Jurnal Infotel
Vol. 16, No. 4, November 2024, pp. 713–728.
doi:10.20895/infotel.v16i3.1206

Research Article

A Systematic Literature Review of BERT-based Models for Natural Language Processing Tasks

Agung Fatwanto[1,*], Fardan Zamakhsyari[2], Rebbecah Ndungi[3], and Norma Latif Fitriyani[4]

[1,2]Informatics Department, UIN Sunan Kalijaga Yogyakarta, Sleman 55281, Indonesia
[3]Faculty of Mathematics and Computer Science, St.Petersburg State University, St.Petersburg, Russia
[4]Department of Artificial Intelligence and Data Science, Sejong University, Seoul, Republic of Korea

*Corresponding email: agung.fatwanto@uin-suka.ac.id

Received: July 11, 2024; Revised: November 19, 2024; Accepted: December 02, 2024.

Abstract: Research area in natural language processing (NLP) domain has made major advances in recent years. The Bidirectional Encoder Representations from Transformers (BERT) and its derivative models have been at the vanguard, gaining notice for their exceptional performance across a variety of NLP applications. As a response to this context, hence, this study aims to conduct a systematic literature review on current research in BERT-based models in order to describe their characteristic variations on three frequently demanded natural language processing (NLP) tasks, i.e. text classification, question answering, and text summarization. This study employed a systematic literature review method as prescribed by Kitchenham. We collected 4,120 papers from publications indexed by Scopus and Google Scholar from which 41 complied with our defined review criteria and finally chosen for further analysis. Our review came up with three conclusions. First, in order to select appropriate models for particular NLP tasks, three primary concerns should be considered: i) the type of NLP problem to be resolved (i.e. NLP task to be served), ii) the specific domain to be handled (such as financial, medical, law/legal or others), and iii) the intended language to be applied (such as English or others). Second, learning rate, batch size, and the type of optimizer were the three most considered hyperparameters to be properly arranged in model training. Third, the most widely used metrics for text classification tasks were F1-score, accuracy, precision, and sensitivity (recall), while question answering, and text summarization tasks were mostly used the Exact Match and ROUGE respectively.

Keywords: bert, nlp, nlp task, systematic literature review, transformer

ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

The area of natural language processing (NLP) has grown in importance within the artificial intelligence (AI) and machine learning areas. Many natural language processing (NLP) tasks, namely text classification (such as sentiment analysis), machine translation, question-answering, text summarization, information extraction, instruction following, machine reading comprehension, and image captioning can currently be handled by deep learning algorithms [1]. Among these various tasks, the demand for text classification, question answering, and text summarization is steadily increasing since their application for particular context requires customized implementation for specific type of usage. It is due to the fact that these three types of NLP tasks, compared to machine translation for example, are apparently context-specific whereas machine translation is more generic in nature (meaning that one can utilize a readily available machine translation tool without having to perform more customization for their specific tasks at hand).

One of the most notable advances on NLP research in recent years was the introduction of the Bidirectional Encoder Representations from Transformers (BERT) model proposed by Google AI Language researchers [2]. Using a transformer architecture, BERT is a pre-trained language model [3] that extracts word contextual representations from vast volumes of unlabeled text data. Through the use of novel training techniques including Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), BERT has demonstrated outstanding outcome on a variety of NLP tasks, assigning a new benchmark in excellence [2]. The success of BERT sparked research interest in the research community, resulting in the development of many versions and model-specific adaptations of BERT [4]. Along with the Generative Pre-trained Transformer (GPT), BERT has become the primary foundational model option among other Transformer-based models to be fine-tuned for various NLP tasks [5]. Its bidirectional framework makes it become context-aware at the word-level (compared to the GPT which only has awareness at token-level). This word-level awareness makes BERT ideal for natural language understanding (NLU) tasks, even though it is not as powerful as GPT for natural language generation (NLG) tasks. Another important thing that makes BERT gain popularity is that, compared to GPT for example, it is categorized as open source hence it is free to be utilized and developed further. Considering these aspects, we therefore decided to focus our literature review on BERT-based models instead of a more general Transformer-based model. Although BERT has been widely studied and applied in various NLP tasks, there is still a need to synthesize and critically analyze existing literature to gain a comprehensive understanding of its performance, limitations, and potential areas for improvement.

BERT's ability to capture contextual information and its impressive performance on a wide range of NLP tasks have made it a game-changer in the natural language processing sector. However, despite its success, BERT is not without limitations. One of the main challenges associated with BERT is its computational complexity and resource requirements, which can make it difficult to deploy in resource-constrained environments or on edge devices. Additionally, like many other language models, BERT can exhibit biases and inconsistencies, which can be problematic in real-world applications [6]. A systematic literature review can provide new insights into the current state of research, identify research gaps, and inform future research directions [7].

This systematic literature review (SLR) intends to structurally analyze current research on BERT-based models in order to describe their characteristic variations on three frequently demanded natural language processing (NLP) tasks, i.e. text classification, question answering, and text summarization. These three NLP tasks represent both "discriminative" (for text classification) and "generative" (for question answering and text summarization) AI. By scrutinizing and combining findings from relevant research studies, this review aspires to present a clear comprehension of the proper implementations of BERT-based models in the field of the three most demanded NLP tasks.

2 Research Method

In recent years, systematic literature review (SLR) has been acclaimed and widely accepted as a type of qualified research by scientific community globally and can easily be found in many respected journal publications [8]. As a type of meta-research, SLR is regarded as a qualified research genre and has become one of the trustworthy sources of knowledge. In regard to this context, hence, this work was conducted as an SLR study employing the framework as suggested by Kitchenham [9]. The purpose of systematic literature reviews (also known as secondary studies) is to search, collect, select (evaluate), analyze, synthesize, extract, and summarize relevant research results on specific and focused topics, hence it is classified as a tertiary review. The stages in conducting a systematic review following the Kitchenham method consist of: Planning, Conducting, and Reporting [9], [10]. For this particular study's stages, the method is depicted in Figure 1.

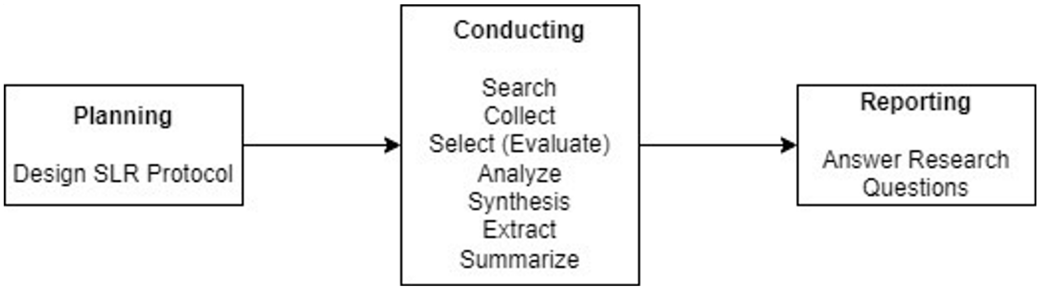

Figure 1: Systematic literature review process.

Following the Kitchenham framework, there are three stages to conduct a SLR study, as shown in Figure 1. The first stage is to plan the systematic review process, in which we designed a review protocol. One of the main elements specified in the protocol is the proposed research questions. These research questions are going to be the platform in conducting the study. The second stage is to conduct the systematic review process, which involves searching, collecting, selecting (evaluating), analyzing, synthesizing, extracting and summarizing the papers according to the defined criteria in order to answer the specified research questions. Finally, the third stage is to report the systematic review, in which the whole processes of the study are reported including the answers and explanations to the proposed research questions.

2.1 Planning

At this stage, a review protocol to conduct the study is designed with the purpose of having a systematic and structured literature review. The protocol consists of research questions, searching tools, searching terms, inclusion, exclusion, and quality assessment criteria. The protocol also establishes a data analysis and synthesis strategy as a guidance in extracting and summarizing the collected papers. The protocol is described in Table 1.

Table 1: Systematic literature review protocol

SLR Protocol	Description
Research Question	1. RQ1. What type of concerns were normally considered when selecting BERT-based model variants for particular NLP tasks? 2. RQ2. What type of hyperparameters were mostly considered to be properly arranged in training the BERT-based models? 3. RQ3. What type of evaluation metrics were normally employed for particular NLP tasks?
Search Tool	Publish or Perish Application.
Search Terms	"Text Classification"; "Question Answer"; "Text Summarization".
Inclusion Criteria	1. The papers are from scientific journals published between 2019-2024. 2. The publications are indexed by Scopus and/or Google Scholar. 3. The publications are not in the form of journals and proceedings. 4. The paper's theme is regarding NLP tasks which employ BERT-based models. 5. The paper is written in English.
Exclusion Criteria	1. Unpublished scientific papers between 2019-2024. 2. The publications are not indexed by Scopus and/or Google Scholar. 3. The publications are not in the form of journals and proceedings. 4. The paper's theme is not regarding NLP tasks which employ BERT-based models. 5. The paper is not written in English.
Quality Assessment Criteria	1. Clarity of research objectives 2. Contains literature review, background, and research results. 3. Provides relevant conclusions. 4. Describes the method of developing or optimizing and evaluating the BERT-based model.
Data Analysis Strategy	Based on the chosen papers, each variable that deemed suitable and relevant to answer the research questions, will be identified, gathered, analyzed, and scrutinized based on the quality of the study.
Data Synthesis Strategy	A data-driven methodology, founded on the data analysis outcomes of the papers, is used in data synthesis. To address the research questions, a table containing comparison list of each variable was composed for every chosen paper using a data-driven approach.

2.2 Conducting

A series of activities to search, collect, select (evaluate), analyze, synthesize, extract, and summarize papers from the publicly available publications following the defined review protocol was performed during this phase. Firstly, papers were searched based on the defined keywords (i.e. text classification, question answer, and text summarization) and the first three inclusion criteria (i.e. the publication range were between 2019-2024, the papers were indexed by Scopus and/or Google Scholar, and the papers were published either in journals or proceedings) from the publicly available publications. Secondly, the collected papers were selected (evaluated) based on the fourth inclusion criteria (i.e. to inspect whether the papers' theme is regarding NLP tasks which employ BERT-based models). Thirdly, the selected papers were further examined to identify the content that was deemed suitable and relevant to answer the proposed research questions based on the quality assesment criteria. Then, variables which deemed suitable and relevant to answer the proposed research questions were analyzed from the chosen papers by considering the quality of the study. Finally, these analyzed variables were then synthesized in which a data-driven approach was employed by composing a table containing comparison list of each variable for every chosen papers.

3 Results

This research was conducted following the defined SLR protocol as described in Table 1. We proposed three research questions to be studied through this review. In order to align with these three specified research questions, we employed three keywords to query the publicly available papers. We also defined inclusion and exclusion criteria to limit the collected papers. A quality assessment criterion was also determined to evaluate and filter out the collected papers so that the final selected papers are clear, credible and relevant to answer the specified research questions. In addition, data analysis and data synthesis strategies were also defined to guide the extraction and summarization of data and information from the final selected papers as the basis to answer the proposed research questions. Following the defined SLR protocol, we performed the process of searching, collecting, selecting (evaluating), analyzing, synthesizing, extracting, and summarizing the papers from publicly available publications. The number of processed papers is described in Table 2.

Table 2: Number of processed papers

Keywords	**Source**	**#of Gathered Papers from Initial Search Action**	**#of Papers with Related Theme within Title & Abstract**	**#of Qualified & Relevant Papers**
Text Classification	Scopus	536	127	**10**
	Google Scholar	882	113	**6**
Question Answer	Scopus	278	117	**9**
	Google Scholar	858	105	**5**
Text Summarization	Scopus	756	118	**8**
	Google Scholar	810	92	**3**
Total		**4,120**	**672**	**41**

Using the Publish or Perish application (https://harzing.com/resources/publish-or-perish), we searched the papers from the publicly available publications by applying each specified keywords (i.e. text classification, question answer, and text summarization) and the first three inclusion criteria (i.e. the publication range were between 2019-2024, the papers were indexed by Scopus and/or Google Scholar, and the papers were published either in journals or proceedings). This action resulted in 4,120 collected papers, in which 1,418 papers found for the keywords of "Text Classification", 1,136 papers found for the keyword of "Question Answer", and followed by 1,566 papers found for the keyword of "Text Summarization".

Based on those collected papers, we then applied the fourth and fifth inclusion criteria which is to select (evaluate) whether the papers' themes are related to NLP tasks which employ BERT-based models and wether they are written in English. We performed this selection (evaluation) process manually by scrutinizing each collected papers. We select (evaluate) each collected papers by identifying their themes which is either described within the title or abstract. This selection (evaluation) process was done through all 4,120 collected papers resulting in 672 papers that deemed meet the inclusion criteria, in which 240 papers were from the "Text Classification" category (previously we found 241 papers but there was one paper written in Bahasa Indonesia hence we excluded it), 222 papers were from the "Question Answer" category, and followed by 210 papers that were from the "Text Summarization" category.

At the final phase of papers selection (evaluation), a thorough review was carried out of all 672 selected papers from previous action. In this phase, each papers were deeply scrutinized to identify the content that deemed suitable and relevant to answer the proposed research questions. During this phase, we scrutinized each selected papers according to the quality assesment criteria. We especially put more efforts to identify and locate any papers which contain data, description, and explanation regarding the method of developing or optimizing BERT-based models for three frequently demanded NLP tasks (i.e. text classification, question answering, and text summarization). It included the identification of any type of BERT-based model variants and implementation of certain hyperparameters arrangement. In addition, we also identified and located any description and explanation about evaluation metrics for particular NLP tasks. A total of 41 papers that met these quality assesment criteria were finally chosen.

The next step was analyzing data and information which deemed suitable and relevant to answer the proposed research questions from these 41 chosen papers. We refered to the specified data analysis strategy as defined in the review protocol in which each variable that deemed suitable and relevant to answer the research questions was identified, gathered, analyzed, and scrutinized based on the quality of the study. The variables that were analyzed were the deployed variants of BERT-based models for three frequently demanded NLP tasks (i.e. text classification, question answering, and text summarization), the implementation of certain hyperparameters arrangement in BERT-based models development or optimization, the deployed variants of BERT-based models for three frequently demanded NLP tasks (i.e. text classification, question answering, and text summarization), and also the employed evaluation metrics for particular NLP tasks.

These analyzed data and information were then synthesized following the specified data synthesis strategy as defined in the review protocol in which a data-driven approach was employed by composing a table containing comparison list of each variable for every chosen papers. Based on this comparison list, we then categorized the data and information according to the specified research questions. These categorized data and information were then extracted and summarized as the basis of answering the proposed research questions.

4 Discussion

The aim of this study is to conduct a systematic literature review on current research in BERT-based models in order to describe their characteristic variations on three frequently demanded natural language processing (NLP) tasks, i.e. text classification, question answering, and text summarization. As has been designed in our SLR protocol, we try to answer three research questions regarding this context. This discussion of this section is hence framed based on these three questions.

RQ 1. What type of concerns were normally considered when selecting BERT-based model variants for particular NLP tasks?

The BERT-based models are commonly used for various types of NLP tasks. Some of the examples are text classification (such as sentiment analysis and information extraction like named entity recognition), machine translation, question-answering, text summarization, information extraction, instruction following, machine reading comprehension, and image captioning. The BERT-based models have also been employed on domain specific applications such as finance, law/legal, medical, etc. Some language-specific models other than English have also been developed and applied. These models are starting to outperform humans on tasks previously thought to be unsolvable by AI, such as question answering and verbal lie detection [11].

This research, however, focuses on the exploration of the BERT-based models for the three frequently demanded NLP tasks, i.e. text classification, question-answering, and text summarization. The first task represents the "discriminative" AI while the last two tasks exemplify the "generative" AI. The exploration was based on the chosen papers that have been examined carefully according to the review protocol. The result of the exploration is described in Table 3.

Table 3: The variants of BERT-based models for particular NLP tasks

Task	Variant	Paper
Text Classification	BERT	[12], [13], [14], [15], [16], [17], [11], [18]
	M-BERT	[19], [20], [21], [22], [23], [24]
	IndoBERT	[25], [19], [26], [27], [14], [23], [28]
	RoBERTa	[12], [19]
Question Answer	BERT	[29], [30], [31]
	M-BERT	[32], [21], [33], [34], [35], [36]
	IndoBERT	[4], [32], [37], [34], [38], [39]
	RoBERTa	[4], [39]
Text Summarization	BERT	[40], [41], [42], [43]
	M-BERT	[44], [45], [46], [47], [48], [49]
	IndoBERT	[44], [50]

Most BERT-based models were developed or fine-tuned with a specific context. Few of the models were even designed for a specific type of NLP task since their inception. For most of the models, their pre-training or especially their subsequent fine-tuning processes employed specific types of datasets (corpus) which obviously will affect the suitability of the produced models. Choosing the appropriate models prior to applying them for specific NLP tasks is therefore critical for achieving optimum performance as expected.

In order to select appropriate models for particular NLP tasks, three primary concerns should be considered: i) the type of NLP problem to be resolved (i.e. NLP task to be served), ii) the specific domain to be handled (such as financial, medical, law/legal or others), and iii) the intended language to be applied (such as English or others). For selecting what type of models should be chosen for a particular NLP task, Table 3 has presented the list of candidates deemed suitable to be applied for text classification, question answering, and text summarization tasks. Text classification apparently has been attracting a lot of attention from researchers as reflected in the number of related publications. It might be due to the fact that research and implementation of text classification have been around longer than question answering and text summarization. Another explanation is that the demand for text classification tasks is relatively high. For highly specific NLP tasks like sentiment analysis or named entity recognition (as part of text classification), task-specific models have shown to outperform generic models hence the demand is increasing. It is therefore more models were developed and supplied for this kind of task. Nevertheless, the demand for question answering tasks has gaining more attention recently. More research and models' constructions specifically for question answering are currently on the way. More domain and language specific models are now even easier to obtain. In a quite different condition, the studies on text summarization do not gain as high attention as text question answering and especially text classification. It is reflected in the number of publications as presented in Table 3.

Some works and studies on domain specific BERT-based models have been conducted. A number of domain specific models, such as for the legal/law domain [48], education [25], finance [18], health especially for biomedical [12], religion [34], customer service for general business [31] and university students' activity [39], and also beauty [37] and general products review [38] are some of the examples. These models were developed on specific corpus in order to cater to a particular implementation domain.

In the context of language consideration, BERT or RoBERTa become a better alternative for English-specific tasks because both are trained on vast volumes of English corpus and perform well on several English-based NLP tasks. M-BERT, on the other hand, is a better option for "low resources" languages-specific tasks in which the local language-specific models are obviously rare. Based on the studies as have been reported in [27], [46], [40], [50], [13], [47], [32], [45], and [23], M-BERT model is the currently available answer to the cross-language problem caused by the original-based BERT models' lack of multilanguage capabilities. Many works and studies on non-English language-specific models have been conducted and still gaining more attention to the researchers recently. Models such as HuBERT for agglutinative language (in this case is Hungarian and Turkish) [43], BERTTurk for Turkish [49], SwedishBERT for Swedish [33], EstBERT for Estonian [46], ItaBERT for Italian [48], and IndoBERT for Bahasa Indonesia [4], [44] are some of the examples. These models are a preferable option compared to BERT or M-BERT for NLP tasks for non-English specific languages. This is because they were specifically trained in those languages-specific corpuses, so it can better capture more nuances and context of that particular type of languages. These languages-specific models have shown to provide more powerful performance when applied on language-specific NLP tasks. For the case of IndoBERT, as an example, it is evidenced by some comparative studies as reported in [24], [44], [49], [39] and [19]. In these studies, IndoBERT was compared with other models for a number of NLP tasks utilizing Indonesian language datasets, and the results show that IndoBERT had outperformed other models.

Aside from type of NLP tasks, working domain, and linguistic considerations, other important attention also needs to be taken into account when choosing the appropriate models for the project at hand, such as model size, available computer resources, and training data availability. Larger models, such as RoBERTa-large, tend to perform better but demand more computational resources. If computational resources are restricted, smaller models like BERT-base may be a better option. The availability of training data (corpus) in a specific language may also be a factor to consider when planning to develop or optimize language-specific models for particular NLP tasks in that language.

Finding the proper model frequently entails conducting empirical experiments and evaluation of a number of candidate models to decide which one performs best. To make an informed decision, it is important to comprehend the available collection of models, particular NLP tasks to handle, and the availability of training materials (corpus) in advance.

RQ 2. What type of hyperparameters were mostly considered to be properly arranged in training the BERT-based models?

BERT is a recent modification of a series of neural models that makes extensive use of pre-training and has shown useful in a wide range of NLP tasks, including text classification and question-answering from datasets. To learn word representations, BERT uses the Transformer architecture, that comprise of several encoder layers. It analyzes sequential input, like words in text, using an attention mechanism. The attention mechanism helps the model better understand the text's context by allowing it to capture the relationship between far-off words [26]. During the training process, numerous parameter variables are applied to setting the process, resulting in many variances in the outcomes.

Certain hyperparameters arrangement plays a crucial role in optimizing the training process and the resulting performance of deep learning models like BERT. The proper arrangement in setting the hyperparameters can have a notable effect on the training process and the resulting models' performance on certain tasks. Learning rate, batch size, and the type of optimizer were the three most considered hyperparameters to be properly arranged in BERT-based model training. Understanding the utility and influence of each hyperparameter is critical for achieving optimal results. Table 4 lists various types of various hyperparameters arrangement applied in past research.

The learning rate was one of the most considered hyperparameters to be properly arranged for the training of deep learning models like BERT. The learning rate defines how much the network's weights change throughout each training iteration. A high value learning rate may speed up the training process but will potentially lead the training process to skip in obtaining models with optimal performance. Meanwhile, too low value of learning rate will potentially lead to obtaining models with optimal performance but at the cost of slow down the training process. Based on the literature, we found that the average applied learning rate was between 2.00E-05 and 3.00E-05. The list of papers that reported the imple-

Table 4: The applied hyperparameters excerpted from the reviewed papers

Hyperparameter	Variation	Paper
Learning Rate	1.00E-05	[12], [4], [26], [11]
	1.00E-04	[41], [42], [43], [36]
	2.00E-03	[44], [40], [47]
	2.00E-05	[25], [19], [32], [20], [26], [15], [29], [30], [16], [17], [34], [28], [31], [50], [49]
	3.00E-05	[25], [20], [26], [13], [27], [33], [22], [39], [35]
	4.00E-05	[21]
	5.00E-05	[25], [20], [37], [14], [24], [38], [45], [46]
Batch Size	8	[32], [33], [23], [38], [45], [35]
	16	[12], [20], [4], [37], [29], [17], [34], [24], [50], [41], [48], [49]
	24	[46]
	32	[25], [19], [20], [26], [27], [15], [30], [39], [42], [47], [43], [28], [11]
	64	[14], [16], [36]
	128	[21]
	256	[31]
	3000	[40]
Optimizer	Adam	[4], [12], [11], [41], [42], [43], [36], [44], [40], [47], [25], [19], [32], [20], [15], [29], [30], [16], [17], [34], [28], [31], [50], [49], [13], [27], [33], [22], [39], [35], [21], [37], [14], [24], [38], [45], [46], [23], [48]
	Stochastic Gradient Descent (SGD)	[26]

mentation of these level of learning rate are [26], [25], [19], [32], [20], [15], [29], [30], [16], [17], [34], [28], [31], [50], [49], [13], [27], [33], [22], [39], and [35]. Most of these studies that apply learning rate levels of 2.00E-05 and 3.00E-05 were perhaps due to the original research on BERT as reported in the paper titled *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"*, in which the authors mentioned that the optimal hyperparameter values are actually specific to a particular training context, nevertheless several ranges of values are generally work well in all type of tasks. For their study, however, the authors specifically set their learning rate value to 2.00E-05, 3.00E-05, and 5.00E-05.

The second most considered hyperparameter to be arranged was batch size, which refers to the amount of data samples handled concurrently in a single iteration of model training. The Greater batch size can expedite the training process. They however may degrade the quality of the trained models toward recognizing patterns in new (unseen) data. Smaller batch sizes, on the other hand, will lead to trained models with higher performance due to their better pattern recognition capability towards new (unseen) data but will need longer time to train. Based on the reviewed papers, the most widely applied batch size is 32. The list of papers that reported the implementation of this level of batch size comprises [32], [20], [29], [26], [22], [30], [17], [21], [36], [19], [44], [50], and [42]. The main consideration for selecting a particular level of batch size for model training is normally to account for the available graphical processing unit (GPU) or central processing unit (CPU) memory to avoid out-of-memory (OOM) issues. Furthermore, we need to examine the trade-off between training speed and the expected performance of the trained model when choosing the batch size. Small batch sizes can be slower per epoch but may lead to produce better models' performance.

The next most considered hyperparameter to be arranged was the type of optimizer, which is an algorithm that updates the network's weights during the model training process. An efficient optimizer can contribute to faster convergence and better models' performance outcomes. Adam is a frequently applied optimizer for BERT-based models training. Adam is an adaptive and efficient optimizer that uses momentum estimates and RMSProp.

Interestingly, all the studies reported within the reviewed papers applied Adam optimizer, except for one study as reported in [26] which applied SGD for the optimizer. As mentioned in their paper, the authors' reason for employing SGD is "that SGD's performance was tuned to identify the ideal parameters for improving pertinence through Grid Search and Random Search". Nonetheless, the application of Adam's optimizer outperforms SGD in several studies as reported in the reviewed papers.

According to several studies, the proper arrangement of hyperparameters has a significant impact on the BERT-based model. Based on our reviewed papers, the three most considered hyperparameters for training models were learning rate, batch size, and the type of optimizer. Most of the studies also applied certain types of arrangement for these three hyperparameters. These kinds of arrangements practiced in several studies have provided valuable insight for future works in BERT-based models development and optimization. It is strongly advised to identify the purpose of the models' development or optimization at hand prior deciding certain type of hyperparameters arrangement in our work. Several factors should be considered when deciding on certain hyperparameter arrangements, including the employed hardware capacity, the size of datasets (corpus), the type of the NLP tasks, and the type of the intended model to be developed or optimized (i.e. generic-foundational or specific-customized models).

RQ 3. What type of evaluation metrics were normally employed for particular NLP tasks?

A number of evaluation metrics to evaluate the BERT-based models for NLP tasks have been mentioned in our reviewed papers. The metrics, as described in Table 5, comprise of F1-Score, Accuracy, Precision, Recall, Receiver Operating Characteristic (ROC)/Area Under the ROC Curve (AUC), Exact Match, and ROUGE. The F1 Score was used in the majority of the studies, in which it is commonly combined with Accuracy, Precision, and Recall. It is due to the fact that most of our reviewed papers were reporting on studies regarding text classification. For the type of binary, multi-class or multi-label text classifications, these types of metrics that were calculated based on the result of confusion matrix have become a de facto standard to evaluate the models' performance. If the previous metrics are normally used for evaluating "discriminative" AI-type of tasks, Exact Match and ROUGE are commonly employed for evaluating models' performance on "generative" AI-type of tasks. Since until recently the number of publications on the topic of "generative" AI is still lesser than its "discriminative" AI counterpart, hence the studies which employed the Exact Match and ROUGE are also lower as reflected in Table 5. Nevertheless, the selection of evaluation metric should be based on the type of NLP tasks at hand (such as it is advised to use ROUGE for text synthesis task) and class distribution of the datasets (e.g. it is suggested to avoid using accuracy for evaluating classification performance on imbalanced datasets).

The F1-Score, Accuracy, Precision, Recall, and ROC (AUC) metrics provide an informative picture of the models' performance in predicting (classifying) the correct label. These scores are widely utilized in various "discriminative" AI NLP tasks, such as sentiment analysis and information extraction. It is reflected in the vast number of studies which employing this type of metrics found in this study as reported in [32], [41], [20], [29], [33], [22], [30], [14], [34], [28], [35], [18], [45], [50], [42], [47], [49], [13], [36], [37], [40], and [11] which address text classification tasks. Interestingly, there was no study employed specificity as the evaluation metric even though one paper mentioned it in their content [27]. As for the

Table 5: The applied evaluation metrics excerpted from the reviewed papers

NLP Task	Evaluation Metric	Paper
Text Classification	F1 Score	[25], [12], [19], [32], [20], [4], [37], [13], [27], [15], [29], [30], [21] [16], [17], [33], [34], [22], [23], [24], [38], [28], [11], [39], [44], [31], [50], [42], [35], [36]
	Accuracy	[25], [20], [37], [26], [13], [27], [14], [21], [16], [17], [22], [28] [50]
	Precision	[25], [19], [37], [13], [27], [22], [24], [28], [11], [31], [42]
	Sensitivity (Recall)	[25], [37], [13], [27], [15], [22], [24], [28], [11], [31], [42]
	ROC (AUC)	[11]
Question Answering	Exact Match	[32], [4], [37], [15], [30], [21], [33], [34], [39], [35], [36]
Text Summarization	ROUGE	[44], [45], [46], [40], [41], [47], [48], [43], [49]

least employed metrics for classification performance evaluation was ROC (AUC) which was employed in [11].

Meanwhile, Exact Match is employed for "generative" AI NLP tasks like question answering and machine reading comprehension. Exact Match is a measure of the percentage of responses predicted by the model that match the actual answers, as used for the evaluation metrics in several studies on question-answering model as reported in [15], [4], [24], [30], [17], [14], [39], [31], [21], [37], and [40].

Finally, the ROUGE metric is commonly applied to a number of other "generative" AI NLP tasks including text summarization, text synthesis, and machine translation. ROUGE measures the difference of the text produced by the model to a reference text that is regarded as valid, taking into account n-grams, word order, and lexical similarity. A number of studies on text summarization as reported in [47], [23], [48], [25], [43], [19], [18], [44], and [27], used this evaluation instrument due to its fitness for the particular task.

5 Conclusion

Based on an extensive as well as intensive study of a sufficient number of papers which complied with our review protocol, we eventually came up with three conclusions. First, each type of BERT-based model variant conveys task-specific suitability due to their inherent training history. In order to select appropriate models for particular NLP tasks, three primary concerns should be considered: i) the type of NLP problem to be resolved (i.e. NLP task to be served), ii) the specific domain to be handled (such as financial, medical, law/legal or others), and iii) the intended language to be applied (such as English or others). Second, certain types of hyperparameters arrangements is suggested for training the BERT-based models for improving their performance. Learning rate, batch size, and the type of optimizer were the three most considered hyperparameters to be properly arranged in BERT-based model training. A number of factors should be considered when deciding on certain hyperparameter arrangements, including the employed hardware capacity, the size of datasets (corpus), the type of the NLP tasks, and the type of the intended model to be developed or optimized (i.e. generic-foundational or specific-customized models). Third, different types of metrics are normally employed for particular NLP tasks. The most widely used metrics for text classification were F1-score, accuracy, precision, and sensitivity (recall). Meanwhile, most of the reviewed papers were reporting using Exact Match for question answering and ROUGE for text summarization tasks.

References

[1] J. Bharadiya, “A comprehensive survey of deep learning techniques natural language processing,” *European Journal of Technology*, vol. 7, no. 1, pp. 58–66, 2023.

[2] M.-W. C. Kenton, Jacob Devlin and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of naacL-HLT*, vol. 1, p. 2, Minneapolis, Minnesota, 2019.

[3] K. Mohiuddin, M. A. Alam, M. M. Alam, P. Welke, M. Martin, J. Lehmann, and S. Vahdati, “Retention is all you need,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4752–4758, 2023.

[4] B. Richardson and A. Wicaksana, “Comparison of indobert-lite and roberta in text mining for indonesian language question answering application,” *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 6, pp. 1719–1734, 2022.

[5] E. Kotei and R. Thirunavukarasu, “A systematic review of transformer-based pre-trained language models through self-supervised learning,” *Information*, vol. 14, no. 3, p. 187, 2023.

[6] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *Science China technological sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[7] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, “Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions,” *IEEE Access*, vol. 10, pp. 17078–17097, 2022.

[8] T. A. Rana, Y.-N. Cheah, and S. Letchmunan, “Topic modeling in sentiment analysis: A systematic review.,” *Journal of ICT Research & Applications*, vol. 10, no. 1, 2016.

[9] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering–a systematic literature review,” *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.

[10] F. Zamakhsyari and A. Fatwanto, “A systematic literature review of design thinking approach for user interface design,” *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 4, pp. 2313–2320, 2023.

[11] E. F. Tsani and D. Suhartono, “Personality identification from social media using ensemble bert and roberta,” *Informatica*, vol. 47, no. 4, 2023.

[12] V. Chandradev, I. M. A. D. Suarjaya, and I. P. A. Bayupati, “Analisis sentimen review hotel menggunakan metode deep learning bert,” *Jurnal Buana Informatika*, vol. 14, no. 02, pp. 107–116, 2023.

[13] A. Kazemi, J. Mozafari, and M. A. Nematbakhsh, “Persianquad: the native question answering dataset for the persian language,” *IEEE Access*, vol. 10, pp. 26045–26057, 2022.

[14] B. Baykara and T. Güngör, “Abstractive text summarization and new large-scale datasets for agglutinative languages turkish and hungarian,” *Language Resources and Evaluation*, vol. 56, no. 3, pp. 973–1007, 2022.

[15] R. Yunanto, E. Wibowo, and R. Rianto, “A bert model to detect provocative hoax,” *Journal of Engineering Science and Technology*, vol. 18, no. 5, pp. 2281–2297, 2023.

[16] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, and M. Gupta, “Cobert: Covid-19 question answering system using bert,” *Arabian journal for science and engineering*, vol. 48, no. 8, pp. 11003–11013, 2023.

[17] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, “Semantic answer similarity for evaluating question answering models,” *arXiv preprint arXiv:2108.06130*, 2021.

[18] M. A. Mutasodirin and R. E. Prasojo, “Investigating text shortening strategy in bert: Truncation vs summarization,” in *2021 international conference on advanced computer science and information systems (icacsis)*, pp. 1–5, IEEE, 2021.

[19] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, “Bert base model for toxic comment analysis on indonesian social media,” *Procedia Computer Science*, vol. 216, pp. 714–721, 2023.

[20] F. Baharuddin and M. F. Naufal, “Fine-tuning indobert for indonesian exam question classification based on bloom’s taxonomy,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, 2023.

[21] N. A. Ranggianto, D. Purwitasari, C. Fatichah, R. W. Sholikah, *et al.*, “Abstractive and extractive approaches for summarizing multi-document travel reviews,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 6, pp. 1464–1475, 2023.

[22] R. Sutoyo, H. Warnars, S. M. Isa, and W. Budiharto, “Emotionally aware chatbot for responding to indonesian product reviews,” *International Journal of Innovative Computing, Information and Control*, vol. 19, no. 03, p. 861, 2023.

[23] D. Licari, P. Bushipaka, G. Marino, G. Comandé, and T. Cucinotta, “Legal holding extraction from italian case documents using italian-legal-bert text summarization,” in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 148–156, 2023.

[24] F. Koto, J. H. Lau, and T. Baldwin, “Liputan6: A large-scale indonesian dataset for text summarization,” *arXiv preprint arXiv:2011.00679*, 2020.

[25] L. B. Hutama and D. Suhartono, “Indonesian hoax news classification with multilingual transformer model and bertopic,” *Informatica*, vol. 46, no. 8, 2022.

[26] M. Mozafari, R. Farahbakhsh, and N. Crespi, “Hate speech detection and racial bias mitigation in social media based on bert model,” *PloS one*, vol. 15, no. 8, p. e0237861, 2020.

[27] R. Vemula, M. Nuthi, and M. Shrivastava, "Tequad: Telugu question answering dataset," in *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pp. 300–307, 2022.

[28] P. Wang, J. Fang, and J. Reinspach, "Cs-bert: a pretrained model for customer service dialogues," in *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pp. 130–142, 2021.

[29] E. P. A. Akhmad, "Analisis sentimen ulasan aplikasi dlu ferry pada google play store menggunakan bidirectional encoder representations from transformers," *Jurnal Aplikasi Pelayaran dan Kepelabuhanan*, vol. 13, no. 2, pp. 104–112, 2023.

[30] A. Jazuli, W. Widowati, and R. Kusumaningrum, "Aspect-based sentiment analysis on student reviews using the indo-bert base model," in *E3S Web of Conferences*, vol. 448, p. 02004, EDP Sciences, 2023.

[31] R. A. Putri and A. Oh, "Idk-mrc: Unanswerable questions for indonesian machine reading comprehension," *arXiv preprint arXiv:2210.13778*, 2022.

[32] G. L. Martin, M. E. Mswahili, and Y.-S. Jeong, "Sentiment classification in swahili language using multilingual bert," *arXiv preprint arXiv:2104.09006*, 2021.

[33] M. I. Rahajeng and P. Ayu, "Indonesian question answering system for factoid questions using face beauty products knowledge graph," *Jurnal Linguistik Komputasional (JLK)*, vol. 4, p. 59, 09 2021.

[34] R. Puri, R. Spring, M. Patwary, M. Shoeybi, and B. Catanzaro, "Training question answering models from synthetic data," *arXiv preprint arXiv:2002.09599*, 2020.

[35] S. Abdel-Salam and A. Rafea, "Performance study on extractive text summarization using bert models," *Information*, vol. 13, no. 2, p. 67, 2022.

[36] X. Yang, C. Zhang, Y. Sun, K. Pang, L. Jing, S. Wa, and C. Lv, "Finchain-bert: A high-accuracy automatic fraud detection model based on nlp methods for financial scenarios," *Information*, vol. 14, no. 9, p. 499, 2023.

[37] Q. Ismail, K. Alissa, and R. M. Duwairi, "Arabic news summarization based on t5 transformer approach," in *2023 14th International Conference on Information and Communication Systems (ICICS)*, pp. 1–7, IEEE, 2023.

[38] R. Calizzano, M. Ostendorff, Q. Ruan, and G. Rehm, "Generating extended and multilingual summaries with pre-trained transformers," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1640–1650, 2022.

[39] F. Dartiko, M. Yusa, A. Erlansari, and S. A. Basha, "Comparative analysis of transformer-based method in a question answering system for campus orientation guides," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 1, pp. 122–139, 2024.

[40] L. L. Maceda, A. A. Satuito, and M. B. Abisado, "Sentiment analysis of code-mixed social media data on philippine uaqte using fine-tuned mbert model," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023.

[41] Q. Li and Y. Zhang, "Improved text matching model based on bert," *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 3, pp. 40–43, 2022.

[42] X. Liu, Y. Li, Y. Shao, A. Li, and J. Liang, "A sentiment analysis model for car review texts based on adversarial training and whole word mask bert," in *China Intelligent Networked Things Conference*, pp. 107–121, Springer, 2022.

[43] M. F. Ullah, A. Saeed, J. Li, T. Mahmood, M. Adeel, *et al.*, "Bert model for roman urdu fake review identification," 2023.

[44] N. K. Nissa and E. Yuliant, "Multi-label text classification of indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Power Electron. Drive Syst*, vol. 13, pp. 5641–5652, 2023.

[45] H. Härm and T. Alumäe, "Abstractive summarization of broadcast news stories for estonian.," *Baltic Journal of Modern Computing*, vol. 10, no. 3, 2022.

[46] T. M. Luu, H. T. Le, and T. M. Hoang, "A hybrid model using the pretrained bert and deep neural networks with rich feature for extractive text summarization," *Journal of Computer Science and Cybernetics*, vol. 37, no. 2, pp. 123–143, 2021.

[47] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[48] F. B. Fikri, K. Oflazer, and B. Yanıkoğlu, "Abstractive summarization with deep reinforcement learning using semantic similarity rewards," *Natural Language Engineering*, vol. 30, no. 3, pp. 554–576, 2024.

[49] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "Qasina: Religious domain question answering using sirah nabawiyah," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pp. 1–6, IEEE, 2023.

[50] H. von Essen and D. Hesslow, "Building a swedish question-answering model," in *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pp. 117–127, 2020.