Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 6 (2024) 748 - 757 e-ISSN: 2580-0760

Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification

Fatma Indriani[1*], Radityo Adi Nugroho[2], Mohammad Reza Faisal[3], Dwi Kartini[4]
[1]Department of Computer Science, Lambung Mangkurat University, Banjarmasin, Indonesia
[1]f.indriani@ulm.ac.id, [2]radityo.adi@ulm.ac.id, [3]reza.faisal@ulm.ac.id, [4]dwikartini@ulm.ac.id

Abstract

Student feedback plays a crucial role in enhancing the quality of educational programs, yet analyzing this feedback, especially in informal contexts, remains challenging. In Indonesia, where student comments often include colloquial language and vary widely in content, effective multilabel classification is essential to accurately identify the aspects of courses being critiqued. Despite the development of several BERT-based models, the effectiveness of these models for classifying informal Indonesian text remains underexplored. Here we evaluate the performance of three BERT variants—IndoBERT, IndoBERTweet, and mBERT—on the task of multilabel classification of student feedback. Our experiments investigate the impact of different sequence lengths and truncation strategies on model performance. We find that IndoBERTweet, with a macro F1-score of 0.8462, outperforms IndoBERT (0.8243) and mBERT (0.8230) when using a sequence length of 64 tokens and truncation at the end. These findings suggest that IndoBERTweet is well-suited for handling the informal, abbreviated text common in Indonesian student feedback, providing a robust tool for educational institutions aiming for actionable insights from student comments.

Keywords: *BERT models; education data; finetuning; multilabel classification; sequence length; student feedback*

How to Cite: F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 6, pp. 748 - 757, Dec. 2024.
DOI: https://doi.org/10.29207/resti.v8i6.6100

1. Introduction

Student feedback is an essential tool for educational institutions aiming to enhance the quality of their programs and services. By analyzing this feedback, institutions can make informed decisions about course content, teaching methods, and overall educational strategies [1]. In the context of higher education, timely and accurate analysis of student feedback can significantly impact the quality of education delivered, ensuring that students' needs and concerns are addressed effectively. However, the manual evaluation of this feedback is often time-consuming and prone to subjectivity, making it challenging for institutions to quickly adapt to the evolving needs of students. Consequently, there is a growing interest in automating the analysis of student feedback using advanced natural language processing (NLP) techniques.

The analysis of student feedback, particularly in informal and diverse linguistic contexts, presents unique challenges. In Indonesia, where student comments are often written in a colloquial style, the variability in language use can complicate the classification process. Traditional machine learning methods, while useful, may not fully capture the nuances of informal language, especially when dealing with multilabel classification tasks where each piece of feedback may pertain to multiple aspects of a course. This necessitates the use of more sophisticated models capable of handling the complexities of natural language, particularly in the Indonesian context.

Several studies have explored the application of machine learning and NLP techniques for classifying student feedback, but the majority have focused on classification models, such as [2]-[6]. For instance, Rusli et al [2] investigated various traditional machine learning algorithms such as Logistic Regression and Support Vector Machines (SVM) for supervised feedback classification in Bahasa Indonesia. However, these studies were primarily limited to multiclass classification, which does not adequately address the

Received: 04-10-2024 | Accepted: 29-11-2024 | Published Online: 27-12-2024

multilabel nature of student feedback where a single comment can be relevant to multiple categories simultaneously. Ruiz Alonso et al. [7] further extended this area by applying multilabel classification techniques such as Binary Relevance and Classification Chains to classify feedback in online courses, showing that Random Forests and SVM performed well for this task.

Several studies have explored the use of neural networks for classifying student feedback, with varying degrees of complexity. Veerachamy [8] applied basic artificial neural networks (ANNs) to automate the analysis of post-course assessments, demonstrating its effectiveness in classifying student satisfaction. More advanced models have also been used, such as Asghar et al [9], who applied a deep neural network using Bi-LSTM to classify emotions in student feedback, outperforming benchmark models in emotion detection. Furthermore, Onan [10] employed a Recurrent Neural Network (RNN) with attention mechanisms to mine opinions from large-scale instructor evaluations, achieving high classification accuracy using GloVe word embeddings.

Recent advancements in NLP, particularly the development of transformer-based models like BERT [11], have shown great promise in improving text classification tasks. BERT, with its bidirectional training approach, allows models to consider the context of words in a sentence, making it particularly effective for understanding nuanced language. A study [12] leveraged BERT to derive word vectors from student feedback data, which were then classified using traditional machine learning methods such as SVM, K-Nearest Neighbors (KNN), and Random Forests (RF). This approach, while innovative, still relied on traditional classifiers rather than directly utilizing the full potential of BERT for end-to-end classification. The study demonstrated the effectiveness of BERT-derived features, but it did not fully explore the capabilities of deep learning models for multilabel classification. Table 1 presents performance comparison of IndoBERT, IndoBERTweet, and mBERT in previous research especially for Indonesian dataset.

In the context of multilabel classification, which is crucial for handling student feedback that addresses multiple aspects of a course, there has been limited research. Nabiilah et al [13] explored the use of IndoBERT and mBERT for multilabel classification of toxic comments in Indonesian, achieving promising results with an F1 score of 0.9032. However, this study focused on a different domain—social media toxic comments—and did not evaluate IndoBERTweet [14], a variant specifically tailored for the Indonesian language. Additionally, the study did not investigate the impact of different sequence lengths and truncation strategies, which are critical factors that can influence the performance of BERT-based models in text classification tasks.

An important consideration in text classification is the handling of sequence length and truncation strategies, particularly in BERT-based models that are sensitive to input lengths. Choosing the appropriate sequence length is crucial for optimizing performance, yet this aspect has been largely overlooked in Indonesian text classification research. For example, Chovanek at al [15] focused on integrating BERT outputs with demographic data for multilabel classification but did not address how sequence length or truncation strategies influence model outcomes. In contrast, studies on long text classification have explored truncation strategies extensively. Mutadosirin and Prasojo [16] found that truncating the beginning of documents often outperformed summarization techniques, while Chen and Lv [17] introduced a method to filter redundant information, preserving key semantic relationships. Yang et al. [18] emphasized the risk of losing crucial context when truncating long texts, yet much of this research ignores shorter, informal texts. This study aims to fill that gap, focusing on how truncation impacts classification in shorter, informal student feedback.

In summary, while there has been significant progress in applying machine learning and NLP techniques to the analysis of student feedback, several gaps remain in the existing research. Most studies have focused on multiclass classification rather than multilabel classification, limiting their ability to fully capture the complexity of student feedback, where a single comment may pertain to multiple aspects of a course. Additionally, traditional machine learning methods have dominated the field, with relatively few studies exploring the full potential of transformer-based models like BERT, particularly within the Indonesian context.

The rationale for using IndoBERT, IndoBERTweet, and mBERT in this study arises from their complementary characteristics. IndoBERT, trained on formal Indonesian text, is well-suited for handling the more structured components of student feedback. However, feedback often contains informal language elements such as abbreviations, relaxed grammar, and colloquial expressions, making IndoBERTweet—which is pre-trained on informal Indonesian social media data—a better choice for capturing these nuances. Though mBERT is not specialized for Indonesian, it serves as a completeness experiment to observe how a multilingual model performs on predominantly Indonesian text. mBERT provides a valuable comparison point for assessing whether it can still capture the subtleties of the Indonesian language as effectively as the more targeted models.

These three models together provide a comprehensive approach to analyzing both formal and informal language in student feedback. Additionally, this study explores the impact of sequence length and truncation strategies on the performance of these BERT-based models, a key aspect that has not been thoroughly examined in previous research.

This study makes several key contributions. Firstly, it evaluates the performance of IndoBERT, IndoBERTweet, and mBERT for multilabel classification of Indonesian student feedback, filling a gap in the literature where these models have not been extensively compared in this context. Secondly, the study investigates the impact of different sequence lengths and truncation strategies on model performance, providing insights into how these factors can be optimized for better classification accuracy. By focusing on informal student feedback, this research offers practical implications for educational institutions seeking to automate the analysis of student feedback, enabling them to respond more effectively to students' needs.

Table 1. Performance comparison of BERT variants on Indonesian datasets in prior studies

Study	Model	Dataset	Task	Evaluation
[19]	BERT, mBERT, IndoBERT	Indonesian Covid-19 articles from Turnbackhoax.id (hoax) and Detik.com (fact)	Hoax Detection	Accuracy (best model achieved over 90%)
[20]	IndoBERT, IndoBERTweet, CNN-LSTM	IndoLEM sentiment data, IndoSMSA sentiment data, crawled Indonesian tweets related to COVID-19 vaccines	Sentiment Analysis	Accuracy, F1-Score (IndoBERTweet achieved the highest accuracy of 0.73 and F1-Score of 0.73)
[21]	LSTM with TF-IDF, IndoBERTweet, Word2Vec	50,000 crawled Indonesian tweets on political, social, and economic topics, manually and system labelled for granularity weight; data from GitHub for corpus building	Sentiment Analysis (Granularity)	Accuracy, F1-Score (manual labelling: 88.97%, system labelling: 97.80%)
[22]	BERT, IndoBERT	1000 Indonesian tweets containing the keyword "covid", labelled as positive or negative	Sentiment Analysis	Accuracy, Specificity, Sensitivity (IndoBERT with preprocessing achieved the highest accuracy of 89.50%)
[23]	IndoBERT	Multi-label, multi-class Indonesian dataset of 21,694 app reviews with sentiment and emotion labels	Sentiment Analysis (Multi-label)	Precision, Recall, F1-score, Accuracy (SMOTE achieved the highest F1-score of 0.86 and accuracy of 0.82)
[24]	IndoBERTweet	1000 Indonesian tweets containing the keyword "covid", labelled for depression detection	Depression Detection	Accuracy, Precision, Recall, F1-score (highest accuracy of 86% achieved with 80:20 data split)
[25]	IndoBERT (base-p1, base-p2, large-p1, large-p2)	1000 crawled Indonesian tweets containing keywords related to the 2024 elections, labelled as positive, negative, or neutral	Sentiment Analysis (Elections)	Accuracy, F1-Score (IndoBERT large-p1 achieved the highest accuracy of 0.8350 and F1-Score of 0.8849)
[26]	IndoBERT, Naïve Bayes, K-NN, Decision Tree	10,000 student reviews from online questionnaires collected from a private university in Indonesia	Aspect-Based Sentiment Analysis (ABSA)	Accuracy, Precision, Recall, F1-score (IndoBERT achieved the highest accuracy of 0.890 and F1-score of 0.897 for aspect extraction)
[27]	IndoBERTweet + BiLSTM	Two public Indonesian hate speech datasets: Alfina et al. and Ibrohim and Budi	Hate Speech Detection	Accuracy, Precision, Recall, F1-score (best performance achieved using the combined model)
[13]	IndoBERT (feature extraction), Multilingual BERT (classification)	Multi-label Indonesian dataset of toxic comments, labelled for pornography, hate speech, radicalism, and defamation	Toxic Comment Classification	Accuracy, F1-score (proposed model achieved an F1-score of 0.9032 on testing data)
[28]	LSTM, IndoBERT	Crawled Indonesian tweets with the keywords "covid-19" and "corona", labelled as hoax or non-hoax	Hoax Detection	Accuracy (IndoBERT achieved higher accuracy than LSTM)
[29]	IndoBERT (feature-based and fine-tuning approaches)	Indonesian hotel reviews dataset from the AiryRooms platform, labelled for aspect-based sentiment analysis	Aspect-Based Sentiment Analysis (ABSA)	Accuracy, F1-score (IndoBERT fine-tuned with single-sentence classification achieved the best F1-score and testing time)
[30]	XLM-R, mBERT	3 Indonesian datasets, 2 English datasets from other research	Sentiment Analysis, Hate Speech Detection	F1-score (adding English data with a feature-based approach improved Indonesian text classification performance)

2. Research Methods

2.1 Dataset

The dataset used in this study was collected from a large public university in Indonesia through its online academic portal, which gathers course feedback from students at the end of each semester. For this study, only one semester's data from one faculty was utilized, capturing feedback from multiple courses. To ensure privacy and confidentiality, the dataset was anonymized by removing identifiable information such as the names of courses, students, and lecturers, leaving only the textual content of the feedback.

The dataset comprises a total of 4,301 instances, each representing a unique piece of feedback from a student. The feedback is labeled manually with up to eight different aspects relevant to the courses, which are: *teaching method, scoring, e-learning, slides and* *resources, schedule, attendance and tardiness, exercises and quizzes, exams, and lab*. Importantly, an instance can have no associated label if the feedback does not explicitly address any particular aspect, such as in comments like "Thank you" or "Great course." The complete distribution of labels per instance is shown in Figure 1.

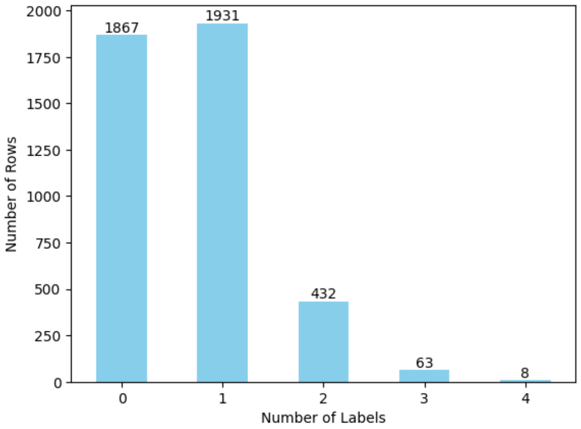

Figure 1. Distribution of labels per instance

The dataset exhibits significant label imbalance, with certain labels being far more frequent than others. For example, the label "teaching method" appears in 1,224 instances, making it the most common label, while "scoring" appears only in 96 instances, making it the least common. This imbalance poses a challenge for the classification model, as it must be trained to accurately recognize and predict less frequent labels without being biased towards more prevalent ones. A breakdown of the label distribution is provided in Table 2.

Table 2. Label counts

Label	Counts
teaching method	1224
scoring	96
elearning	366
lecture slide / resources	99
schedule	122
attendance or tardiness	200
exercises and quizzes	524
exams	239
lab	146

The language of the feedback is Indonesian, and the text is characterized by an informal style. This informality is reflected in the lack of capitalization and punctuation, the use of abbreviations, the absence of full sentences, and a generally relaxed approach to grammar. There are also many misspellings due to the students filling out the online feedback survey in a hurry. Despite these characteristics, the feedback remains polite, reflecting the cultural norms of the student population. Sample feedback text is shown in Table 3.

Preprocessing of the dataset was minimal, limited to the removal of punctuation, while further steps such as case normalization, stopword removal, or stemming were intentionally avoided. This decision was based on the fact that BERT-based models are pre-trained to handle raw text, including stopwords and varied word forms, by leveraging their contextual embeddings. Removing stopwords or altering word forms through stemming could disrupt the semantic richness and informal nuances present in the feedback, which is important in student responses. Moreover, retaining these elements allows BERT to learn contextual relationships naturally, which can improve model performance, especially in handling the diverse and informal language of student feedback. Additionally, comments with no associated labels were kept in the dataset to preserve the full distribution of feedback.

Table 3. Sample feedback text

Feedback Text	Label(s)
Dimohon ppt berbahasa indonesia dan tidak berbahasa inggris karena di ppt bingris susah diterjemahkan dan ada beberapa huruf tidak sesuai, dimohon Nilai akhir dari ujian atau kuiz atau nilai-nilai lain bisa lebih transparent (I wished the presentation (PPT) could be in Indonesian instead of English because the slides are hard to translate, and some letters aren't quite right. I also hoped the final grades from exams, quizzes, and other scores can be more transparent.)	lecture slides, scoring, exercises and quizzes, exams
cara mengajar bu D sangat jelas tetapi mohon tepat waktu saat kulaih selesai terkadang molor, akibatnya kami yang shbs kuliah ada pretest dan dilanjutkan dengan ada praktikum. kami terburu-buru untuk sholat serta makan siang (Ms. D's teaching is very clear, but I wished the class would finish on time. Sometimes it ran over, and as a result, those of us who had a pretest and practical session afterwards were in a rush for prayer and lunch)	teaching method, tardiness
di harap kan untuk materi kuliah libih bisa di bagikan sebelum perkuliahan di mulai sehingga mahasiswa bisa memahami terlebih dahulu, dan di harap kan untuk praktikum nya lebih baik lagi dan jangan memilih asisten dosen yg mengekkang karena kami hanya mencari ilmu tidak mencari yang tidak-tidak (I wished the lecture materials were shared before the class starts, so students can understand them ahead of time. I also hoped the practical sessions improve, and that the lab assistants aren't too strict, as we are here to gain knowledge, not to deal with unnecessary stress.)	elearning, lecture slides, lab

2.2 BERT models

Transformer-based models have revolutionized natural language processing (NLP) by enabling machines to understand and generate human language with high accuracy. At the forefront of this advancement is the Bidirectional Encoder Representations from Transformers (BERT) model, which has set a new standard for many NLP tasks. BERT [11] is a transformer-based model designed to pre-train deep bidirectional representations by jointly conditioning on both left and right contexts in all layers. This bidirectional approach allows BERT to capture intricate patterns in language, understanding the context of a word based on its surrounding words in a sentence.

The BERT architecture consists of multiple transformer layers (commonly 12 or 24), each composed of self-attention mechanisms and feedforward neural networks. Each of these layers is structured around two primary components: Multi-Head Attention and Feed feed-forward networks, supplemented by residual connections and layer normalization to facilitate training deep networks (Figure 2).

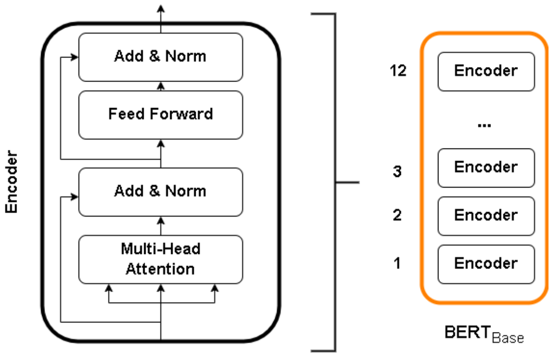

Figure 2. BERT architecture

The model is pre-trained using two tasks, Masked Language Modeling (MLM): Where a percentage of the input tokens are masked, and the model must predict the original token based on the context provided by the unmasked tokens; Next Sentence Prediction (NSP): Where the model is trained to predict whether two sentences follow each other in the original text, helping it understand the relationships between sentences.

This study employs three BERT variants: IndoBERT, IndoBERTweet, and mBERT. Each variant is tailored to different linguistic contexts and datasets. A comparison of these models is summarized in Table 4

Table 4. BERT models comparison

Feature	IndoBERT	IndoBERTweet	mBERT
Primary Objective	General Indonesian text	Informal Indonesian social media text	Multilingual text processing across 104 languages
Pre-training Corpus	Indonesian Wikipedia, news articles	Indonesian tweets	Wikipedia articles in 104 languages
Target Language	Indonesian	Indonesian	Multiple languages (including Indonesian)
Text Style	Formal and semi-formal	Informal	social media
Architecture	12 layers, 12 attention heads, hidden size of 768	12 layers, 12 attention heads, hidden size of 768	12 layers, 12 attention heads, hidden size of 768
Training Strategy	Pre-trained on formal texts using MLM	Pre-trained on informal social media text using MLM	Pre-trained on multilingual corpus using MLM and NSP
Maximum Sequence Length	512 tokens	128 tokens	512 tokens

IndoBERT [31] is a BERT variant pre-trained specifically on Indonesian text, including sources such as Wikipedia and news articles. It retains the standard BERT architecture but is optimized for formal and semi-formal Indonesian text.

IndoBERTweet [14] is designed to handle informal Indonesian text, particularly from social media platforms like Twitter. It shares the same architecture as IndoBERT but is pre-trained on a large corpus of Indonesian tweets. This training equips IndoBERTweet to process informal, abbreviated, and noisy text, which is common in social media and relevant to the student feedback analyzed in this study.

mBERT [11], or Multilingual BERT, is pre-trained on text from 104 languages, including Indonesian. While it is not tailored to any specific language, its multilingual training makes it versatile and capable of handling cross-lingual tasks.

2.3. Experimental Setup

This study’s experimental process involved fine-tuning BERT variants for multilabel classification on the student feedback dataset. The key steps include data preprocessing, hyperparameter tuning, and evaluation, all of which are detailed in the accompanying flowchart (see Figure 3).

The first step was data preprocessing, where feedback text was cleaned and prepared for model input. After preprocessing, the dataset was split into training and testing sets, with an 80%-20% ratio. The split was done using a stratified approach, specifically the *MultilabelStratifiedShuffleSplit* method from the *iterative-stratification* Python package [32], ensuring that the label distribution in both the training and test sets was proportional to the full dataset. This strategy helps ensure that both the training and test sets reflect the imbalanced nature of the labels.

Following the train-test split, we fine-tuned three BERT variants: IndoBERT, IndoBERTweet, and mBERT. Initial pre-experiments through trial and error revealed that training for 10 epochs, with a batch size of 16 and a learning rate of 3*e*-5, provided a good balance between training time and performance. While other values (e.g., learning rates of 5*e*-5 and 1*e*-5, batch sizes of 8 and 32) were briefly tested during this process, these hyperparameters were selected based on their consistency in achieving optimal results. They were applied across all models to maintain uniformity. Additionally, the *EqualWeightBCEWithLogitsLoss* function was used as the loss function, suitable for multilabel classification problems. This function assigns equal importance to each label, addressing the label imbalance of the dataset.

The core of the experimentation involved adjusting key hyperparameters, particularly the maximum sequence length and truncation strategy. Four different maximum sequence lengths were evaluated: 128, 96, 64, and 32 tokens. For sequences shorter than the maximum length, padding was applied. When sequences were longer than the defined length, truncation was employed. For the 64 and 32 token lengths, three truncation strategies were tested: Beginning Truncation (Removes tokens from the start of the sequence); Middle Truncation (Removes tokens from the middle of the sequence); End Truncation (Removes tokens from the end of the sequence).

These truncation strategies were applied only for sequence lengths of 64 and 32 tokens, as the longest sequence in the dataset was 76 tokens. No truncation was necessary for sequences with maximum lengths of 128 and 96 tokens.

Once fine-tuning was complete, model performance was evaluated using the macro-average F1-score as the primary metric. This metric was selected due to the imbalance in the dataset's label distribution. By focusing on the macro-average F1-score, the evaluation gives equal weight to each label, regardless of how frequently it appears, ensuring that even less common labels are considered. In addition to the F1 score, precision and recall metrics were also reported to provide a more detailed analysis of the model's ability to correctly classify the labels while accounting for false positives and false negatives.

After identifying the best model based on macro-average F1 scores, we further evaluated its performance by breaking down the precision, recall, and F1 scores for each label. This analysis highlighted the lower-performing labels, providing insights into specific challenges with less frequent categories and areas for improvement. The formula for precision, recall, F1-score, and macro-average F1-scores are presented in Equations 1 to 4.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

$$Macroaverage\ F1 = \frac{1}{n}\sum_{i=1}^{n} F1_i \quad (4)$$

TP is true positives, FP is false positives, FN is false negatives, n is the total number of classes, and $F1_i$ is the F1-score for each class.

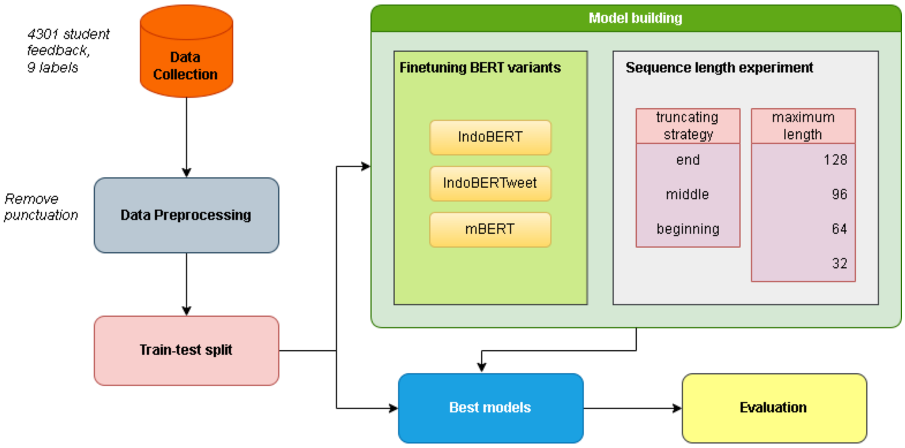

Figure 3. Experimental setup workflow

3. Results and Discussions

3.1 Overall Performance of BERT Variants

Table 5 presents the detailed macro-average F1-scores for IndoBERT, IndoBERTweet, and mBERT across all sequence lengths and truncation strategies. The best model is from IndoBERTweet with macro-average F1-score of 0.8462, while the best models from indoBERT and mBERT respectively 0.8243 and 0.8230. Additionally, Figure 4(a) provides a visualization of the average F1-scores for each BERT variant, aggregated over all experiments, offering a clear comparative overview.

The results show that IndoBERTweet consistently outperforms both IndoBERT and mBERT in terms of macro-average F1-scores. IndoBERTweet achieves the highest overall average F1-score, demonstrating its superior ability to handle the multilabel classification task effectively.

IndoBERT, while behind IndoBERTweet, generally performs better than mBERT across all sequence lengths and truncation strategies. This can be explained by IndoBERT's pre-training on a large corpus of formal Indonesian language, which gives it a deeper understanding of Indonesian linguistic patterns compared to mBERT. mBERT, being a multilingual model, is trained on text from many languages, which dilutes its ability to capture the nuances of any single language, including Indonesian.

The performance of IndoBERTweet can be attributed to its pre-training on informal Indonesian text, including social media data. This makes it well-suited for handling the informal language present in student feedback, which includes abbreviations, slang, relaxed grammar, and colloquial expressions. These characteristics of the dataset are more easily managed by IndoBERTweet compared to models trained on formal or multilingual data.

Table 5. Macro-average F1 scores across all experiments

max length, truncating	IndoBERT	IndoBERTweet	mBERT
128	0.7849	0.8276	0.8230
96	0.8041	0.8340	0.7982
64, end	0.7671	0.8462	0.7998
64, middle	0.8243	0.8262	0.7884
64, beginning	0.7940	0.8431	0.7882
32, end	0.7610	0.8065	0.7016
32, middle	0.7941	0.8180	0.7251
32, beginning	0.7918	0.8188	0.6971

3.2 Effect of Sequence Length and Truncation Strategy

The impact of sequence length on model performance is illustrated in Figure 4b, while the impact of truncation strategy is shown in Figure 4c. The results indicate that a sequence length of 64 tokens provides the best balance between efficiency and performance for both IndoBERT and IndoBERTweet (Figure 4b). IndoBERTweet achieves its highest F1-score of 0.8462 at 64 tokens with the end truncation strategy. Increasing the sequence length to 128 tokens does not significantly improve performance, suggesting that longer sequences introduce diminishing returns. Conversely, reducing the sequence length to 32 tokens leads to a noticeable decline in performance across all models, particularly for mBERT, which reaches its lowest score of 0.6971.

The choice of truncation strategy, as shown in Figure 4c, plays a role but is less impactful than sequence length. End truncation delivers the best results, especially for IndoBERTweet, where truncating the end retains the most critical context from the feedback. Middle truncation produces competitive but generally lower results, while beginning truncation yields the poorest performance across all models. This highlights the importance of retaining the beginning of the text, which contains key information for classification.

While both sequence length and truncation strategies influence performance, the most critical factor in these experiments is the choice of BERT variant. IndoBERTweet consistently outperforms both IndoBERT and mBERT across all sequence lengths and truncation strategies, demonstrating that the model's pre-training on informal Indonesian text makes it far more suited to this task. This suggests that selecting the right BERT variant tailored to the dataset's language style is more important than fine-tuning sequence length or truncation strategy.

3.3 Performance of IndoBERTweet Best Model

The overall performance of the best model configuration, IndoBERTweet with a sequence length of 64 and end truncation, is strong across most labels, achieving a macro-average precision of 0.8567, recall of 0.8391, and F1-score of 0.8462 (Table 6). High-performing labels include *exercise_and_quiz* and *lab*, with F1-scores of 0.9108 and 0.9259, respectively. These labels benefit from a relatively larger dataset size (524 and 146 instances, respectively), allowing the model to capture their patterns more effectively. Additionally, despite having only 96 instances, *scoring* performs well with a high recall of 0.9474, demonstrating the model's ability to handle even smaller label sets.

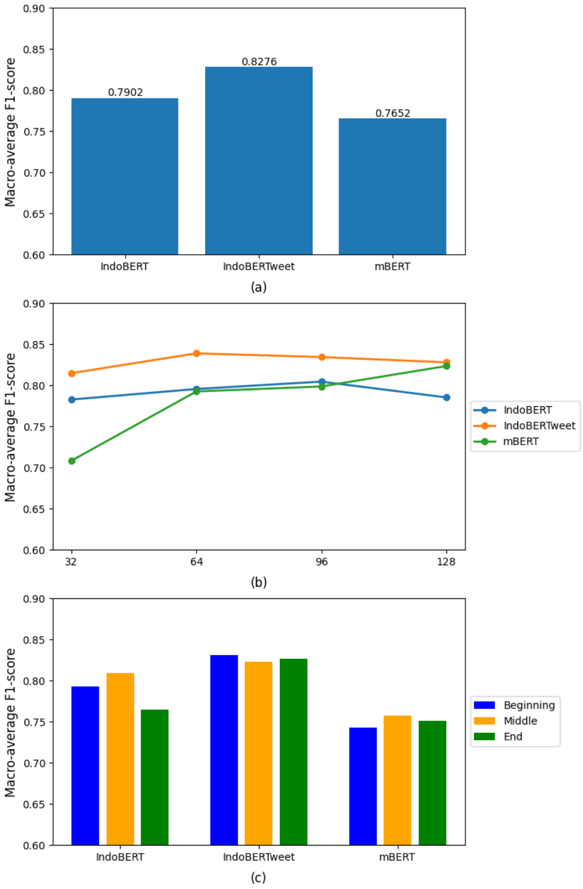

Figure 4. (a) Average performance of BERT variants (b) BERT performance by maximum sequence length (c) Comparison of truncation strategies across BERT variants

Table 6. Precision, Recall, and F1-score of the Best Model

Label	Precision	Recall	F1-score
teaching method	0.8552	0.7714	0.8112
scoring	0.8571	0.9474	0.9000
elearning	0.9375	0.8219	0.8759
lecture slides and resources	0.7895	0.7500	0.7692
schedule	0.7200	0.7500	0.7347
attendance or tardiness	0.8293	0.8500	0.8395
exercise and quiz	0.8981	0.9238	0.9108
exam	0.8235	0.8750	0.8485
lab	1.0000	0.8621	0.9259
Macro-average	0.8567	0.8391	0.8462

However, some labels pose challenges for the model, particularly *schedule* and *lecture_slides_and_resources*, which have the lowest F1-scores of 0.7347 and 0.7692, respectively. These labels, with fewer instances (122 and 99), highlight the model's difficulty in predicting less frequent labels, as both show lower precision and recall compared to the more common labels. The effect of label frequency is clear: larger datasets tend to yield better performance, while smaller datasets require the model to generalize with less available information. This pattern show IndoBERTweet's robustness in handling imbalanced multilabel classification, though further optimization could improve predictions for infrequent labels.

Table 7. Examples of errors by IndoBERT and mBERT correctly handled by IndoBERTweet

Feedback text	True labels	Predicted labels IndoBERTweet prediction	IndoBERT prediction	mBERT prediction
Better understanding of informal language and misspellings				
saran agar dosen tidak terlambat lagi sehingga waktu untuk perkuliahan dapat di mulai tepat waktu dan waktu untuk memahami materi lebih panjang. saran lainnya yaitu untuk slide di buat dengan kata yang lbh mudah di pahami dan dihafal (Suggestion for the lecturer not to be late so that class time can start on time and allow more time to understand the material. Another suggestion is to make the slides with simpler language that is easier to understand and memorize)	lecture slides, attendance/tardiness	lecture slides, attendance/tardiness	teaching methods, attendance/tardiness	teaching methods
Semoga kedepannya lebih baik lagi dan kalo bisa gak usah ada ujian pakai elearning karena itu semua menurut saya sangat tergantung pada jaringan dan sering error (Hopefully, things will be better in the future, and if possible, there should be no exams via e-learning because I feel it is too dependent on the network and often has errors.)	elearning, exams	elearning, exams	exams	teaching methods
untuk bapak ikhwan tolong jgn memberikan tugas saat mendekati uas atau uts karena mahasiswa kesulitan untuk fokus ke uts atau uas. hampir semua matakuliah yg bapak ampu diberikan tugas. Jdi kalau bisa tugasnya diberikan tidak di akhir-akhir pertemuan. Untuk semua dosen sarannya agar memanfaatkan fasilitas elearning (To Mr. Ikhwan, please don’t give assignments close to the midterms or finals because students find it difficult to focus on them. Almost every course you teach has assignments. So if possible, give the assignments not at the end of the term. For all lecturers, the suggestion is to make use of e-learning facilities.)	elearning, exercises, exams	elearning, exercises, exams	exercises	(none)
Better understanding of contextual information				
Dosennya sudah baik tetapi terlalu cepat dalam menjelaskan (The lecturer is good but explains too quickly.)	teaching methods	teaching methods	(none)	(none)
Sudah bagus tapi mungkin untuk tugas presentasi kelompok dikurangi karena mahasiswa banyak yg tidak memperhatikan (It’s good, but maybe group presentation assignments should be reduced because many students are not paying attention.)	teaching methods	teaching methods	exercises	(none)
Banyak simbol-simbol yang asing dikenal. Lebih baik setiap simbol dijelaskan kembali dan diberikan contoh soal (There are many unfamiliar symbols. It would be better if each symbol is explained again and example problems are provided)	exercises	exercises	(none)	teaching methods

Error analysis reveals two key advantages of IndoBERTweet over IndoBERT and mBERT: a better understanding of informal language and misspellings, and a better understanding of contextual information. Table 7 presents six examples from the test data highlighting these advantages. Examples 1-3 demonstrate IndoBERTweet's ability to accurately label feedback despite informal language and misspellings. IndoBERT manages to identify some labels correctly, while mBERT has difficulty identifying most of the labels. This advantage stems from IndoBERTweet's pre-training on a large corpus of Indonesian tweets, which exposed it to a wide variety of informal language patterns, abbreviations, and misspellings common in online communication. This pre-training makes IndoBERTweet more robust and adaptable to the casual language style prevalent in student feedback.

Examples 4-6 showcase IndoBERTweet's better ability to grasp contextual information within the feedback, enabling it to accurately assign labels even when keywords aren't explicitly stated. For instance, one feedback states "Dosennya sudah baik tetapi terlalu cepat dalam menjelaskan" (The lecturer is good but explains too quickly). This implicitly criticises the teaching pace, but neither IndoBERT nor mBERT classify it as relating to "teaching_method". This strength likely arises from its pre-training on text where meaning is often conveyed through context and subtext. This contrasts with IndoBERT and mBERT, which often struggle to classify feedback that relies on implicit meaning or indirect phrasing, particularly when those labels appear less frequently in the dataset. These examples demonstrate the value of pre-training on data closely aligned with the target task, as IndoBERTweet's success in capturing both informal language and contextual nuances makes it a more reliable model for analysing student feedback.

4. Conclusions

This study investigates the performance of IndoBERT, IndoBERTweet, and mBERT for multilabel classification of informal Indonesian student feedback. The dataset consists of 4,301 comments from an academic institution, with feedback assigned into nine predefined labels. We applied a fine-tuning approach to each BERT variant, experimenting with varying sequence lengths (128, 96, 64, and 32 tokens) and three truncation strategies (beginning, middle, end). IndoBERTweet consistently outperformed IndoBERT and mBERT across all configurations, with the best performance achieved using a sequence length of 64 tokens and end truncation. The model achieved a macro-average F1-score of 0.8462, outperforming IndoBERT (best model 0.8243) and mBERT (best model 0.8230). IndoBERTweet's better performance in this study can be attributed to its pre-training on Indonesian social media data, which includes informal language, abbreviations, misspellings, and colloquial expressions similar to those found in online student feedback. This pre-training enables IndoBERTweet to better interpret and classify informal text, making it particularly suited for tasks that involve casual language and varied expression styles. Additionally, IndoBERTweet's ability to capture nuanced context allows it to excel in multilabel classification, where feedback often addresses multiple aspects of a course within a single comment. These strengths make IndoBERTweet an effective model for analyzing informal, nuanced feedback in educational settings. In the future, several directions could be explored to improve the model's robustness and applicability. Firstly, future research could explore advanced data augmentation techniques to create synthetic data for underrepresented categories, enhancing model performance and generalisation. Secondly, given the relative strengths of different BERT variants, exploring ensemble techniques that combine predictions from IndoBERT, IndoBERTweet, and mBERT could potentially improve overall accuracy and address performance gaps in specific labels. This approach could combine the unique strengths of each model for a more robust and comprehensive analysis. Finally, these findings have potential applications in educational data mining, such as automating real-time feedback analysis, adjusting course content based on feedback, and enhancing personalized learning experiences. These tools could help educators better understand student sentiment and engagement, leading to more informed and targeted educational strategies.

References

[1] A. S. Sunar and M. S. Khalid, “Natural Language Processing of Student’s Feedback to Instructors: A Systematic Review,” *IEEE Trans. Learning Technol.*, vol. 17, pp. 741–753, 2024, doi: 10.1109/TLT.2023.3330531.

[2] A. Rusli, A. Suryadibrata, S. B. Nusantara, and J. C. Young, “A Comparison of Traditional Machine Learning Approaches for Supervised Feedback Classification in Bahasa Indonesia,” *International Journal of New Media Technology*, vol. 7, no. 1, pp. 28–32, Jul. 2020, doi: 10.31937/ijnmt.v1i1.1485.

[3] C. A. Haryani, W. Daicy, A. E. Widjaja, A. Aribowo, K. Prasetya, and Hery, “Educational Data Mining: The Application in The University’s Feedback Survey Analysis using Classification and Clustering Techniques,” in *2022 International Conference on Science and Technology (ICOSTECH)*, Batam City, Indonesia: IEEE, Feb. 2022, pp. 01–08. doi: 10.1109/ICOSTECH54296.2022.9829148.

[4] F. K. Khaiser, A. Saad, and C. Mason, “Analysis Of Students’ Feedback on Institutional Facilities Using Text-Based Classification and Natural Language Processing (NLP),” *JLC*, vol. 10, no. 1, pp. 101–111, Mar. 2023, doi: 10.47836/jlc.10.01.06.

[5] M. Edalati, A. S. Imran, Z. Kastrati, and S. M. Daudpota, “The Potential of Machine Learning Algorithms for Sentiment Classification of Students’ Feedback on MOOC,” in *Intelligent Systems and Applications*, vol. 296, K. Arai, Ed., in Lecture Notes in Networks and Systems, vol. 296. , Cham: Springer International Publishing, 2022, pp. 11–22. doi: 10.1007/978-3-030-82199-9_2.

[6] O. Rakhmanov, “A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments,” *Procedia Computer Science*, vol. 178, pp. 194–204, 2020, doi: 10.1016/j.procs.2020.11.021.

[7] D. Ruiz Alonso, C. Zepeda Cortés, H. Castillo Zacatelco, J. L. Carballido Carranza, and J. L. García Cué, “Multi-label classification of feedbacks,” *IFS*, vol. 42, no. 5, pp. 4337–4343, Mar. 2022, doi: 10.3233/JIFS-219224.

[8] V. Veerachamy, A. George, and J. Beulah, “Intelligent Analysis of Student Feedback in Post-course Assessment Using a Multiclass Classification Model,” in *Intelligent Systems Design and Applications*, vol. 1052, A. Abraham, A. Bajaj, and T. Hanne, Eds., in Lecture Notes in Networks and Systems, vol. 1052. , Cham: Springer Nature Switzerland, 2024, pp. 376–387. doi: 10.1007/978-3-031-64776-5_36.

[9] M. Z. Asghar *et al.*, “An Efficient Classification of Emotions in Students’ Feedback using Deep Neural Network,” in *2022 13th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan: IEEE, Jun. 2022, pp. 186–191. doi: 10.1109/ICICS55353.2022.9811152.

[10] A. Onan, “Mining opinions from instructor evaluation reviews: A deep learning approach,” *Comp Applic In Engineering*, vol. 28, no. 1, pp. 117–138, Jan. 2020, doi: 10.1002/cae.22179.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[12] H. Setiawan, C. Fatichah, and A. Saikhu, “Multilabel Classification of Student Feedback Data Using BERT and Machine Learning Methods,” in *2023 14th International*

Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia: IEEE, Oct. 2023, pp. 147–152. doi: 10.1109/ICTS58770.2023.10330849.

[13] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, “Indonesian multilabel classification using IndoBERT embedding and MBERT classification,” *IJECE*, vol. 14, no. 1, p. 1071, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[14] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[15] K. Chovanec, J. Fields, and P. Madiraju, “Combining Demographic Tabular Data with BERT Outputs for Multilabel Text Classification in Higher Education Survey Data,” in *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy: IEEE, Dec. 2023, pp. 1403–1409. doi: 10.1109/BigData59044.2023.10386843.

[16] M. A. Mutasodirin and R. E. Prasojo, “Investigating Text Shortening Strategy in BERT: Truncation vs Summarization,” in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia: IEEE, Oct. 2021, pp. 1–5. doi: 10.1109/ICACSIS53237.2021.9631364.

[17] J. Chen and S. Lv, “Long Text Truncation Algorithm Based on Label Embedding in Text Classification,” *Applied Sciences*, vol. 12, no. 19, p. 9874, Sep. 2022, doi: 10.3390/app12199874.

[18] Z. Yang, J. Li, H. Song, and X. Du, “Global Semantic Information Extraction Model for Chinese long text classification based on fine-tune BERT,” in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Chongqing, China: IEEE, Jun. 2022, pp. 209–213. doi: 10.1109/ITAIC54216.2022.9836921.

[19] L. H. Suadaa, I. Santoso, and A. T. B. Panjaitan, “Transfer Learning of Pre-trained Transformers for Covid-19 Hoax Detection in Indonesian Language,” *Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, p. 317, Jul. 2021, doi: 10.22146/ijccs.66205.

[20] S. Saadah, Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, and Aniq Atiqi Rohmawati, “Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia,” *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 6, no. 4, pp. 648–655, Aug. 2022, doi: 10.29207/resti.v6i4.4215.

[21] N. M. Azahra and E. B. Setiawan, “Sentence-Level Granularity Oriented Sentiment Analysis of Social Media Using Long Short-Term Memory (LSTM) and IndoBERTweet Method,” vol. 9, no. 1, 2023.

[22] I. Budiman *et al.*, “Classification Performance Comparison of BERT and IndoBERT on SelfReport of COVID-19 Status on Social Media,” *J. Comput. Sci. Inst.*, vol. 30, pp. 61–67, Mar. 2024, doi: 10.35784/jcsi.5564.

[23] L. D. Cahya, A. Luthfiarta, J. I. T. Krisna, S. Winarno, and A. Nugraha, “Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model,” *TEKNOSI*, vol. 9, no. 3, pp. 290–298, Jan. 2024, doi: 10.25077/TEKNOSI.v9i3.2023.290-298.

[24] M. Fadhel and W. Maharani, “Depression Detection of Users in Social Media X using IndoBERTweet,” *SinkrOn*, vol. 8, no. 2, pp. 885–891, Mar. 2024, doi: 10.33395/sinkron.v9i2.13354.

[25] L. Geni, E. Yulianti, and D. I. Sensuse, “Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models,” vol. 9, no. 3, 2023.

[26] A. Jazuli, Widowati, and R. Kusumaningrum, “Aspect-based sentiment analysis on student reviews using the Indo-Bert base model,” *E3S Web Conf.*, vol. 448, p. 02004, 2023, doi: 10.1051/e3sconf/202344802004.

[27] J. F. Kusuma and A. Chowanda, “Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter,” *JOIV : Int. J. Inform. Visualization*, vol. 7, no. 3, pp. 773–780, Sep. 2023, doi: 10.30630/joiv.7.3.1035.

[28] Muhammad Ikram Kaer Sinapoy, Yuliant Sibaroni, and Sri Suryani Prasetyowati, “Comparison of LSTM and IndoBERT Method in Identifying Hoax on Twitter,” *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 7, no. 3, pp. 657–662, Jun. 2023, doi: 10.29207/resti.v7i3.4830.

[29] E. Yulianti and N. K. Nissa, “ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches,” *Bulletin EEI*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[30] I. F. Putra and A. Purwarianti, “Improving Indonesian Text Classification Using Multilingual Language Model,” in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, Tokoname, Japan: IEEE, Sep. 2020, pp. 1–5. doi: 10.1109/ICAICTA49861.2020.9429038.

[31] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[32] *iterative-stratification: Package that provides scikit-learn compatible cross validators with stratification for multilabel data*. Python. Accessed: Sep. 23, 2024. [Online]. Available: https://github.com/trent-b/iterative-stratification