International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.153, November 2026

Cross-University Functional Requirements Elicitation from Nigerian University Academic Staff Promotion Policy Documents using ELMo and BERT

Babatunde I. Ayinla
Data Science Department
University of Ibadan, Nigeria

Micheal Tomiwa Oyebamiji
Data Science Department
University of Ibadan, Nigeria

ABSTRACT

The task of identifying functional requirements (FR) and non-functional requirements (NFR) from natural language institutional documents is challenging, particularly across institutions with different policy vocabularies and conventions. This paper presents a cross-university evaluation framework using policy documents from five Nigerian universities: University of Ibadan (UI), Obafemi Awolowo University (OAU), University of Ilorin (UNILORIN), Ahmadu Bello University (ABU), and University of Port Harcourt (UNIPORT). The manual annotation was employed for 3,201 sentences (approximately 640 per university) extracted from academic staff promotion guidelines and institutional policies. An evaluation of two model types was done: classical machine learning classifiers (Logistic Regression, Random Forest, SVM) trained on ELMo embeddings, and a fine-tuned BERT model. Additionally, two evaluation settings were carried out: (1) within-university evaluation with 80/20 train/test splits per institution, and (2) cross-university evaluation where models trained on one university were evaluated on others. Within-university results were strong across all models, with BERT achieving weighted F_1 scores between 0.9610 and 0.9879. Cross-university performance was substantially lower (mean F_1: 0.7910), revealing significant domain transfer challenges. Notably, models trained on ABU data exhibited the poorest cross-institutional transfer, a pattern not fully explained by class distribution alone and suggesting that stylistic and structural differences in ABU's policy documents impact generalization.

Keywords

functional requirements, non-functional requirements, requirements elicitation, ELMo, BERT, cross-university evaluation, Natural Language Processing, Nigerian Universities, Nigeria Academic staff promotion.

1. INTRODUCTION

Requirements elicitation is the process of discovering, gathering, and articulating the functions that a system or process is supposed to perform [20]. Among the basic distinctions made during this process is the separation between functional requirements and non-functional requirements. Functional requirements specify the operations, actions, and behaviours that a system must perform, whereas non-functional requirements describe quality attributes, constraints, and conditions under which those behaviours must be maintained [19]. This distinction matters because the two requirement types inform different aspects of system design, verification, and policy compliance.

In institutional settings such as universities, requirement laden documents appear in a variety of administrative and regulatory texts. Promotion guidelines, gender policies, data protection policies, IT governance documents, quality assurance policies, and sustainability frameworks all contain sentences that either prescribe specific actions and obligations or describe properties and standards the institution must uphold. The manual effort required to read through these documents, identify requirement statements, and classify them correctly is substantial [1]. It is also prone to inconsistency, since the boundary between what a system must do and what properties it must have is not always obvious from the surface form of a sentence. As Cleland-Huang et al. [7] noted, requirements documents tend to be structured around functional requirements, which means non-functional requirements become intertwined with functional ones, making extraction time consuming and error prone [27].

Automated methods for requirement classification based on Natural Language Processing (NLP) are not new. Early work relied on rule-based methods and keyword matching, which do not perform well with domain variation [15, 17]. Later studies explored statistical machine learning classifiers trained on bag-of-words or TF-IDF representations [4, 21]. More recently, dense word embeddings and pre-trained language models have shown that semantic relationships between words can be captured in ways that sparse representations cannot [10, 23].

A key practical question that existing studies have not fully addressed is whether a classifier trained on the institutional documents of one university can reliably elicit requirements from the documents of a different university it has never seen. Most requirement classification studies evaluate models on held-out data from the same corpus they were trained on, which does not test whether the model has learned anything generally applicable about how requirement language works or whether it has simply learned the vocabulary of its training institution. Hey et al. [13] addressed this for software project requirements by evaluating their NoRBERT model on unseen projects, and Ivanov et al. [16] tested transfer from an-

notated requirements datasets to informal Request for Information documents. The present study applies the same logic to university institutional documents by introducing a cross-university generalisation evaluation across five Nigerian institutions.
The rest of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the methodology, covering data collection, preprocessing, feature extraction, and model development. Section 4 reports experimental results. Section 5 discusses the findings and limitations, and Section 6 concludes.

2. LITERATURE REVIEW

2.1 Related Works

Requirements elicitation is the beginning of the software development lifecycle, and if it is done correctly can impact all aspects of the lifecycle. Kotzé and Smuts [20] define it as the process of identifying and communicating the actions and interactions that a system must perform. What has changed considerably over the years is how much of this work is done manually. Umar and Lano [30] and Ayinla and Adeniniji [3] performed a systematic literature review on automated requirement engineering from 1996 to 2022, Dual-Use Research (DUR) 2018–2023, and found that the manual execution of RE tasks or manual classification of DUR is time-consuming and prone to errors. It is also emphasised that NLP-based tools are the most widely used approach to support automated RE. Their analysis of 85 papers revealed that the most automated phase in the tools (49.53%) is the review phase, NLP techniques like POS tagging and tokenisation are the most commonly used techniques, and most of the tools developed focus on requirement identification and quality defect detection. Hussain et al. [15] were among the earliest contributors to this field. They proposed a linguistic approach to classify NFRs in requirements specification documents. They used part-of-speech tagging and semantic patterns to identify NFR-indicative constructs in the PROMISE benchmark dataset. Although linguistic structure provided valuable features for the classification task, rule-based approaches were found to be domain-sensitive, a property that has been noted in other NLP tasks on different types of institutional and domain-specific texts [17]. Later, Kurtanović and Maalej [21] tackled this by developing a supervised machine learning pipeline, based on lexical, syntactic, and metadata characteristics of the RE17 Quality Attributes dataset. By applying undersampling and oversampling to handle class imbalance, they achieved precision and recall values of around 92% for FR and NFR binary classification using SVM, and found part-of-speech tags and cardinal numbers to be the most informative features. This is still one of the most explicit examples that shows that vocabulary-level features are good discriminative indicators for the FR/NFR distinction, the same assumption that is used in the present study with dense embeddings.
Around the same time, Darwish et al. [9] took a different approach to requirements elicitation: rather than classifying requirements text, they asked which elicitation technique should be used for a given project in the first place. They elaborated their three-component method and selected 17 influential attributes identified from literature, developed a multiple regression model to determine the most critical attributes, and finally trained an Artificial Neural Network to select the elicitation techniques with 81% accuracy. This work is relevant here because it establishes that even the selection of how to elicit requirements can benefit from machine learning, which reinforces the broader case for automation across the RE process.
Returning to classification specifically, Binkhonain and Zhao [4] conducted a survey of machine learning methods on 16 different algorithms and found that SVM is the most commonly used method. They found two ongoing difficulties: absence of shared datasets, and absence of a common definition and classification schema for NFRs. They also pointed out that NFRs and FRs tend to be intertwined in the same document [7], which makes sentence-level classification necessary and hard. This observation applies directly to the institutional documents used in the present study, where promotion criteria and policy standards appear in the same administrative text. Rahman et al. [26] then extended the classification approach into recurrent neural networks by feeding Word2Vec features into LSTM and GRU architectures for classifying NFRs in the PROMISE dataset. They attained precision, recall, and F1 score of 0.973, 0.967, and 0.966 respectively with their LSTM model, which outperforms CNN and GRU variants, and showed that sequence-sensitive architectures can be used to capture the requirement language patterns that are not captured by bag-of-words and even by static embedding approaches.
The embedding choice and the classifier choice are not always independent, and Kabootari et al. [18] looked into this aspect directly. They evaluated the performance of the various combinations of TF-IDF, Word2Vec, GloVe, and BERT embeddings, along with multiple deep learning classifiers on the PROMISE, FNFC, and Fault-prone SRS datasets. To the authors' surprise, GloVe combined with Bidirectional LSTM performed better than all other combinations, even BERT-based ones, due to the high number of parameters in BERT and the need for fine-tuning, which may not be possible with smaller datasets. It is a nice counterpoint to the common belief that more expressive models perform better, and it is also of interest to this study since the within-university datasets are of moderate size, which could have a similar effect on the extent to which BERT can be fine-tuned, as compared to simpler embedding models.
Moving to larger datasets and finer-grained classification, Tahir et al. [29] constructed a dataset of 9,529 functional requirements from 315 software requirements specification documents using the EARS boilerplate in order to reduce natural language ambiguity. When they compared classical ML with CNN and LSTM for subcategory-level FR classification, CNN was found to be the most accurate with an accuracy of 0.93, followed by LSTM with an accuracy of 0.92. They stated that only a handful of previous studies had dealt with FR subcategory classification at all. Their work shows that large, structured requirement datasets can be used to train deep learning models that will perform well on fine-grained classification tasks, which points to a direction the present study has not explored, but could be realized by cross-university data pooling.
One of the more significant methodological changes in this regard has been in the use of static versus contextual embeddings. Peters et al. [25] proposed ELMo, which generates various vector representations for the same word in different sentences, a property that is important for policy language, in which the same word may have different meanings (operational or attributive) based on the surrounding context. Ivanov et al. [16] used mean-pooled ELMo representations, along with a baseline of fastText, for the problem of extracting software requirements from unstructured documents. They manually annotated 7,745 sentences from requirements specifications in the PURE dataset and fine-tuned BERT on this dataset and compared it with fastText and ELMo using an SVM classifier. For the PURE test set, BERT achieved the highest F1 score (0.86), with a precision of 0.92 and a recall of 0.80, whereas ELMo with SVM achieved an F1 score of 0.83, and a more balanced precision of 0.78 and recall of 0.88. The models were then evaluated on a different set of Request for Information documents that the

models were not trained on. While all models suffered from performance loss under this transfer, the performance drop varied by model: BERT maintained the highest precision (0.90), but had the lowest recall (0.71) of the three, while fastText had the highest recall (0.82) at a lower precision (0.54). Their central finding, that fine-tuned BERT transfers with a precision advantage but a recall cost when moved to a document type the training data did not represent, is one of the direct comparison points for the present study's own cross-institutional results, discussed further in Section 5.

Devlin et al. [10] introduced BERT, which extended the contextual representation idea by using deep bidirectional pretraining and has become one of the dominant architectures for NLP classification tasks. In particular, Hey et al. [13] fine-tuned BERT for requirements classification (NoRBERT) and evaluated it on projects not seen during training, which is the evaluation logic adopted in the present study.

3. METHODOLOGY

3.1 Data Collection

The dataset was manually constructed from publicly available institutional documents obtained from the websites of five Nigerian universities. The framing of functional and non-functional requirements follows a document type logic: promotion guidelines are treated as the source of functional requirements because they specify the concrete actions, criteria, and procedural obligations that govern staff advancement, while other institutional policy documents are treated as the source of non-functional requirements because they describe qualities, standards, and constraints rather than operational steps. In all five institutions, promotion guidelines were used as the FR source. The NFR documents were selected per institution to cover a range of policy categories. Table 1 shows the specific documents used for each university.

Each sentence was manually extracted from these documents and labelled as FR or NFR based on whether it described a specific action, operation, or procedural obligation (FR) or a quality attribute, policy constraint, or non-operational standard (NFR). Following the manual extraction, the sentences were compiled into a CSV file per university.

The overall architecture of the system is shown in Figure 1. Each data entry in the five CSV files passes through several preprocessing stages which include de-duplication, lowercasing, stopword removal, lemmatisation, and label encoding. The preprocessed sentences are then passed to the modelling stage where ELMo embeddings are computed and used as feature inputs for both the three classical ML classifiers and the BERT model. In the within-university evaluation setting, the models are trained and evaluated on data from the same institution using an 80/20 stratified split which is represented by the self-loop, while in the cross-university setting, a model trained on one university's data is evaluated directly on the held-out data of each of the remaining four universities. To assess the performance of the models, the output is evaluated using four classification metrics together with a Friedman test and post-hoc Wilcoxon tests. The coefficient of variation is also calculated to measure the generalisation stability of the models.

3.2 Exploratory Data Analysis

The composition of the dataset at each of the five universities is summarized in Table 2 after the dataset has been de-duplicated. The total number of sentences is 3,201, of which 944 are FR and 2,257 are NFR. NFR sentences are more common than FR sentences at all institutions, which is not a sampling artefact, but an inherent characteristic of the source documents. This is also not surprising because positive classes are usually relatively few [22].

There are different degrees of class imbalance in different universities. Four of the five universities are in the moderate-imbalance range (30% to 37%). ABU is an exception to this group, as only 16.8% of its sentences are FR sentences. It is worth pointing out that UNILORIN has the highest FR proportion of the four universities in the moderate-imbalance range, even though it has the fewest FR sentences in absolute terms (95) of any university. This distinction between raw count and proportion is revisited in Section 4, since the two do not turn out to predict cross-university performance in the same way.

Algorithm 1 Sentence Preprocessing and Tokenisation

Require: Set of raw sentences S from dataset CSV file
Ensure: Set of preprocessed sentence strings S_{clean}
Initialize an empty list S_{clean}
function CLEANTEXT(s)
    $s \leftarrow$ lowercase(s)
    $s \leftarrow$ regex_substitute("[^a-zA-Z\s]", "", s) ▷ Remove punctuation and non-alphabetic characters
    $T \leftarrow$ word_tokenize(s) ▷ Split string into individual tokens
    $T_{\text{processed}} \leftarrow [\,]$
    for each token $t \in T$ **do**
        if $t \notin$ NLTK_stopwords **then**
            $t_{\text{lemma}} \leftarrow$ WordNetLemmatizer(t)
            Append t_{lemma} to $T_{\text{processed}}$
        end if
    end for
    return join($T_{\text{processed}}$, " ")
end function
for each sentence $s \in S$ **do**
    Append CLEANTEXT(s) to S_{clean}
end for
return S_{clean}

3.3 Text Preprocessing

All sentences were preprocessed using the NLTK library [5] before feature extraction. The preprocessing procedure is formalised in Algorithm 1.

3.4 ELMo Feature Extraction

ELMo [25] was used to obtain sentence representations by applying a bidirectional language model to the entire input sequence. Unlike static embedding approaches like Word2Vec, ELMo is able to provide a different vector for each word based on its context, which makes it more suitable to capture the variation in meaning of words in policy text that can have different implications across sentence types.

For each sentence, the ELMo layer from TensorFlow Hub returned a tensor of shape $(\text{sequence length}, 1024)$, corresponding to a 1024-dimensional contextual embedding for each token. Mean pooling was applied across the sequence dimension to produce a fixed-length 1024-dimensional vector per sentence. This mean-pooled ELMo representation strategy was borrowed from Ivanov et al. [16], who applied it as one of their baselines to extract software requirements from unstructured documents. For each university, the feature matrix was formed by stacking the sentence vectors row-wise, giving matrices of the shape $(n, 1024)$, where n is the number of sentences in the university's dataset.

Table 1. Source Documents by University

University	FR Source	NFR Sources
UI	Promotion Guidelines and Regulations	ITeMS Policy (including eLearning), Gender Policy
OAU	Promotion Guidelines and Regulations	Data Protection Policy, Gender Policy
UNILORIN	Promotion Guidelines and Regulations	Drug and Alcohol Policy, Sustainable Development Policy (2024)
ABU	Promotion Guidelines and Regulations	Quality Assurance Policy, Teaching and Learning Policy
UNIPORT	Promotion Guidelines and Regulations	Quality Assurance and Quality Control Policy, eLearning Policy, Public Relations and Information Policy

Table 2. Dataset Summary by University

University	Total	FR	NFR	FR (%)	NFR (%)
UI	823	262	561	31.8	68.2
OAU	609	212	397	34.8	65.2
UNILORIN	256	95	161	37.1	62.9
ABU	642	108	534	16.8	83.2
UNIPORT	871	267	604	30.7	69.3

3.5 Classifier Development

3.5.1 Classical Machine Learning Models. The ELMo sentence vectors were used to train three supervised classifiers. The three models were tuned using five-fold cross-validated grid search on the weighted F1 score. Hyperparameter search was done independently for each university in each of the evaluation settings.

Logistic Regression models class membership probability as a logistic function of a linear combination of the input features. It was trained with L2 regularisation, with the regularisation strength C searched over $\{0.01, 0.1, 1, 10, 100\}$ and solvers over {`lbfgs`, `liblinear`} [14].

Random Forest aggregates predictions across multiple decision trees, each trained on a bootstrap sample of the data and a random subset of features [6]. The search covered 100 and 200 trees, maximum depths of None, 10, and 20, and minimum samples per internal node split of 2 and 5.

SVM finds the hyperplane that maximises the margin between class boundaries in the feature space, with a kernel function used to handle non-linear separation [8]. The search covered C values of $\{0.1, 1, 10\}$, kernels of {`rbf`, `linear`}, and gamma values of {`scale`, `auto`}. SVM has been the most frequently used algorithm

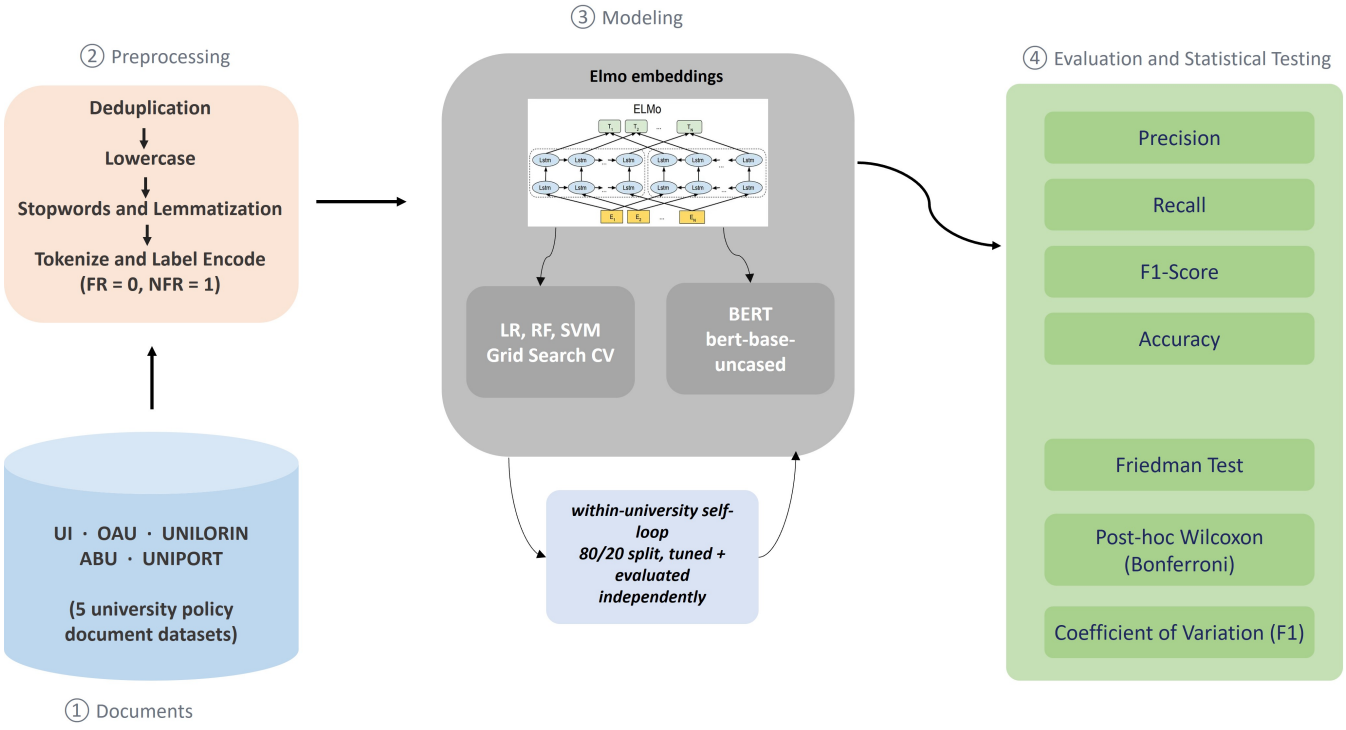

Fig. 1. System Architecture for Cross-University Functional Requirement Elicitation

in prior NFR classification work [4] and has shown strong performance specifically on FR/NFR classification tasks [21].

3.5.2 BERT. The pre-trained `bert-base-uncased` model was loaded with its weights unchanged and a two-class classification head was added on top of the `[CLS]` token representation following the approach of Hey et al. [13]. The model was optimized for 3 epochs, with a learning rate of 2×10^{-5} and the Adam optimiser, using sparse categorical cross-entropy loss over the FR and NFR label classes. The input sentences were tokenised with a WordPiece tokeniser from BERT with a maximum sequence length of 128 tokens, using padding and truncating where necessary.

3.6 Evaluation Design

Two evaluation settings were used.
In the within-university setting, each university's data was split into a training set (80%) and a test set (20%) using stratified random sampling to preserve the class ratio. All four models were trained and evaluated independently within each university. This setting measures how well each model performs when the training and test data come from the same institutional context.
In the cross-university setting, each university served in turn as the sole training source, and each of the remaining four universities served as a separate held-out test set, producing 20 evaluation pairs in total. This protocol directly measures how well a model trained on one institution's documents can classify requirement sentences from institutions it has never encountered. The design borrows its overall structure from the cross-project generalisation protocol of Hey et al. [13] and the cross-domain transfer setting of Ivanov et al. [16].
Accuracy, weighted precision, weighted recall, and weighted F1 score were used to measure model performance. The coefficient of variation (CV) of the F1 score per model was used to quantify the generalisation stability across the pairs of universities; the lower the CV, the more consistent the results will be when using a different pair of universities. To test whether observed performance differences between models were statistically meaningful, a Friedman test was applied to the 20-block cross-university F1 matrix, followed by post hoc Wilcoxon signed-rank tests with Bonferroni correction where the Friedman test reached significance ($p < 0.05$).

4. RESULTS

4.1 Model Configuration and Training

Table 3 lists the hyperparameter settings selected by the five-fold cross-validated grid search for each classical classifier in both evaluation settings. The selected configurations were broadly similar across institutions. Logistic Regression retained the `lbfgs` solver in nine of the ten searches, with `liblinear` chosen only in the ABU cross-university search, and $C = 100$ was selected in five of the ten searches. Random Forest favoured unrestricted tree depth in eight searches, 200 trees in eight searches, and a minimum split size of 2 in nine searches. SVM selected the linear kernel in five searches and the RBF kernel in the other five, with the `scale` setting for gamma in all ten searches, although gamma has no effect under a linear kernel.
BERT was fine-tuned afresh for every evaluation pair, so the four cross-university BERT results that share a training university come from separate fine-tuning runs and not from a single trained model. The final epoch training accuracy ranged from 0.9862 to 1.0000 across the 20 cross-university runs and from 0.9666 to 1.0000 across the five within-university runs. Every BERT model therefore fitted its own training data almost perfectly, and the differences between the two settings reported below arise on held-out data and not from a failure to fit the training data.

4.2 Within-University Performance

Table 4 reports the full classification results for all four models under the within-university evaluation.
Several patterns stand out. First, BERT is the best model at four of the five universities, the only exception being ABU, where Logistic Regression ($F1 = 0.9845$) edges it out by a small margin (BERT $= 0.9695$). ABU's heavily skewed class distribution within its own training set, 83.2% NFR, means the training set contains very few FR examples, and gradient-based fine-tuning appears no more effective than a well-regularised linear classifier in that data-scarce minority-class situation. Second, Random Forest is the weakest model at three of the five universities (UI, OAU, and UNIPORT) in the within-university setting, where it performed noticeably below SVM and Logistic Regression even though all three are trained on the same ELMo features. Third, UNILORIN produces the lowest within-university scores for Logistic Regression and BERT, in line with it being the smallest dataset at 256 sentences after deduplication.
The overall picture from the within-university setting is that all four models are able to learn the FR/NFR distinction reliably when trained and tested on the same institution. The F1 scores for all models at all universities are in the range of 0.8748 to 0.9879, which is similar to the range reported in previous studies for within-corpus requirement classification [13, 21]. The real question for this study is whether that competence applies to another institution that the model has never encountered.
To complement the aggregate scores, Table 5 converts the within-university accuracy values into the number of misclassified test sentences, where the test sets contained 165, 122, 52, 129, and 175 sentences for UI, OAU, UNILORIN, ABU, and UNIPORT respectively. Summed over the five universities, BERT misclassified 16 test sentences, compared with 29 for Logistic Regression, 33 for SVM, and 53 for Random Forest. The counts also show how few sentences separate the models at the smaller test sets. At UNILORIN, a single misclassified sentence changes accuracy by about 0.0192, so the difference between BERT (2 errors) and SVM (3 errors) corresponds to one sentence. At ABU, Logistic Regression made 2 errors against 4 for both SVM and BERT, so its reported advantage at that university rests on two sentences.
Averaged over the five universities, the within-university F1 was 0.9734 for BERT, 0.9499 for SVM, 0.9467 for Logistic Regression, and 0.9171 for Random Forest. The ordering of the classical classifiers is not preserved across settings: SVM ranked above Logistic Regression on average within universities but below it in the cross-university setting (Table 6).

4.3 Cross-University Performance

Table 6 summarises the average metrics per model across all 20 cross-university evaluation pairs.
The drop from within-university to cross-university performance is substantial for every model. BERT falls from an average within-university F1 near 0.97 to a cross-university mean of 0.7910. The classical classifiers go even lower, with Logistic Regression dropping to 0.7254, Random Forest to 0.7167, and SVM to 0.7035. As shown in Table 6, BERT leads on all four average metrics in the cross-university setting, with the gap between it and the classical classifiers being most visible on accuracy and F1.

Table 3. Hyperparameters Selected by Grid Search for the Classical Classifiers

Setting	**University**	**Logistic Regression** (C, solver)	**Random Forest** (depth, split, trees)	**SVM** (C, kernel)
Cross-university	UI	1, lbfgs	None, 2, 200	0.1, linear
	OAU	100, lbfgs	10, 2, 200	10, rbf
	UNILORIN	100, lbfgs	None, 2, 200	10, linear
	ABU	1, liblinear	None, 2, 200	1, rbf
	UNIPORT	10, lbfgs	None, 2, 200	1, linear
Within-university	UI	100, lbfgs	None, 2, 200	1, rbf
	OAU	100, lbfgs	None, 2, 100	10, rbf
	UNILORIN	1, lbfgs	None, 2, 200	1, linear
	ABU	100, lbfgs	None, 2, 200	10, rbf
	UNIPORT	10, lbfgs	10, 5, 100	1, linear

Table 4. Within-University Classification Results

University	**Model**	**Accuracy**	**Precision**	**Recall**	**F1**
UI	Logistic Regression	0.9455	0.9460	0.9455	0.9447
	Random Forest	0.8970	0.9060	0.8970	0.8917
	SVM	0.9394	0.9394	0.9394	0.9387
	BERT	**0.9879**	**0.9883**	**0.9879**	**0.9879**
OAU	Logistic Regression	0.9754	0.9763	0.9754	0.9752
	Random Forest	0.9426	0.9472	0.9426	0.9412
	SVM	0.9754	0.9763	0.9754	0.9752
	BERT	**0.9836**	**0.9840**	**0.9836**	**0.9835**
UNILORIN	Logistic Regression	0.8846	0.8899	0.8846	0.8811
	Random Forest	0.9038	0.9067	0.9038	0.9018
	SVM	0.9423	0.9424	0.9423	0.9420
	BERT	**0.9615**	**0.9637**	**0.9615**	**0.9610**
ABU	Logistic Regression	**0.9845**	**0.9845**	**0.9845**	**0.9845**
	Random Forest	0.9767	0.9774	0.9767	0.9761
	SVM	0.9690	0.9686	0.9690	0.9684
	BERT	0.9690	0.9708	0.9690	0.9695
UNIPORT	Logistic Regression	0.9486	0.9490	0.9486	0.9478
	Random Forest	0.8800	0.8822	0.8800	0.8748
	SVM	0.9257	0.9251	0.9257	0.9251
	BERT	**0.9657**	**0.9662**	**0.9657**	**0.9653**

Table 5. Misclassified Test Sentences in the Within-University Setting

University	**LR**	**RF**	**SVM**	**BERT**
UI (165)	9	17	10	2
OAU (122)	3	7	3	2
UNILORIN (52)	6	5	3	2
ABU (129)	2	3	4	4
UNIPORT (175)	9	21	13	6
Total (643)	29	53	33	16

An interesting thing to note from Table 6 is the relationship between weighted precision and weighted recall in the classical classifiers. In the cross-university setting, Logistic Regression and SVM both have noticeably higher precision (0.8377 and 0.8016, respectively) than recall (0.7401 and 0.7264, respectively). Because weighted recall is equal to accuracy, this gap reflects overall misclassification across both classes and does not by itself show how many FR sentences are missed. The class-level errors are examined in Section 4.5.

Table 7 breaks down the mean cross-university F1 by training university, averaged across the four held-out test universities.

The results of two institutions are worth highlighting for different reasons. UNIPORT-trained BERT achieves the highest mean cross-university F1 of any (training institution, model) combination at 0.8861, and UNILORIN-trained SVM reaches 0.8296 despite UNILORIN contributing the fewest training sentences of the five institutions. At the other end, models trained on ABU record the lowest mean cross-university F1, ranging from 0.5329 (SVM) to 0.5888 (Logistic Regression). ABU's class distribution, in which FR sentences make up only 16.8% of its training data, is one plausible contributor to this pattern, since a model exposed to comparatively fewer FR examples has less material from which to build a representation of the FR class. Class distribution alone does not fully account for the gap, however: UNILORIN contributes fewer FR sentences in absolute terms (95) than ABU (108), yet UNILORIN-trained models transfer considerably better across the other four universities by a wide margin. A more likely contribu-

Table 6. Average Cross-university Metrics per Model

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.7401	0.8377	0.7401	0.7254
Random Forest	0.7377	0.7971	0.7377	0.7167
SVM	0.7264	0.8016	0.7264	0.7035
BERT	**0.8089**	**0.8600**	**0.8089**	**0.7910**

Table 7. Mean Cross-university F1 by Training University

Training University	Logistic Regression	Random Forest	SVM	BERT
UI	0.6298	0.7006	0.5852	**0.8540**
OAU	0.7953	0.7846	**0.8083**	0.7979
UNILORIN	0.8185	0.7996	0.8296	**0.8489**
ABU	0.5888	0.5408	0.5329	**0.5683**
UNIPORT	0.7946	0.7576	0.7613	**0.8861**

tor is that ABU's promotion guidelines are composed with different sentence patterns and phrasing conventions than the promotion guidelines used at the other four universities. Because the classifiers used here, both the ELMo-based classical models and fine-tuned BERT, identify FR sentences from statistical regularities in wording and sentence structure rather than from any semantic understanding of promotion criteria, a training source whose phrasing conventions differ from those of the target institutions will transfer its learned patterns less readily, independent of how clearly or completely those requirements are expressed within ABU's own institutional context.

The full cell-level F1 values are shown in Figure 2 as heatmaps across all four models. Looking at the BERT heatmap specifically, the UNIPORT and UI training rows each have three cells above 0.85, while the ABU row shows uniformly lower values, between 0.503 and 0.646. The heatmaps for Logistic Regression and SVM show a similar pattern but with more variability in the higher-scoring rows, indicating that these models are more sensitive to which specific test university they are evaluated on, even when trained on a favourable source. The UI training row is comparatively low for the classical classifiers (Logistic Regression $F1 = 0.2661$ when trained on UI and tested on ABU) but considerably higher for BERT ($F1 = 0.8538$ for the same pair), which illustrates one specific case where BERT's contextual representations provide transfer that the static ELMo plus classifier pipeline does not.

4.4 Performance by Test University and Pair-Level Comparison

Table 8 is the counterpart of Table 7 which groups the same 80 results by the university used for testing instead of the one used for training. Each cell averages the four pairs that share a test university. BERT obtained the highest mean F1 at four of the five test universities, the exception being OAU, where Logistic Regression (0.7668) was marginally ahead of BERT (0.7582). ABU was the hardest test university for the three classical classifiers, with mean F1 values of 0.5995 (Logistic Regression), 0.6536 (Random Forest), and 0.5692 (SVM), whereas BERT reached 0.7944 on ABU, a lead of 0.1408 over the best classical classifier there. ABU was therefore difficult both as a source and as a target for the classical models, while BERT was affected mainly when ABU was the source.

The ABU column pools four training universities whose effects differ sharply for the classical classifiers. When trained on OAU or UNILORIN, they scored between 0.7549 and 0.8267 on ABU, whereas when trained on UI or UNIPORT they scored between 0.2336 and 0.5876. BERT was much less sensitive to the choice of source for the same target, with F1 values of 0.6983 (OAU), 0.7701 (UNILORIN), 0.8538 (UI), and 0.8556 (UNIPORT).

Table 9 summarises the distribution of the 20 pair-level F1 values per model. BERT achieved the highest F1 in 10 of the 20 pairs, against 4 for Logistic Regression, 4 for Random Forest, and 2 for SVM. It also reached an F1 of at least 0.80 in 12 pairs, compared with 8 for Logistic Regression, 7 for Random Forest, and 8 for SVM, and its median F1 (0.8450) was higher than that of the classical classifiers (0.7627 to 0.7744). The lowest pair-level value for each classical classifier occurred when UI was the training university and ABU the test university (0.2661, 0.3891, and 0.2336 for Logistic Regression, Random Forest, and SVM), while the lowest BERT value (0.5033) occurred for ABU as the training university and UNILORIN as the test university. BERT trailed the best classical classifier in the other 10 pairs, with the largest deficit at OAU as training university and ABU as test university (0.6983 against 0.8267 for Random Forest). Its largest advantage was the UI to ABU pair (0.8538 against 0.3891). This split in pair-level outcomes is consistent with the Wilcoxon result in which BERT did not differ significantly from Logistic Regression after correction.

Because every model suffered when ABU was the training university, the four ABU-trained pairs weigh heavily on the averages in Table 6. Restricting the comparison to the 16 pairs with a training university other than ABU, the mean F1 was 0.8467 for BERT, 0.7606 for Random Forest, 0.7596 for Logistic Regression, and 0.7461 for SVM. BERT remains first and SVM remains last under this restriction, while Random Forest and Logistic Regression exchange places, and all four means rise, by 0.0557 for BERT, 0.0439 for Random Forest, 0.0342 for Logistic Regression, and 0.0426 for SVM.

4.5 Error Analysis from Confusion Matrices

The weighted precision, recall, and F1 values reported so far combine both classes, and weighted recall is equal to accuracy. To examine the errors on the FR class directly, the confusion matrices of all 20 pairs were pooled for each model, with FR treated as the positive class. Table 10 reports the outcome. Across the 20 pairs, the test sets contained 3,776 FR sentences and 9,028 NFR sentences.

BERT produced the fewest false positives, 1,126 against 1,734 for Random Forest, 2,165 for Logistic Regression, and 2,185 for SVM, and consequently the highest FR precision (0.6825) and the high-

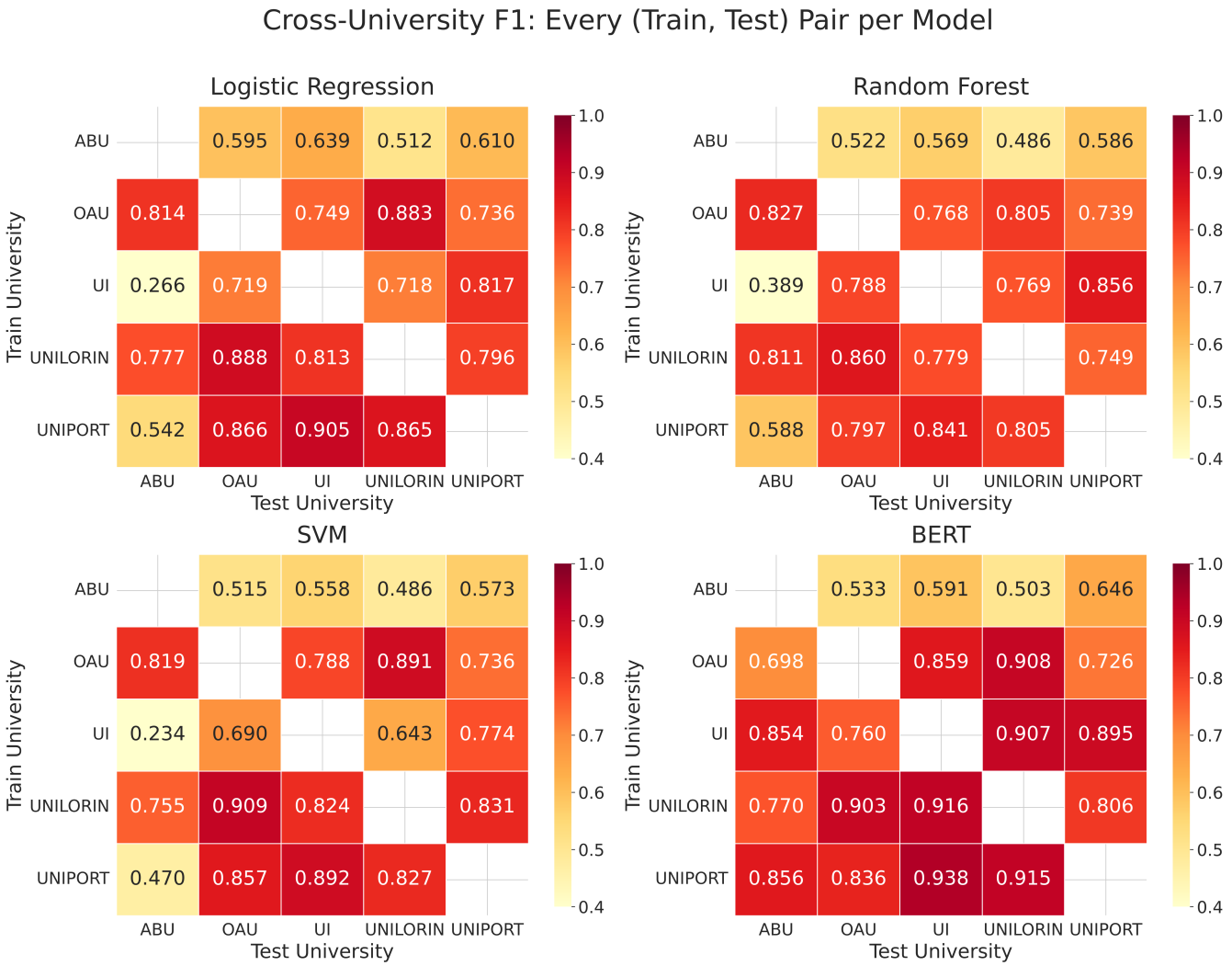

Fig. 2. Cross-university F1 Heatmap for Every (Train, Test) Pair per Model

Table 8. Mean Cross-university F1 by Test University

Test University	Logistic Regression	Random Forest	SVM	BERT
UI	0.7764	0.7393	0.7654	**0.8259**
OAU	**0.7668**	0.7416	0.7425	0.7582
UNILORIN	0.7445	0.7163	0.7115	**0.8082**
ABU	0.5995	0.6536	0.5692	**0.7944**
UNIPORT	0.7397	0.7325	0.7286	**0.7684**

est NFR recall (0.8753). Its FR recall (0.6409) was not the highest, however. Logistic Regression (0.6909) and SVM (0.6608) identified a larger share of the FR sentences, and Random Forest identified the smallest share (0.5744). The higher cross-university accuracy of BERT therefore comes mainly from fewer NFR sentences being wrongly labelled as FR, and not from a larger number of FR sentences being found. The pooled FR precision of the three classical classifiers lay between 0.5331 and 0.5557.

The direction of the error depends on whether ABU is the training or the test university, as shown in Table 11. When ABU was the training university, the pooled test sets held 836 FR and 1,723 NFR sentences. The classical classifiers recovered only 0.0933 (Logistic Regression), 0.0179 (Random Forest), and 0.0048 (SVM) of the FR sentences, and BERT recovered 0.0634, while false positives were almost absent (at most 6 in total). These models labelled nearly every sentence in the target universities as NFR, which is consistent with the 16.8% FR share of the ABU training data. When ABU was the test university, the pattern reversed. The pooled test sets held 432 FR and 2,136 NFR sentences, FR recall was high for every model (between 0.8958 and 0.9977), but the classical clas-

Table 9. Summary of Pair-Level Cross-university F1 Values (20 Pairs per Model)

Model	Mean	Median	Minimum	Maximum	Pairs with Highest F1	Pairs with F1 $\geq$ 0.80
Logistic Regression	0.7254	0.7627	0.2661	0.9052	4	8
Random Forest	0.7167	0.7744	0.3891	0.8596	4	7
SVM	0.7035	0.7647	0.2336	0.9085	2	8
BERT	0.7910	0.8450	0.5033	0.9379	10	12

Table 10. Pooled Confusion Counts over the 20 Cross-university Pairs (FR as Positive Class)

Model	TP	FN	FP	TN	FR Recall	FR Precision	NFR Recall
Logistic Regression	2609	1167	2165	6863	0.6909	0.5465	0.7602
Random Forest	2169	1607	1734	7294	0.5744	0.5557	0.8079
SVM	2495	1281	2185	6843	0.6608	0.5331	0.7580
BERT	2420	1356	1126	7902	0.6409	0.6825	0.8753

sifiers labelled between 0.4363 and 0.5332 of the ABU NFR sentences as FR, against 0.2570 for BERT. For example, the UI-trained Logistic Regression and SVM models labelled 456 and 469 of the 534 NFR sentences in the ABU data as FR while recovering all 108 FR sentences, which produced accuracy values of 0.2897 and 0.2695, whereas the UI-trained BERT model produced 102 false positives for the same test set.

This direction dependence qualifies the gap between weighted precision and weighted recall of the classical classifiers in Table 6. The widest gaps in the pair-level results occurred when UI or UNIPORT was the training university and ABU the test university, for example 0.8640 against 0.2897 for UI-trained Logistic Regression on ABU, and these gaps arise from false positives on the large ABU NFR class and not from missed FR sentences.

4.6 Within-University vs Cross-University Comparison

Figure 3 places the within-university and cross-university F1 scores side by side for each university and model combination for better comparison. The gap between the two settings is not uniform across institutions or models. At UI, BERT drops from 0.9879 within-university to a mean cross-university F1 of 0.8540 when UI is the training institution, a fall of about 13 percentage points. The gap is much larger at ABU: BERT goes from 0.9695 within-university down to 0.5683 as a cross-university training source, a drop of over 40 points. This is the largest gap observed in the comparison, and it shows that within-university performance can be a misleading proxy for a model's usefulness once it is applied to a different institution. A model that achieves a high F1 score of up to 97% in its home institution may not be able to recognize FR sentences in other institutions if the training data does not cover the diversity of the FR sentences in the broader set of target institutions.

UNILORIN tells a different story. Its within-university scores are the lowest among the five institutions for Logistic Regression and BERT, which is expected because of its small dataset size. But surprisingly, its cross-university performance as a training source is competitive with or better than that of much larger institutions such as UI. UNILORIN-trained SVM achieves a mean cross-university F1 of 0.8296, higher than UI-trained SVM at 0.5852. This indicates that the size of the dataset is not the only factor that determines a model's ability to generalise, and that the content and wording of the training examples are at least as important.

Table 12 expresses the same comparison numerically as the difference between the within-university F1 and the mean cross-university F1 obtained when the university served as the training source. ABU produced the largest gap for every model, between 0.3957 (Logistic Regression) and 0.4355 (SVM), while the smallest gaps were 0.0626 for Logistic Regression at UNILORIN and 0.0792 for BERT at UNIPORT. Averaged over the five universities, the gap was smallest for BERT (0.1824) and largest for SVM (0.2464), with Random Forest at 0.2005 and Logistic Regression at 0.2213. The smaller average gap of BERT comes mainly from UI (0.1339 against 0.1911 to 0.3535 for the classical classifiers) and UNIPORT (0.0792 against 0.1172 to 0.1638). At OAU, BERT had the largest gap of the four models (0.1856), and at UNILORIN its gap (0.1121) was close to that of SVM (0.1124).

4.7 Generalisation Stability

BERT not only has the highest mean F1 but also the lowest CV, which indicates that it is the best performing and most predictable model among all the combinations of universities as shown in Figure 4. SVM is the opposite: the mean F1 is the lowest of the four and the CV is the highest at 25.50%. This is a meaningful finding. The performance of SVM is highly variable and whether it performs well or poorly on a specific target institution is highly dependent on which university's data it was trained on, making its behaviour in new institutional settings less predictable. Random Forest achieves a slightly lower CV than Logistic Regression (19.43% vs 21.99%) even though it has a lower mean F1, which indicates that the predictions of Random Forest are at least more consistent in their pattern of errors for the cross-university pairs.

4.8 Statistical Testing

A Friedman test applied to the 20-block cross-university F1 matrix yielded a statistic of 10.2663 and a p-value of 0.0164. This confirms that the observed performance differences between models are statistically significant and not attributable to chance variation across the evaluation pairs.

Post-hoc Wilcoxon tests with Bonferroni correction identified BERT as significantly better than both Random Forest ($p = 0.0042$) and SVM ($p = 0.0064$). No significant pairwise differences were found among the three classical classifiers. The interpretation of this is that Logistic Regression, Random Forest, and SVM perform statistically comparably on cross-university FR elicitation when trained on ELMo features, even though their mean F1 scores differ by up to two percentage points. The Logistic Regression versus BERT comparison did not survive the Bonferroni threshold ($p = 0.0583$, threshold = 0.0083), which means the data does not provide conclusive evidence that BERT is better than Lo-

Table 11. Error Behaviour of Each Model when ABU Is the Training University and when ABU Is the Test University

	ABU as Training University		ABU as Test University	
Model	**FR Recall**	**NFR False Positive Rate**	**FR Recall**	**NFR False Positive Rate**
Logistic Regression	0.0933	0.0035	0.9977	0.5037
Random Forest	0.0179	0.0000	0.9213	0.4363
SVM	0.0048	0.0000	0.9931	0.5332
BERT	0.0634	0.0000	0.8958	0.2570

Table 12. Drop in F1 from the Within-University Setting to the Cross-university Setting (Training Source)

University	**Logistic Regression**	**Random Forest**	**SVM**	**BERT**
UI	0.3149	0.1911	0.3535	0.1339
OAU	0.1799	0.1566	0.1669	0.1856
UNILORIN	0.0626	0.1022	0.1124	0.1121
ABU	0.3957	0.4353	0.4355	0.4012
UNIPORT	0.1532	0.1172	0.1638	0.0792
Mean	0.2213	0.2005	0.2464	0.1824

Table 13. Post-hoc Wilcoxon Signed-Rank Test Results (Bonferroni corrected $\alpha = 0.0083$)

Model A	**Model B**	**Wilcoxon Statistic**	**p-value**	**Significant**
Logistic Regression	Random Forest	81.0	0.3884	No
Logistic Regression	SVM	44.0	0.0215	No
Logistic Regression	BERT	54.0	0.0583	No
Random Forest	SVM	81.0	0.5732	No
Random Forest	BERT	31.0	0.0042	Yes
SVM	BERT	34.0	0.0064	Yes

gistic Regression after accounting for multiple comparisons, even though its mean F1 is about 6.5 points higher.

4.9 Interpretation of Results

The main result of this study is that fine-tuned BERT is a more trustworthy tool for eliciting functional requirements across universities than ELMo with classical classifiers, but the amount of trust depends significantly on the training documents used. This isn't just a tale of BERT outperforming older approaches. In all four models, the within-university results are high enough to indicate that the FR/NFR distinction is learnable from institutional documents and that the ELMo embeddings contain enough semantic signal to aid in that learning. The problem is manifested specifically when the learned model is trained to process documents from another institution and the magnitude of the problem depends on how representative the training institution's data is relative to the wider spectrum of requirement language used in Nigerian universities.

The ABU training scenario is the most obvious example of this. Models trained on ABU data generalise far less well to the other four universities than models trained on OAU, UNILORIN, or UNIPORT data, even though ABU's dataset is of comparable size to OAU's. Class imbalance is one contributor: ABU's training set contains only 108 FR sentences against 534 NFR sentences, which limits how much the model can learn about FR wording from that institution alone. But imbalance is not the whole story, since UNILORIN contributes even fewer FR sentences in absolute terms (95) and still produces some of the strongest cross-university transfer results in the study. A more direct explanation lies in how ABU's promotion guidelines are written.

The classifiers used here, whether the ELMo-based classical models or fine-tuned BERT, learn to identify FR sentences from recurring wording and sentence-level patterns present in the training data, not from any deeper understanding of institutional promotion criteria. If ABU's promotion guidelines express requirements using conventions that differ from those used at the other four universities, a model trained only on ABU data will have learned a set of patterns that do not map cleanly onto the phrasing used elsewhere, regardless of how clearly or completely those requirements are expressed within ABU's own documents. This is a property of cross-institutional pattern matching rather than a judgement on the ABU documents themselves, and it points to a practical recommendation: when selecting training data for a cross-institutional classifier, institutions whose FR sentences are phrased in ways that overlap with the broader convention used across the target institutions should be preferred as primary training sources, alongside considering class balance.

The observation that the smallest dataset, UNILORIN, produces competitive cross-university results, and that UNIPORT-trained BERT has the highest mean cross-university F1 of any (training institution, model) combination at 0.8861 further proves this point. The distributions of FR/NFR examples are fairly even for both institutions (37.1% and 30.7% respectively), so the models trained on them have been exposed to sufficient examples of FR to create good representations of what FR promotion requirement language looks like. Dataset balance and phrasing overlap with other institutions both appear to matter more than raw dataset size for the quality of cross-university transfer.

The error profile of the models in the cross-university setting is also worth noting. The weighted precision of Logistic Regression (0.8377) is higher than its weighted recall (0.7401), but weighted

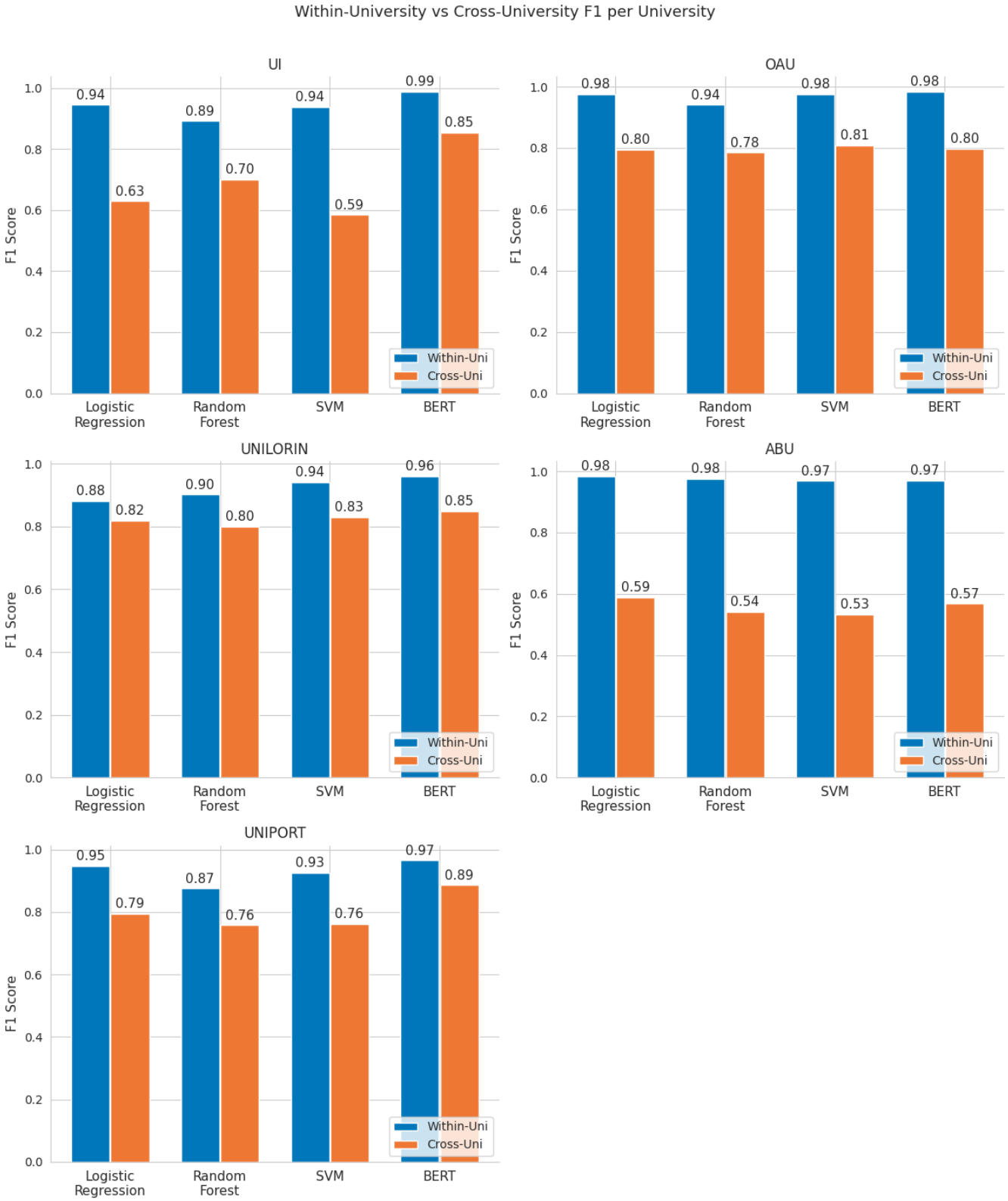

Fig. 3. Within University vs Cross-university F1 per University

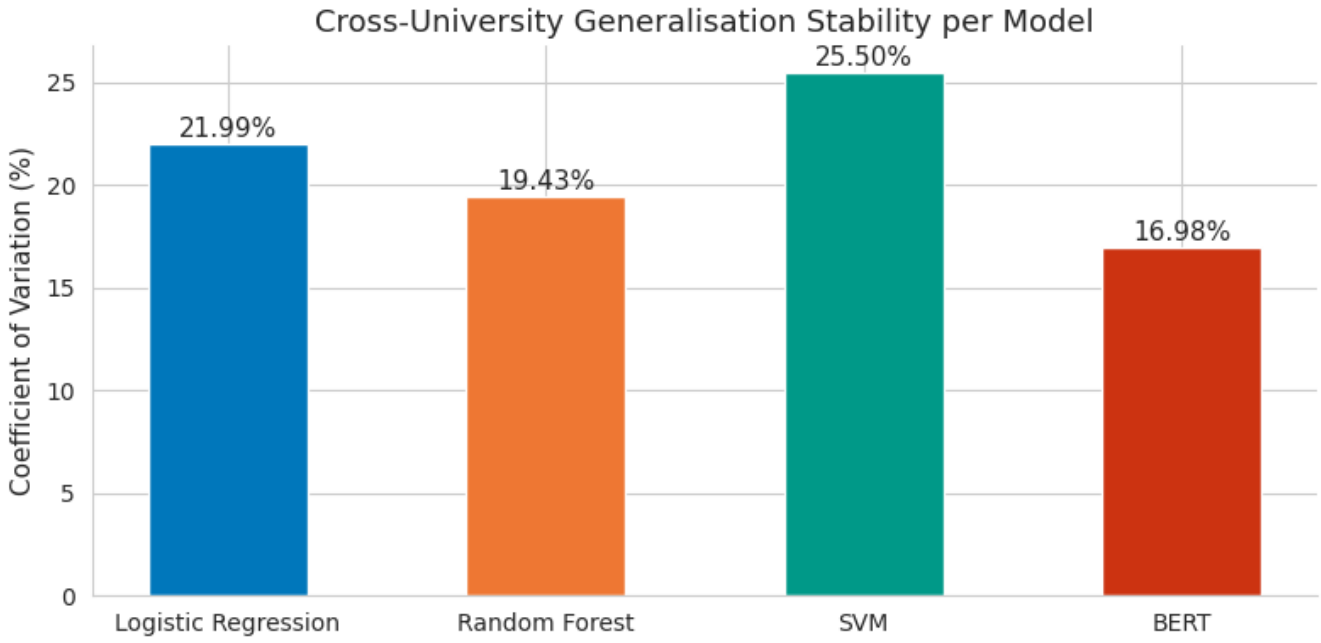

Fig. 4. Cross-university Generalisation Stability per Model

recall is equal to accuracy, so this gap does not by itself describe how many FR sentences are missed. The pooled confusion counts give a more direct picture. BERT has the highest FR precision (0.6825) and the fewest false positives, while its FR recall (0.6409) is below that of Logistic Regression (0.6909) and SVM (0.6608). In an FR elicitation context, recall on the FR class is the more operationally important metric: a missed FR sentence is a requirement that gets dropped from the analysis entirely, while a false positive can be filtered out by a human reviewer. On this basis, the advantage of BERT is not that it finds more FR sentences but that it wrongly flags fewer NFR sentences as FR, which reduces the manual filtering effort. The direction of the error also depends on the source and target universities: models trained on ABU find almost none of the FR sentences in the other universities, whereas the classical classifiers tested on ABU flag between 0.4363 and 0.5332 of its NFR sentences as FR.

The results of the statistical tests are an additional qualification to the model comparisons. The Friedman test establishes that the four models are not equivalent across the 20 cross-university pairs, but the post-hoc analysis shows that the significant differences are specifically between BERT and the two lowest-performing classical classifiers. The Logistic Regression versus BERT comparison, which would arguably be the most interesting pairwise question given Logistic Regression's competitive mean F1, does not clear the Bonferroni threshold. This means there is uncertainty about whether BERT's advantage over Logistic Regression is a stable property of these models on this task, or whether it reflects variation across the specific university combinations tested here.

5. DISCUSSION

5.1 Connection to Prior Work

These results are interpreted with the help of two studies with directly comparable evaluation designs. Hey et al. [13] fine-tuned BERT to classify functional and non-functional requirements using the PROMISE NFR dataset and tested it on projects not seen during training. They reported F1 scores of up to 0.94 on both the seen and unseen project splits. The difference in the current study is significantly larger: BERT trained within a university reaches F1 scores above 0.96, but its mean F1 across the 20 cross-university pairs falls to 0.7910. Software requirement specifications across different projects, as used by Hey et al. [13], tend to follow shared templates and a comparatively narrow technical vocabulary, whereas the promotion guidelines and policy documents used here were drafted independently by five different institutional bodies with no shared template constraining sentence structure or wording. This larger difference in this study is not a sign of BERT's poor performance in the task but rather a reflection of the difference in the origin of the documents.

The second point of comparison is provided by Ivanov et al. [16]. They fine-tuned BERT to distinguish between requirement and non-requirement sentences in the PURE dataset and then evaluated the same model on a collection of Request for Information documents it was not fine-tuned on. BERT achieved the best precision for all three methods tested (0.90), but the lowest recall (0.71) for the transferred documents, while a fastText baseline achieved a lower precision (0.54) but the highest recall (0.82) for the transferred documents. The present study finds a similar direction for BERT under domain transfer, although the gap between precision and recall is much smaller. In the cross-university setting, BERT recorded the highest weighted precision (0.8600) and the highest weighted recall (0.8089) among the four models, but weighted recall is equal to accuracy and combines both classes. On the FR class alone (Table 10), the pooled precision of BERT (0.6825) was higher than its pooled recall (0.6409), which matches the precision advantage and recall cost that Ivanov et al. [16] reported for BERT. The three classical classifiers showed the opposite direction, with pooled FR recall (0.5744 to 0.6909) above pooled FR precision

(0.5331 to 0.5557), which resembles the pattern of higher recall at lower precision reported for the fastText baseline. A likely explanation for the smaller gap is task framing. Ivanov et al. [16] tested a general requirement versus non-requirement judgement across documents drawn from mixed genres, structured SRS documents on one side and free-form RFI text on the other, while the present task separates two requirement types drawn from a single genre category, institutional policy prose, across institutions. A narrower task of this kind may leave BERT's fine-tuned representations less exposed to the kind of structural mismatch that produced the low recall pattern in the RFI transfer setting reported by Ivanov et al. [16].

5.2 The ABU Pattern and What It Suggests

To recall, the results discussion so far has focused on ABU's weak cross-university transfer. This section takes a closer look at this pattern. The obvious factor which can be attributed to this pattern is class distribution. ABU's training set contains only 108 FR sentences and 534 NFR sentences, which is the most severe imbalance among the five universities. But as noted in Section 4.6, class distribution alone does not account for the gap, since UNILORIN contributes fewer FR sentences in absolute terms and still produces considerably stronger cross-university transfer.

No definitive claim is made about what caused ABU's poor transfer performance. However, there are contextual details about ABU's promotion guidelines that are worth noting as possible contributors. The dataset for this study was constructed from ABU's 2024 promotion guidelines, but in August 2025, ABU publicly announced that it would set aside those 2024 guidelines and revert to its 1988 guidelines for the 2025 promotion exercise. They stated that this was done in order to avoid delays while two committees continued reviewing the 2024 version [24]. This is not an isolated observation. El-Yakubu [11], a professor within ABU, had already noted that the university was operating under promotion guidelines that were 34 years old at that point, referring to the same 1988 document, and described departments where there were eighteen professors and eleven readers but not a single entry-level academic staff member. He attributed this situation directly to the absence of clear performance-based promotion criteria. A promotion guideline that was still being reviewed by committees at the time this dataset was produced and later shelved in favour of a guideline written 37 years earlier may not be a final or settled statement of the requirements of promotion at ABU.

Also, more recently, ABU began a verification process of a professor whose academic profile, which was made public, attributed to him an unusually large number of qualifications [2]. This research is not drawing a direct line from any of this to the model performance numbers. The researchers were conscious of the very thin line between correlation and causation. It is rather suggested that the promotion guideline text used as ABU's FR source may be less internally consistent than the corresponding documents from the other four universities, and that this is one plausible reason, among others, why models trained on ABU data transfer less well across institutional boundaries. This pattern should be highlighted for future work that will be using ABU documents as a training source.

6. CONCLUSION

This paper evaluated the use of ELMo embeddings and fine-tuned BERT for the cross-university elicitation of functional requirements from Nigerian university institutional documents. A dataset of 3,201 manually labelled sentences was constructed from the promotion guidelines and policy documents of five universities: UI, OAU, UNILORIN, ABU, and UNIPORT. Under within-university evaluation, all four models performed well. BERT achieved F1 scores between 0.9610 and 0.9879 and Logistic Regression performed comparably or better than BERT on ABU. Under cross-university evaluation across 20 (train, test) pairs, performance dropped for every model, with BERT achieving the highest mean F1 of 0.7910 and the lowest coefficient of variation at 16.98%, which shows it as both the most accurate and the most predictable model across institutional boundaries. A Friedman test confirmed that the model differences were statistically significant, and post-hoc Wilcoxon tests with Bonferroni correction showed that BERT was significantly better than Random Forest and SVM. ABU produced the least transferable training signal of the five universities. This pattern was linked to both its class distribution and the unsettled state of its promotion guidelines which were still under committee review at the time the dataset was constructed. On the technical path, these findings show that contextual language representations can support FR elicitation across different university contexts, and that BERT fine-tuning is preferable to ELMo-based classical classifiers when cross-institutional generalisation is the goal. Closing the gap between within-university and cross-university performance will require either broader training data drawn from a wider range of institutional phrasing conventions or domain adaptation techniques that can leverage small amounts of labelled data from the target institution.

6.1 Limitations

There are some restrictions on this work. The results of the five universities are not necessarily representative of other universities in Nigeria or any other educational setting. The documents used are also a particular type of institutional requirement documents and the findings may not be generalised to other categories of institutional requirement documents like government regulatory documents or corporate policy manuals.

The document-level labelling approach, where a sentence from promotion guidelines is assigned FR and a sentence from policy documents is assigned NFR, introduces noise at the sentence level. Not every sentence in a promotion guideline strictly describes an operational obligation, and not every sentence in a gender policy strictly describes a non-operational constraint. Sentence-level annotation by trained requirements engineers would produce a cleaner ground truth, though it would also require considerably more annotation effort across 3,201 sentences.

UNILORIN's small dataset size (256 sentences after deduplication) means that results involving UNILORIN as a training source, while competitive, rest on a narrow base. The fine-tuned BERT model trained on UNILORIN's 204 training sentences completes three epochs very quickly and may not have had sufficient exposure to develop stable representations for all the linguistic patterns it will encounter in the larger test universities.

6.2 Future Work

The most direct extension of this work is to include university documents from universities across different regions of Nigeria and even beyond in order to test how well the model generalises. In terms of modelling, pre-training BERT on a collection of institutional and policy documents before fine-tuning on the task at hand, as found to be beneficial in other domain-specific contexts [12], may mitigate the gap between universities. The FR/NFR distinction is prototype matching in nature and may be better modelled by

a similarity-based encoder, such as Sentence-BERT [28], which is worth considering for evaluation.
Another direction to go is active learning. The time-consuming nature of annotating sentences from new universities makes it very useful to have a method that uses a small seed of labelled examples from the target institution to update the model incrementally. It is also a natural choice for few-shot learning methods in which a small number of labelled FR examples from a new institution might be used to adapt the classifier without full retraining.

References

[1] J. Abbas, A. Ahmad, S. M. Shaheed, R. Fatima, S. Shah, M. Elaffendi, and G. Ali. Classification and comprehension of software requirements using ensemble learning. *Computers, Materials & Continua*, 80(2):2839–2855, 2024.

[2] S. Akhaine. Abu begins verification of professor's multiple academic credentials. The Guardian Nigeria, August 2026.

[3] B. I. Ayinla and D. Adeniniji. A comprehensive systematic review of dual use research classification using artificial intelligence. *International Journal of Computer Applications*, 187(135):20–27, 2026.

[4] M. Binkhonain and L. Zhao. A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Systems with Applications: X*, 1:100001, 2019.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[6] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[7] Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, and Peter Solc. Automated classification of non-functional requirements. *Requirements Engineering*, 12(2):103–120, 2007.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] N. R. Darwish, A. A. Mohamed, and A. S. Abdelghany. A hybrid machine learning model for selecting suitable requirements elicitation techniques. *International Journal of Computer Science and Information Security*, 14(6):380–391, 2016.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[11] B. El-Yakubu. Abu at 60: Why it is not yet great and what to do about that. Daily Trust, November 2022.

[12] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

[13] Tobias Hey, Jan Keim, Anne Koziolek, and Walter F. Tichy. Norbert: Transfer learning for requirements classification. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 169–179, 2020.

[14] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 3rd edition, 2013.

[15] Ibrar Hussain, Leila Kosseim, and Olga Ormandjieva. Using linguistic knowledge to classify non-functional requirements in srs documents. In *Natural Language Processing and Information Systems*, volume 5039 of *Lecture Notes in Computer Science*, pages 287–298. Springer, 2008.

[16] Vladimir Ivanov, Andrey Sadovykh, Alexandr Naumchev, Alessandra Bagnato, and Kirill Yakovlev. Extracting software requirements from unstructured documents. *arXiv preprint arXiv:2202.02135*, 2022.

[17] O. Julius, S. F. Verkijika, and I. C. Obagbuwa. Hybrid semantic-syntactic nlp framework for intelligent grading of short answers and cloze questions. *Applied Sciences*, 16(7):3191, 2026.

[18] M. Kabootari, Y. Abdeahad, Y. Kheirkhah, and E. Kheirkhah. Enhancing software requirements classification: a comparative study of deep learning model integration with embedding techniques. *Computing*, 107:191, 2025.

[19] Gerald Kotonya and Ian Sommerville. *Requirements Engineering: Processes and Techniques*. John Wiley & Sons, 1998.

[20] T. Kotzé and H. Smuts. Model for knowledge capturing during system requirements elicitation in a high reliability organization. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2018)*, pages 305–312, 2018.

[21] Zijad Kurtanović and Walid Maalej. Automatically classifying functional and non-functional requirements using supervised machine learning. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*, pages 490–495, 2017.

[22] M. Limaylla-Lunarejo, N. Condori-Fernandez, M. Rodríguez Luaces, and O. Karras. Improving the multi-class classification of non-functional requirements in spanish: A study of dataset balancing and performance. *Empirical Software Engineering*, 31:6, 2026.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS 26)*, volume 26, pages 3111–3119, 2013.

[24] MySchoolNews. Ahmadu bello university clarifies use of 1988 guidelines for 2025 promotions exercise. MySchoolNews, August 2025.

[25] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[26] Md. A. Rahman, Md. A. Haque, Md. N. A. Tawhid, and Md. S. Siddik. Classifying non-functional requirements using rnn variants for quality software development. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation*, pages 25–30, 2019.

[27] A. Rashwan, O. Ormandjieva, and R. Witte. Ontology-based classification of non-functional requirements in software specifications: A new corpus and svm-based classifier. In *2013 IEEE 37th Annual Computer Software and Applications Conference*, pages 381–386, 2013.

[28] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[29] T. Tahir, H. Jahankhani, K. Tasleem, and B. Hassan. Cross-project multiclass classification of ears-based functional requirements utilizing natural language processing, machine learning, and deep learning. *Systems*, 13(7):567, 2025.

[30] M. A. Umar and K. Lano. Advances in automated support for requirements engineering: a systematic literature review. *Requirements Engineering*, 29(2):177–207, 2024.