

Fine-tuning language models for social sentiment: insights from startup user feedback

Li-Hua Li¹, Agus Cahyo Nugroho^{1,2}

¹Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan

²Department of Information System, Faculty of Computer Science, Soegijapranata Catholic University, Semarang, Indonesia

Article Info

Article history:

Received Jun 25, 2025

Revised Jul 20, 2026

Accepted Aug 12, 2026

Keywords:

Large language models

Machine learning

Sentiment analysis

Small language models

Startup

ABSTRACT

Startup growth and sustainability are heavily impacted by user comments. This research has done a comparative study between large language models (LLMs), small language models (SLMs), and traditional machine learning (ML) methods in classifying user comments on startups. The performances of bidirectional encoder representations from transformers (BERT), Indonesian bidirectional encoder representations from transformers (IndoBERT), distilled bidirectional encoder representations from transformers (DistilBERT), and ML models; logistic regression, support vector machine (SVM), random forest, and extreme gradient boosting (XGBoost) were evaluated with a diverse dataset and beyond extensive preprocessing. Out of all the models this research trained, IndoBERT outperformed with an accuracy and recall 90.99% and F1-score of 89.94%, due to its intermediate pretraining for Indonesian language task. DistilBERT, from the SLM family of models and only requiring a fraction of resources to make predictions performed significantly better (over 90% in both accuracy and recall). These outcomes indicate that LLMs, particularly those targeting specific languages, are significantly superior in performance for sentiment analysis. The results seem to imply that models like these can be utilized to study the phenomenon and further enhance user engagement with startups. Future work will involve the further development of more language specific BERT versions and expanding LLMs in other ways within startup spaces.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Agus Cahyo Nugroho

Department of Information Management, Chaoyang University of Technology

Taichung, Taiwan

Email: agus.nugroho@unika.ac.id

1. INTRODUCTION

Sentiment analysis [1] without a doubt one of the most important method/techniques for companies in Indonesia ranging from start-up to corporate and even government. As new companies grow, it becomes essential to analyze sentiment [2] from many sources—from news stories and social media to consumer evaluations. In contrast, despite the successful application of sentiment analysis in many cases, there is not much research works that combines traditional natural language processing (NLP) techniques and modern transformer-based model specifically for Indonesian startup companies. This tech gap is considerable—particularly with the scale of linguistic fragmentation, rapidly increasing digital adoption and ever-growing economic potential that Indonesian startups represent. Closing this gap can deliver better, more actionable insights for these new kinds of businesses. Yet, there are also challenges regarding Indonesian language sentiment analysis such as linguistic variations, informal expressions and lack of labeled data. Standard NLP methods such as machine learning (ML)-based sentiment classifiers [3] are commonly employed for

sentiment analysis. Recently, novel small language models (SLM) and large language models (LLMs) have allowed new opportunities for sentiment analysis [4]. LLMs use a large amount of computational power, but they provide better contextual understanding than SLMs; on the other hand, SLMs offer a more efficient tradeoff between performance and resource utilization.

This paper evaluates the performance of traditional ML models, SLM, and LLMs for sentiment analysis on Indonesian startups. The objective of this research is to evaluate them based on their performance, efficiency and use as a practical aspect for sentiment analysis over some textual texts. In any traditional sentiment analysis, the ML models [3], for instance, naïve Bayes [3], support vector machines (SVM) [3], or logistic regression [3], extract handcrafted features such as term frequency and use them to classify text into positive, negative or neutral sentiments. These are light weight and interpretable models though they often fail to understand context, sarcasm or the nuances of expression.

Now, Indonesia's startup ecosystem has been growing quickly through various rounds of funding, especially for types like fintech, e-commerce and edtech. With increasing competition, actual user feedback is proving more paramount for startups to help reduce the time taken to reach product-market fit, improve user experience and achieve growth in a constantly changing market. This research provides a reference into the best practice of sentiment analysis in Indonesian startup ecosystem by connecting classical NLP approach to modern language model so that startup can make data driven decision based on public sentiment.

This research aims to find out the best and most feasibility methods for analyzing user sentiment on Indonesian startups. In this way, the study intends to help start-ups going through different sentiment analysis tools based on their technical capacity and business needs as well as to use user feedback in decision-making. The following are this study's primary contributions: i) a comprehensive comparison of ML-based, SLM-based, and LLM-based sentiment analysis approaches; and ii) evaluation of the impact of different model architectures on sentiment classification accuracy, precision, recall, and F1-score.

This paper is organized as follows: the introduction of SLM, LLM, ML, startup information, and the review of the literatures are covered in section 2. Section 3 methods, details about data collection, data preprocessing, fine tuning, evaluation metrics, and experimental design. Section 4 results and discussions. The key findings are concluded in section 5 along with recommendations for future research.

2. LITERATURE REVIEW

2.1. Overview startups in Indonesia

An Indonesian e-commerce site called Blibli at <https://www.blibli.com/> was established in 2011 and provides a large selection of goods and services, including quick shipment and safe payment methods. Established in 2009, Tokopedia at <https://www.tokopedia.com/> is a prominent e-commerce platform that facilitates business-to-business and consumer-to-consumer transactions, thereby empowering small and local enterprises. Bukalapak at <https://www.bukalapak.com/> (founded in 2010) connects small and medium enterprises (SMEs) with consumers, selling diverse products and offering financial technology solutions like digital payments. DANA at <https://www.dana.id/> (founded in 2018) a digital wallet providing cashless transactions, bill payments and person-to-person (P2P) transfers. Gojek started in 2010 as a courier and transportation company, becoming a mega app with more than 20 services across Southeast Asia. OVO at <https://www.ovo.id/> (established in 2017)-a payment platform that offers digital payments, mobile top-up and bills, and loyalty-based rewards program by working with merchants and digital player partners. Ruangguru at <https://www.ruangguru.com/> founded in 2014 to improve the accessibility of education for students in Indonesia. Traveloka at <https://www.traveloka.com/en-en> (established in 2012) is an online travel and lifestyle platform, offers services on flight and hotel bookings throughout Southeast Asia.

These were chosen because they not only represent a broad cross section of Indonesia's digital economy including e-commerce, fintech, edtech, and travel, but like quality, also have enough active users giving feedback on their platforms. Due to their significant roles in the Indonesian startup ecosystem, these platforms make great case studies for applying and assessing sentiment analysis models because comments made by users on these platforms can reveal consumer expectations, satisfaction, and issues. Sentiment analysis from these firms provides this research with an opportunity to investigate practical uses of ML, SLM, and LLM approaches in various business situations.

2.2. Sentiment analysis

Sentiment analysis (also known as opinion mining) is a key area of NLP [5] that focuses on finding, extracting, and classifying subjective information in textual data to determine whether the expressed opinion is neutral, negative, or positive. This process is essential for deciphering consumer sentiments, market trends, and public sentiment to make better decisions through a variety of industries. It is the use of computational techniques to discover and extract opinions from text data regarding products, services, organizations, individuals, or events [6]. Several studies 6 review sentiment analysis methods in detail; they cover both

classical ML [7] like naïve Bayes, SVM, and logistic regression, as well as state-of-the-art deep learning or transformer-based methods using bidirectional encoder representations from transformers (BERT) and its variants. These studies overview a diverse set of use cases across sectors from e-commerce (Tokopedia, Bukalapak, and Blibli), fintech (DANA and OVO), transportation and lifestyle services (Gojek and Traveloka) to education (Ruangguru). All these platforms get tons of user-generated feedback like product reviews, comments on social media posts and complaints; therefore, sentiment analysis becomes a great assistant in getting insights about user satisfactions, service gaps and strategic decision making. This study focused on categorizing those diverse startup areas to show the many shapes: it encompasses how sentiment analysis-based research can be applied in a complex multi-dimensional space for real-world contexts, that both enables its applicable potential and exposes the types of challenges faced when deploying such models in Indonesia digital economy.

2.3. Machine learning methods

ML is a branch of artificial intelligence, which allows machines to learn from the data and make predictions without being explicitly programmed. In sentiment analysis, classical ML models (logistic regression [3], naïve Bayes [3], SVM [3], random forests [3], and extreme gradient boosting (XGBoost)) are trained on positively or negatively labeled text to predict whether a new instance of text is positive or negative and label it accordingly. Due to their simple, interpretable and low numeric cost, these models are both widely used. ML methods are useful to summarize as they provide good performance and comparisons with more advanced SLMs and LLMs. Additionally, they still make sense for start-ups or organizations with limited resources. This study covers approaches of ML to benchmark sentiment analysis methods and their implementation in the context of Indonesian startups.

2.3.1. Logistic regression

Logistic regression [8] is one of the most used algorithms for binary classification and it is widely applied in sentiment analysis to determine whether a text falls into positive, neutral or negative categories. Due to its simplicity and effectiveness, it is very commonly used in both research and real-world applications, such as the TripAdvisor review analysis [9]. Danyal *et al.* [10] found that logistic regression is better than naïve Bayes in analyzing 104,215 movie reviews.

Logit is the function that underlies the formula of logistic regression which is used as an approach to solve binary classification problem. The fundamental formula of logit function [11] is shown in (1).

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_nX_n)}} \quad (1)$$

Where $P(Y = 1 | X)$ is the probability that the dependent variable Y equals 1, given the independent variable X , β_0 is the intercept (bias term), $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients for each independent variable $X_1, X_2, \dots, X_n$, and e is the base of the natural logarithm (Euler's number, approximately 2.718).

2.3.2. Support vector machine

In sentiment analysis, SVM [12] are efficient supervised learning techniques that categorize text as neutral, negative, or positive. They work by finding an optimal hyperplane in high-dimensional space for classification. SVM has proven successful in analyzing tweets, handling short and ambiguous text. Studies show SVM [13] often outperforms models like logistic regression, naïve Bayes, and random forest, especially in tasks like classifying sentiments in United States airline X (formerly Twitter) data.

SVM is a supervised learning technique that is applied to tasks involving regression and classification. Finding the best hyperplane to optimize the margin between two classes is its aim.

- i) Equation of the decision boundary (hyperplane): for a linear SVM [14], the decision boundary is defined as shown in (2).

$$f(X) = w^T X + b = 0 \quad (2)$$

Where X is the input feature vector, w is the weight vector (coefficients), and b is the bias (intercept). A point X belongs to class +1 if $f(X) > 0$ and class -1 if $f(X) < 0$.

- ii) Margin maximization condition: the SVM [15] optimizes the margin by solving as shown in (3).

$$\min_{w,b} = \frac{1}{2} \|w\|^2 \quad (3)$$

Subject to the constraint as shown in (4).

$$y_i(w^T X_i + b) \geq 1, \forall_i \quad (4)$$

Where y_i is the class label (+1 or -1) and X_i are the training points.

2.3.3. Random forest

An ensemble learning technique called random forest [16] is used in sentiment analysis to categorize text as neutral, negative, or positive. To increase accuracy and decrease overfitting, it constructs numerous decision trees and use majority voting. Munshi *et al.* [17] have applied it to sentiment classification in support of social risk management, while Wang [18] highlights its effectiveness alongside SVM and neural networks in both academic and real-world applications.

- i) Individual decision tree formula: as demonstrated by (5), every random forest [19] decision tree corresponds to a recursive splitting rule.

$$G = \sum p_k (1-p_k) \quad (5)$$

Where G is the Gini impurity (used for classification), p_k is the proportion of class k in a node, and the split aims to minimize the Gini impurity. For regression, decision trees use the mean squared error (MSE) as shown in (6).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2 \quad (6)$$

Where y_i is the computed output and $\hat{y}$ is the average of target values in the node.

2.3.4. Extreme gradient boosting

Chen and Guestrin [20] created the quick and effective gradient boosting technique known as XGBoost in 2016. It is frequently utilized in challenges and real-world projects because of its excellent performance. In sentiment analysis, studies have shown improved results by combining XGBoost with reviewer reliability and the whale optimization algorithm [21], enhancing product recommendation systems.

- i) Prediction formula: in XGBoost, the model consists of K decision trees, and the final prediction [22] is the sum of all tree outputs as shown in (7).

$$\hat{y}_i = \sum_{k=1}^K f_k(X_i) \quad (7)$$

Where $\hat{y}_i$ is the predicted value for instance i , $f_k(X_i)$ is the output of the k -th decision tree, and K is the total number of trees. Each tree learns from the residuals (errors) of the previous trees.

- ii) Loss function (objective function): XGBoost [23], [24] minimizes a regularized loss function as shown in (8).

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (8)$$

Where $l(y_i, \hat{y}_i)$ is the loss function (e.g., squared error for regression, log loss for classification), and $\Omega(f_k)$ is the regularization term to prevent overfitting as shown in (9).

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_j w_j^2 \quad (9)$$

Where T is the number of leaves in the tree, w_j are the leaf weights, and γ and λ are hyperparameters for regularization.

2.4. Large language model

LLMs [25] are sophisticated deep learning models that have been trained on enormous text corpora to comprehend and produce language that is like that of humans. LLMs like BERT [26] (launched by Google in 2018), GPT (first published by OpenAI in 2018), and Indonesian bidirectional encoder representations from transformers (IndoBERT) [27], which are based on transformer architectures, are excellent at capturing contextual relationships between words, allowing them to carry out challenging NLP tasks like sentiment analysis, text classification, and question answering. LLMs have been around since about 2018, when they showed an extraordinary, pre-trained generalization ability from massive natural language data at high accuracy to many different domains and languages. LLMs perform better than traditional models in sentiment analysis because they can do more fine-grained text understanding, which makes them a good

candidate for extracting insights from user feedback, especially when there is informal or multilingual corpus as would be the case with many startup companies.

2.4.1. Bidirectional encoder representations from transformers

A new language representation approach by Google researchers called BERT was introduced in 2018 [28]. Such a paradigm helps us to understand the uses of words based on their context in a sentence. This capability has benefits that have revolutionized NLP.

Architecture and training; BERT [29] is a transformer-based language model that only makes use of the encoder. It creates embeddings by integrating token, location, and segment information, tokenizes input into subwords, and adds special tokens like [CLS] and [SEP]. For BERT to learn rich and bidirectional context, the model uses 12 transformer layers with feed-forward networks and multi-head self-attention. For classification or other NLP tasks, the final output is fed into a task-specific layer. BERT performs well in a variety of language processing applications thanks to its deep architecture and bidirectional nature as seen in Figure 1.

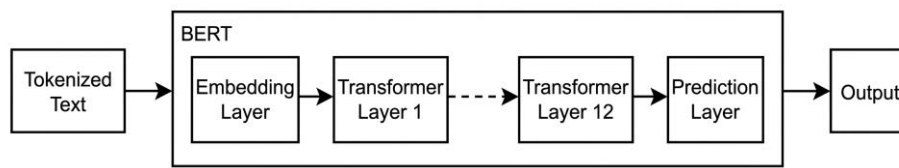


Figure 1. BERT architecture

There are two main tasks involved in BERT training:

- i) Masked language modeling (MLM): the process involves randomly masking some words in a sentence so that the model may learn to anticipate the masked words based on their context. In this instance, the model is urged to gain a thorough comprehension of the word relationships and language context, as indicated by (10).

$$L_{MLM} = -\sum_{i \in M} \log P(y_i | X_{masked}) \quad (10)$$

Where M is the set of masked tokens, y_i is the correct token, X_{masked} is the input with some tokens replaced by [MASK], and L_{MLM} is the loss function for MLM. Example: input: “The cat sat on the [MASK]” and model learns to predict “mat”.

- ii) Next sentence prediction (NSP): this task determines whether two sentences follow each other logically. By practicing on this task, BERT learns the relationship between sentences, which is very important in tasks such as question answering and natural language inference as shown in (11).

$$L_{NSP} = -\sum [y \log P(A|S_1, S_2) + (1 - y) \log P(B|S_1, S_2)] \quad (11)$$

Where S_1 and S_2 are two consecutive text segments, $P(A|S_1, S_2)$ the predicted probability that S_2 is the next sentence after S_1 , $P(B|S_1, S_2)$ the predicted probability that S_2 is not the next sentence after S_1 , $y = 1$ if S_2 is the next sentence, otherwise $y = 0$, and L_{NSP} the loss function for the NSP task. It is a binary cross-entropy loss that measures how well the model predicts whether S_2 follows S_1 or not. Example: input: (“I love deep learning.”, “It is very powerful.”) → Label: next and input: (“I love deep learning.”, “Dogs are friendly.”) → Label: Not Next.

- iii) BERT fine tuning formula: For downstream tasks, a classification head is added as shown in (12).

$$P(y|X) = \text{softmax}(Wh_{[CLS]} + b) \quad (12)$$

Where, $h_{[CLS]}$ is the representation of the [CLS] token, and W and b are trainable parameters.

2.4.2. Indonesian BERT

IndoBERT [30] is a transformer-based language model tailored specifically for the Indonesian language, inspired by BERT. It was developed to address the underrepresentation of Indonesian in NLP

research, despite the language being spoken by over 200 million people. IndoBERT was trained on Indo4B, a large-scale corpus containing 4.3 billion tokens collected from diverse sources, including news articles, social media content, Indonesian Wikipedia, and various Indonesian web pages. This model was proposed with the IndoLEM benchmark at <https://github.com/indolem/indolem>, a set of evaluation tasks for seven core NLP from 7 tasks: part-of-speech (POS) tagging [31], named entity recognition (NER) [31], dependency parsing [31], constituency parsing [31], semantic role labeling (SRL), and natural language inference (NLI) document classification [31]. These tasks characterize a model's syntactic knowing (POS tagging and parsing) and semantic thinking (SRL and NLI), yielding a complete portrait of language understanding. IndoBERT yielded the overall best performance on most of these tasks achieving state-of-the-art results for both measuring syntactic structure and semantic nuances of the Indonesian language.

2.5. Small language model

Researchers develop compact versions of LLMs to address NLP tasks (SLMs) [32], consuming less computing power. Even though SLMs [33] have fewer parameters compared to large models such as GPT-4 or BERT-large, they maintain respectable performance on text classification, summarization and question answering tasks. Their compactness allows for faster inference, decreased memory consumption, and deployment in low-resource contexts or edge devices. While SLMs usually do not reach comparable levels of understanding as larger models, they are trained on a specific task—within a domain with high efficiency and accuracy for targeted tasks.

2.5.1. Distilled BERT

BERT was distilled into a small-BERT with 60% less parameters and around 97% of BERT's performance, known as distilled bidirectional encoder representations from transformers (DistilBERT) [34]. It performs well with sentiment analysis on Huge social media data and is best in low-resource environment. Trained by mimicking a larger BERT model [35], DistilBERT has six transformer layers as seen in Figure 2, half of BERT-base, using self-attention and feed-forward networks. It simplifies architecture by removing token type embeddings and NSP, offering a balance of speed, simplicity, and accuracy for practical NLP applications.

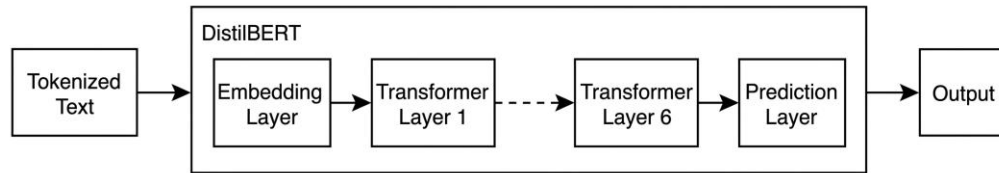


Figure 2. DistilBERT architecture

Transformer encoder formula; like BERT and DistilBERT [27] is based on the transformer encoder architecture. Each encoder layer consists of self-attention as in (13) and feedforward layers as in (14).

- i) Self-attention mechanism (scaled dot-product attention)

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

Where Q, K, V are the query, key, and value matrices from the input embeddings, d_k is the dimensionality of keys (scaling factor), softmax ensures that the attention scores sum to 1, and each token attends to all other tokens in the input sequence.

- ii) Feedforward network: after attention, a fully connected feedforward network is applied, as in (14).

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (14)$$

Where W_1, W_2 and b_1, b_2 are learnable weights and biases, and $max(0, \cdot)$ is the rectified linear unit (ReLU) activation function.

- iii) Knowledge distillation loss: DistilBERT is trained using a distillation loss, which combines:
- Cross-entropy loss (LCE) for label prediction as shown in (15).

$$LCE = -\sum y \log(\hat{y}) \quad (15)$$

Where y is the true label and $\hat{y}$ is the predicted probability.

- Distillation loss (L_{KL}) for soft targets as shown in (16).

$$L_{KL} = T^2 \sum p_{teacher} \log \left(\frac{p_{teacher}}{p_{student}} \right) \quad (16)$$

Where, T is the temperature (scaling factor), and $p_{teacher}$ and $p_{student}$ are probability distributions from BERT and DistilBERT. The final combined loss is as shown in (17).

$$L = \alpha L_{CE} + (1 - \alpha) L_{KL} \quad (17)$$

Where α balances hard and soft labels.

DistilBERT also uses transformers like BERT therefore it has the same formula regarding self attention, feedforward network, and classification loss.

3. METHOD

This section outlines the process for creating sentiment analysis models based on startup user reviews and ratings using ML, SLM, and LLM. The overall workflow of this study is structured into several systematic phases. Figure 3 shows how these phases are organized for data collection, preprocessing, training, and evaluation.

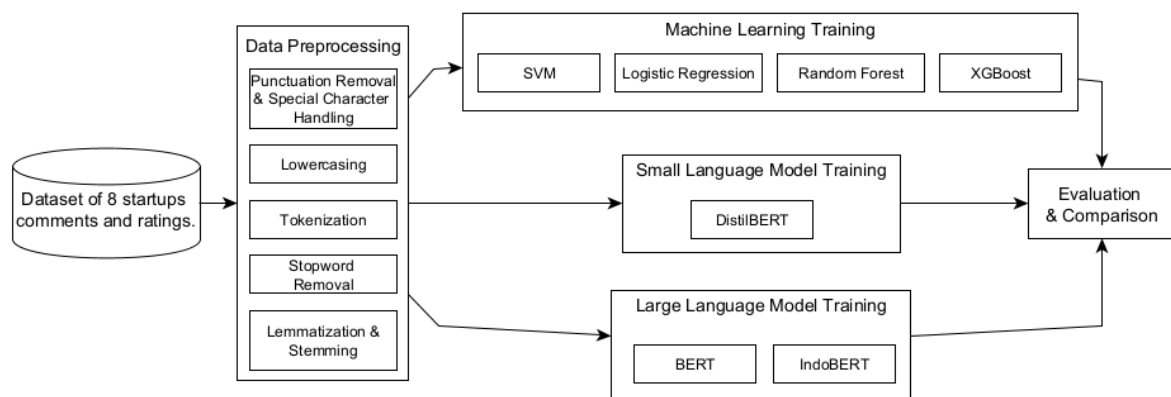


Figure 3. Research workflow

3.1. Data collection

Dataset is collected from Kaggle at <https://www.kaggle.com/datasets/rezkyayang/reviews-of-indonesian-app-startups-on-playstore> and the license is CC0: public domain means no copyright. The listed Indonesian startups are notable for their diverse services and significant user engagement as seen in Table 1. This dataset includes reviews from popular Indonesian apps as seen in Table 2.

Table 1. Summary of startups app reviews and downloads

Startups	Main services	Reviews	Downloads
Gojek	Ride-hailing, food delivery, and online payments	116,000+	100,000,000+
Tokopedia	Online marketplace and supports local entrepreneurs	71,000+	100,000,000+
Dana	Digital wallet, payments, and money transfers	106,000+	100,000,000+
Traveloka	Travel booking (flights, hotels, and entertainment)	145,000+	50,000,000+
Bukalapak	Digital marketplace and supports micro, small, and medium enterprises (MSMEs)	90,000+	50,000,000+
OVO	Digital wallet, payments, transfers, and investments	83,000+	50,000,000+
Blibli	E-commerce platform	116,000+	10,000,000+
Ruangguru	Educational platform, online learning, and live classes	111,000+	10,000,000+

Table 2. Example dataset indonesia startups

Example	Score	Original content	English translation
Ex1	5	<i>"Aplikasi yang sangat membantu, puas banget dengan respon cepatnya. Semoga tambah sukses dan bagus terus untuk traveloka."</i>	"A very helpful app, very satisfied with the quick response. I wish Traveloka continued success and success."
Ex2	4	<i>"knp bintang 4? sebenarnya sistemnya udh bagus, and semoga suatu hr nnti.... disistem Traveloka mampu "membaca" kebiasaan user nya (via medsos), sehingga secara otomatis memberikan "promo/diskon" sesuai dg minat/kebiasaan user nya. Terimakasih Traveloka, hope for long life and the best thing ever ðŸ™ ðŸ'ž"</i>	"Why 4 stars? Actually, the system is already good, and hopefully one day.... the Traveloka system will be able to 'read' the habits of its users (via social media), so that it automatically provides 'promos/discounts' according to the interests/habits of its users. Thank you Traveloka, hope for long life and the best thing ever ðŸ™ ðŸ'ž"
Ex3	3	<i>"cepat proses beli tiketnya. tp akhir2 ini harga tiket ke luar negeri yg ditampilkan suka pake harga hoax. Tampil 2 juta pas mau bayar 3 juta."</i>	"The ticket purchase process is fast. However, recently, the displayed international ticket prices have been hoaxed. It shows 2 million when you're trying to pay 3 million."
Ex4	2	<i>"saya sangat kecewa karna sampai sekarang paylater nya tidak aktif, padahal saya untuk pembayaran selalu tepat waktu dan tidak pernah sekalipun terlambat, saya hanya minta aktifkan kembali paylaternya, jadi saya kasih bintang 2 dulu untuk sementara waktu. thanks"</i>	"I'm very disappointed because PayLater is still inactive, even though I always make payments on time and have never been late. I just asked for PayLater to be reactivated, so I'm giving it 2 stars for the time being. Thanks."
Ex5	1	<i>"Aplikasi makin susah dan sulit utk di pahami di bandingkan aplikasi Traveloka sebelum masa COVID lbh mudah utk di pahami"</i>	"The application is getting more difficult and harder to understand compared to the Traveloka application before the COVID period, which was easier to understand."

3.2. Data preprocessing

In this study, data pretreatment is essential since it directly affects how well sentiment categorization works for text about Indonesian startups. To lessen textual noise that does not add semantic meaning and could adversely impact feature extraction in ML models, punctuation reduction and special character handling are used. Lowercasing is performed to keep lexically consistent and does not allow case changes to lead to duplicate feature representations. Tokenization: the first step in downstream-language analysis, this piece can be typically defined as a meaningful piece of text. Removing stop words enables the model to focus on emotionally rich words, removing high-frequency sentiment-neutral words. Finally, lemmatization and stemming were both used to normalise the word forms, reducing vocabulary sparsity presented in these morphologically rich languages (e.g., Indonesian) so they generalised better.

This further preprocessing pipeline helps to improve the data quality while reducing noise and redundancy, facilitating better, and more accurate sentiment classification. This also means that both SLM and regular ML use the same preprocessing procedures to ensure consistency in methodology. On the other hand, LLMs use internal subword tokenization and contextual representations that are designed to work well with raw text, so they need very little preparation other than minimal text cleaning.

3.3. Fine tuning process

The human fine-tuning technology on pretrained representations through language features and contextual nuances from the case study of sentiment analysis conducted by eight companies in Indonesia for both LLMs and SLMs such as BERT and DistilBERT. In supervised learning, a labeled dataset obtained from the score column is used. For ML and the SLM models, we perform a lot of preprocessing to improve data quality and reduce feature sparsity. Such as lowercasing, tokenisation, stop word removal, punctuation and special character removal, and lemmatization or stemming. In contrast, LLMs involve less preprocessing and rely on contextual embeddings and internal subword tokenization.

Lightweight SLMs are evaluated in conjunction with heavier LLMs to provide a balance between computational efficiency and predictive ability. This research employs a learning rate equals to 5×10^{-1} , three epochs of training, batch size equals to 16 for training and 64 for evaluation, 500 warm-up steps and weight decay equal to 0.01 during fine-tuning. These are designed to provide stable convergence while still being able to adapt to changing input distributions. They are evaluated using accuracy, precision, recall, and F1-score, with similar good results for all models after fine-tuning.

3.4. Evaluation metrics

The performance output following the completion of data training and production—which leads to evaluation. The accuracy, precision, recall and F1-scores of the models evaluated are shown in (18) to (21), respectively. Figure 4 explains each completed step. The results and discussion sections give detailed insight into this.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

$$Precision = \frac{TP}{TP+FP} \quad (19)$$

$$Recall = \frac{TP}{TP+FN} \quad (20)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (21)$$

Input: 8 Startup Comments Dataset

Output: Evaluation metrics.

Models: M1(BERT), M2(DistilBERT)

For Mi Do ($i \in [1,2]$)

- (1) Load the Startup Comments dataset in CSV format.
- (2) Punctuation removal & special character handling lowercasing, tokenization, stopword removal, lemmatization and stemming.
- (3) Split data into 70% for training, 15% for testing, and 15% for validation.
- (4) Convert tokenized text into numerical format using embedding techniques.
- (5) Load the pre-trained models Mi.
- (6) Send tokenized text to Mi model for classifying.
- (7) Fine-tune the model using supervised learning.
- (8) Optimize using AdamW optimizer with 0.00005 learning rate.
- (9) Use CrossEntropy for classification loss.
- (10) Train the models with n epochs.
- (11) Obtain the training loss and derive:
 - Accuracy
 - Precision
 - Recall
 - F1-Score

End Mi

Return (Evaluation Metrics based on each Mi)

Figure 4. Comments classifier using LLM/SLM

3.5. Experimental design

Positive, negative, or neutral comments refer to the dataset after preprocessing. It indicates whether the comments are categorized as positive, negative, or neutral sentiment. These details are outlined in Table 3. The training parameters as seen in Table 4 for BERT and DistilBERT were optimized for efficiency, with training conducted for three epochs to ensure quick convergence. The learning rate was set to 0.00005, with a 500-step warm-up for gradual adjustment. A weight decay of 0.01 was used to prevent overfitting. The dataset was split into training, validation, and testing subsets to enhance model training and evaluation.

Table 3. Description of the dataset

Dataset	Feature	Positive	Negative	Neutral	Total
Blibli	3	85,771	19,941	11,190	116,902
Tokopedia	3	35,599	31,058	4,585	71,242
Bukalapak	3	62,549	23,473	4,125	90,147
DANA	3	31,163	66,584	8,718	106,465
Gojek	3	49,331	56,659	10,027	116,017
OVO	3	24,075	53,580	5,527	83,182
Ruangguru	3	97,065	10,684	4,089	111,838
Traveloka	3	118,977	18,490	8,481	145,948

Table 4. Training parameters

Model parameters	BERT and DistilBERT
Epochs	3
Learning rate	0.00005
Warmup_steps	500
Weight Decay	0.01

The dataset will be split into training (70%), testing (15%), and validation (15%) subsets for reliable results. After acquisition and preprocessing, the dataset is divided, followed by training, testing, and validation. Next, fine tuning of the SLM and LLM will be performed, followed by model training. Finally, the sentiment analysis performance of ML, SLM, and LLM will be evaluated and compared to assess the effectiveness of fine tuning.

A comparison is made between BERT (represent LLM) and DistilBERT (represent SLM), along with four other conventional ML models to measure how they perform. Preprocessing: removal of punctuation, lowercasing, and tokenization (the process of breaking a sentence into tokens), stopwords removal, lemmatizing or stemming. For that experiment, the hardware is an Intel i9 14900 K CPU, NVIDIA GeForce RTX 4060 GPU and 32 GB of RAM. Programming used is python on Anaconda environment.

4. RESULT AND DISCUSSION

4.1. Result

Traditional ML methods, as illustrated in Table 5, only the logistic regression model for Ruangguru dataset gave the highest F1-score (87.54%). IndoBERT (represent LLM) with the highest F1-score = 89.94% for Ruangguru as shown in Table 6, and DistilBERT (represent SLM) gains higher score than Ruangguru dataset which is equal to 89.53% as shown in Table 7. In Table 8, this research compares the F1-scores of all models with respect to both datasets, where BERT (represent LLM) achieves superior performance compared to traditional ML techniques. DistilBERT (represent SLM) provides similar performance with a smaller amount of computational power, and IndoBERT (represent LLM), that is specifically trained for Indonesian outperforms all the other models at the same sentiment analysis task.

This research takes one sample from the highest performance using ML model which is logistic regression on Ruangguru startup dataset. Then visualizes precison-recall curves of the model from Figure 5. Besides F1-score this research also analyzes deeper using receiver operating characteristic-area under the curve (ROC-AUC). The summary of the performance is F1-score 87.54%, ROC-AUC (OvR): 90.59%, average precision (AP) (macro): 58.56%, and AP (weighted): 91.73%. Logistic regression succeeds well for polarized sentiments, especially positive reviews, according to the precision-recall analysis, but neutral sentiment is still difficult because of its semantic ambiguity, underscoring the significance of imbalance-aware evaluation metrics.

Table 5. The performance of using ML traditional models

Startups	Models	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Blibli	Logistic regression	82.17	79.76	82.17	80.14
	SVM	81.96	78.97	81.96	78.85
	Random forest	80.63	77.22	80.63	77.83
	XGBoost	80.50	77.47	80.50	76.71
Tokopedia	Logistic regression	84.19	81.36	84.19	81.88
	SVM	84.25	82.56	84.25	81.63
	XGBoost	82.73	79.50	82.73	80.22
	Random forest	82.75	80.32	82.75	80.14
Bukalapak	Logistic regression	86.93	85.13	86.93	85.26
	SVM	86.79	83.35	86.79	84.99
	Random forest	85.55	83.13	85.55	83.83
	XGBoost	85.37	82.57	85.37	83.54
DANA	Logistic regression	81.23	77.85	81.23	77.92
	SVM	81.10	77.61	81.10	77.43
	XGBoost	79.97	76.72	79.97	76.31
	Random forest	79.88	76.85	79.87	76.10
Gojek	Logistic regression	79.10	75.64	79.10	75.99
	SVM	79.29	79.04	79.29	75.81
	Random forest	77.69	73.72	77.69	74.27
	XGBoost	77.21	74.67	77.21	73.87
OVO	Logistic regression	83.75	80.92	83.75	80.83
	SVM	83.68	78.55	83.68	80.62
	Random forest	82.62	77.77	82.62	79.51
	XGBoost	82.39	79.78	82.39	79.27
Ruangguru	Logistic regression	89.51	86.33	89.51	87.54
	SVM	89.44	86.23	89.44	87.14
	XGBoost	88.96	85.81	88.96	86.28
	Random forest	88.39	84.69	88.39	85.64
Traveloka	Logistic regression	88.16	85.52	88.16	85.62
	SVM	88.16	85.63	88.16	85.54
	Random forest	87.11	83.90	87.11	84.76
	XGBoost	86.99	84.23	86.99	83.99

Table 6. The performance of using BERT (LLM)

Startups	Accuracy (%)		Precision (%)		Recall (%)		F1-score (%)	
	BERT	IndoBERT	BERT	IndoBERT	BERT	IndoBERT	BERT	IndoBERT
Bibli	84.49	85.17	83.28	83.89	84.49	85.17	83.75	84.35
Tokopedia	86.15	85.62	84.55	83.90	86.15	85.62	84.21	84.60
Bukalapak	88.55	88.91	86.65	87.39	88.55	88.91	87.04	87.95
DANA	83.38	83.80	80.94	81.49	83.38	83.80	81.08	81.83
Gojek	81.70	81.68	78.69	79.45	81.70	81.68	79.33	80.18
OVO	86.09	85.76	84.40	83.45	86.09	85.76	83.66	84.03
Ruangguru	89.27	90.99	85.68	89.56	89.27	90.99	87.44	89.94
Traveloka	89.44	89.62	87.58	87.86	89.44	89.62	87.81	88.40

Table 7. The performance of using DistilBERT (SLM)

Startups	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bibli	84.58	83.30	84.58	83.78
Tokopedia	85.99	84.00	85.99	84.50
Bukalapak	88.76	86.92	88.76	87.32
DANA	83.33	80.78	83.33	81.03
Gojek	81.75	78.89	81.75	79.29
OVO	85.81	83.62	85.81	83.65
Ruangguru	90.81	89.36	90.81	89.39
Traveloka	89.39	87.71	89.39	88.12

Table 8. The F1-score comparison of using ML, SLM, and LLM

Startups	Logistic regression (%) (Highest ML score)	BERT (%) (LLM)	DistilBERT (%) (SLM)	IndoBERT (%) (LLM)
Bibli	80.14	83.75	83.78	84.35
Tokopedia	81.88	84.21	84.50	84.60
Bukalapak	85.26	87.04	87.32	87.95
DANA	77.92	81.08	81.03	81.83
Gojek	75.99	79.33	79.29	80.18
OVO	80.83	83.66	83.65	84.03
Ruangguru	87.54	87.44	89.39	89.94
Traveloka	85.62	87.81	88.12	88.40

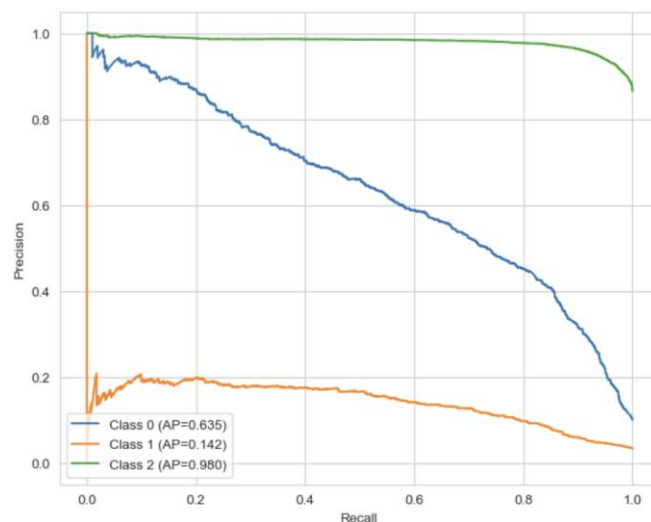


Figure 5. Precision-recall curves for logistic regression on Ruangguru dataset

In summary, the IndoBERT model on Ruangguru startup dataset performs well overall with an F1-score of 89.94% and ROC-AUC of 92.17%, however it is significantly biased because of class imbalance. There is a significant difference between the model's weighted (92.86%) and macro (65.48%) AP scores since it achieves almost perfect precision for class 2 ($AP = 0.983$) but is unable to identify class 1 ($AP = 0.280$) as seen in Figure 6. This implies that the model needs to be optimized to better generalize across minority classes, even despite high aggregate metrics.

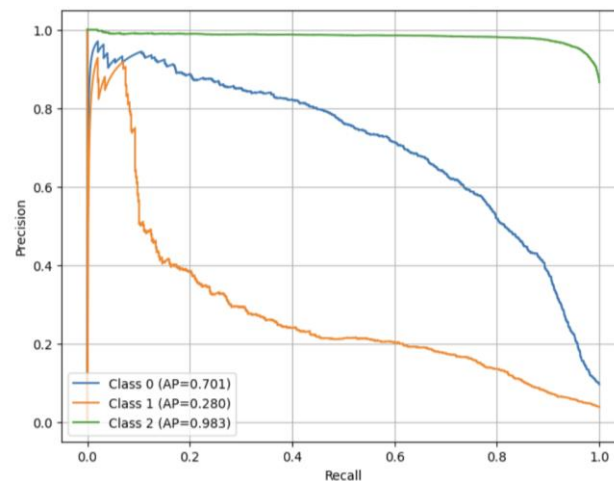


Figure 6. Precision-recall curves for IndoBERT on Ruangguru dataset

In conclusion, DistilBERT on Ruangguru startup dataset has a significant class imbalance despite nearly same IndoBERT's predictive power with 90.81% accuracy and 89.39% F1-score. With only a ~0.7% decrease in macro-AP (64.76%) and weighted (92.66%), DistilBERT's performance is nearly comparable to that of the heavier IndoBERT variant. The lighter DistilBERT model maintains the predictive power of the complete model, but it still has the same bias problems with the "neutral" class as seen in Figure 7, which is a good finding for efficiency. As a result, without data rebalancing strategies like resampling or class weighting, the model's high aggregate scores conceal its inability to recognize minority opinions, making it unreliable for negative reviews.

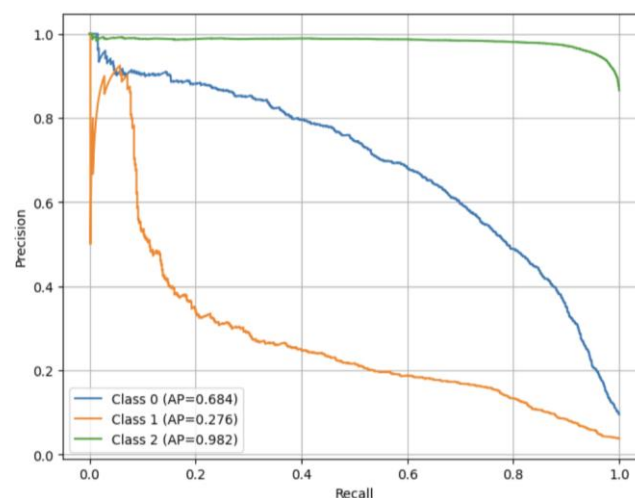


Figure 7. Precision-recall curves for DistilBERT on Ruangguru dataset

4.2. Discussion

Traditional ML models like logistic regression, SVM, random forest, and XGBoost are very limited as they rely on hand-crafted (term frequency-inverse document frequency (TF-IDF) or n-grams) features that do not represent contextual semantics and long-range dependencies. Although the models are suitable for problems with explicit lexical indicators of sentiments, it struggles with implicit sentiment and Indonesian text features like negation a contextual polarity shifts that often presented in user-generated content. On the contrary, nowadays technologies like IndoBERT primarily utilize greatly deeper transformer architectures with hundreds of additional self-attention layers and aim to learn more hierarchical or structure-based language representation. IndoBERT architecture has 12 transformer layers, and each layer contains 12 heads to learn local and global contextual dependencies. This allows us to have better modelling on the syntactic structure, semantic nuance and word sense disambiguation compared with shallow or feature based classifiers. Hence, IndoBERT usually performs better than large models for such tasks as sentiment analysis.

This research had already shown in previous works that it can identify fake news with 98% accuracy [31] and classify disaster sentiment at 91% [32].

However, there is a trade-off linked to this speed gain, and it comes at an extra computation cost. Thus, it involves a lot more memory, training time and requires much more powerful hardware when compared with light-weight alternatives like DistilBERT (represent SLMs) or even other traditional ML models. DistilBERT, which maintains most of BERT's representational power while reducing inference latency and resource usage (it has 6 transformer layers and ~40% fewer parameters), sits well between the two extremes. This is why SLMs are most effective in constrained environments, where accuracy vs. computational economy must be balanced in their antecedent prediction.

Error analysis detects systematic misclassification trends in models that are not captured by quantitative measures. Traditional ML models fail to classify sarcastic comments, implicit sentiment, or emotional expressions that are intermixed as they operated on superficial lexical features. For example, statements made in roundabout or gentle phrases to express displeasure tend to be construed as impartial or sure. Sure, they perform better at these situations than LLMs, but they still struggle with code-mixed Indonesian English expressions, emerging startup languages and domain-specific slangs. In addition, vague sentiment and highly contextual references combined to influence residual classification errors across all model categories.

The result of these findings indicate that classical modelling is still valid in contexts with low data/computational burden or speed-requirements given its efficiency and interpretability, despite LLMs clearly capturing far richer linguistic events. Future work should move towards making LLM robust to low-resource and code-mixed languages, domain-adaptive pretraining and better interpretable LLM. Future work could aim at an even less misclassification of mixed sentiments such as through explainable artificial intelligence (XAI) methods and targeted data augmentation.

5. CONCLUSION

This study compares traditional ML models (logistic regression, SVM, random forest, and XGBoost) with SLMs (DistilBERT) and LLMs (BERT and IndoBERT) for sentiment classification of user input at eight Indonesian startups. It is also noticeable in the results that LLMs provide higher performance consistently compared to traditional models, with IndoBERT representing the highest accuracy (89.94%) and F1-score. That likely since IndoBERT pretrained on a large scale Indonesian text data that it can capture context and semantic property of Indonesian language better. These results have real-world applications. DistilBERT is a more compress model, find same basic principles for LLMs but do not by big computational needs, so many startups can provide an LLM based sentiment classifier in their customer feedback pipeline that automatically filter reviews marked as negative, identify trends over time and understand where to spend developer's resources doing product enhancements. Later, another variation was suggested however associated with drawbacks like domain adaptation issues or informal/code-mixed language and large model computational cost. This study also focused on future work directions, mentioning the need for multilingual and zero-shot sentiment models, domain-adapted pretraining, and XAI methods in sentiment analysis that you can find here! It will allow for much better utilization of sentiment analysis by startups and make it way more effective in utilizing the data to make decisions and gain an edge over competitors.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Li-Hua Li	✓	✓		✓	✓	✓			✓	✓		✓		
Agus Cahyo Nugroho	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**xperimentation

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This section not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

This section not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely on computational models, publicly available datasets, or non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or animal ethics committee was not necessary for this study.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

- [1] A. Patel, P. Oza, and S. Agrawal, "Sentiment analysis of customer feedback and reviews for airline services using language representation model," *Procedia Computer Science*, vol. 218, pp. 2459–2467, 2022, doi: 10.1016/j.procs.2023.01.221.
- [2] C.-H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data*, vol. 10, no. 1, May 2023, doi: 10.1186/s40537-023-00782-9.
- [3] E. M. Mercha and H. Benbrahim, "Machine learning and deep learning for sentiment analysis across languages: a survey," *Neurocomputing*, vol. 531, pp. 195–216, Apr. 2023, doi: 10.1016/j.neucom.2023.02.015.
- [4] N. C. Hellwig, J. Fehle, and C. Wolff, "Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings," *Expert Systems with Applications*, vol. 261, Feb. 2025, doi: 10.1016/j.eswa.2024.125514.
- [5] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: a survey," *Natural Language Processing Journal*, vol. 2, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.
- [6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
- [7] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: a systematic literature review," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 4, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.
- [8] V. L. Xavier, "Hyperbolic regression: a new regression model with applications to the binary classification problem," *Information Sciences*, vol. 609, pp. 15–25, Sep. 2022, doi: 10.1016/j.ins.2022.07.038.
- [9] C. G. Soh, Y. Zhu, and T. L. Toh, "A regularised logistic regression model with structured features for classification of geographical origin in olive oils," *Chemometrics and Intelligent Laboratory Systems*, vol. 237, Jun. 2023, doi: 10.1016/j.chemolab.2023.104819.
- [10] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF-IDF and count vectorizer," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, doi: 10.1007/s13278-024-01250-9.
- [11] R. R. Ehrman and J. A. Kline, "Primer on logistic regression for emergency care researchers," *The Journal of Emergency Medicine*, vol. 63, no. 5, pp. 683–691, Nov. 2022, doi: 10.1016/j.jemermed.2022.09.032.
- [12] H. Cam, A. V. Cam, U. Demirel, and S. Ahmed, "Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers," *Heliyon*, vol. 10, no. 1, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23784.
- [13] S. Wu and Y. Gao, "Machine learning approach to analyze the sentiment of airline passengers' Tweets," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2678, no. 2, pp. 48–56, Feb. 2024, doi: 10.1177/03611981231172948.
- [14] P. Berg, L. Buecher, B. Michele, M.-T. Pham, L. Chapel, and N. Courty, "Multi-prototype hyperbolic learning guided by class hierarchy," *International Journal of Computer Vision*, vol. 133, no. 11, pp. 7969–7984, Nov. 2025, doi: 10.1007/s11263-025-02571-8.
- [15] Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan, "How two-layer neural networks learn, one (giant) step at a time," *Journal of Machine Learning Research*, vol. 25, pp. 1–65, 2024.
- [16] M. M. R. Mamun, O. Sharif, and M. M. Hoque, "Classification of textual sentiment using ensemble technique," *SN Computer Science*, vol. 3, no. 1, Jan. 2022, doi: 10.1007/s42979-021-00922-z.
- [17] A. Munshi, M. Arvindhan, and K. Thirunavukkarasu, "Random forest application of twitter data sentiment analysis in online social network prediction," in *Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models*, Hoboken, United States: Wiley, 2021, pp. 299–313, doi: 10.1002/9781119792345.ch12.
- [18] C. K. Wang, "Sentiment analysis using support vector machines, neural networks, and random forests," in *2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Aug. 2023, pp. 23–34, doi: 10.2991/978-94-6463-300-9_4.

- [19] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Systems with Applications*, vol. 237, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.
- [20] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [21] N. Ramshankar and P. M. Joe Prathap, "Reviewer reliability and XGboost whale optimized sentiment analysis for online product recommendation," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 1, pp. 1547–1562, Jan. 2023, doi: 10.3233/JIFS-221633.
- [22] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: an updated survey," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4765–4800, May 2023, doi: 10.1007/s10462-022-10275-5.
- [23] J. Tian *et al.*, "Synergetic focal loss for imbalanced classification in federated XGBoost," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 647–660, Feb. 2024, doi: 10.1109/TAI.2023.3254519.
- [24] M. A. Mosa, "Optimizing text classification accuracy: a hybrid strategy incorporating enhanced NSGA-II and XGBoost techniques for feature selection," *Progress in Artificial Intelligence*, vol. 14, no. 2, pp. 275–299, Jun. 2025, doi: 10.1007/s13748-025-00365-0.
- [25] H. Naveed *et al.*, "A comprehensive overview of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, Oct. 2025, doi: 10.1145/3744746.
- [26] Y. Annepaka and P. Pakray, "Large language models: a survey of their development, capabilities, and applications," *Knowledge and Information Systems*, vol. 67, no. 3, pp. 2967–3022, Mar. 2025, doi: 10.1007/s10115-024-02310-4.
- [27] Y. Aliyu, A. Sarlan, K. U. Danyaro, A. S. B. A. Rahman, and M. Abdullahi, "Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources," *IEEE Access*, vol. 12, pp. 66883–66909, 2024, doi: 10.1109/ACCESS.2024.3398635.
- [28] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artificial Intelligence Review*, vol. 58, no. 6, Mar. 2025, doi: 10.1007/s10462-025-11162-5.
- [29] A. Al Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung, "Challenges and opportunities of text-based emotion detection: a survey," *IEEE Access*, vol. 12, pp. 18416–18450, 2024, doi: 10.1109/ACCESS.2024.3356357.
- [30] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," *Social Network Analysis and Mining*, vol. 15, no. 1, Mar. 2025, doi: 10.1007/s13278-025-01444-9.
- [31] M. Y. Ridho and E. Yulianti, "From text to truth: leveraging IndoBERT and machine learning models for hoax detection in Indonesian news," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 3, pp. 544–555, 2024, doi: 10.26555/jiteki.v10i3.29450.
- [32] F. Wang *et al.*, "A comprehensive survey of small language models in the era of large language models: techniques, enhancements, applications, collaboration with LLMs, and Trustworthiness," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, 2025, doi: 10.1145/3768165.
- [33] Q. Zhang, Z. Liu, and S. Pan, "The rise of small language models," *IEEE Intelligent Systems*, vol. 40, no. 1, pp. 30–37, Jan. 2025, doi: 10.1109/MIS.2024.3517792.
- [34] R. S. Gargees, "Scholarly article classification leveraging DistilBERT transformer and transfer learning," in *Intelligent Systems, Blockchain, and Communication Technologies-(ISBCom 2024)*, Cham, Switzerland: Springer, 2025, pp. 554–566, doi: 10.1007/978-3-031-82377-0_45.
- [35] C.-S. Lin, C.-N. Tsai, J.-S. Jwo, C.-H. Lee, and X. Wang, "Heterogeneous student knowledge distillation from BERT using a lightweight ensemble framework," *IEEE Access*, vol. 12, pp. 33079–33088, 2024, doi: 10.1109/ACCESS.2024.3372568.

BIOGRAPHIES OF AUTHORS



Li-Hua Li    is a distinguished researcher in the Department of Information Management at Chaoyang University of Technology, Taiwan. She holds a Ph.D. in Computer Science from the University of Alabama and specializes in artificial intelligence, recommendation systems, neural networks, and e-commerce. Her interdisciplinary research bridges technology and business, focusing on intelligent systems for practical applications such as mobile commerce and digital learning. An experienced educator, she teaches courses ranging from neural networks to advanced information systems, fostering both technical and communicative skills in her students. She can be contacted at email: lhli@gm.cyut.edu.tw.



Agus Cahyo Nugroho    is a lecturer in the Department of Information Systems at Soegijapranata Catholic University, Indonesia, and a Ph.D. candidate in Information Management at Chaoyang University of Technology, Taiwan. His research focuses on information systems, machine learning, deep learning, self-directed learning, and technology readiness among students. He can be contacted at email: agus.nugroho@unika.ac.id.