Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 1 (2026) 150 - 159 e-ISSN: 2580-0760

Detecting Language Anxiety in Indonesian Students: Deep Learning and Traditional Classification Methods for English Learning Anxiety

Kemas Muslim Lhaksmana[1], Muhammad Sya'bani Falif[2], Iis Kurnia Nurhayati[3], Muhammad Yudhi Rezaldi[4,5], Esa Prakasa[5], Rikman Roedavan[6]

[1]Department of Informatics, School of Computing, Telkom University, Bandung, Indonesia
[2]Department of Data Science, School of Computing, Telkom University, Bandung, Indonesia
[3]Department of Communication Science, School of Communication and Social Sciences, Telkom University, Bandung, Indonesia
[4]Department of Visual Communication Design, School of Creative Industry, Telkom University, Bandung, Indonesia
[5]Research Center for Data and Information Sciences, National Research and Innovation Agency (BRIN), Bandung, Indonesia
[6]Department of Multimedia Engineering Technology, School of Applied Sciences, Telkom University, Bandung, Indonesia

[1]kemasmuslim@telkomuniversity.ac.id, [2]msfalif@student.telkomuniversity.ac.id, [3]iiskurnia@telkomuniversity.ac.id, [4]myud001@brin.go.id, [5]esap001@brin.go.id, [6]rikman@telkomuniversity.ac.id

Abstract

Mastery of the English language represents a fundamental determinant of professional achievement, particularly for individuals seeking to develop their careers and participate in international contexts. However, when learning a foreign language such as English, Indonesian students may experience language anxiety that cause them to hesitate to practicing English in the class, both in orally and in writing. For English teachers, it is crucial to identify students experiencing language anxiety early on, so that they may provide appropriate teaching strategies and interventions from the first class meeting. To address this issue, this study compares machine learning methods to provide a solution for early detection of students experiencing language anxiety. Furthermore, these methods are classification models, including LSTM, GRU, decision tree, naïve Bayes, logistic regression, and SVM. The implementation of each of these models is combined with different text representation techniques, such as Word2Vec, BERT, FastText, Glove, and TF-IDF. The advantage of our model is that despite the imbalance, limited, and smaller than baseline dataset size, this research finds that GRU with focal loss achieves the highest F_1 score of 0.89. This result outperforms our baseline and thus suggests that this method is effective in detecting students who experience language anxiety.

Keywords: classification; deep learning; english learning; GRU; language anxiety

How to Cite: K. M. Lhaksmana, M. S. Falif, I. K. Nurhayati, M. Y. Rezaldi, E. Prakasa, and R. Roedavan, "Detecting Language Anxiety in Indonesian Students: Deep Learning and Traditional Classification Methods for English Learning Anxiety", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 1, pp. 150 - 159, Feb. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i1.7109

Received: August 21, 2025
Accepted: February 6, 2026
Available Online: February 25, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

As globalization and cross-cultural interaction become more common, effective communication skills are increasingly becoming more important. Those who seek to improve their career abroad are expected to be proficient in foreign languages, especially English. One of the main obstacles in learning and using non-native language is language anxiety, which is the uncomfortable emotional and feeling of worry when communicating using a new non-native language. The problem of foreign language anxiety (language anxiety for brevity) has been studied in foreign language learning as one of the issues to achieve language proficiency. Such anxiety is often experienced by language learners that causes difficulty to interact and communicate with speakers of their foreign languages [1], [2]. Therefore, educators need to identify language anxiety among the students as early as possible to provide appropriate teaching strategies and it is necessary to find the fastest method to solve this issue.

The concept of language anxiety was firstly addressed by Horwitz et al. in their seminal work [1]. The work distinguishes language anxiety with the other types of

anxiety to investigate its effect on foreign language learning. Such anxiety, which is indicated by feelings of discomfort and psychological tension [3], has been proven to interfere with the language learning process and negatively impact academic and social performance [4]. The discourse on language anxiety has become more relevant since COVID-19 pandemic, where people are exposed to new sources of stress due to massive online learning and their working environment at home [5]. Responding to these challenges, much research has been done to study different kinds of intervention, from creating a supportive class environment [4] to implementing gamification learning style to reduce stress [6]. In general, this research emphasizes the importance of continuous intervention to assist students to handle multifaceted challenges of language anxiety.

To address these multifaceted challenges of language anxiety, educators are increasingly turning to technology. The advancement of AI has driven and redefined pedagogical innovation both in theory and practice. The role of AI in education research shows its significance in personalized learning, instructional support, and administrative work efficiency [7]. AI technology has also been applied to improve language learning, such as chat bots for language learning, EnglishBot, that is designed to facilitate English speaking practice, as well as to solve the issue of limited interaction in traditional learning environments [8]. Various AI technology and applications have been explored to be implemented in language learning, including the use of conversational AI to the collaboration between AI-based applications and educators in improving the quality of language learning [9]. Such systems have demonstrated their capabilities to support intelligence amplification and to assist educators in managing their classroom activities.

While most of the research on AI in education focuses on improving learning quality, there is a growing interest in utilizing machine learning for detecting and overcoming affective factors, including anxiety and motivation that affect learning performance. Machine learning has been implemented to diagnose mental health conditions and various anxieties among learners. For example, traditional machine learning algorithms, such as support vector machines (SVM), Naïve Bayes classifiers, and decision trees, demonstrate acceptable performance in detecting programming anxiety using AutoML, especially using logistic regression of which accuracy achieves 97.8% [10].

In addition to traditional machine learning algorithms, deep learning algorithms have been explored to improve language learning. For example, long short-term memory (LSTM) has been proven to outperform traditional algorithms in classifying text to detect mental health at 91% accuracy [11]. The comparison between LSTM and gated recurrent unit (GRU) on Yelp Review corpus has concluded that GRU tends to be more efficient and achieve slightly higher accuracy on smaller dataset and longer text sequence [6]. These findings motivate researchers in the field to further explore other algorithms for detecting language anxiety from text corpora.

Despite the fact that language anxiety has been studied for more than 40 years, machine learning is still a relatively new tool for identifying and comprehending language anxiety. Previous studies have examined language anxiety in various contexts, such as German language learning [12] and English language learning among Turkish students [13], but these studies are often limited to specific cultural or demographic settings. For Indonesian students, research in this area is still limited, especially those using computational methods.

The potential of machine learning models for this purpose has been investigated recently. The gradient boosting (GBT) and SVM approaches showed the best results, with an accuracy rate of 0.843, according to a comparison of deep learning and standard learning approaches applied to the performance and demographic profiles of 292 Indonesian students [14]. In a different experiment involving 324 English essay assignments, several machine learning algorithms were applied and compared. The results showed that the GRU architecture performed best, with an accuracy rate of 0.82 and an F_1 metric of 0.82 [15]. These findings suggest that although current models demonstrate a fairly good level of effectiveness, there is still room for improvement, particularly in addressing data imbalance and reducing bias caused by low-quality or AI-generated text.

This research aims to address the challenges of early identification of language anxiety, particularly among Indonesian students learning English. Problems such as manual identification methods similar to questionnaires are time-consuming, require linguistic and psychological expertise, and are quite demanding to implement on a large scale. Furthermore, questionnaires are susceptible to various types of bias, such as social desirability bias, response bias, and subjectivity from both respondents and researchers [16]. The application of machine learning to identify language anxiety in foreign language learners is crucial to overcome these limitations [17-19], however existing models still struggle with data imbalance and unfiltered AI generated contents. Therefore, this study introduces a novel approach that filters AI-generated content and utilizes a GRU model optimized with focal loss to address class imbalance. This method achieves a state-of-the-art F_1 score of 0.89, significantly outperforming existing baselines in early language anxiety detection.

Traditional questionnaire and data analysis took some time to identify language anxiety [12], [13], [16]. Meanwhile, current faster methods using machine learning have issues with data and AI generated contents. Gozali et al. achieved accuracy of 0.84 using GBT and SVM, however the dataset is not text and based on student's performance and demographic data

[14]. A research by Falif et al. shows that GRU with binary cross entropy resulted in F_1 score of 0.82, however there is still issue with data imbalance and unfiltered AI generated content [15].

The contribution of this paper is demonstrating that despite the limited dataset size, language anxiety in students can be successfully detected by appropriately implementing preprocessing steps to improve data quality and selecting the correct loss function. These steps and findings distinguish this research from the previous [14], [15]. For the former, unlike the standard classification, this research validates the student essays by excluding those that are potentially generated by AI.

This step is important to make sure that the detection of language anxiety shows the real student's writing and not the fake pattern. The other issues that need to be handled are imbalance class and limited size of the data. For the former, the GRU and LSTM model is optimized with focal loss. As for the latter, the implemented model demonstrates higher than baseline performance despite smaller dataset size.

Machine learning offers several advantages in identifying language anxiety:

Passive non-intrusive assessment: Language anxiety can be identified from students' written essays without requiring them to complete self-report questionnaires. This approach minimizes response bias and participant fatigue, allowing a more natural and objective observation of anxiety-related patterns.

Unconscious expression in writing: Since essays have been commonly used as one of the learning methods in language learning, such datasets can be easily acquired and processed further. Subtle anxiety can be detected in the text by language experts and thus its pattern can be captured in machine learning classification.

Scalability and reusability: Although developing machine learning models takes time and resources, once implemented, they can quickly analyze large amounts of text with little human effort. In addition, the models can be reused or updated for different groups of learners, making them suitable for scalable and continuous monitoring of language anxiety in educational settings.

In this research, multiple machine learning algorithms were employed and evaluated to identify signs of language anxiety from student essays. The models include LSTM, GRU, decision tree, naïve Bayes, logistic regression, SVM, and GBT. The classifiers were evaluated together with diverse feature extraction approaches, including TF-IDF as well as embedding techniques like Word2Vec, FastText, GloVe, and BERT. The models were selected because they represent both deep learning and traditional machine learning paradigms and have demonstrated effectiveness in classifying language anxiety from text in previous research [14], [15].

Building upon these prior works, our proposed GRU-based implementation includes preprocessing to exclude potentially AI-generated essays and applies focal loss to handle class imbalance [20], achieving an F_1 score of 0.89. The F_1 score was used as an evaluation metric because it provides a balanced assessment of precision and recall, two important aspects in identifying anxious learners, and allows for direct comparison with previous research in this area [21]. These findings highlight the novelty and contribution of our research in the early detection of language anxiety using machine learning.

This research is part of a larger project aimed at developing a metaverse-based language learning platform that integrates machine learning to minimize language anxiety. In the related research, we have developed MetaSpeak, an application that enables foreign language learners to practice and enhance their language skills within a virtual environment, thereby reducing the potential for language anxiety

2. Research Methods

2.1 Language Anxiety Detection

Early detection of language anxiety that is conducted in this research is part of a broader project in the form of class action research (Figure 1). The class action research generally follows the plan, act, observe, and reflect cycle [22]. The research has been conducted on two English classes: an experiment class and a control class. In the experiment class, machine learning algorithms will be implemented to detect language anxiety at an early stage, whereas the control class solely uses traditional language anxiety assessment using questionnaires. This stage is the focus of the current research, while the remaining stages will be addressed in the related ongoing research. After the students are assessed, the process of English learning in this class will use MetaSpeak, a 3D animated application offering a metaverse style and immersive experience, in which students can practice speaking English with virtual characters. The successful implementation of technology in language education requires collaboration between human teachers and technological tools, rather than relying solely on the technology itself [9].

Measuring language anxiety among students via self-report questionnaires has become a common practice. However, the quality of such self-report instruments should be improved in several aspects, including sample and research design, survey design, scale design, reliability, content validity, and construct validity [23]. In this context, machine learning approaches offer advantages over self-report questionnaires since they are inherently passive, non-intrusive, capturing unconscious expression in writing, and providing greater scalability and reusability.

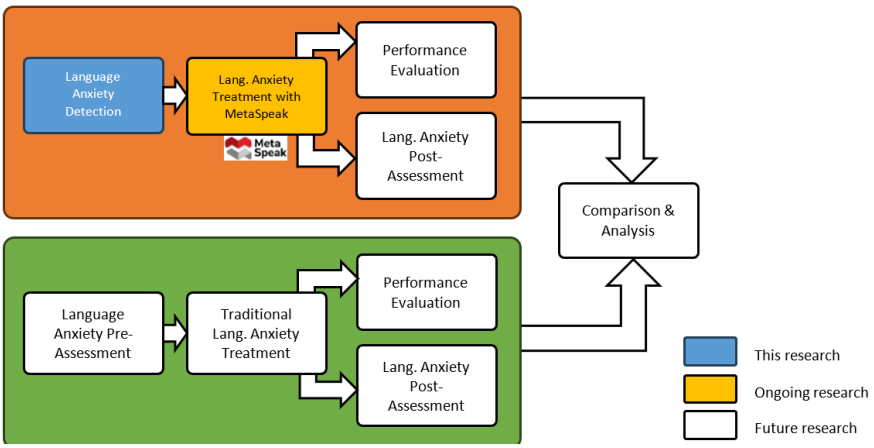


Figure 1. In this research, we investigate machine learning algorithms for early detection of language anxiety among Indonesia students who enrolled in English courses (blue rectangle). This research is part of a research project to develop a self-paced multilingual learning platform, MetaSpeak.

As illustrated in Figure 2, our research begins with exploring the dataset, which is the English essays. The data exploration is performed to understand its composition and identify potential issues such as imbalance, noise, or missing values. Subsequently, the research workflow proceeds with several preprocessing stages, including text cleaning (eliminating redundant symbols and characters), case normalization (transforming all words into lowercase), tokenization, lemmatization, and stopword elimination to ensure data consistency.

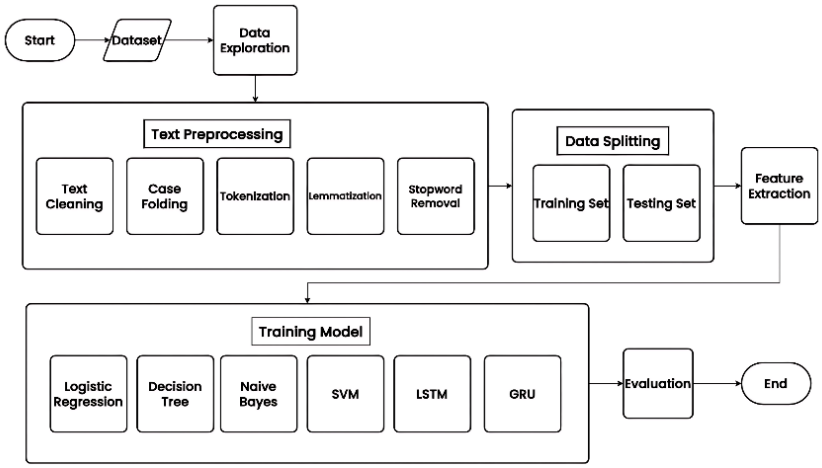


Figure 2. Research workflow of the machine learning models implementation for language anxiety detection. The dataset consists of English essays written by students, labeled based on their anxiety levels obtained from questionnaire responses. The process starts from data exploration, text processing, data splitting, feature extraction, training, until evaluation.

For clarity, the complete steps to perform early detection of language anxiety among the Indonesian students enrolled in the English courses are outlined as follows:

Collecting and preparing English essays: Essay assignment is usually part of English courses, making the essay dataset relatively easy to obtain, or has been already available as in the case of our research. Teachers can use previously submitted essay assignments or request the students to work on new essay assignments as a pretest. After applying the necessary preprocessing, these essays are ready as the dataset to be processed further.

Labelling the English essays based on the result of English language anxiety scale (ELAS) questionnaires: The student essays collected from the earlier phase were labeled to make them suitable for training machine learning models. Here, students are requested to fill in the ELAS questionnaire, which is a standardized instrument validated in previous research [24]. Although self-report data has a risk of subjectivity, language anxiety is fundamentally an internal emotional state. Therefore, the student's self-assessment is the most direct and authentic measure of their condition. Using external experts to label every student is not practical for large-scale implementation in schools. The labeling phase is carried out solely for model training. In deployment scenarios, after the learning algorithms are trained and operational, there is no further requirement for questionnaire data gathering or manual labeling.

Questionnaires are distributed containing several questions related to the participants' experiences in

learning a foreign language. In the control class, levels of language anxiety are evaluated through questionnaires, while in the experimental class, these instruments act as benchmark data to validate the predictive performance of machine learning systems.

The questions investigate situations such as feeling anxious during foreign language use or nervous when performing communicative tasks in front of others. Participants' responses were scored using a Likert scale ranging from 1 to 5, indicating increasing levels of anxiety. The detailed criteria for the scoring and the dataset distribution are explained in the following Section 2.2.

After completing the questionnaire, each participant's English essay was classified using a machine learning model, which served as the core of this research's analysis process. These essays served as training data for the model. Through textual analysis, linguistic cues such as sentence construction, word usage, and emotional tone were examined as potential indicators of anxiety, consistent with findings in the Natural Language Processing (NLP) literature [25].

Following the preparation stage, the dataset was split into training and testing subsets. Both sets' features were extracted using text representation approaches such TF-IDF and pre-trained embeddings (Word2Vec, FastText, GloVe, and BERT). During training, the deep learning model used an embedding layer to learn word vectors directly. By combining standard and deep learning methodologies, this work enables a comprehensive comparison to find the best strategy for detecting linguistic anxiety [26-28]. The purpose of this procedure is to convert textual input into numeric vectors appropriate for a classification algorithm.

Finally, the machine learning models were trained using a variety of algorithms, including traditional techniques, logistic regression, decision tree, naïve Bayes, and SVM, as well as deep learning architectures such as LSTM and GRU. Model performance was then evaluated using the F_1 score to assess the extent to which the models were able to detect language anxiety from students' essays.

2.2. Questionnaire Dataset

This research builds on previous work [15] that uses the same English essay. To maintain the integrity and authenticity of the data, essays which were created by AI were filtered using a pre-trained random forest model. This model is available on Hugging Face and has been trained with corpus that included both human-generated and AI-generated texts, using linguistic and stylistic cues to distinguish them [29], [30]. The model accuracy is > 98%, which is claimed by the company who created the model. The filtering process finally excludes the potentially AI-generated content, and thus only leaving human-written essays for analysis.

The final dataset consists of 216 valid samples from a total of 359 essay submissions. Participants had completed 31 questions to assess their level of foreign language anxiety. The survey has a five-point Likert scale, with higher scores of greater anxiety. Since the focus of this research is to detect the potential of language anxiety, the questionnaire, which changed from five levels into two levels. The mean score from the questionnaire is translated as follows. If scores above 2.5, then they were classified as “anxious,” while scores equal to or below 2.5 were considered “not anxious.” This binary classification is used as the target variable used by the model.

Table 1 shows several excerpts from student essays together with their anxiety labels. These excerpts, taken from longer essays, illustrate linguistic common for both anxious and non-anxious learners. On the Likert scale of 1 to 5, the score of 3 is neutral. However, for the purpose of early detection, it is critical to capture the student who has moderate anxiety level. Therefore, the cutoff is set at 2.5 to include this moderate group. This approach prioritizes the sensitivity of the model. It is safer to classify the moderate student as 'anxious' rather than missing them (false negative). This ensures that the teacher can pay attention to all students who potentially have language anxiety problems. The method of anxiety assessment using questionnaires as the above has been widely used, including in ELAS [24], which is used in this research to obtain the data.

Table 1. Sample excerpts from students' essays with corresponding language anxiety labels

Essay excerpt	Label
Currently, social networking, especially in the case of sharing digital content, has become the order of the day for millions...	Anxiety
The innovation of technology has transformed almost every aspect of our lives, even the way we…	Anxiety
With the Increasing focus on Generative Al, 2024 marks an important tipping point for Al ethics and Law...	No Anxiety
In recent years, more adolescents are experiencing mental health issues. According to the World Health Organization (WHO)...	No Anxiety

2.3 Data Exploration

The dataset in Figure 3 shows that the value between anxious students (168) and non-anxious students (48) is imbalance.

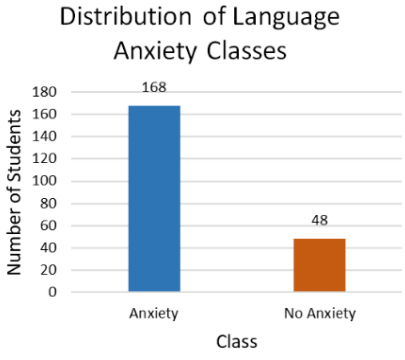


Figure 3. Class distribution of students experiencing language anxiety (168) and those not experiencing anxiety (48).

This distribution created a challenge in model training, because classifiers tend to be biased towards the majority class. To fix this problem, traditional models apply class weighting to balance the representation. This fix can reduce the influence of correctly classified cases and emphasize more difficult cases. With this method, a dataset for analyzing the comparative performance for different text and classification will have more balanced value to detecting language anxiety.

2.4 Classification Algorithms

This research uses machine learning to identify foreign language anxiety from the texts submitted by students. Each classification method used in this research has their advantages and disadvantages. Especially in interpretability function, computational efficiency, and the ability to get linguistic or sequential patterns. The models used in this research include LSTM, GRU, logistic regression, naive Bayes, decision tree, and SVM. Combining these traditional and deep learning classifiers, this research attempts to assess the comparative performance of each model under various schemes. This comparative analysis provides an evaluation of the model's effectiveness in identifying foreign language anxiety from students' text.

The LSTM model is an extension of the recurrent neural network (RNN). This model is used to get long-term dependencies in sequential data results. The result can be used to model the linguistic context and temporal relationships. This model ability makes it well-suited for detecting language anxiety. However, LSTMs are computationally intensive and require large amounts of annotated data to achieve optimum results [31].

Meanwhile, the GRU offers a more solid RNN design while retaining much of the sequential learning capabilities. Similar to the previous one, but with fewer parameters. The GRU typically trains faster and requires less memory. Although sometimes, it is less effective than the LSTM when handling very long contexts [32].

One method that can be used to solve binary classification problems is the logistic regression. This method offers advantages in terms of efficiency, ease of implementation, and ease of interpretation. The logistic (sigmoid) transformation is used as the basic approach in this method, with the aim of predicting the probability of an instance belonging to a particular class. However, its performance is undeniable, as the linear relationship between features and the target variable can limit its performance in complex classifications [33].

The goal of designing a SVM model is to find a separating plane (hyperplane) to maximize the margin between several different classes. It is highly effective in high-dimensional feature spaces and is well-suited for sparse text representations. However, when used for large-scale datasets, the SVM method becomes very laborious and less than optimal when dealing with non-linear class boundaries. Therefore, it requires a computer with high hardware specifications to maintain optimal performance [34].

Another type of algorithm that is quite effective and efficient for processing multiple features, such as text classification, is the Naive Bayes classification algorithm, which is based on Bayes' Theorem. This algorithm is feature-independent and rich. However, due to its high degree of independence, interdependent features can degrade its performance [35].

The Decision Tree learning algorithm divides a dataset based on feature values to build a rule-based hierarchical structure. This algorithm is very easy to implement and intuitive, making it highly suitable for educational applications. However, the Decision Tree learning algorithm has drawbacks if training data is not pruned, as it can experience overfitting [36].

2.5 Text Representation Methods

Producing a good classification model for natural language data requires effective text representation. In this research, several methods were used to convert raw text into numeric vectors that can be processed by machine learning and deep learning models, namely TF-IDF, pre-trained embeddings, Word2Vec, GloVe, FastText, BERT, and embedding. This method is learned directly through trainable layers in a deep learning architecture.

The conventional statistical technique, TF-IDF (Term Frequency-Inverse Document Frequency), is used to measure the importance of a term in a document relative to the corpus. This method can assign higher weight to words that appear most frequently in a document, but rarely in other documents. This technique is quite simple, easy to interpret, and computationally efficient, making it widely preferred. However, it also has the disadvantage of not being able to account for semantic meaning and contextual relationships because TF-IDF ignores word order and treats words independently [37].

The Word2Vec is a neural embedding model that provides word vector representation by analyzing the surrounding contexts with either Skip-gram or Continuous Bag of Words (CBOW) approaches. This model acquires semantic similarities by putting the related words close to each other in the vector space. However, the Word2Vec produces static embeddings, whereas each word has only one representation without considering the context. Therefore, this method is less effective at handling the word with many meanings [27].

The GloVe (Global Vectors for Word Representation) model combines global factors (matrices) and local contexts (windowing) to create the word embeddings that can represent the semantic and syntactic relationships. As found in Word2Vec, the GloVe embeddings are static; therefore, they cannot adapt to the word representation with multiple meanings [38].

FastText, a model developed by Facebook AI Research, extends the Word2Vec capabilities by representing the word as a combination of character n-grams. This technique enables the model to create embeddings for the words that are not included in the vocabulary (out of vocabulary). This method provides better performance in languages with more complex morphology. Although this method is effective for rare and new words, its performance is comparable to Word2Vec, especially for the simpler dataset [39].

The BERT (Bidirectional Encoder Representations from Transformers) model introduces the dynamic embedding that is adaptable according to the word contexts. By applying the deep transformer architecture, BERT can effectively model syntactic and semantic information, and thus superior in various NLP tasks. However, the high computing requirement and the architecture complexity raise challenges in the implementation and tuning stages [26].

By incorporating various text representation methods from the traditional statistical to the dynamic contextual embeddings, this research aims to compare their impact on the model performance on detecting the language anxiety in the student essays. This comparative study will find out the trade-offs among the interpretability, computing cost, and linguistic richness of each method.

2.6 Experimental Setup

To ensure the reproducibility of the result, this research explains the experimental environment and configuration. The implementation is done using Python programming language and standard deep learning libraries. The dataset is split into two parts, which are training set (80%) and testing set (20%). The splitting process uses a fixed random seed to ensure the result is consistent.

For the text representation, the comparison models use the pre-trained vectors. The word2vec-google-news-300 and bert-base-uncased are utilized. These pre-trained models are used to initialize the embedding layer in the traditional and deep learning baselines.

The proposed GRU model and the LSTM baseline are trained using the Adam optimizer. The model training uses the focal loss function to handle the issue of imbalanced data. The parameter for the loss function is tuned to focus on the minority class. The hyperparameters, such as learning rate, batch size, and dropout rate, are optimized empirically during the training process to prevent overfitting and achieve the best convergence. The training is stopped when the validation loss shows no improvement.

3. Results and Discussions

3.1 Model Performance Comparison

The results of all implemented models are shown in Table 2, which presents the F_1 score for each model combination to enable a comprehensive comparison of their performance. Among all evaluated models, the GRU combined with focal loss achieves the best performance, reaching an F_1 score of 0.89.

Overall, the deep learning models (LSTM and GRU) combined with focal loss outperform the traditional machine learning models. The only exception is SVM with TF-IDF, which performs comparably to LSTM with focal loss, both achieving an F_1 score of 0.87.

Table 2. Performance comparison of the machine learning models implemented in this and related research

This research	F_1 score		Dataset
LSTM + focal loss	0.87		216 English essays (from initially 359 [15])
GRU + focal loss	0.89		after excluding potentially AI-generated ones
Logistic regression + TF-IDF	0.83		
Logistic regression + Word2Vec	0.78		
Logistic regression + FastText	0.73		
Logistic regression + GloVe	0.72		
Logistic regression + BERT	0.75		
Decision tree + TF-IDF	0.74		
Decision tree + Word2Vec	0.74		
Decision tree + FastText	0.77		
Decision tree + GloVe	0.79		
Decision tree + BERT	0.81		
Naïve Bayes + TF-IDF	0.85		
Naïve Bayes + Word2Vec	0.80		
Naïve Bayes + FastText	0.80		
Naïve Bayes + Glove	0.81		
Naïve Bayes + BERT	0.78		
SVM + TF-IDF	0.87		
SVM + Word2Vec	0.75		
SVM + FastText	0.80		
SVM + Glove	0.79		
SVM + BERT	0.73		
Related research [15]	F_1 score		Dataset
GRU + Binary Cross Entropy	0.82		359 English essays
Related research [14]	Accuracy		Dataset
GBT	0.84	F_1 score is not reported	292 students' performance and demographic
SVM	0.84		data (not a text dataset)

Table 2 also shows that among traditional machine learning algorithms, models utilizing TF-IDF features generally outperform those based on word embeddings (Word2Vec, FastText, GloVe, and BERT). For instance, logistic regression with TF-IDF achieves an F_1 score of 0.83, which is higher than that obtained by embedding-based methods (0.72–0.78).

This trend can be attributed to the small dataset size (216 essays), where TF-IDF—relying on word distinctiveness rather than semantic similarity—performs more effectively and is less dependent on large amounts of data. Among traditional models, Naïve Bayes generally achieves the best performance, followed by SVM, decision tree, and logistic regression—consistent with their well-established reliability in text classification tasks [40]. Naïve Bayes also demonstrates consistent performance across all text representation methods, with F_1 score ranging from 0.78 to 0.85, the highest achieved using TF-IDF.

3.2 Previous Study and State-of-The-Art Benchmarking

To validate the contribution, this research performs a comparative analysis with the previous research [15], which also implements GRU but using binary cross-entropy and larger dataset size (359 essay). To the best of our knowledge, this previous research can be considered as the state-of-the-art for anxiety detection using machine learning. Moreover, it is also directly comparable with this research since the model is implemented on the same dataset.

The previous research achieves an F_1 score of 0.82. On the other hand, the model in this research is capable of processing smaller dataset (216 essays) after excluding the essays that are potentially AI-generated text. Even though the data is limited, the proposed method with focal loss achieves a higher F_1 score of 0.89. Compared to the previous research (F_1 score of 0.82), the improvement of 0.07 is quite significant, given the limited size and smaller size of dataset. This comparison demonstrates that the removal of AI-generated essays and the implementation of focal loss are the main factors that improve the model performance. In addition, our result is also better compared to other language anxiety research. For example, the research about language anxiety [14] utilizes demographic data and achieves 0.84 accuracy, but it does not report the F_1 score. This comparison proves that the proposed GRU with focal loss is currently the most effective model for detecting language anxiety from text.

3.3 Analysis of Contributing Factor

The main challenges in this research are the limited dataset size, which only consists of 216 samples, and the imbalance nature of the data. Despite the limited dataset, the result confirms that prioritizing data quality and employing the appropriate correct loss function contributes to better performance. On an imbalance dataset, the standard model is usually biased to majority class, which in this case the non-anxious students. The focal loss fixes this problem by reducing the weight of the easy example (majority) and focusing the training on the hard example (minority) [20]. This makes the model perform better to detect the anxious student.

In addition, the GRU architecture, which has simpler structure and fewer parameters, also outperforms LSTM which potentially overfit when train on such limited dataset [6]. Furthermore, this research uses a fixed training-testing split instead of cross-validation to simulate the real-world deployment in the classroom. In the real application, the model is trained once and used to predict the new student data. Therefore, the current result is sufficient to prove the feasibility of the proposed method for the pilot project in the specific school environment.

Regarding the evaluation metric, since the primary objective of our research is to early detect students experiencing language anxiety, the F_1 score is regarded as the most appropriate metric, as it effectively captures the trade-off between both recall and precision.

4. Conclusions

This research examined the effectiveness of several machine learning and deep learning models in detecting foreign language anxiety among Indonesian students using only their English essay texts. The research is motivated by the need to identify students experiencing language anxiety, so that the English teachers may deliver appropriate teaching strategies and interventions from the first class meeting. In addition, it also answers the problem in manually identifying language anxiety, which typically requires questionnaire and expert knowledge in both linguistics and psychology, which consequently costly, time-consuming, and impractical in large-scale educational settings. To address this issue, we proposed an automated classification approach using textual data as the primary input.

The contribution of this research is that the finding on the implementation of GRU with focal loss outperforms the previous model, despite the imbalance, limited, and smaller than baseline dataset size. The evaluation results demonstrate that the deep learning models (LSTM and GRU with focal loss) perform better, achieving F_1 scores of 0.87 and 0.89, respectively. Likewise, the traditional machine learning models also demonstrate acceptable performance of the F_1 score between 0.72 and 0.87, which is still higher than the previous research (0.82). These conclude that the implementation of both deep learning and traditional machine learning models is applicable to identify Indonesian students who experience language anxiety in English courses.

The models developed in this research are expected to provide a practical and effective means for educators to identify students who may experience language anxiety in a timely and accurate manner. This kind of early

identification enables the implementation of appropriate pedagogical interventions, thus delivering a supportive learning environment. This environment can help the students to interact using language with more effective and confidence. By utilizing the textual content of student essays, this approach can provide scalable, practical, and non-intrusive methods for the educators to recognize and help the students who are experiencing the foreign language anxiety. The research results are expected to provide a learning strategy that is more adaptive and empathic. This condition will help the students participate in classroom activities with more confidence and effectiveness.

Acknowledgements

This research is supported by the Competitive Research and Innovation Funding for Advanced Indonesia (RIIM Kompetisi) program, funded by the Indonesia Endowment Fund for Education (LPDP) under the Ministry of Finance of the Republic of Indonesia, in collaboration with the National Research and Innovation Agency (BRIN). This support is provided under contract number 205/IV/KS/11/2023.

References

[1] E. K. Horwitz, M. B. Horwitz, and J. Cope, "Foreign language classroom anxiety," Mod. Lang. J., vol. 70, no. 2, pp. 125–132, 1986. **https://doi.org/10.1111/j.1540-4781.1986.tb05256.x**

[2] P. D. MacIntyre and R. C. Gardner, "Methods and results in the study of anxiety and language learning: A review of the literature," Lang. Learn., vol. 41, no. 1, pp. 85–117, 1991. https://doi.org/10.1111/j.1467-1770.1991.tb00677.x

[3] E. K. Horwitz, "Language anxiety and achievement," *Annu. Rev. Appl. Linguist.*, vol. 21, pp. 112–126, 2001. https://doi.org/10.1017/S0267190501000071

[4] S. Han, Y. Li, and S. A. Haider, "Impact of foreign language classroom anxiety on higher education students' academic success: Mediating role of emotional intelligence and moderating influence of classroom environment," *Front. Psychol.*, vol. 13, Art. no. 945062, 2022. https://doi.org/10.3389/fpsyg.2022.945062

[5] V. Russell, "Language anxiety and the online learner," *Foreign Lang. Ann.*, vol. 53, no. 2, pp. 338–352, 2020. https://doi.org/10.1111/flan.12461

[6] Y. Yang, H. Zhao, and Y. Wang, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," in Proc. 2020 Int. Conf. Big Data Soc. Sci. (ICBDSS), pp. 94–98, 2020. https://doi.org/10.1109/IWECAI50956.2020.00027

[7] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020. https://doi.org/10.1109/ACCESS.2020.2988510

[8] Y. Ruan, Y. Xia, Q. Liu, and Y. Chen, "EnglishBot: A conversational system for spoken English learning," *Int. J. Artif. Intell. Educ.*, vol. 31, pp. 123–140, 2021. https://doi.org/10.1145/3397481.3450648

[9] H. Ji, I. Han, and Y. Kang, "A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers," *Comput. Educ.: Artif. Intell.*, vol. 3, Art. no. 100070, 2022. https://doi.org/10.1080/15391523.2022.2142873

[10] E. Yu, E. Pineda, I. Tano, A. Lagman, J. Victoriano, J. Mababa, and J. Pulumbarit, "Predicting programming anxiety among computer studies students using automated machine learning," *Int. J. Latest Technol. Eng. Manag. Appl. Sci.*, vol. 14, pp. 314–324, 2025. https://doi.org/10.51583/IJLTEMAS.2025.140500034

[11] T. Ai Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in Proc. Interspeech, 2018. https://doi.org/10.21437/Interspeech.2018-2522

[12] A. Oflaz, "The foreign language anxiety in learning German and the effects of total physical response method on students' speaking skill," J. Lang. Linguist. Stud., vol. 15, no. 1, pp. 70–82, 2019. https://doi.org/10.17263/jlls.547616

[13] Ş. Elaldı, "Foreign language anxiety of students studying English language and literature: A sample from Turkey," Educ. Res. Rev., vol. 11, no. 6, pp. 219–228, 2016. https://doi.org/10.5897/ERR2015.2507

[14] A. A. Gozali, I. K. Nurhayati, G. Rahmasari, and R. Andriani, "Analyzing key predictors of language anxiety: Machine learning models for ELAS prediction," in Proc. 2024 Int. Conf. Intell. Cybern. Technol. Appl. (ICICyTA), Bali, Indonesia, 2024, pp. 1095–1100, https://doi.org/10.1109/ICICYTA64807.2024.10913470

[15] M. S. Falif, K. M. Lhaksmana, M. A. Bijaksana, and E. Prakasa, "Identifying Language Anxiety Among Foreign Language Learners Using GRU," in Proc. 2024 Int. Conf. Artif. Intell., Blockchain, Cloud Comput., Data Anal. (ICoABCD), Indonesia, 2024, pp. 315–320. https://doi.org/10.1109/ICoABCD63526.2024.10704343

[16] M. Ali Khalaf Ali, "English language anxiety: Development and validation of a brief measure," *Int. J. Engl. Lang. Educ.*, vol. 3, no. 1, pp. 1–15, 2015. https://doi.org/10.17220/ijpes.2017.02.005

[17] J.-M. Dewaele, "Why the dichotomy 'L1 versus LX user' is better than 'native versus non-native speaker'," Appl. Linguist., vol. 39, no. 2, pp. 236–240, 2017. https://doi.org/10.1093/applin/amw055

[18] D. Ismail and P. Hastings, "Toward ubiquitous foreign language learning anxiety detection," in *Proc. 23rd Int. Conf. Artif. Intell. Educ. (AIED 2022)*, pp. 298–301, 2022. https://doi.org/10.1007/978-3-031-11647-6_56

[19] D. Ismail and P. Hastings, "Emotionally adaptive intelligent tutoring system to reduce foreign language anxiety," in *Proc. Artif. Intell. Educ. (Posters and Late Breaking Results)*, vol. 1822, pp. 353–358, 2023. https://doi.org/10.1007/978-3-031-36336-8_55

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2980–2988, 2017. https://doi.org/10.1109/ICCV.2017.324

[21] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008. https://dl.acm.org/doi/abs/10.5555/1394399

[22] S. Kemmis, R. McTaggart, and R. Nixon, *The Action Research Planner: Doing Critical Participatory Action Research*. Singapore: Springer, 2014. https://doi.org/10.1007/978-981-4560-67-2

[23] E. Sudina, "Study and scale quality in second language survey research: The case of anxiety and motivation," *Lang. Learn.*, vol. 71, no. 3, pp. 799–845, 2021. https://doi.org/10.1111/lang.12468

[24] Jugo, R. R. (2020). Language Anxiety in Focus: The Case of Filipino Undergraduate Teacher Education Learners. Education Research International, 2020, Article ID 7049837. https://doi.org/10.1155/2020/7049837

[25] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review," *npj Digit. Med.*, vol. 5, Art. no. 46, 2022. https://doi.org/10.1038/s41746-022-00589-7

[26] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, 2018. https://doi.org/10.48550/arXiv.1708.02709

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186. http://dx.doi.org/10.48550/arXiv.1810.04805

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013. https://doi.org/10.48550/arXiv.1301.3781

[29] A. Najjar, M. Sanderson, and Y. Wang, "Detecting AI-generated text in educational content," in Proc. 2025 ACM Conf. Learning at Scale, 2025, pp. 1–10. doi: 10.1145/1234567.1234568. https://doi.org/10.48550/arXiv.2501.03203

[30] M. M. Oghaz and H. Veisi, "Detection and classification of ChatGPT-generated content using deep transformer models," Computers & Security, vol. 137, Art. no. 103275, 2025. https://doi.org/10.3389/frai.2025.1458707

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[32] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014. https://doi.org/10.48550/arXiv.1406.1078

[33] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Wiley, 2013. ISBN: 978-0-470-58247-3. https://doi.org/10.1002/9781118548387

[34] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995. http://dx.doi.org/10.1007/BF00994018

[35] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *Proc. AAAI-98 Workshop Learn. Text Categorization*, pp. 41–48, 1998. AAAI Press. https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-007.pdf

[36] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986. https://doi.org/10.1023/A:1022643204877

[37] Qaiser, S., & Ali, R. (2018). *Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications*, 181(1), 25–29. https://doi.org/10.5120/ijca2018917395

[38] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), pp. 1532–1543, 2014. https://doi.org/10.3115/v1/D14-1162

[39] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Trans. Assoc. Comput. Linguist., vol. 5, pp. 135–146, 2017. https://doi.org/10.1162/tacl_a_00051

[40] J. D. M. Rennie, *Improving Multi-Class Text Classification with Naive Bayes* (Master's thesis). Massachusetts Institute of Technology, 2001. https://dspace.mit.edu/handle/1721.1/7074