Journal of Engineering Science and Technology
Vol. 18, No. 5 (2023) 2281 - 2297
© School of Engineering, Taylor's University

A BERT MODEL TO DETECT PROVOCATIVE HOAX

RIO YUNANTO[1,*], ERI PRASETYO WIBOWO[2], R. RIANTO[3]

[1]Faculty of Computer Science, Universitas Komputer Indonesia, Indonesia
[2]Faculty of Postgraduate, Universitas Gunadarma, Indonesia
[3]Faculty of Information Technology, Universitas Teknologi Yogyakarta, Indonesia
*Corresponding Author: rio.yunanto@email.unikom.ac.id

Abstract

The information flood makes social media users vulnerable to becoming victims of provocative hoaxes or even spreading hoaxes themselves. This research examines the capabilities of two variants of Bidirectional Encoder Representations from Transformers (BERT) models for the Indonesian language (IndoBERT Base Model and Indonesian BERT base model 522M) in developing the detection of provocative hoaxes in the Indonesian language. The proposed method used two variants of the monolingual BERT model for the Indonesian language from the Huggingface library. The proposed method's architectural flow starts with data collection and labelling from community hoax collector websites, followed by pre-processing. The cleaned data is then divided into training and test data to proceed to the fine-tuning stage, where several layers and weights of the BERT model are adjusted to fit the desired classification task. The experimental results of the study show that the recommended Indonesian BERT variant for the detection of provocative hoaxes is the IndoBERT Base Model with a learning rate of 1e-5, a batch size of 32, and a maximum length limit of 128 tokens, achieving an average training accuracy of 99,22%, with a training time of 21min 52s. The research findings also indicate that a learning rate 1e-5 can produce better test accuracy than a learning rate of 2e-5 or 3e-5. The detection model of provocative hoaxes using Indonesian BERT variants needs to be improved, especially in terms of collecting a large amount of hoax data, to enhance the accuracy of the provocative hoax detection model.

Keywords: Accuracy, Classification, Hoax, Provocative, Transformer.

1. Introduction

Provocative hoaxes refer to intentional actions aimed at eliciting reactions from others using false narratives or misinformation, with various specific purposes, including creating panic, controversy, or attracting public attention [1]. Advances in information technology and the internet have driven societal behavioural changes, significantly impacting community social changes [2]. Communication through mobile technology has evolved from point-to-point to point-to-many and can even become viral quickly. The current condition has made us inundated with news and information through group chat applications and social media. Information overload has become a significant problem known as information pollution, social media fatigue, or communication overload [3]. The constant and poorly filtered flood of information makes social media users vulnerable to becoming victims of provocative hoaxes or even spreading hoaxes due to the confusion in distinguishing between hoaxes and truthful information.

Previous research findings have shown a significant correlation between education level, income, and the behaviour of spreading false news without prior verification. The lower the respondents' education level and income, the higher the likelihood of them spreading fake news [4]. The tendency to share news on social media without verification can be measured by the extent of a person's internet experience. The more experienced someone uses the internet, the higher their ability to search for, share, and verify information. In other words, someone who has just started using the internet is more prone to sharing information without checking it first [5]. There still needs to be a universally agreed-upon definition for fake news, even within journalism. A clear and accurate definition can help differentiate and analyse various types of fake news, necessitating the presentation of broad and narrow definitions for the various terms associated with fake news. Based on how these terms and concepts are defined, they can be distinguished from one another based on three main characteristics: 1) authenticity (containing non-factual or false statements), 2) intent (aimed at misleading or entertaining the public), and 3) whether the content is considered news or not, as shown in the categories of fake news in Fig. 1 [6]. This research examines the capabilities of two variants of Bidirectional Encoder Representations from Transformers (BERT) models for the Indonesian language (IndoBERT Base Model and Indonesian BERT base model 522M) in developing the detection of hoaxes in the Indonesian language.

The development of information technology has also transformed business activities in society. E-commerce users are familiar with mobile marketplace applications that offer various attractive products from multiple sellers within a single app. E-commerce is claimed to be one of the economy's most dynamic and essential sectors [7]. One impact of mobile marketplace applications is that users often feel disappointed with the app's services or the products they receive, leading them to write negative reviews on the Google Play Store. These negative reviews are often composed of sentences expressing disappointment or anger. User reviews can influence the opinions and decisions of other users when determining which products to purchase by reading reviews about those products [8].

The negative reviews in this research are labelled as non-hoax for several reasons that serve as arguments. Negative user reviews can potentially provoke other users, trigger emotions, or create dissatisfaction among other users. However, it is crucial to understand that not all users will be provoked by negative reviews, as some users may choose to form their own opinions based on their own objective experiences. Negative user reviews may stem from subjective user experiences, and although they can influence others' opinions, they are not classified as manipulative hoaxes or fake news. Negative reviews can be categorized as non-hoax because "hoax" refers to false, misleading information deliberately spread to manipulate people and create distrust or uncertainty. Categories of fake news are shown in Fig. 1.

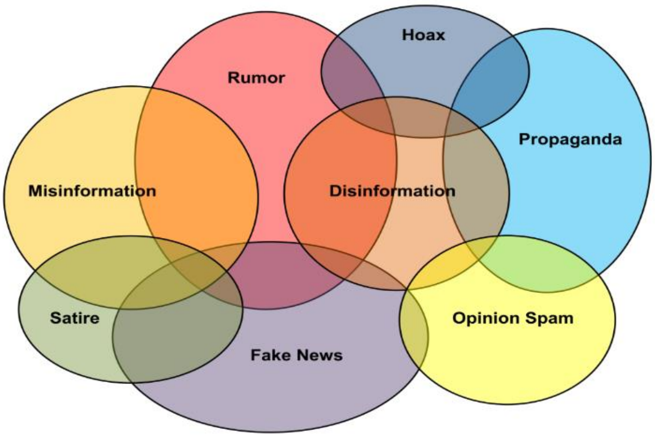


Fig. 1. Categories of fake news [6].

Provocative hoaxes can shape negative perceptions of the subjects targeted by the hoaxes or of specific groups. For example, spreading a provocative hoax accusing someone of engaging in immoral behaviour can lead to negative perceptions of the accused person. Negative reviews can also create perceptions of app service providers or product sellers, but they do not constitute legal violations or pose a danger to others. Therefore, provocative hoaxes and negative reviews are similar, making them interesting as training data for designing hoax detection systems.

This research differs from previous studies as the raw text data were derived from negative user reviews of mobile marketplace applications and combined with hoax texts from the turnbackhoax.id website. The study also uses the BERT (Bidirectional Encoder Representations from Transformers) model, composed of a transformer architecture for natural language processing tasks and supports the Indonesian language. Previous research on hoax detection used 251 textual data consisting of 151 non-hoax and 100 hoax data obtained from online news articles. These were classified using Multilayer Perceptron, Naïve Bayes, SVM, Random Forest, and Decision Tree algorithms [9]. Another study on hoaxes used comparison were conducted between two feature extraction techniques, namely Term Frequency and Term Frequency-Inverted Document Frequency, on the SVM, Linear SVM, KNN, Stochastic Gradient Descent (SGD), Decision Trees (DT), and Logistic Regression (LR) algorithms. This evaluation resulted in the highest accuracy score of 93.5%. Performance tuning was applied to SGD to find the best parameters for enhancing accuracy using Grid Search CV and Random Search CV [10]. Previous research focused on hoaxes or fact news about the coronavirus in Indonesia. The data comprised 535 SMS messages containing facts and 425 texts containing hoaxes. This data was classified using an SVM algorithm, which achieved the highest test accuracy score of 83.82% [11].

2. Literature Review

2.1. Hoax detection

Research related to hoax detection by researchers has shown significant progress in recent years. In the previous research, a web-based system to detect hoaxes and non-hoaxes in Indonesian-language news links has been developed [12]. They collected training and testing data from the *Forum Anti Fitnah Hasut dan Hoax* (FAFHH) archive site, performed TF-IDF weighting, and classified the data using SVM. In another research describing the frequent themes in fake news content, including religion, politics, health, ethnicity, and technology [13]. Term Document Matrix (TDM) weighting was applied to 74 hoax news and 74 genuine news articles, then modelled using the K-Nearest Neighbor (KNN) classification algorithm. Then another research, conducted a series of experiments in hoax detection using various models, including Deep Neural Network (DNN) architectures such as Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), Bidirectional GRU (Bi-GRU), and 1-Dimensional Convolutional Neural Network (1D-CNN) [14]. The results of the training and testing process showed that the DNN algorithm had a higher accuracy score than the others.

Since its introduction, the Transformer architecture, an advancement of DNN, introduced a self-attention mechanism that allows models to better understand the relationships between words in the text without relying on recurrence as in traditional DNN models. The transformer is the foundation for BERT (Bidirectional Encoder Representations from Transformers), developed by many researchers, including the Google AI team, to improve and develop BERT [15]. Research and development on BERT have been highly effective in Natural Language Processing (NLP) as it utilizes the transformer approach to generate contextual representations of words in the text. With this approach, BERT can understand the relationships between words in a sentence and consider the surrounding context to interpret the current word being processed [16]. This aids in capturing more accurate meanings within a broader sentence context. Researcher then utilized the BERT model to develop a hoax detection system for classifying Covid-19 themed hoaxes [17]. They revealed that the experimental results of the IndoBERT model trained on the Indonesian monolingual corpus had better accuracy than fine-tuned multilingual BERT.

This study is closely related to text classification using Indonesian BERT, so we have reviewed several relevant studies implementing BERT or similar transformer models. The work that has been done shows that a sophisticated pre-training model, IndoBERT, is used in classifying NLP news and making news recommendations based on categories. It was compared with other pre-training models such as XLNET, BERT multilingual, and XLMRoberta. To ensure the relevance and up-to-datedness of this research, we examined the dataset that has been used by researcher in developing hoax news classification, which consists of Indonesian news articles from two previous studies, labelled as either valid or hoax news, with a total of 1,100 data [18]. We also searched for recent research conducted in hoax classification using BERT [19], where automatic hoax news classification was performed using a multilingual transformer model combined with the BERTopic model as a topic distribution model. The proposed model has been proven to improve the performance of each multilingual model.

2.2. BERT model

BERT (Bidirectional Encoder Representations from Transformers) is one of the models in the Transformer family that possesses various capabilities in Natural Language Processing (NLP) and tasks such as word understanding, sentence understanding, and document understanding. BERT utilizes the Transformer architecture, which consists of multiple encoder layers to learn word representations. Transformers are considered remarkable in NLP because they leverage a powerful attention mechanism to process sequential data, such as words in a text. The attention mechanism allows the model to effectively capture relationships between distant words, enabling a better understanding of the context within the text [15]. Transformers can also process sequences in parallel, enabling fast text processing, as shown in Fig. 2. Transformers, like BERT, can learn rich language representations through pre-training and fine-tuning. The model can be customized to achieve excellent performance in various NLPtasks by fine-tuning specific labelled tasks [20].

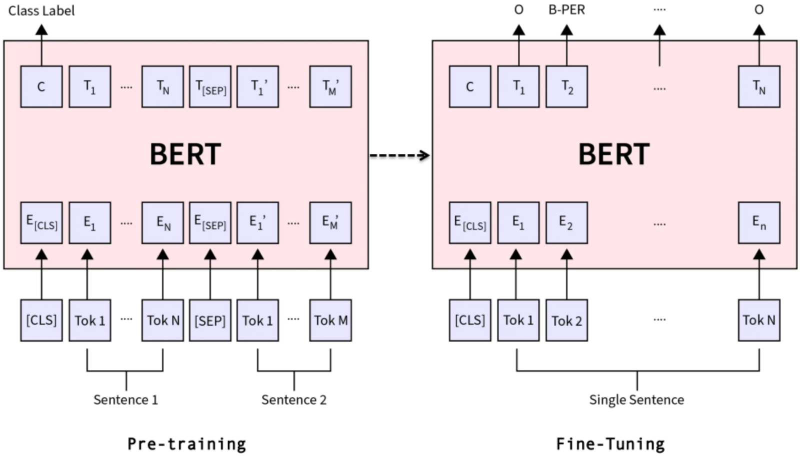


Fig. 2. Overall pre-training and fine-tuning for BERT.

This research utilizes two variants of BERT models specifically developed for the Indonesian language, namely IndoBERT Base Model ("indobenchmark/indobert-base-p1") and Indonesian BERT base model 522M ("cahya/bert-base-indonesian-522M"), which are the results of fine-tuning BERT on a large Indonesian dataset. Both BERT model variants have better understanding and processing capabilities for Indonesian texts, compatible with Indonesian grammar, vocabulary, and linguistic characteristics [21, 22]. These Indonesian BERT models were initially trained on a large dataset during pre-training. To meet the research requirements for developing hoax detection, the models were further trained on a relatively smaller dataset specific to the domain. This process of retraining the models is known as fine-tuning BERT. Therefore, we utilize these pre-trained models to develop provocative hoax detection.

3.Method

The proposed BERT model for hoax classification utilizes two types of monolingual BERT models for the Indonesian language: 1) IndoBERT Base Model (phase1 - uncased) and 2) Indonesian BERT base model (uncased), both from the Huggingface library, which also provides natural language processing resources and integrates AI language models for researchers, developers, and practitioners in the NLP field. The workflow and architecture of the proposed method can be seen in Fig. 3, where the collected hoax and non-hoax text, labelled accordingly, go through a text-cleaning process called pre-processing. There are sub-processes within pre-processing aimed at facilitating and improving the machine learning model's performance, including text normalization, tokenization, removal of common words (stop-words), and stemming [23]. The cleaned data is then divided into training and testing data to be passed into BERT classification, which generally consists of two main stages: pre-training and fine-tuning. In the pre-training stage, the BERT model is trained on tasks such as "masked language modelling" and "next sentence prediction" using many unlabelled text data; this helps BERT understand rich contextual representations of words. Then, in the fine-tuning stage, the BERT model is fine-tuned with classification data consisting of pairs of text and labels. In this process, some layers and weights of the BERT model are adjusted to fit the desired classification task [24].

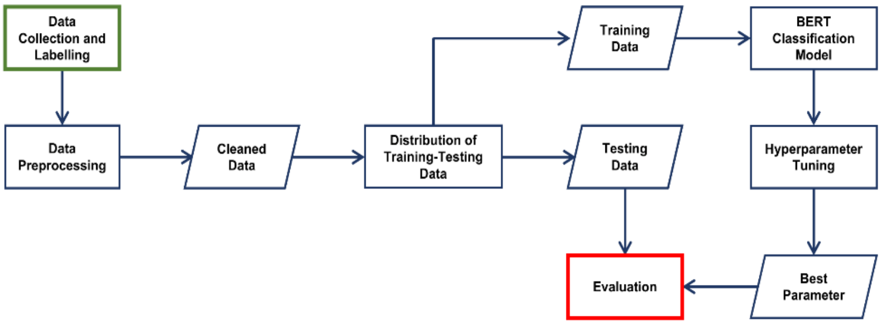


Fig. 3. Flowchart of proposed method and experiment.

3.1. Data collection

Text data collection is necessary to train and test the BERT model in the classification task. Identifying data sources is required to gather the text data. This research has two primary data sources: 1) hoax text data obtained from an Indonesian-language community website for collecting hoaxes [25] and 2) negative user reviews from Google Play Store. The collected hoax text data from the website turnbackhoax.id consists of Indonesian-language hoaxes from 2016 to 2021, totalling six thousand data. The hoax data will be used as a dataset, divided into training and testing data. Examples of hoax texts can be seen in Table 1.

The non-hoax text was collected from negative user reviews of mobile applications found in the comment section of the Google Play Store. We chose this data as non-hoax text because although some negative review texts may contain offensive and provocative language, they are non-hoaxes. Some mobile application users write complaints or comments about the application on the Google Play Store to help manage positive or negative user opinions [26]. Two mobile marketplace applications were sampled, and then user review data was selected that provided the lowest rating, 1 and 2. We needed six thousand non-hoax texts to balance the previously collected hoax data, and an example of non-hoax text can be seen in Table 2.

Table 1. Sample of hoax text.

No.	Text	Translation in English	Label
1	*Kejadian tadi pagi di riau ga boleh sholat berjamaah sungguh biadab aparat itu membabi buta entah apa yang sudah merasuk di alam pikir anak2 PKI*	There was an incident this morning in Riau. Congregational prayer was prohibited; what a barbaric act by the officers. It's unclear what has influenced the mindset of the PKI children.	1
2	*Waw heboh!!! Berita pagi ini: pernyataan terbaru panglima TNI: kami akan habisi semua pki di indonesia walau presiden jokowi melarang, Gatot: Tetap saya lakukan walau resiko dipecat!!!*	Wow, sensational!!! Morning news update: The latest statement from the commander of the Indonesian national armed forces: We will eradicate all PKI (communist party of Indonesia) members in Indonesia, even if President Jokowi opposes; Gatot: I will still proceed despite the risk of being fired!	1

Table 2. Sample of non-hoax text.

No.	Text	Translation in English	Label
1	*Sangat mengecewakan... Dapat voucher potongan harga untuk pengguna baru, ketika selesai pembayaran malah di batalkan begitu saja... Aplikasi ini banyak voucher menipu. gratis ongkir mesti belanja 50.000 dulu. Ga mau rugi banget tarik dana ke rekening saja butuh waktu yang lama banget.*	Very disappointing... Received a discount voucher for new users, but after payment was made, it was suddenly cancelled... This app has many deceptive vouchers. Free shipping requires a minimum purchase of 50,000. Don't want to lose too much; withdrawing funds to a bank account takes a very long time.	0
2	*Maaf Saya Kasih Bintang Satu... Sangat Kecewa, Proses Vocer Gratis Gagal, padahal pengguna baru. Aplikasi Lemot. Maaf saya Un Instal di HP saja... :) Oh ya untuk pengguna baru... Hati2 jangan Sampe Top Up saldo, karna kalau pesanan anda gagal kirim. Uangnya bakalan di balikin berupa saldo. Saldonya kalo di tarik ada batas minimumnya. Takutnya gak bisa d tf lagi ke rek kalian.*	I'm sorry, I give one star... Very disappointed, the process for the free voucher failed, even though I'm a new user. The app is slow. Sorry, I'm uninstalling it from my phone... :) Oh, and for new users... Be careful not to top up your balance because if your order fails to be delivered, the money will be refunded as a balance. There's a minimum limit for withdrawing the balance. You might not be able to transfer it back to your bank account.	0

3.2. Pre-processing

The collected hoax and non-hoax text are then concatenated to form a large dataset of twelve thousand entries. The dataset is then reformatted through the pre-processing stage to prepare the text for training or testing the classification model. Proper pre-processing can improve the quality and performance of the classification model by removing noise and preparing the input text as needed. A series of pre-processing steps are carried out, as shown in Fig. 4.

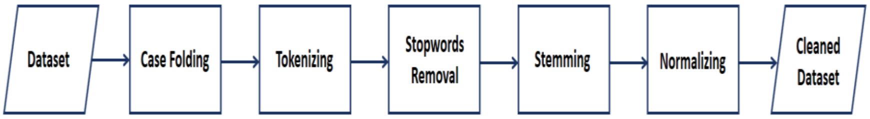


Fig. 4. Flowchart of pre-processing.

The pre-processing stage can begin with case folding, where all texts undergo letter alignment. This is done to facilitate searchability, as the available text data is currently inconsistent in using capital letters. Next is the tokenizing process, which involves dividing long texts into smaller parts called tokens. At the same time, tokenizing also removes certain characters considered punctuation marks. Stop-word removal eliminates words that do not contribute significantly to the text content, including conjunction words. Words that belong to the stop-word list are removed because they harm the subsequent processes. Following that is the stemming process, a technique in NLP to transform words into their base or root form, aiming to reduce word variations so that words with the same root are considered equivalent. Then comes the text normalization stage, a necessary process to reduce variations in the text.

The collected raw text often lacks structure. Texts collected from social media and mobile applications often use non-standard abbreviations or slang words known as "alay." This has been a challenge in previous research to convert slang words into formal word forms. Related research on non-formal sentences or slang words by comparing Indonesian language stemming techniques, including "Sastrawi," with a new technique known as "Incorbiz" has been conducted [27]. The research results showed that the "Incorbiz" technique is more effective in finding base words in non-formal sentences, and the "Sastrawi" and "Incorbiz" techniques can work well together to achieve better-stemming results.

3.3. Evaluation criteria

This study uses accuracy as a performance evaluation measure for all models. Accuracy is one of the metrics that are easy to understand and interpret. It is an evaluation metric measuring how well a classification model can correctly predict the overall test data. The confusion matrix, as shown in Fig. 5. is a table used to depict the performance of the classification model by comparing the model's predictions with the actual values in the test data. The confusion matrix consists of four main elements: True Positive (TP) represents the number of samples correctly predicted as positive by the model. True Negative (TN) represents the number of samples correctly predicted as negative by the model, False Positive (FP) represents the number of samples incorrectly predicted as positive by the model, and False Negative (FN) represents the number of samples incorrectly predicted as negative by the model. Using these elements, accuracy is calculated by dividing the sum of True Positive (TP) and True Negative (TN) by the total number of predictions, as shown in Fig. 5 and Eq. (1).

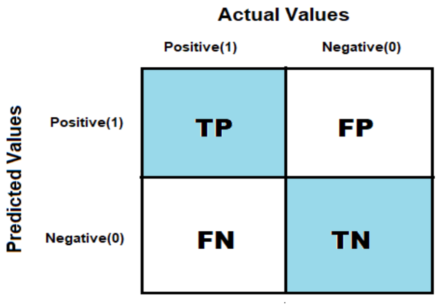


Fig. 5. Confusion matrix with the formulas of accuracy.

In many cases, accuracy indicates how well the model can perform classification correctly. Accuracy also has an intuitive interpretation that can depict the overall proportion of correct predictions in the dataset. It can demonstrate how well the model can classify samples correctly within the context of classification.

$$Accuracy \ = (TP + TN)/\,(TP + TN + FP + FN) \qquad (1)$$

4.Results and Discussion

The collected hoax data shows an increase in the number of hoaxes occurring in Indonesia in 2020 compared to previous years. This increase in hoaxes coincided with the onset of the Covid-19 pandemic in Indonesia. The peak of hoax dissemination in 2020 occurred in September, which coincided with the government's vaccination program during the Covid-19 pandemic, as shown in Fig. 6. This is supported by other studies that also depict the pandemic in 2020 as a momentum for the spread of hoaxes through the internet and chat applications. Indonesia has not been spared from hoax attacks carried out by irresponsible parties. Other researchers have conducted studies that measure hoax trends for six months, from January to June 2020, by analysing content from the website turnbackhoax.id, which verifies hoaxes circulating on social media [28].

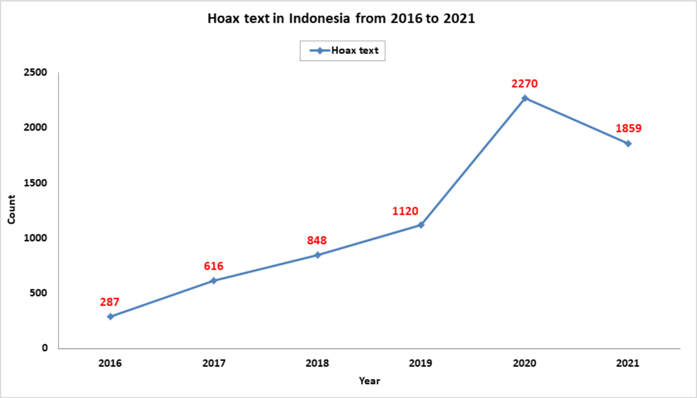


Fig. 6. Hoax text in Indonesia from 2016 to 2021.

The Covid-19 pandemic has spread worldwide and affected Indonesia, with a very large population. Reporting and analysing data on the government and public response to handling Covid-19 between January and March 2020 became the main focus of this submission. Social media has a positive potential for advocating economic solidarity. Community members created short videos that raise awareness of self-isolation, and even videos promoting hand washing and mask-wearing were made in regional languages (Javanese, Sundanese, and Sasak). However, it is common for social media to invade by dangerous hoaxes. The government quickly responded to the spread of hoaxes by developing a hoax buster (www.covid10.go.id/hoaks); at the same time, civil society, such as anti-hoax community groups, also developed anti-hoax measures such as turnbackhoax.id [29]. Hoaxes spreading during the pandemic peaked when hoaxes about vaccines started to emerge and spread. Research on vaccine-related hoaxes has been circulating since mid-2020, with various content claiming that vaccines contain harmful substances. Spreading hoaxes also claimed that the composition of the Covid-19 vaccine contained dangerous ingredients, including borax and formalin. There were even hoaxes stating that the vaccine was made from baby fetuses. Hoaxes about vaccines also claimed that the side effects of vaccination were death, infertility, and could even alter human DNA. Hoaxes stating that the Indonesian Medical Association (IDI) rejected the Covid-19 vaccination significantly impacted public trust in the government's vaccination program. Vaccine-related hoaxes circulated in Indonesian society from November 2020 to January 2021 [30].

Some characteristics of hoax texts and negative reviews can be considered to have emotional similarities that can be reflected in their respective raw texts, including 1) in perspective, both of them convey information from a subjective point of view or tend to be non-objective, 2) emotionally, both of them result in negative perceptions of the target party, 3) in terms of social impact, both of them raise the perception of distrust towards the intended party, 4) textually, both of them often use non-standard sentences, words, or slang. However, they also have differences, as shown in Table 3, so it is expected that the hoax detection model can distinguish between them well to improve the accuracy of hoax detection when implemented.

Table 3. Differences between a hoax and negative user reviews.

No.	Hoax	Negative User Reviews
1	Influence the reader to reject the information submitted by the authorities.	Customer disappointment opinions can influence other customers to be careful when buying/using the product.
2	Some spread hoaxes are written systematically and structured but use the wrong facts.	Often found written spontaneously and using facts from personal experiences.
3	Many hoaxes are made to attack or damage the good name of a particular person or institution, and they can even trigger mass unrest that is not controlled.	Not to damage the product's good name but to convey the facts that happened so that the product/service can be even better in the future.

The development of provocative hoax detection using the BERT model does not guarantee better results than manual detection (using human resources). However, with the high volume of hoaxes and their rapid propagation, manual methods are no longer effective in terms of cost, time, and scalability. BERT uses a transformer architecture that allows it to understand the context of the text better. The strong representational capabilities of BERT and its transfer learning ability enable it to be trained on specific tasks (such as hoax detection), leveraging the knowledge learned from large text datasets previously [31]. BERT also exhibits better flexibility and adaptability, making it customizable and expandable to meet different hoax detection needs. The model can be modified, retrained, or combined with other techniques to enhance detection performance.

This research aims to assess the capabilities of two BERT model variants for the Indonesian language (IndoBERT Base Model and Indonesian BERT base model 522M) in developing Indonesian hoax detection. This includes determining hyperparameters, parameters set before model training that remain unchanged during training. Proper hyperparameter selection can significantly impact the model's performance and convergence [32]. One aspect of our concern is the influence of the learning rate on both BERT model variants since the learning rate controls the magnitude of the optimizer's steps in updating the model weights based on the computed gradients during training. Training stability is also affected by the learning rate, which can cause significant fluctuations in loss and gradients, leading to instability and difficulties in convergence.

The development of hoax detection in this research requires a fine-tuning process to obtain the best model that suits the needs. However, before proceeding to the BERT model's fine-tuning process, we used traditional machine learning algorithms as a basis for comparison in previous research. We selected six machine learning classification algorithms: K-Nearest Neighbor (KNN), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) because these algorithms have been used for text classification in Indonesian language research [33, 34]. We also used CountVectorizer as a feature extraction method, showing good classification performance compared to other features [35, 36]. In the experiments using the same dataset as this hoax detection research, we encountered overfitting, where the machine learning model became too biased towards the training data and could not generalize well to new data. We suspect this occurred because the model memorized the training data well but failed to recognize general patterns in new data, as shown in Table 4.

Table 4. The result of using ML as a basis for comparison.

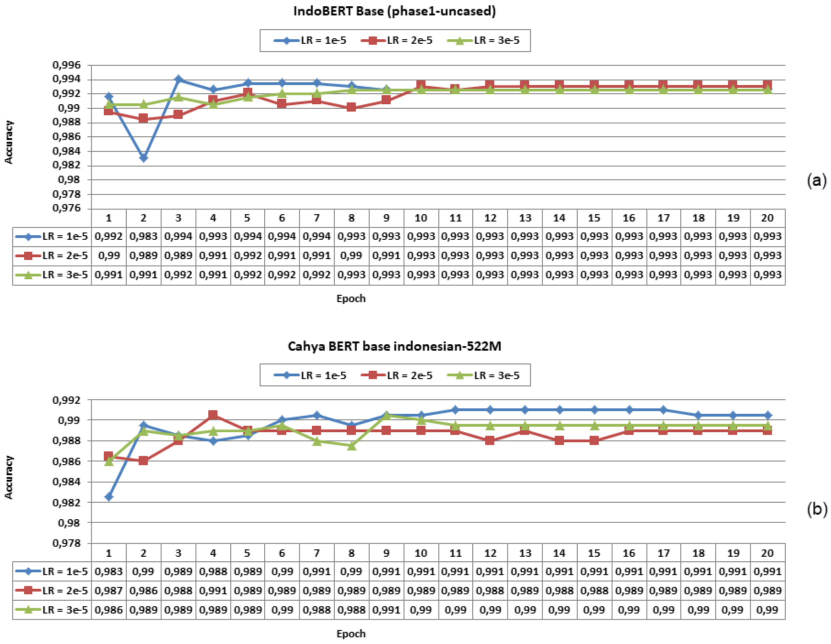
ML Model	Training Accuracy	F1-score	Test Accuracy
KNN	0.940	0.911	0.672
DT	0.945	0.927	0.777
LR	0.988	0.980	0.719
SVM	0.989	0.980	0.784
RF	0.987	0.982	0.778
NB	0.949	0.942	0.724

The training and testing of the hoax detection models using BERT required a fine-tuning process on both Indonesian BERT variants. Some hyperparameters used were a batch size of 32, a learning rate range of (1e-5, 2e-5, 3e-5), a maximum length limit of tokens set to 128, and 20 epochs [37]. The fine-tuning process was performed on a GPU: A100-SXM4-40GB MIG 3g.20 GB, with 1.0 TB RAM. A batch size that is too small can result in unstable gradients, while a batch size that is too large requires more GPU memory. For fine-tuning BERT, the common batch size is around 16 to 32. The learning rate determines the magnitude of the steps taken when updating the weights during training. The typical learning rate used for fine-tuning BERT is around 1e-5 to 5e-5. Sequence length refers to the maximum number of tokens allowed in a single input sequence. The commonly used sequence length ranges from 128 to 512 tokens. An epoch represents one complete pass through the entire training dataset. Fine-tuning BERT requires epochs ranging from 2 to 10, depending on the task complexity and dataset size [38].

The results of fine-tuning both Indonesian BERT variants for developing provocative hoax detection models achieved a training accuracy of 99,22% and the highest test accuracy of 88,7%, as shown in Table 5 and Fig. 7. There was a slight decrease in test accuracy compared to the training accuracy, suggesting some degree of overfitting. The best training time achieved was 21min 51s, and the longest training time required was 23min 39s. Although consistent training time can provide stability in model development, it does not directly indicate the quality of the resulting model. Longer training times have the potential to produce better models by allowing more opportunities to learn from the training data. However, increased training time only sometimes results in a significant improvement in model quality. The recommended Indonesian BERT model for hoax detection is the IndoBERT Base Model ("indobenchmark/indobert-base-p1") with a learning rate of 1e-5, a batch size of 32, and a maximum length limit of tokens set to 128.

Table 5. Fine-tuning result of BERT models on 20 epochs.

Pre-trained	LR	Avg. Training Accuracy	Avg. Training Loss	Training Time	Test Accuracy
IndoBERT Base (phase1-uncased)	1e-5	0.9922	0.0717	21min 52s	0.887
IndoBERT Base (phase1-uncased)	2e-5	0.9918	0.0688	22min 55s	0.873
IndoBERT Base (phase1-uncased)	3e-5	0.9921	0.0630	21min 51s	0.875
Cahya BERT base indonesian-522M	1e-5	0.9898	0.0819	23min 33s	0.874
Cahya BERT base indonesian-522M	2e-5	0.9886	0.0939	23min 35s	0.858
Cahya BERT base indonesian-522M	3e-5	0.9891	0.0829	23min 39s	0.861



Epoch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
LR = 1e-5	0,992	0,983	0,994	0,993	0,994	0,994	0,994	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993
LR = 2e-5	0,99	0,989	0,989	0,991	0,992	0,991	0,991	0,99	0,991	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993
LR = 3e-5	0,991	0,991	0,992	0,991	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993	0,993

Epoch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
LR = 1e-5	0,983	0,99	0,989	0,988	0,989	0,99	0,991	0,99	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991	0,991
LR = 2e-5	0,987	0,986	0,988	0,991	0,989	0,989	0,989	0,989	0,989	0,989	0,989	0,988	0,989	0,988	0,988	0,989	0,989	0,989	0,989	0,989
LR = 3e-5	0,986	0,989	0,989	0,989	0,989	0,99	0,988	0,988	0,991	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Fig. 7. Accuracy curves for BERT model.

The results of fine-tuning on both BERT variants, as shown in Fig. 7, exhibit fluctuations or fluctuations in accuracy values. This may be due to the model not fully converging or reaching an optimal point at the initial stage of training, as well as suboptimal text processing. Increasing the number of epochs may help improve the overall accuracy of the BERT model but could potentially lead to overfitting. Therefore, further research is needed to follow up on this.

BERT is one of the crucial components in developing the Large Language Model (LLM), where the use of the BERT architecture is one of the components in the construction. LLM has a wide potential to detect and address hoaxes more efficiently than manual detection. LLMs can process vastly larger amounts of information in less time than humans. LLMs can also understand context, recognize patterns, and identify linguistic features that may indicate hoaxes when trained with data spanning various texts to recognize patterns and characteristics that may be difficult for humans to recognize. An LLM can provide consistency in the hoax detection approach because emotional or personal prejudices do not influence it. However, the LLM requires extensive and representative training data. LLMs must be rigorously tested and evaluated using independent and diverse data sets to ensure their accuracy and reliability in detecting hoaxes. There is an expanding risk of possible bias in the training data or decisions generated by an LLM, so it needs to be combined with human efforts, such as professional verification of facts or curation of good news.

This study can be used as a reference for current issue in the social media and how to control hoax or incorrect news [39, 40] as well as its control on education performance [41-50].

5. Conclusions

The development of provocative hoax detection remains important because spreading fake news or hoaxes has significant and detrimental impacts. The rapid dissemination of provocative hoaxes through social media and online platforms causes panic, influences public opinion, and triggers conflicts or distrust in certain institutions. The provocative hoax detection model using Indonesian BERT variants needs to be improved, particularly in collecting many hoax data sources, to enhance the accuracy of hoax detection models. The IndoBERT Base Model ("indobenchmark/indobert-base-p1") performs exceptionally well with a learning rate 2e-5, achieving the highest training accuracy and the lowest average loss. It requires a training time almost comparable to other models, not the worst, but not the best. Several aspects should be noted for further research, including the impact of imbalanced datasets on accuracy results for each class and the effect of mixed-language test data. These aspects are worth exploring as they can affect the accuracy of the provocative hoax detection model when implemented.

Acknowledgement

The authors would like to thank the Ministry of Education and Culture of Indonesia for its research grant award through Penelitian Pascasarjana Penelitian Disertasi Doktor (PPS-PDD) 2022 grant, contract number 417/LL3/AK.04/2022, 11.4/LP/UG/VI/2022, on June 20, 2022. The authors would like to thank Universitas Gunadarma, which has supported this research by providing access to take advantage of the Nvidia DGX-1 and DGX A100 supercomputers. Any opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the sponsorship.

References

1. Wahyuni, N.; and Nguyet, D.M. (2022). Hoax and provocative content by Muslim cyber army (MCA) and its enforcement in Indonesia. *Indonesian Journal of Counter Terrorism and National Security*, 1(1), 67-90.
2. Bai, C.; Dallasega, P.; Orzes, G.; and Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International Journal of Production Economics*, 229(1), 1-15.
3. Kocabiyik, O. (2021). Social media usage experiences of young adults during the COVID-19 pandemic through social cognitive approach to uses and gratifications. *International Journal of Technology in Education and Science*, 5(3), 447-462.
4. Buchanan, T.; and Benson, V. (2019). Spreading disinformation on Facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"?. *Social Media Society*, 5(4), 1-9.
5. Kusuma, R. (2020). Implementation of counseling by Bhabinkamtibmas in preventing the spread of hoaxes in Kebumen police station. *Journal of Law and Legal Reform*, 1(3), 395-414.
6. Zhou, X.; and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1-40.
7. Kawa, A.; and Walesiak, M. (2019). Marketplace as a key actor in e-commerce value networks. *LogForum*, 15(4), 521-529.
8. Sohail, S.S.; Siddiqui, J.; and Ali, R. (2016). Feature extraction and analysis of online reviews for the recommendation of books using opinion mining technique. *Perspectives in Science*, 8(1), 754-756.
9. Putri, T.T.; Sitepu, I.Y.; Sihombing, M.; and Silvi, S. (2019). Analysis and detection of hoax contents in indonesian news based on machine learning. *Journal of Informatic Pelita Nusantara*, 4(1), 19-26.
10. Asha, J.; and Meenakowshalya, A. (2021). Fake news detection using n-gram analysis and machine learning algorithms. *Journal of Mobile Computing, Communications and Mobile Networks*, 8(1), 33-43.
11. Utomo, W.H.; and Prayoga, K.J. (2021). Hoax classification corona virus (Covid-19) news in Indonesian using the support vector machine (SVM) method. *Journal of Computer Science*, 17(8), 692-708.
12. Satyawati, N.P.; Utari, P.; and Hastjarjo, S. (2019). Fact checking of hoaxes by masyarakat antifitnah Indonesia. *International Journal of Multicultural and Multireligious Understanding*, 6(6), 159-172.
13. Dwivedi, Y.K.; Hughes, L.; Ismagilova, E.; Aarts, G.; Coombs, C.; Crick, T.; and Williams, M.D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57(1), 1-47.
14. Nayoga, B.P.; Adipradana, R.; Suryadi, R.; and Suhartono, D. (2021). Hoax analyzer for Indonesian news using deep learning models. *Procedia Computer Science*, 179(1), 704-712.

15. Ma, T.; Wang, W.; and Chen, Y. (2023). Attention is all you need: An interpretable transformer-based asset allocation approach. *International Review of Financial Analysis*, 90(1), 1-10.
16. Huang, Y.; Li, Z.; Deng, W.; Wang, G.; and Lin, Z. (2021). D-BERT: Incorporating dependency-based attention into BERT for relation extraction. *CAAI Transactions on Intelligence Technology*, 6(4), 417-425.
17. Suadaa, L.H.; Santoso, I.; and Panjaitan, A.T.B. (2021). Transfer learning of pre-trained transformers for Covid-19 hoax detection in Indonesian language. *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, 15(3), 317-326.
18. Juarto, B. (2023). Indonesian news classification using IndoBert. *International Journal of Intelligent Systems and Applications in Engineering*, 11(2), 454-460.
19. Hutama, L.B.; and Suhartono, D. (2022). Indonesian hoax news classification with multilingual transformer model and BERTopic. *Informatica*, 46(8), 81-90.
20. Rogers, A.; Kovaleva, O.; and Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8(1), 842-866.
21. Richardson, B.; and Wicaksana, A. (2022). Comparison of indobert-lite and roberta in text mining for Indonesian language question answering application. *International Journal of Innovative Computing, Information and Control*, 18(6), 1719-1734.
22. Jayadianti, H.; Kaswidjanti, W.; Utomo, A.T.; Saifullah, S.; Dwiyanto, F.A.; and Drezewski, R. (2022). Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN. *ILKOM Jurnal Ilmiah*, 14(3), 348-354.
23. Bagui, S.; Fang, X.; Kalaimannan, E.; Bagui, S.C.; and Sheehan, J. (2017). Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features. *Journal of Cyber Security Technology*, 1(2), 108-126.
24. Zaman, B.; Justitia, A.; Sani, K.N.; and Purwanti, E. (2020). An Indonesian hoax news detection system using reader feedback and naïve bayes algorithm. *Cybernetics and Information Technologies*, 20(1), 82-94.
25. Kesumawati, A.; and Thalib, A.K. (2018). Hoax classification with term frequency-inverse document frequency using non-linear SVM and naïve bayes. *International Journal of Advances in Soft Computing and its Applications*, 10(3), 115-128.
26. Saputri, Y.R.; and Februariyanti, H. (2022). Sentiment analysis on shopee e-commerce using the naïve bayes classifier algorithm. *Jurnal Mantik*, 6(2), 1349-1357.
27. Rianto, R.; Mutiara, A.B.; Wibowo, E.P.; and Santosa, P.I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, 8(1), 1-16.
28. Muzykant, V.L.; Muqsith, M.A.; Pratomo, R.R.; and Barabash, V. (2021). Fake news on COVID-19 in Indonesia. *Pandemic Communication and Resilience*, 6(1), 363-378.
29. Djalante, R.; Lassa, J.; Setiamarga, D.; Sudjatma, A.; Indrawan, M.; Haryanto, B.; and Warsilah, H. (2020). Review and analysis of current responses to

COVID-19 in Indonesia: Period of January to March 2020. *Progress in Disaster Science*, 6(1), 1-9.

30. Rahayu, R.N. (2021). Vaksin COVID-19 di Indonesia: Analisis berita hoax. *Jurnal Ekonomi, Sosial and Humaniora*, 2(7), 39-49.
31. Prottasha, N.J.; Sami, A.A.; Kowsher, M.; Murad, S.A.; Bairagi, A.K.; Masud, M.; and Baz, M. (2022). Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors*, 22(11), 1-19.
32. Sai, A.B.; Mohankumar, A.K.; Arora, S.; and Khapra, M.M. (2020). Improving dialog evaluation with a multi-reference adversarial dataset and large scale pre-training. *Transactions of the Association for Computational Linguistics*, 8(1), 810-827.
33. Setifani, N.A.; Fitriana, D.N.; and Yusuf, A. (2020). Perbandingan algoritma naïve bayes, SVM, dan decision tree untuk klasifikasi sms spam. *Jurnal Sistem Informasi Musirawas* (*JUSIM)*, 5(2), 167-174.
34. Alita, D.; and Isnain, A.R. (2020). Pendeteksian sarkasme pada proses analisis sentimen menggunakan random forest classifier. *Jurnal Komputasi*, 8(2), 50-58.
35. Priyambodo, A.; and Prihati, P. (2020). Evaluasi ekstraksi fitur klasifikasi teks untuk peningkatan akurasi klasifikasi menggunakan naive bayes. *Elkom: Jurnal Elektronika dan Komputer*, 13(1), 159-175.
36. Negara, A.B.P.; Muhardi, H.; and Sajid, F. (2021). Perbandingan algoritma klasifikasi terhadap emosi tweet berbahasa indonesia. *Jurnal Edukasi dan Penelitian Informatika* (*JEPIN)*, 7(2), 242-249.
37. Rodrawangpai, B.; and Daungjaiboon, W. (2022). Improving text classification with transformers and layer normalization. *Machine Learning with Applications*, 10(1), 1-9.
38. Widianto, M.H.; and Cornelius, Y. (2023). Sentiment analysis towards cryptocurrency and NFT in Bahasa Indonesia for twitter large amount data using BERT. *International Journal of Intelligent Systems and Applications in Engineering*, 11(1), 303-309.
39. Gunawan, B.; Ratmono, B.M.; Abdullah, A.G.; Sadida, N.; and Kaprisma, H. (2022). Research mapping in the use of technology for fake news detection: Bibliometric analysis from 2011 to 2021. *Indonesian Journal of Science and Technology*, 7(3), 471-496.
40. Pablo, A.M.M.; Corpuz, J.L.G.; Deypalan, P.C.; Musa, H.Z.; and Asoy, V.C. (2022). 21st century watchdogs: The credibility of news media outlets in the Philippines. *ASEAN Journal of Community Service and Education*, 1(2), 95-102.
41. Hashim, S.; Masek, A.; Abdullah, N.S.; Paimin, A.N.; and Muda, W.H.N.W. (2020). Students' intention to share information via social media: A case study of covid-19 pandemic. *Indonesian Journal of Science and Technology*, 5(2), 236-245.
42. Haristiani, N.; and Rifa'i, M.M. (2020). Combining chatbot and social media: Enhancing personal learning environment (PLE) in language learning. *Indonesian Journal of Science and Technology*, 5(3), 487-506.
43. Mojaveri, H.S.; Daftaribesheli, M.; and Allahbakhsh, A. (2016). The relationship between social performance and corporate financial performance. *Indonesian Journal of Science and Technology*, 1(2), 216-231.

44. Bedua, A.B.S.V.; Bengan, C.V.B.; Ea, E.P.; Goleng, D.J.G.; Posanso, D.G.D.; Pueblo, C.T.; and Abusama.H.P. (2021). Social media on the students' academic performance. *Indonesian Journal of Educational Research and Technology,* 1(2), 41-44.
45. Abubakar, B.D.; Kayode, F.E.; Abiodun, M.H.; Samson, A.B.; and Abdulrasaq, A. (2022). Social media efficacy on prevention and control of covid-19 pandemic in Ilorin south local government area, Kwara state. *Indonesian Journal of Educational Research and Technology*, 2(3), 195-204.
46. Suroto, S.; and Nandiyanto, A.B.D. (2021). The effectiveness of using whatsapp social media as learning media at elementary school. *Indonesian Journal of Multidiciplinary Research*, 1(1), 79-84.
47. Ramdhani, T.; and Nandiyanto, A.B.D. (2021). The use of whatsapp social media as reinforcement online learning during the Covid-19 pandemic. *Indonesian Journal of Multidiciplinary Research,* 1(1), 107-112.
48. Sopian, A.; Nandiyanto, A.B.D.; Kurniawan, T.; and Bilad, M.R. (2022). The influence use of social media on the learning motivation of junior high school students. *Indonesian Journal of Multidiciplinary Research*, 2*(*1), 137-142.
49. Prabowo, T.T.; and Suroso, D.J. (2022). Indonesian public response to online learnings during the covid-19 pandemic: An analysis of social media. *Indonesian Journal of Teaching in Science*, 2(2), 193-206.
50. Aladesusi, G.A.; Issa, A.I.; Abodunrin, S.O.; Boris, O.A.; Babalola, E.O.; and Nuhu, K.M. (2021). Perception of undergraduate students on the utilization of social media to enhance learning in University of Ilorin. *ASEAN Journal of Science and Engineering Education*, 2(1), 183-192.