitability for subsequent analysis and modeling tasks. Tokenization and removal of stopwords by implementing the Natural Language Toolkit (NLTK) library were necessary pre-processing steps. Responses were tokenized into individual words or tokens, making analyzing and processing text data easier. Stopwords like "as," "one," "of," "the," "it," "me," "a," and "in" are common in the responses but typically lack significant meaning alone. Eliminating these enhanced the quality, interpretability, and efficiency of the generated topics within the UAQTE framework by reducing noise and emphasizing content words that conveyed the core themes.

Furthermore, domain experts play a key role in collaborating on crafting and refining predefined categories for qualitative responses. Leveraging their expertise ensures that the categories accurately represent the diverse dimensions of

student experiences within the framework of the UAQTE program. Through close collaboration with domain experts, the study ensures that the labeling of the dataset reflects the nuanced perspectives relevant to the research domain. Table I presents the categories derived through this collaborative effort.

TABLE I. DOMAIN-EXPERTS IDENTIFIED CATEGORIES

Categories	Description
Financial Support	Responses that refer to the financial assistance provided, alleviation of financial burdens, and support with tuition fees, allowances, and other expenses.
Educational Opportunity	Responses that describe students' gratitude towards the program, enabling them to pursue their preferred courses, continue their studies, and access to quality education.
Family Support	Responses express families' gratitude for the support provided by the program and the ease it brings to their lives, as it relieves financial burdens, allowing them to save money and allocate resources to other expenses.
Academic Focus and Personal Development	Responses describe students being more focused on studying, becoming more responsible, and having more chances to invest in school projects due to the financial support received. They also attribute personal growth, increased enthusiasm, and improved class standings to being part of the program.
Program Implementation	Responses encompass a range of perspectives regarding the implementation of the program, reflecting both positive and negative viewpoints.

For data splitting, an 80-20 train-validation split is implemented. 80% of the entire dataset is allocated for training, with 80% of this training set used for actual model training and the remaining 20% reserved for validation. This partitioning strategy ensures that the model is trained on a sufficiently large portion of the dataset while allowing validation to monitor model performance and prevent overfitting. The remaining 20% of the entire dataset is held out for testing the model's performance, providing an independent evaluation of its generalization capabilities.

B. Tokenization and Formatting

Tokenization and formatting play crucial roles in preparing text data for transformer-based machine learning models such as BERT, RoBERTa, and ALBERT. These models rely on specific input structures to effectively process textual information. In the context of multiclass text classification using the UAQTE student responses dataset, tokenization involves breaking down the text into smaller units, typically words or subword units. This task is simplified by specialized tokenizers available in the Hugging Face's transformers library. For instance, the BERT tokenizer utilizes a WordPiece tokenizer to decompose words into subword units based on a predetermined vocabulary. Similarly, RoBERTa and ALBERT leverage WordPiece tokenization for the same purpose.

Once tokenization is completed, the tokenized text data needs to be formatted into an appropriate input structure for the transformer models. This formatting process includes adding special tokens like [CLS] (classification token) at the beginning of each sentence and [SEP] (separator token) between sentences. Additionally, sequences are padded to a fixed length, and attention masks are created to differentiate actual words from padding tokens. These formatting steps ensure uniform input lengths and assist the model in focusing on relevant tokens during the training and inference phases.

C. Model Training

Model training with BERT, RoBERTa, and ALBERT involves several steps to adapt these transformer-based models for multiclass text classification tasks using the UAQTE student responses dataset. Pre-trained BERT, RoBERTa, and ALBERT models are initially loaded from the Hugging Face's transformers library. These models possess an extensive contextual understanding of language, making them suitable for diverse natural language processing tasks, including multiclass text classification.

Subsequently, defining an optimizer and a loss function is crucial for training efficiency. Optimizers like Adam or SGD and loss functions such as Cross Entropy Loss are commonly employed. These components contribute to the model's ability to adjust parameters and minimize the loss during training, ensuring convergence towards accurate predictions. Data loaders are then created to handle the pre-processed text data efficiently, converting it into PyTorch or TensorFlow datasets. These loaders manage tasks like shuffling, batching, and loading data onto the GPU, streamlining the training process by optimizing resource utilization.

The training loop iterates through batches of data from the training set. The input data is passed through the BERT, RoBERTa, or ALBERT model to obtain predictions in each iteration. Subsequently, the loss is calculated by comparing the model's predictions with the true labels, followed by a backward pass to compute gradients and update model parameters using the chosen optimizer. This iterative process continues for multiple epochs, with the model's weights adjusted iteratively to improve performance.

D. Model Evaluation

The model evaluation phase is critical to comprehensively assess the performance of the trained multiclass text classification models. This evaluation encompasses a range of metrics to gauge different aspects of the model's effectiveness in handling the UAQTE student responses dataset. Initially, the evaluation considers training accuracy, which reflects how well the models have learned from the training data by measuring the proportion of correctly classified instances within this dataset. Subsequently, validation accuracy is examined to understand the models' generalization performance on unseen data, offering insights into their ability to perform accurately on examples beyond the training set. In addition, test accuracy serves as a critical metric in evaluating the overall performance of the models on entirely novel and unobserved instances, thereby offering a practical indication of their efficacy.

Furthermore, the evaluation procedure includes precision, recall, and F1-score metrics to offer a more comprehensive assessment of the models' performance with respect to the accuracy of classification by class. Precision quantifies the proportion of true positive predictions among all positive predictions made by the model, while recall calculates the proportion of true positive predictions among all actual positive instances. By calculating the harmonic mean of precision and

recall, the F1-score provides an equitable evaluation of the performance of the models in all classes.

In summary, the confusion matrix provides a comprehensive breakdown of the errors committed by the models. It serves as a tabular representation of the discrepancies between the predicted and actual class labels. This aids in identifying particular domains that require enhancement within the UAQTE program context.

E. Hyperparameter Tunning

Tuning hyperparameters is a crucial component in maximizing the efficiency of a model. By conducting experiments involving hyperparameters such as learning rate, sample size, and number of epochs, it is possible to attain optimal performance. Additionally, monitoring model performance on a validation set during training facilitates the adjustment of hyperparameters to ensure convergence toward accurate predictions. The following are the hyper-parameters used:

1) Batch Size. Determines the number of training examples processed in one iteration during training. By experimenting with batch sizes ranging from 16 to 64, the impact of different batch sizes on training dynamics and model convergence was observed. A larger batch size could accelerate the training process but could lead to memory constraints, while a smaller batch size could result in more noise during optimization.

2) Epochs. The number of epochs denoted how often the model iterated through the training dataset. Altering the number of epochs, ranging from 1 to 15, impacted the training duration and the model's evaluation. Increasing the number of epochs enabled the model to extract more insights from the data, yet excessive epochs could lead to overfitting on the training set.

3) Learning Rates. Controls the size of the step taken during optimization. Adjusting learning rates from 1e-5 to 5e-5 allowed for the evaluation of the sensitivity of the model's performance. A higher learning rate might have led to faster convergence but risked overshooting the optimal solution, while a lower learning rate could have resulted in slower convergence but more stable training.

4) Epsilon. It is a small value added to the denominator of the AdamW optimizer to prevent division by zero. A default value 1e-8 was typically used to ensure numerical stability during training. However, exploring the impact of adjusting epsilon to a value of 8 allowed for observing any changes in training dynamics or model performance.

5) Max-length. Refers to the maximum number of tokens allowed in each input sequence. By varying the max-length between 128 and 256, the effect of considering different amounts of context on model performance could be investigated. A larger max-length allowed the model to capture more contextual information but might have required more computational resources and memory.

Hyperparameter tuning involves systematically adjusting these parameters and evaluating their impact on the model's performance metrics, such as accuracy, precision, recall, and F1-score. This process can determine the optimal configuration of hyperparameters to improve the model's generalization and performance on unseen data.

F. Inference

In the final inference phase, domain experts validate predictions made by trained models and provide the predicted dataset's true labels. Their role is crucial in ensuring the accuracy and reliability of the model's predictions, as they verify the alignment of these predictions with the ground truth they have provided. This validation process is a robust quality control mechanism, guaranteeing that the model's predictions accurately reflect the qualitative responses within the UAQTE program context.

Moreover, domain experts' involvement fosters a collaborative environment for continuous improvement and refinement of the classification system. Valuable insights and feedback are exchanged through a feedback loop between domain experts and machine learning models based on domain knowledge and expertise. Experts guide the iterative optimization of the models' performance, enhancing their predictive capabilities.

During the inference phase, fine-tuned models like BERT, RoBERTa, and ALBERT classify qualitative responses from the UAQTE dataset, with domain experts providing true labels as the validation benchmark. Predictions are generated automatically, and the predictions made by the models and the true labels are saved for future reference or analysis. This collaborative effort between domain experts and machine learning models ensures that the insights derived from the predictions are accurate and trustworthy, contributing to informed decision-making in education policy analysis.

Additionally, the inference phase evaluates models' performance on new data, validating their generalization capabilities. Deployment in real-world applications may require integrating existing systems, ensuring compatibility, and addressing technical challenges. Overall, domain experts' involvement, who validate predictions and provide true labels and fine-tuned models, advances natural language processing techniques and facilitates informed decision-making in education policy analysis.

IV. Results and Discussion

The multiclass text classification task employing BERT, RoBERTa, and ALBERT architectures provided insights into their performance dynamics across various hyperparameter configurations. Both BERT and RoBERTa consistently exhibited improved accuracy with smaller batch sizes and higher numbers of epochs, as seen in Table II and Table III, suggesting the importance of detailed updates during training. Optimal learning rates, particularly 1e-5 and 3e-5, consistently yielded superior accuracy across different experimental settings, indicating their significance in facilitating effective model learning.

However, it is noteworthy that larger maximum sequence lengths did not consistently enhance accuracy, revealing complexities in balancing sequence length and model performance.

Similarly, as seen in Table IV, ALBERT demonstrated consistent performance trends with smaller batch sizes and increased epochs, improving accuracy metrics. Notably, the impact of maximum sequence length on model performance varied across experiments, suggesting the need for careful consideration in adjusting sequence length for optimal accuracy.

Instances of overfitting were observed beginning in five epochs, where training accuracy exceedingly surpassed validation and test accuracy, emphasizing the importance of early stopping or regularization techniques to prevent performance degradation. Overall, RoBERTa emerges as a strong choice due to its balanced performance, stability, and efficiency in training, making it a recommended option for practitioners aiming for reliable results.

TABLE II. BERT HYPERPARAMETERS AND ACCURACY SCORES

Batch Size	Epoch	Learning rate	Max-length	Training Accuracy	Validation Accuracy	Test Accuracy
16	3	1e-5	128	73%	72%	66%
16	3	1e-5	256	77%	73%	68%
32	3	1e-5	128	65%	65%	65%
32	3	1e-5	256	70%	70%	67%
16	5	1e-5	128	84%	73%	71%
16	5	1e-5	256	85%	75%	70%
32	5	1e-5	128	76%	72%	68%
32	5	1e-5	256	79%	72%	69%
16	**3**	**3e-5**	**128**	**85%**	**76%**	**73%**
16	3	3e-5	256	87%	75%	71%
32	3	3e-5	128	82%	75%	71%
32	3	3e-5	256	81%	75%	70%
16	5	3e-5	128	95%	75%	73%
16	5	3e-5	256	96%	75%	73%
32	5	3e-5	128	93%	74%	72%
32	5	3e-5	256	92%	74%	72%
16	3	5e-5	128	89%	76%	73%
16	3	5e-5	256	88%	76%	72%
32	3	5e-5	128	85%	76%	72%
32	3	5e-5	256	85%	76%	73%
16	5	5e-5	128	93%	74%	72%
16	5	5e-5	256	97%	76%	73%
32	5	5e-5	128	96%	75%	73%
32	5	5e-5	256	96%	74%	72%

TABLE III. ROBERTA HYPERPARAMETERS AND ACCURACY SCORES

Batch Size	Epoch	Learning rate	Max-length	Training Accuracy	Validation Accuracy	Test Accuracy
16	3	1e-5	128	80%	75%	72%
16	3	1e-5	256	79%	75%	71%
32	3	1e-5	128	75%	73%	68%
32	3	1e-5	256	76%	74%	71%
16	5	1e-5	128	86%	76%	73%
16	5	1e-5	256	86%	74%	72%
32	5	1e-5	128	82%	73%	70%
32	5	1e-5	256	83%	74%	72%
16	3	3e-5	128	84%	75%	72%
16	3	3e-5	256	85%	76%	72%
32	3	3e-5	128	83%	75%	73%
32	3	3e-5	256	82%	75%	72%
16	5	3e-5	128	92%	76%	74%
16	5	3e-5	256	93%	76%	72%
32	5	3e-5	128	90%	75%	72%
32	5	3e-5	256	91%	75%	73%
16	**3**	**5e-5**	**128**	**84%**	**76%**	**74%**
16	3	5e-5	256	87%	76%	71%
32	3	5e-5	128	85%	76%	74%
32	3	5e-5	256	84%	76%	73%
16	5	5e-5	128	93%	76%	72%
16	5	5e-5	256	93%	76%	73%
32	5	5e-5	128	91%	75%	73%
32	5	5e-5	256	91%	76%	73%

TABLE IV. ALBERT HYPERPARAMETERS AND ACCURACY SCORES

Batch Size	Epoch	Learning rate	Max-length	Training Accuracy	Validation Accuracy	Test Accuracy
16	3	1e-5	128	77%	72%	71%
16	3	1e-5	256	75%	72%	69%
32	3	1e-5	128	50%	42%	44%
32	3	1e-5	256	68%	69%	63%
16	5	1e-5	128	87%	73%	72%
16	5	1e-5	256	83%	74%	70%
32	5	1e-5	128	77%	68%	65%
32	5	1e-5	256	77%	70%	66%
16	3	3e-5	128	30%	33%	28%
16	**3**	**3e-5**	**256**	**83%**	**75%**	**72%**
32	3	3e-5	128	67%	70%	64%
32	3	3e-5	256	76%	73%	70%
16	5	3e-5	128	85%	74%	72%
16	5	3e-5	256	94%	75%	74%
32	5	3e-5	128	90%	74%	72%
32	5	3e-5	256	85%	72%	71%
16	3	5e-5	128	22%	19%	21%
16	3	5e-5	256	74%	73%	70%
32	3	5e-5	128	71%	72%	67%
32	3	5e-5	256	70%	70%	68%
16	5	5e-5	128	84%	73%	72%
16	5	5e-5	256	68%	69%	63%
32	5	5e-5	128	89%	72%	72%
32	5	5e-5	256	86%	74%	73%

The performance metrics of BERT, RoBERTa, and ALBERT, outlined in Table V, provided insights into their effectiveness across different categories. BERT demonstrated strong precision, recall, and F1-score for "Family Support," while "Financial Support" also performed well, albeit with room for precision improvement. Similarly, "Academic Focus & Personal Development" showed balanced precision-recall metrics, while "Educational Opportunity" and "Program Implementation" exhibited lower scores, particularly in the recall.

TABLE V. PERFORMANCE METRICS BY ARCHITECTURE AND CATEGORY

Categories (BERT)	**Precision**	**Recall**	**F1-Score**
Academic Focus & Personal Development	70%	78%	74%
Educational Opportunity	54%	60%	57%
Family Support	95%	96%	96%
Financial Support	73%	77%	75%
Program Implementation	80%	55%	65%
Weighted Average	**74%**	**73%**	**73%**
Categories (RoBERTa)	**Precision**	**Recall**	**F1-Score**
Academic Focus & Personal Development	71%	76%	74%
Educational Opportunity	53%	59%	56%
Family Support	95%	98%	97%
Financial Support	75%	83%	79%
Program Implementation	79%	52%	63%
Weighted Average	**74%**	**74%**	**74%**
Categories (ALBERT)	**Precision**	**Recall**	**F1-Score**
Academic Focus & Personal Development	69%	77%	72%
Educational Opportunity	53%	59%	55%
Family Support	95%	95%	95%
Financial Support	74%	78%	76%
Program Implementation	74%	52%	61%
Weighted Average	**73%**	**72%**	**72%**

RoBERTa consistently performed well across categories, with outstanding performance in "Family Support" and "Financial Support." However, "Educational Opportunity" and "Program Implementation" still showed room for improvement, mirroring BERT's findings. While competitive, ALBERT showed slightly lower scores than BERT and RoBERTa. "Family Support" and "Financial Support" demonstrated strong performance, yet "Educational Opportunity" and "Program Implementation" again presented areas for refinement, particularly in recall.

Overall, while all models showed effectiveness in specific categories, improvements were needed, especially in those with lower recall scores. Analyzing misclassified instances and adjusting model parameters could enhance performance across all categories. Addressing the classification of similar terms into multiple categories was also crucial to improve overall accuracy and mitigate confusion.

The heatmap visualization of the confusion matrix, depicted in Fig. 2, provided nuanced insights into the classification performance of BERT, complementing the precision, recall, and F1-score metrics. In the "Academic Focus & Personal Development" (AF&PD) category, for instance, BERT accurately classified 103 instances (true positives), with 17 instances misclassified (false negatives), aligning with its recall of 78%. Similar observations were made across other categories such as "Educational Opportunity" (EO), "Family Support" (FaS), "Financial Support" (FiS), and "Program Implementation" (PI). Notably, categories with lower recall scores, like PI, exhibited a higher number of false negatives, indicating potential areas for improvement. Conversely, categories with high precision and recall, like FaS, demonstrated fewer misclassifications. This in-depth analysis, in conjunction with precision, recall, and F1-score metrics, provided a comprehensive understanding of BERT's classification performance across diverse categories, thereby guiding optimization strategies for enhanced accuracy.

Fig. 2. BERT confusion matrix heatmap.

Similarly, the confusion matrix heatmap for RoBERTa, shown in Fig. 3, confirmed its accuracy, recall, and F1-score metrics across several categories. For example, in the "Academic Focus & Personal Development" (AF&PD) category, RoBERTa correctly categorized 100 occurrences (true positives) and misclassified 17 instances (false negatives), corresponding to a recall of 76%. Similar trends were seen in other categories, including "Educational Opportunity" (EO), "Family Support" (FaS), "Financial Support" (FiS), and "Program Implementation" (PI).

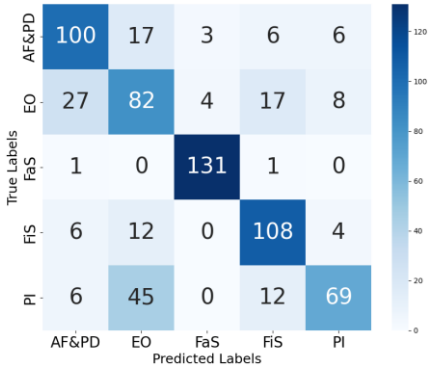

Fig. 3. RoBERTa confusion matrix heatmap.

Categories with higher recall scores, like FaS, exhibited fewer false negatives, indicating robust classification performance. Conversely, categories with lower recall scores, such as PI, demonstrated a higher number of false negatives, suggesting potential areas for improvement. This detailed examination, combined with precision, recall, and F1-score metrics, facilitated a comprehensive evaluation of RoBERTa's classification performance, guiding targeted enhancements for optimal accuracy.

Finally, Fig. 4 shows the confusion matrix heatmap for ALBERT, which provides insights into its accuracy, recall, and F1-score metrics across several categories. In the "Academic Focus & Personal Development" (AF&PD) category, for example, ALBERT accurately categorized 101 occurrences (true positives) while misclassifying 13 instances (false negatives), resulting in a 77% recall. This pattern continued in other areas, including "Educational Opportunity" (EO), "Family Support" (FaS), "Financial Support" (FiS), and "Program Implementation" (PI). Categories with greater recall scores, such as FS, produced fewer false negatives, suggesting strong categorization ability. Conversely, categories with lower recall scores, such as PI, had a larger incidence of false negatives, indicating possible areas for improvement. By integrating precision, recall, and F1-score metrics with the confusion matrix, a comprehensive assessment of ALBERT's classification performance was achieved, facilitating targeted enhancements for optimal accuracy.

Fig. 4. ALBERT confusion matrix heatmap.

Our findings are consistent with several prior investigations highlighting the effectiveness of transformer-based models in multiclass text classification tasks. For example, using pre-trained language models, Lee et al. [70] conducted a comparative study on multiclass text classification within research proposals. Their research demonstrated exceptional performance in natural language understanding (NLU) tasks, showcasing the robust capabilities of transformer-based models in handling complex textual data. Similarly, Prabhu et al. [70] applied a BERT-based active learning approach to classify customer transactions into multiple categories, aiming to discern market needs across diverse customer segments. Furthermore, the study conducted by Chen et al. [71] observed significant enhancements in long-text classification performance when employing transformer-based models compared to traditional methods such as Convolutional Neural Networks (CNNs).

In terms of model performance, RoBERTa consistently demonstrates superior performance compared to BERT and ALBERT in multiclass text classification tasks, a trend also observed in other studies. This aligns with the research conducted by Zhao et al. [72], who leveraged the RoBERTa base model to conduct financial text mining and public opinion analysis within social media contexts. The enhanced performance of RoBERTa can be attributed to its more extensive pre-training and modifications to the architecture, enabling it to capture more nuanced linguistic features and contextual information. Moreover, the investigation by Angin et al. [73] underscores the efficacy of fine-tuned RoBERTa-based classification models for automating the processing of large document collections to detect relevance. Fine-tuning RoBERTa involves adjusting model parameters and hyperparameters to adapt the pre-trained RoBERTa model to specific tasks or datasets, enhancing its performance for the targeted classification task. This process allows the model to learn domain-specific features and nuances [73], [74], improving classification accuracy and relevance detection.

While the study provided valuable insights into the performance of BERT, RoBERTa, and ALBERT in multiclass text classification, several constraints were encountered. Achieving an optimal balance between sequence length and model efficacy posed challenges, with inconsistencies in the impact of maximum sequence length on accuracy across different experiments. Additionally, addressing the classification of similar terms into multiple categories remained a limit, impacting overall accuracy and potentially leading to confusion in classification. These underscore the need for further research and refinement to enhance the effectiveness of transformer-based models in multiclass text classification tasks.

The research has several limitations and deficiencies that should be acknowledged. Firstly, its narrow focus solely on evaluating transformer-based machine learning models (BERT, RoBERTa, and ALBERT) within the context of multiclass text classification in the Universal Access to Quality Tertiary Education (UAQTE) program restricts the generalizability of the findings beyond this specific domain. Secondly, while the research explores various hyperparameter configurations for model training, it may not comprehensively cover all possible combinations or consider other factors, such as optimization algorithms or regularization techniques. This limitation could be partly attributed to hardware requirements, as exhaustive exploration of hyperparameters may be computationally intensive.

Lastly, while the results offer actionable recommendations, it is crucial to acknowledge that these suggestions serve as guidance rather than mandates, potentially limiting their enforceability and practical implementation within educational policy. Overcoming these limitations and deficiencies would strengthen the reliability and practical relevance of the research findings, providing a more thorough understanding of the performance of transformer-based machine learning models in educational settings.

V. Conclusions and Recommendations

The study's comprehensive evaluation of BERT, RoBERTa, and ALBERT in multiclass text classification tasks revealed nuanced insights into their performance dynamics. Hyperparameter configurations played a crucial role, with smaller batch sizes and increased epochs consistently enhancing model accuracy. Optimal learning rates, particularly in the range of 1e-5 to 3e-5, significantly contributed to superior accuracy across experimental settings. However, the impact of larger maximum sequence lengths on accuracy was inconsistent, indicating the complexity of balancing sequence length and model performance. Moreover, instances of overfitting, particularly observed beyond five epochs, underscored the necessity of early stopping or regularization techniques to prevent performance degradation.

Interpreting the classification results provided valuable insights into students' experiences within the UAQTE program. Categories like "Family Support" and "Financial Support" demonstrated high precision, recall, and F1 scores, indicative of the program's effectiveness in addressing student needs in these areas. Conversely, categories such as "Educational Opportunity" and "Program Implementation" exhibited lower scores, suggesting potential areas for improvement. The study's findings highlight the importance of selecting appropriate model architectures and hyperparameters tailored to the specific classification task. RoBERTa emerged as a robust choice due to its balanced performance, stability, and efficiency in training, making it a recommended option for similar classification tasks in educational contexts.

For future works, researchers are encouraged to delve deeper into hyperparameter tuning, exploring alternative configurations to optimize model performance further. Addressing overfitting remains a critical concern, necessitating ongoing monitoring of training processes and fine-tuning regularization strategies. Continuous review and refining of educational programs, guided by evidence-based decision-making and stakeholder feedback, is critical for effectively fulfilling students' changing needs.

Future research approaches may also include looking into the interpretability of model predictions and researching socio-cultural aspects that influence students' experiences to understand educational interventions' effectiveness better. By embracing these recommendations, researchers and practitioners can advance the multiclass text classification field and contribute to enhancing educational programs to support student success.

Acknowledgment

The researchers would like to thank the Philippine Commission on Higher Education (CHED) and the Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, specifically the eParticipation 2.1 initiative, "Harnessing Natural Language Processing (NLP) for Community Participation." We appreciate the substantial financial assistance offered by this initiative, which has been vital in making our research pursuits feasible. The researchers congratulate CHED and the LAKAS Project for their significant contributions, acknowledging their critical role in expanding our quest for knowledge.

References

[1] M. Kristina et al., "Process Evaluation of the Universal Access to Quality Tertiary Education Act (RA 10931): Status and Prospects for Improved Implementation," 2019. [Online]. Available: https://www.pids.gov.ph

[2] M. Beerkens, "Evidence-based policy and higher education quality assurance: progress, pitfalls and promise," European Journal of Higher Education, vol. 8, no. 3, pp. 272–287, Jul. 2018.

[3] G. Ferguson-Cradler, "Narrative and computational text analysis in business and economic history," Scandinavian Economic History Review, vol. 71, no. 2, pp. 103–127, 2023.

[4] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification," J Healthc Eng, 2022.

[5] H. Wang, K. C. Haudek, A. D. Manzanares, C. L. Romulo, and E. A. Royse, "Extending a Pretrained Language Model (BERT) using an Ontological Perspective to Classify Students' Scientic Expertise Level from Written Responses," 2024.

[6] B. Xie, M. J. Davidson, B. Franke, E. McLeod, M. Li, and A. J. Ko, "Domain Experts' Interpretations of Assessment Bias in a Scaled, Online Computer Science Curriculum," in L@S 2021 - Proceedings of the 8th ACM Conference on Learning @ Scale, Association for Computing Machinery, Inc, pp. 77–89, Jun. 2021.

[7] O. Awujoola, Philip O Odion, Martins E Irhebhude, and Halima Aminu, "Performance Evaluation of Machine Learning Predictive Analytical Model for Determining the Job Applicants Employment Status," Malaysian Journal of Applied Sciences, vol. 6, no. 1, pp. 67–79, Apr. 2021.

[8] V. Kuleto et al., "Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions," Sustainability (Switzerland), vol. 13, no. 18, Sep. 2021.

[9] M. Tanveer, S. Hassan, and A. Bhaumik, "Academic policy regarding sustainability and artificial intelligence (Ai)," Sustainability (Switzerland), vol. 12, no. 22, pp. 1–13, Nov. 2020.

[10] H. Luan et al., "Challenges and Opportunities for Sustainable Development Education Sector United Nations Educational, Scientific and Cultural Organization," 2019. [Online]. Available: https://en.unesco.org/themes/education-policy-

[11] I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A systematic review of teaching and learning machine learning in K-12 education," Educ Inf Technol (Dordr), vol. 28, no. 5, pp. 5967–5997, May 2023.

[12] H. Luan et al., "Challenges and Future Directions of Big Data and Artificial Intelligence in Education," Frontiers in Psychology, vol. 11. Frontiers Media S.A., Oct. 19, 2020.

[13] A. Androutsopoulou and Yannis Charalabidis, "A framework for evidence based policy making combining big data, dynamic modelling and machine intelligence," 11th International Conference on Theory and Practice of Electronic Governance, pp. 575–583, 2018.

[14] M. A. EL-Omairi and A. El Garouani, "A review on advancements in lithological mapping utilizing machine learning algorithms and remote sensing data," Heliyon, vol. 9, no. 9. Elsevier Ltd, Sep. 01, 2023.

[15] M. El Hajj and J. Hammoud, "Unveiling the Influence of Artificial Intelligence and Machine Learning on Financial Markets: A Comprehensive Analysis of AI Applications in Trading, Risk Management, and Financial Operations," Journal of Risk and Financial Management, vol. 16, no. 10, Oct. 2023.

[16] H. Yue and S. Huang, "Min-Max Machine Learning Estimation Model with Big Data Analytics in Industry-Education Fusion," International Journal of Intelligent Systems and Applications in Engineering, 2024.

[17] S. E. Bibri, J. Krogstie, A. Kaboli, and A. Alahi, "Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review," Environmental Science and Ecotechnology, vol. 19. Editorial Board, Research of Environmental Sciences, May 01, 2024.

[18] J. Jia, W. Liang, and Y. Liang, "A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.05589

[19] P. J. Worth, "Word Embeddings and Semantic Spaces in Natural Language Processing," Int J Intell Sci, vol. 13, no. 01, pp. 1–21, 2023.

[20] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation," Natural Language Processing Journal, vol. 6, p. 100056, Mar. 2024.

[21] K. Khan, "A Large Language Model Classification Framework (LLMCF)," International Journal of Multidisciplinary Research and Publications, 2023.

[22] T. Ahmed, N. R. Ledesma, and P. Devanbu, "SYNFIX: Automatically Fixing Syntax Errors using Compiler Diagnostics," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.14671

[23] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," Journal of Biomedical Informatics, vol. 126. Academic Press Inc., Feb. 01, 2022.

[24] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-Trained Language Models and Their Applications," Engineering, vol. 25. Elsevier Ltd, pp. 51–65, Jun. 01, 2023.

[25] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," Journal of Biomedical Informatics, vol. 126. Academic Press Inc., Feb. 01, 2022.

[26] X. Luo, Y. Xue, Z. Xing, and J. Sun, "PRCBERT: Prompt Learning for Requirement Classification using BERT-based Pretrained Language Models," in ACM International Conference Proceeding Series, Association for Computing Machinery, Sep. 2022.

[27] M. Azam Khan Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," IEEE Access, 2024.

[28] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," J Big Data, vol. 11, no. 1, p. 25, Feb. 2024.

[29] M. Islam and S. Basu, "Tunable persistent currents in a spin-orbit coupled pseudospin-1 fermionic quantum ring," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.06804

[30] O. E. Ojo, O. O. Adebanji, A. Gelbukh, H. Calvo, and A. Feldman, "MedAI Dialog Corpus (MEDIC): Zero-Shot Classification of Doctor and AI Responses in Health Consultations," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.12489

[31] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, "Bangla-BERT: Transformer-based Efficient Model for Transfer Learning and Language Understanding," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3197662.

[32] K. Taha, P. D. Yoo, C. Yeun, and A. Taha, "Text Classification: A Review, Empirical, and Experimental Evaluation," arXiv preprint arXiv:2401.12982, 2024.

[33] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "Survey on Text Classification Algorithms: From Text to Predictions," Information (Switzerland), vol. 13, no. 2, Feb. 2022.

[34] M. O. Kenneth, F. Khosmood, and A. Edalat, "Systematic Literature Review: Computational Approaches for Humour Style Classification," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2402.01759

[35] M. Garg, "WellXplain: Wellness Concept Extraction and Classification in Reddit Posts for Mental Health Analysis," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.13710

[36] S. Bansal, K. Gowda, and N. Kumar, "Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network," Expert Syst Appl, vol. 236, Feb. 2024.

[37] D. Zaikis and Loannis Vlahavas, "From Pre-Training to Meta-Learning: A Journey in Low-Resource-Language Representation Learning," IEEE, 2023.

[38] D. Ali, M. M. S. Missen, and M. Husnain, "Multiclass Event Classification from Text," Sci Program, 2021.

[39] F. Gargiulo, S. Silvestri, M. Ciampi, and G. De Pietro, "Deep neural network for hierarchical extreme multi-label text classification," Applied Soft Computing Journal, vol. 79, pp. 125–138, Jun. 2019.

[40] J. Briskilal and C. N. Subalalitha, "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa," Inf Process Manag, 2022.

[41] A. Sukhov, A. Sihvonen, J. Netz, P. Magnusson, and L. E. Olsson, "How experts screen ideas: The complex interplay of intuition, analysis and sensemaking," Journal of Product Innovation Management, vol. 38, no. 2, pp. 248–270, Mar. 2021.

[42] Y. Mao et al., "How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right qestion?," Proc ACM Hum Comput Interact, vol. 3, no. GROUP, Dec. 2019.

[43] A. Saka et al., "GPT models in construction industry: Opportunities, limitations, and a use case validation," Developments in the Built Environment, vol. 17. Elsevier Ltd, Mar. 01, 2024..

[44] F. Li, X. Wang, B. Li, Y. Wu, Y. Wang, and X. Yi, "A Study on Training and Developing Large Language Models for Behavior Tree Generation," Jan. 2024.

[45] D. Te'eni et al., "Reciprocal Human-Machine Learning: A Theory and an Instantiation for the Case of Message Classification," Manage Sci, Nov. 2023.

[46] E. H. Weissler et al., "The role of machine learning in clinical research: transforming the future of evidence generation," Trials, vol. 22, no. 1. BioMed Central Ltd, Dec. 01, 2021.

[47] L. Von Rueden et al., "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," IEEE Trans Knowl Data Eng, vol. 35, no. 1, pp. 614–633, Jan. 2023.

[48] S. Amershi et al., "Software Engineering for Machine Learning: A Case Study," 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/devops/learn/devops-at-microsoft/

[49] E. Hassan, T. Abd El-Hafeez, and M. Y. Shams, "Optimizing classification of diseases through language model analysis of symptoms," Sci Rep, vol. 14, no. 1, Dec. 2024.

[50] E. Shnarch et al., "Label Sleuth: From Unlabeled Text to a Classifier in a Few Hours," arXiv preprint arXiv:2208.01483, Aug. 2022.

[51] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," Applied Sciences (Switzerland), vol. 12, no. 11. MDPI, Jun. 01, 2022.

[52] M. Beseiso and S. Alzahrani, "An Empirical Analysis of BERT Embedding for Automated Essay Scoring," International Journal of Advanced Computer Science and Applications, 2020.

[53] A. K. Durairaj and A. Chinnalagu, "Transformer based Contextual Model for Sentiment Analysis of Customer Reviews: A Fine-tuned BERT A Sequence Learning BERT Model for Sentiment Analysis," International Journal of Advanced Computer Science and Applications, vol. 12, no. 11, pp. 474–480, 2021.

[54] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," International Journal of Electrical and Computer Engineering, vol. 13, no. 5, pp. 5641–5652, Oct. 2023.

[55] R. Desai, A. Shah, S. Kothari, A. Surve, and N. Shekokar, "TextBrew: Automated Model Selection and Hyperparameter Optimization for Text Classification," International Journal of Advanced Computer Science and Applications, 2022.

[56] T. S. Alharbi and F. Fkih, "Building and Testing Fine-Grained Dataset of COVID-19 Tweets for Worry Prediction," International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, pp. 645–652, 2022.

[57] A. A. Jalil and A. H. Aliwy, "Classification of Arabic Social Media Texts Based on a Deep Learning Multi-Tasks Model," Al-Bahir Journal for Engineering and Pure Sciences, vol. 2, no. 2, May 2023.

[58] E. T. Luthfi, Z. Izzah, M. Yusoh, and B. M. Aboobaider, "BERT based Named Entity Recognition for Automated Hadith Narrator Identification," International Journal of Advanced Computer Science and Applications, 2022.

[59] B. Omarov and Zhandos Zhumanov, "Bidirectional Long-Short-Term Memory with Attention Mechanism for Emotion Analysis in Textual Content," International Journal of Advanced Computer Science and Applications, 2023.

[60] S. Saleem and Sapna Kumarapathirage, "AutoNLP: A Framework for Automated Model Selection in Natural Language Processing," IEEE, 2023.

[61] B. K. Jha, C. M. V. Srinivas Akana, and R. Anand, "Question Answering System with Indic multilingual-BERT," in Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, Institute of Electrical and Electronics Engineers Inc., Apr. 2021.

[62] X. Jiang et al., "On the Evolution of Knowledge Graphs: A Survey and Perspective," arXiv preprint arXiv:2310.04835, Oct. 2023.

[63] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," International Journal of Intelligent Systems and Applications in Engineering, 2023.

[64] S. Joshi and E. Abdelfattah, "Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews," in 2021 IEEE World AI IoT Congress, AIIoT 2021, Institute of Electrical and Electronics Engineers Inc., May 2021.

[65] S. Riyanto, T. D. Sukaesih Sitanggang Imas, and Tika Dewi Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," International Journal of Advanced Computer Science and Applications, 2023.

[66] A. Toktarova, D. Syrlybay, G. Anuarbekova, and G. Rakhimbayeva, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," International Journal of Advanced Computer Science and Applications, 2023.

[67] K. A. Binsaeed and Alaaeldin M. Hafez, "Enhancing Intrusion Detection Systems with XGBoost Feature Selection and Deep Learning Approaches," International Journal of Advanced Computer Science and Applications, 2023.

[68] S. Joshi and E. Abdelfattah, "Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews," in 2021 IEEE World AI IoT Congress, AIIoT 2021, Institute of Electrical and Electronics Engineers Inc., May 2021.

[69] A. Chauhan, A. Agarwal, and R. Sulthana, "Genetic Algorithm and Ensemble Learning Aided Text Classification using Support Vector Machines," International Journal of Advanced Computer Science and Applications, 2021.

[70] S. Prabhu, M. Mohamed, and H. Misra, "Multi-class Text Classification using BERT-based Active Learning," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.14289

[71] X. Chen, P. Cong, and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN," IEEE Access, vol. 10, pp. 34046–34057, 2022.

[72] L. Zhao, L. Li, and X. Zheng, "A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts," 2021.

[73] M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," IEEE Access, pp. 1–1, Feb. 2024.

[74] B. V. P. Kumar and M. Sadanandam, "A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms," International Journal of Computing and Digital Systems, vol. 15, no. 1, pp. 51–66, 2024.