INTENSIF, Vol.8 No.1 February 2024
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971

Comparative Analysis of Transformer-Based Method In A Question Answering System for Campus Orientation Guides

Received:
9 January 2024
Accepted:
2 February 2024
Published:
10 February 2024

[1*]Fedryanto Dartiko, [2]Mochammad Yusa, [3]Aan Erlansari, [4]Shaikh Ameer Basha
[1-2]Informatics, Universitas Bengkulu
[3]Information System, Universitas Bengkulu
[4]Mathematics, Bearys Institute of technology
E-mail: [1]fedryanto2007@gmail.com, [2]mochammad.yusa@unib.ac.id, [3]aan_erlanshari@unib.ac.id, [4]shaikhameerbasha@gmail.com
*Corresponding Author

Abstract—The campus introduction process is a stage where new students acquire information about the campus through a series of activities and interactions with existing students. However, the delivery of campus introduction information is still limited to conventional methods, such as using guidebooks. This limitation can result in students having a limited understanding of the information needed during their academic period. The one of solution for this case is to implement a deep learning system with knowledge-based foundations. This research aims to develop a Question Answering System (QAS) as a campus introduction guide by comparing two transformer methods, namely the RoBERTa and IndoBERT architectures. The dataset used is processed in the SQuAD format in the Indonesian language. The collected SQuAD dataset in the Indonesian language consists of 5046 annotated data. The result shows that IndoBERT outperforms RoBERTa with EM and F1-Score values of 81.17 and 91.32, respectively, surpassing RoBERTa with EM and F1-Score values of 79.53 and 90.18.

Keywords— Question Answering; NLP; Transformer; IndoBERT; RoBERTa

Corresponding Author:

Fedryanto Dartiko,
Informatics,
Universitas Bengkulu,
Email: fedryanto2007@gmail.com,
Orchid ID: http://orcid.org/ 0000-0002-4126-9247

I. INTRODUCTION

A university campus can be defined as a tangible complex or locale comprised of diverse structures and amenities devoted to the facilitation of higher education activities. It functions as the locus wherein institutions of higher learning, namely universities or colleges, undertake academic, research, and administrative pursuits. Within the campus infrastructure are inclusive provisions such as lecture halls, laboratories, libraries, cafeterias, dormitories, and ancillary facilities [1]. The acquisition of information transpires through the mechanism of data transformation, a process by which raw data is converted into a meaningful and valuable format for the benefit of information recipients, thereby constituting the foundational basis for decision-making [2]. The articulation of campus introduction information confronts numerous noteworthy challenges. Foremost among these challenges is the impediment faced by students in accessing essential information encompassing academic particulars, facilities, classroom locations, faculty profiles, and other relevant details. Regrettably, campus information lacks integration and proves inaccessible to new students. Furthermore, the optimal dissemination of campus information remains unrealized due to a paucity of comprehensive references. Presently, the predominant references exist in print format, an approach that may not fully satisfy the informational requisites of students. Despite efforts to leverage social media as an alternative avenue for information dissemination, the execution of this solution is hindered by administrative constraints, resulting in an incapacity to respond promptly and accurately to the multitude of inquiries posed via social media platforms. Consequently, there exist impediments to furnishing adequate responses to prospective students seeking comprehensive insights into the campus.

In response to the aforementioned challenges, a discernible imperative emerges for the implementation of a system that can expeditiously and precisely address inquiries. Henceforth, a resolution in the guise of a Knowledge-Based Question Answering System (QAS) is conceived [3]. The QAS fashioned as a knowledge-driven question-answering system, is orchestrated to empower computational systems to discern the intent and purpose underlying user queries through the employment of Natural Language Processing (NLP). This advanced cognitive framework consequently formulates responses to user inquiries by proffering answers derived from extant information reservoirs [4].

Employing a methodology rooted in transformer architecture, our Question Answering System (QAS) demonstrates exceptional proficiency in scrutinizing inter-sentence relationships, a pivotal facet for achieving accurate natural language comprehension. The inherent self-attention mechanism within transformers endows the system with the capability to attend to pertinent information dispersed across multiple sentences, thereby surpassing methodologies confined to individual sentence analysis. This heightened contextual awareness equips the QAS with the capacity to render more precise predictions at the sentence level and extract deeper semantic nuances from intricate textual inputs [5]. The transformer model operates through the utilization of layered self-attention and fully connected connections between the encoder and decoder components [6]. The instantiation of the transformer model in this investigation arises from a meticulous comparative assessment between the RoBERTa [7] and IndoBERT [8] models.

A considerable body of scholarly inquiry has delved into the deployment of diverse models within the domain of Question Answering Systems (QAS). Noteworthy contributions include the investigation conducted by [5], wherein the efficacy of Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Neural Network (CNN), and Transformer models was systematically evaluated for answer selection in QAS, highlighting their aptitude in supporting the selection process within the transformer encoder. Furthermore, the work of [9] centered on assessing the performance of BERT-based models, encompassing Latent Dirichlet Allocation (LDA)-based, Bert-base, and transformer-base architectures, within the realm of QAS applications. Significantly, their focus extended to the comprehensive evaluation of diverse QAS frameworks against established datasets. Concurrently, the pragmatic study by [10] adopted a utilitarian stance, delineating the development of a chatbot leveraging the Artificial Intelligence Markup Language (AIML) as an information service catering to student registration. Hanifah and Kusumaningrum conducted a developmental study on Non-Factoid answering cases using the LSTM method [11] Noraset et al. extended their research on QAS based on Wikipedia in the Thai language [12] Aurpa et al. conducted research on the utilization of transformer methods in the case of Reading Comprehension based on QAS in the Bengali language [13]. Identifying a conspicuous lacuna within the scholarly landscape, it is discerned that there exists a notable dearth of dedicated research endeavors focusing on the development of a Question Answering System (QAS) tailored specifically for facilitating campus information acquisition, serving as an indispensable guide for campus orientation, especially within the Indonesian language domain. The adoption of Transformer methodologies emerges as a strategic pursuit in response to this identified gap, underscored by their innovative nature and their inherent capacity to ameliorate shortcomings inherent in traditional methods delineated within antecedent scholarly inquiries.

II. RESEARCH METHOD

The present study undertakes a comparative analysis of two transformer models, followed by the development of a Question Answering System (QAS) grounded in the superior model. This inquiry aligns itself within the realm of applied research, positioned strategically to augment performance within the practical domain and contribute substantively to the advancement of scientific knowledge. Applied research, as elucidated by the primary objective of this investigation, is characterized by its targeted focus on addressing specific issues, with outcomes designed for pragmatic implementation, thereby conferring tangible benefits upon humanity [14].

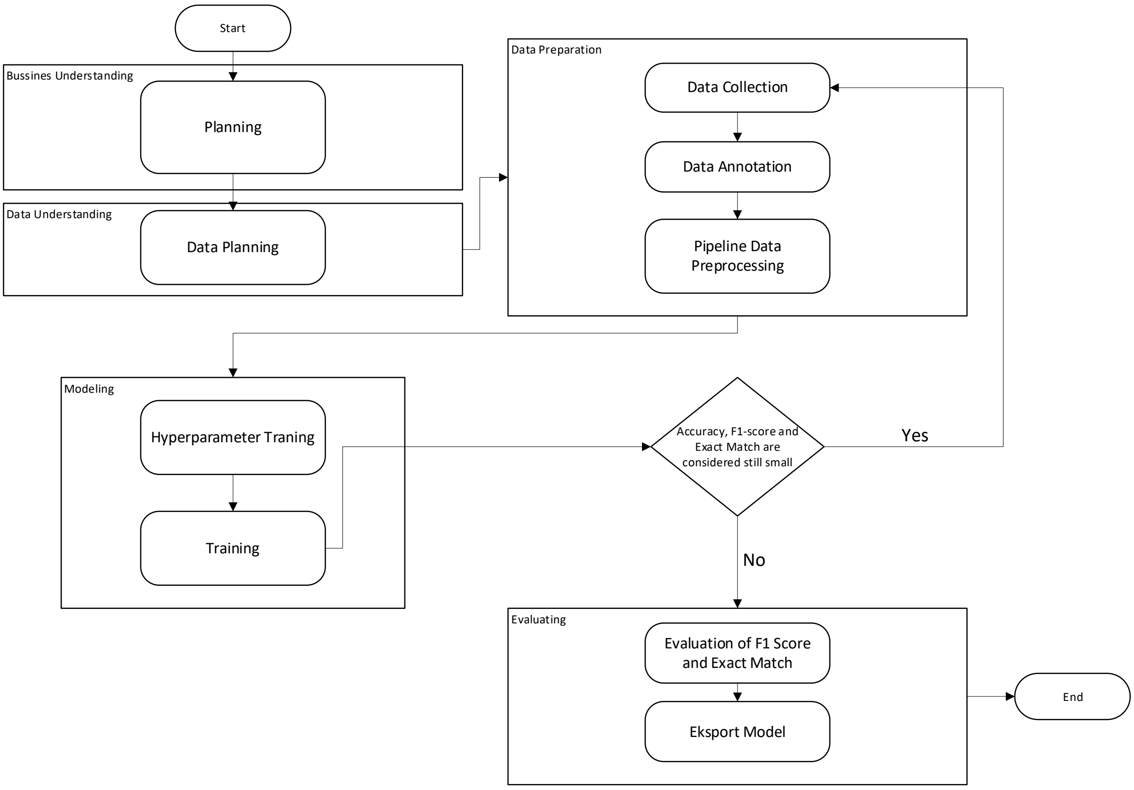


Fig 1. Research Flow

A. Planning

The initial stage of this research involves planning and formulating the topic and issues. Observations are conducted to identify relevant issues, and a literature review is undertaken to strengthen the knowledge base and determine the research topic and methodology. At this stage, the research needs are analyzed, potential challenges are identified, and preventive measures and solutions are formulated.

B. Data Planning

The data planning process involves a literature review to evaluate the types, formats, and methods of data collection and formation. This stage aims to understand the data requirements, data processing strategies, as well as potential data issues and their solutions

C. Data Collection

The dataset utilized in this study was meticulously curated from samples drawn from established higher education institutions in Indonesia, with the data meticulously extracted from academic guidebooks corresponding to the 2022/2023 academic year. Compliant with the standard format for the construction of question answering systems, the dataset is formatted in accordance with the Stanford Question Answering Dataset (SQuAD) [15]. Subsequently, the Question Answering System (QAS) engenders textual responses to interrogative prompts. The instantiation of this QAS involves the deployment of a model derived from the outcome of a rigorous comparative analysis between two transformer models, specifically RoBERTa and IndoBERT. The efficacy of the training and prediction processes is contingent upon the availability of a meticulously annotated dataset.

The amassed data in this investigation assumes a pivotal role as the foundational substrate for the inception of a Question Answering System (QAS) specifically tailored for campus guideline applications. This dataset is of paramount significance, wielding an indispensable function in both the training and evaluative phases of the QAS model. It serves as the cornerstone for assessing and refining the model's adeptness in generating articulate and contextually appropriate responses to inquiries originating from users [16].

D. Data Preparation

The meticulous preparation of data constitutes an imperative phase in the intricate developmental trajectory of the Question Answering System (QAS). This preparatory phase encompasses multifaceted activities of paramount importance, encompassing data annotation, judicious data splitting, meticulous data normalization, and rigorous preprocessing procedures. Each of these activities contributes significantly to the foundational underpinnings of the QAS, ensuring the integrity and coherence of the dataset harnessed for subsequent model training and evaluation.

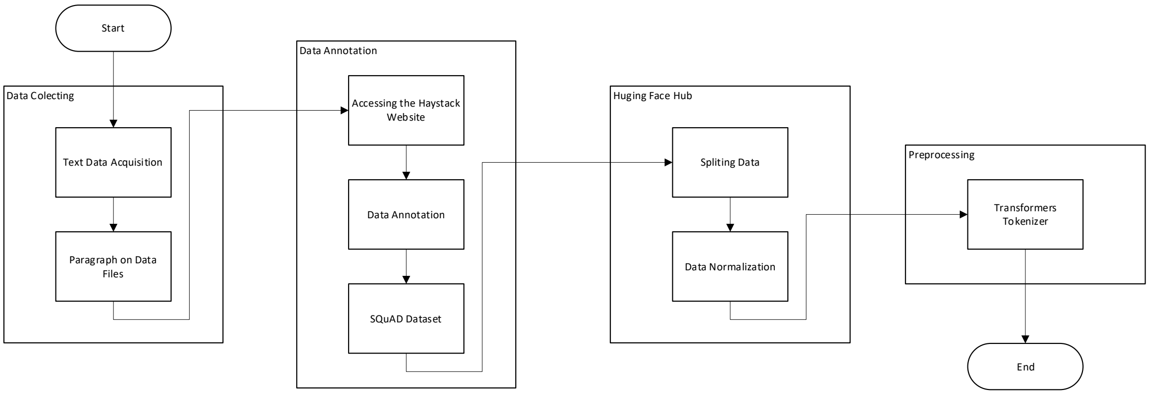


Fig 2. Data Collection and Processing Workflow

Figure 2 delineates the comprehensive procedural flow encompassing the trajectory from data collection to data processing, ultimately culminating in the formulation of a meticulously crafted dataset tailored for training purposes. This sequential orchestration encompasses the pivotal stages of data extraction, organizational structuring, and normalization to engender a prepared dataset aptly suited for utilization in the ensuing model training process. The subsequent elucidation provides a detailed explication of each constituent activity:

1. Data Annotation: The data annotation process integrates Haystack Annotation methodology, facilitating the transformation of collected paragraph data into a dataset structure characterized by fields such as ID, context, question, and answer [17]. This process involves the judicious labeling of answers corresponding to individual questions and the judicious pairing of questions with their pertinent answers.
2. Data Splitting: The data splitting procedure is executed leveraging the hub system accessible through Hugging Face. Equipped with code for partitioning the data into distinct subsets—namely, training data and validation data—the system achieves this segmentation [18]. Notably, the preferred data splitting ratio, often recommended by scholars, adheres to the 80/20 paradigm for training and validation datasets to establish an optimal configuration [18]. The partitioned data, thus prepared, is primed for integration into the modeling phase.
3. Data Normalization: A critical stride in the preparatory process, data normalization is instrumental in configuring the dataset to conform to the requisite structure for effective utilization in the model training process. This transformation is indispensable, particularly when the initial annotation dataset structure assumes a singular array format, necessitating alignment for the seamless integration of question and answer data [19].
4. Data Preprocessing: The data preprocessing phase involves tokenization, a transformative process wherein text is converted into tokens. This intricate procedure is enacted through the application of the Transformers Tokenizer, which, based on a pre-trained vocabulary, translates words into unique identifiers (IDs) [20]. Tokens are subsequently formatted to align with the architectural specifications of the designated model. Executing this preprocessing entails the instantiation of an AutoTokenizer_from_pretrained instance, generating model-specific tokens and vocabulary derived from the pre-trained model [21].

   Collectively, the entirety of the data preparation process is engineered to ensure the dataset attains a format conducive to the effective training of the Question Answering System (QAS) model, employing a transformer model. The annotated, partitioned, normalized, and preprocessed dataset collectively serves as the substratum for the subsequent training regimen, aimed at endowing the QAS model with the capacity to discern and respond efficaciously to user-generated inquiries.

E. Modeling

During the pivotal modeling phase, nuanced adjustments to parameters are methodically executed to ascertain optimal performance in question response leveraging transformer architectures. The hyperparameters delineated for the training process encompass epochs, learning rate, batch size, max length, and doc stride [7]. The determination of epochs initiates at 10, and in instances where early manifestations of overfitting materialize, a judicious reduction in the epoch count ensues. Conversely, in the absence of overfitting indications, the epochs undergo a doubling augmentation, culminating in a finalized epoch value of 5. The "learning_rate" parameter is meticulously set to 3e-5, a selection grounded in its efficacy in facilitating effective model learning and attaining peak results. The "batch_size" parameter embarks upon the training process at a magnitude of 16 and undergoes a systematic reduction until optimal outcomes are achieved, culminating in a conclusive value of 8. This choice strikes an equilibrium, avoiding an undue diminution that might impede the training process and a magnitude that could strain the device's Random Access Memory (RAM) capacity [22]. The "max_length" parameter, defining the model's proficiency in processing questions and answers based on a stipulated number of tokens, is judiciously set to 500. This selection ensures congruence with the model's token capacity, constrained to 512 tokens [23]. Concomitantly, the "doc_stride" parameter, integral to the preprocessing phase, facilitates the model's capacity to traverse the original text in overlapping segments during the slicing process. A deliberate choice of 128 for the "doc_stride" parameter is substantiated by antecedent research recommendations [24], [25], thereby contributing to the seamless integration of the model within the existing discourse on parameter tuning in transformer-based question answering systems.

F. Evaluation

The assessment phase constitutes an integral juncture in the maturation of the Question Answering System (QAS) model, as elucidated by the scholarly discourse on the development of question answering systems [26]. During this evaluative stage, the pre-trained model undergoes rigorous testing using a dedicated validation dataset, thereby facilitating a comprehensive appraisal of its efficacy in responding to diverse inquiries. The evaluation process is systematically executed for each of the models under consideration, specifically RoBERTa and IndoBERT, engendering a meticulous examination of their performance nuances. The evaluation metrics employed encompass both Exact Match and F1-Score values, indicative of the models' precision and overall accuracy in generating responses to posed questions [27]. A visual representation of the intricate evaluation process flow is encapsulated in Figure 3.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971

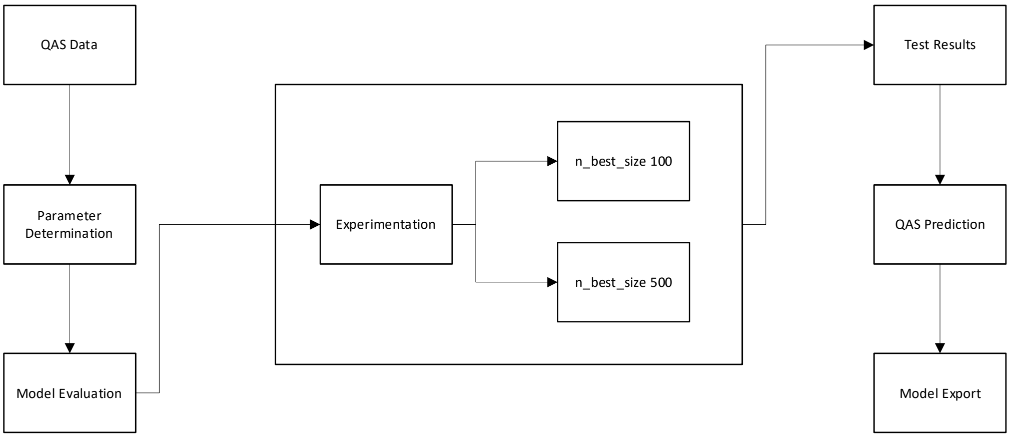


Fig 3. EVALUATION PROCESS FLOW

Figure 3 illuminates the intricate procedural dynamics characterizing the model evaluation process, wherein a systematic assessment is conducted by varying the max_length parameters. The max_length parameter within the transformer model assumes a pivotal role, dictating the upper threshold for the length of text that the model is capable of processing. The deliberate examination of the model across an array of max_length values serves as a methodological approach to discerning the nuanced impact of this parameter on the model's performance in the domain of text comprehension [28].

III. RESULT AND DISCUSSION

A. Data Collection

The acquisition of data is meticulously executed through a comprehensive literature review methodology, predicated upon the perusal of university academic guidebooks. The garnered data, presented in a textual format, is systematically extracted from these guidebooks, organized into paragraphs, and meticulously cataloged into 1023 distinct files, each adopting the .txt format. Subsequently, a judicious categorization is implemented, segregating the amassed data into discrete training and validation datasets to ensure the robustness and integrity of the subsequent analytical processes [29]. The collected textual data was subsequently labeled, resulting in a total of 5046 data pairs consisting of questions and answers

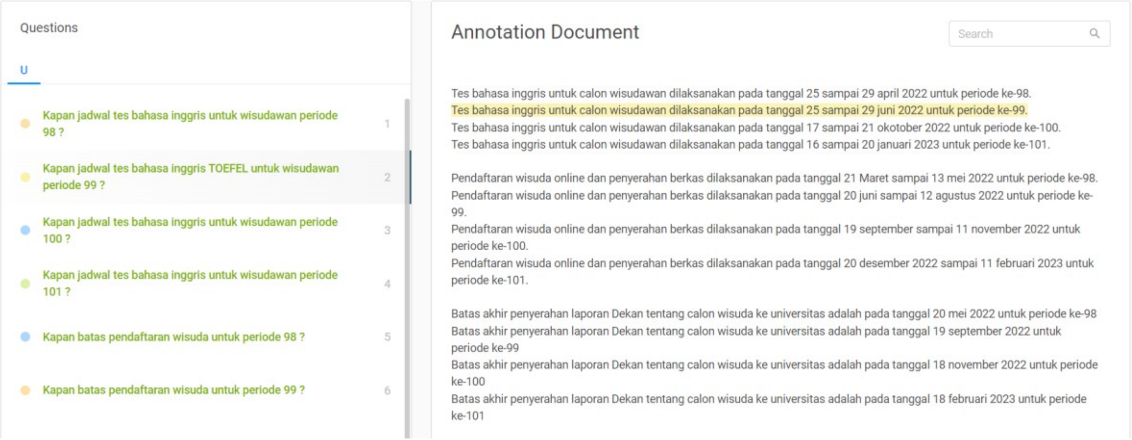
B. Data Preparation

The meticulous preparation of data is an intricate and indispensable facet of the research methodology, encompassing several discerning stages: Data Annotation, Data Splitting, Data Normalization, and Data Preprocessing. Each of these stages is methodically orchestrated to ensure the integrity and suitability of the dataset for subsequent model training and evaluation.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971

1. Data Annotation:

The data annotation process is conducted through the adept utilization of the Haystack Annotation Tools, a resource provided by Deepset [30]. This process involves the extraction of answers from potential questions formulated on the basis of raw paragraph data. A meticulous annotation of 1023 data files is executed, illustrated comprehensively in Figure 4. This annotation process entails the judicious selection or marking of answers corresponding to the generated questions, culminating when all questions are systematically annotated. The resultant datasets, totaling 5046 after labeling, can be conveniently stored in the standardized Stanford Question Answering Dataset (SQuAD) format, denoted by the .json extension.



- Kapan jadwal tes bahasa inggris untuk wisudawan periode 98 ? 1
- Kapan jadwal tes bahasa inggris TOEFEL untuk wisudawan periode 99 ? 2
- Kapan jadwal tes bahasa inggris untuk wisudawan periode 100 ? 3
- Kapan jadwal tes bahasa inggris untuk wisudawan periode 101 ? 4
- Kapan batas pendaftaran wisuda untuk periode 98 ? 5
- Kapan batas pendaftaran wisuda untuk periode 99 ? 6

Tes bahasa inggris untuk calon wisudawan dilaksanakan pada tanggal 25 sampai 29 april 2022 untuk periode ke-98.
Tes bahasa inggris untuk calon wisudawan dilaksanakan pada tanggal 25 sampai 29 juni 2022 untuk periode ke-99.
Tes bahasa inggris untuk calon wisudawan dilaksanakan pada tanggal 17 sampai 21 okotober 2022 untuk periode ke-100.
Tes bahasa inggris untuk calon wisudawan dilaksanakan pada tanggal 16 sampai 20 januari 2023 untuk periode ke-101.

Pendaftaran wisuda online dan penyerahan berkas dilaksanakan pada tanggal 21 Maret sampai 13 mei 2022 untuk periode ke-98.
Pendaftaran wisuda online dan penyerahan berkas dilaksanakan pada tanggal 20 juni sampai 12 agustus 2022 untuk periode ke-99.
Pendaftaran wisuda online dan penyerahan berkas dilaksanakan pada tanggal 19 september sampai 11 november 2022 untuk periode ke-100.
Pendaftaran wisuda online dan penyerahan berkas dilaksanakan pada tanggal 20 desember 2022 sampai 11 februari 2023 untuk periode ke-101.

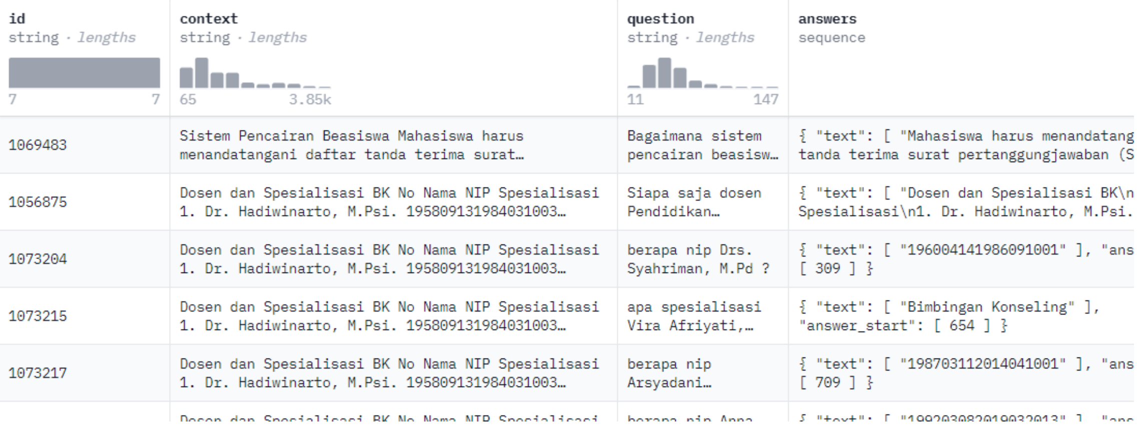
Batas akhir penyerahan laporan Dekan tentang calon wisuda ke universitas adalah pada tanggal 20 mei 2022 untuk periode ke-98
Batas akhir penyerahan laporan Dekan tentang calon wisuda ke universitas adalah pada tanggal 19 september 2022 untuk periode ke-99
Batas akhir penyerahan laporan Dekan tentang calon wisuda ke universitas adalah pada tanggal 18 november 2022 untuk periode ke-100
Batas akhir penyerahan laporan Dekan tentang calon wisuda ke universitas adalah pada tanggal 18 februari 2023 untuk periode ke-101

Fig 4. Data Annotation Process

2. Data Splitting and Normalization:

The strategic partitioning of data into training and validation sets is undertaken to appraise the model's performance rigorously. The annotated data undergoes a division, allotting 80% to training data and 20% to validation data, yielding a sum of 4010 instances for training and 1036 instances for validation [31]. Post partitioning, the data undergoes uploading to the Huggingface Hub for normalization, a crucial step given the single array format inherent in the annotated data. The normalization process ensures the extraction of data for subsequent utilization, with the visual representation of the dataset post the data splitting and normalization procedures elucidated in Figure 5.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971



id string · *lengths* 7 7	context string · *lengths* 65 3.85k	question string · *lengths* 11 147	answers sequence
1069483	Sistem Pencairan Beasiswa Mahasiswa harus menandatangani daftar tanda terima surat…	Bagaimana sistem pencairan beasisw…	{ "text": ["Mahasiswa harus menandatang tanda terima surat pertanggungjawaban (S
1056875	Dosen dan Spesialisasi BK No Nama NIP Spesialisasi 1. Dr. Hadiwinarto, M.Psi. 195809131984031003…	Siapa saja dosen Pendidikan…	{ "text": ["Dosen dan Spesialisasi BK\n Spesialisasi\n1. Dr. Hadiwinarto, M.Psi.
1073204	Dosen dan Spesialisasi BK No Nama NIP Spesialisasi 1. Dr. Hadiwinarto, M.Psi. 195809131984031003…	berapa nip Drs. Syahriman, M.Pd ?	{ "text": ["196004141986091001"], "ans [309] }
1073215	Dosen dan Spesialisasi BK No Nama NIP Spesialisasi 1. Dr. Hadiwinarto, M.Psi. 195809131984031003…	apa spesialisasi Vira Afriyati,…	{ "text": ["Bimbingan Konseling"], "answer_start": [654] }
1073217	Dosen dan Spesialisasi BK No Nama NIP Spesialisasi 1. Dr. Hadiwinarto, M.Psi. 195809131984031003…	berapa nip Arsyadani…	{ "text": ["198703112014041001"], "ans [709] }
	Dosen dan Spesialisasi BK No Nama NIP Spesialisasi	berapa nip Anna	{ "text": ["199203082019032013"], "ans

Fig 5. Annotated Dataset

3. Data Preprocessing:

Antecedent to the incorporation of data into the model training regimen, the dataset undergoes a pivotal preprocessing stage [32]. This transformative phase entails the conversion of textual data into vectors or numerical representations, thereby enabling the model to interpret textual data based on prevailing vector sequences. The preprocessing focus is specifically directed towards question-answer pairs, resulting in the generation of vector pairs for each dataset [33]. This methodical progression reinforces the foundational rigor and academic exactitude embedded within the data preparation stage of the research endeavor.

Table 1. Preprocessed Vector Data

Vectorized Question-Answer Pairs							
0	1	36	49	138	145	233	235
0	0	50	56	146	151	235	237
0	3	57	62	151	152	237	238
4	8	63	72	154	161	239	246
9	14	73	77	162	170	247	254
15	25	78	89	171	182	255	260
26	35	90	98	183	191	261	270
36	40	99	104	192	193	271	277
41	43	105	113	193	195	278	279
44	48	114	117	195	197	279	281

Table 1 elucidates illustrative outcomes derived from the preprocessing stage, leveraging a transformer tokenizer within the model architecture. The transformative process entails the dissection of textual data into discrete tokens, thereby rendering the information intelligible to the model. Each token, encompassing entire words, sub-words, or fragments thereof, undergoes

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971

representation as a numeric vector through the embedding procedure. This intricate transformation facilitates the model's capacity to interpret textual data through numeric representations. The resultant token values encapsulate critical information pertaining to the actual locations of answers within the text, thereby enhancing the model's proficiency in discerning precise answer positions. The tokenizer additionally generates pairs of Input IDs and Attention Mask. Input IDs manifest as numeric representations of tokens in the text post-tokenization, with each word or sub-word assigned a corresponding number or ID—effectively an index within the model's acquired vocabulary. Serving as the primary input for the model, these Input IDs encapsulate the numerical structure of the text. Concurrently, the Attention Mask assumes a pivotal role in directing the model's focus by indicating which portions of the input are pertinent and which should be disregarded. Typically presented as binary pairs (0 and 1), where 1 designates positions meriting attention and 0 designates positions to be ignored, this binary pair becomes input for the transformer model. This input configuration empowers the model to process and comprehend text structure based on the stipulated tokenization parameters. Consequently, this numerical pair facilitates the model's adept understanding of word or sub-word sequences, harnessing attentional elements to glean precise contextual information within the text.

C. Model Comparison

Transformer models, within their operational framework, proficiently execute both encoding and decoding processes on input sentences, wherein the systematic analysis encompasses the comprehension of sentences, the conversion of words into tokens, and the calculation of token spans from initiation to termination. This intricate orchestration is instrumental in delineating potential question candidates and answer candidates through a meticulous evaluation of token sequences. The synthesized information thereby enables the model to engage in predictive endeavors, culminating in the formulation of responses to posed questions [34].

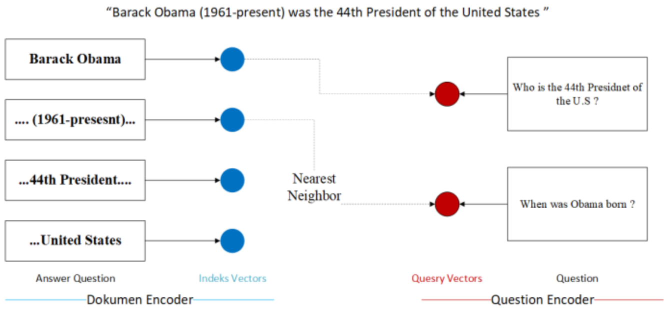


Fig 6. Transformer Encoder And Decoder Processes [5]

Figure 5 provides a visual representation of the encoder and decoder processes within the transformer architecture when processing question-answer pairs, a salient elucidation within the broader context of transformer-based models employed for natural language understanding [5].

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.21971

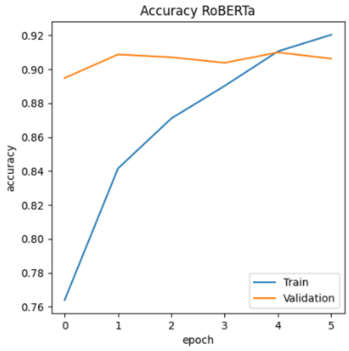
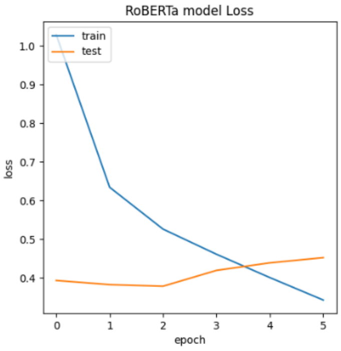
The training phase, utilizing a dataset comprising 5046 distinct entries, was systematically executed under a consistent set of hyperparameters for both the RoBERTa and IndoBERT models. The consequential outcomes of this training regimen, encapsulating the models' proficiency in encoding and decoding question-answer interactions, are meticulously cataloged and presented in a tabulated format in Tables 1 and 2. This analytical exposition serves to enhance the scholarly discourse surrounding the practical implementation and comparative evaluation of transformer models in the domain of question answering systems

Table 2. Comparison of Model Accuracy

No.	Architecture	*Star_logits_accuracy*	*End_logits_accuracy*
1	RoBERTa	Training	
		0,8526	0,8168
		Validation	
		0,8253	0,8052
2	IndoBERT	Training	
		0.8911	0.8671
		Validation	
		0.8715	0.8876

Table 2 offers a comprehensive comparative analysis of the outcomes derived from the model training process, furnishing a nuanced examination of the performance metrics of the RoBERTa and IndoBERT models. Noteworthy is the commendable accuracy exhibited by both models, with an overarching average value attaining a commendable 0.8. Significantly, IndoBERT emerges as the superior performer, demonstrating superiority over RoBERTa across critical metrics, notably eclipsing both start_logit_accuracy and end_logit_accuracy for both the training and validation datasets.

Table 3. Comparison Of Model Training Charts

Model	accuracy/epoch	Loss/epoch
RoBERTa		

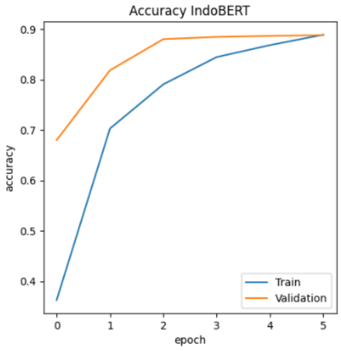
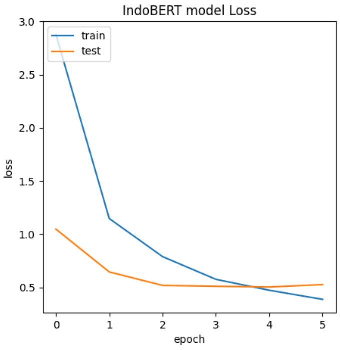
Model	accuracy/epoch	Loss/epoch
IndoBERT		

Table 3 delineates the training graph of the model encompassing 5 epochs, a strategic adjustment made subsequent to an initial utilization of 10 epochs. The decision to reduce the epoch value arose from an observed presence of overfitting, necessitating a calibrated approach to enhance model generalizability. Proficient training processes are characterized by graphs demonstrating a parallel trajectory between training and validation datasets. An insightful comparison of the training graphs, as documented in Table 2, underscores the suboptimal performance of the RoBERTa model. Manifesting challenges in comprehending previously unseen or validation data, the RoBERTa model's limitations are evidenced in the training graph for validation data, which exhibits negligible improvement and inclines toward decline. In stark contrast, the IndoBERT model attains superior results vis-à-vis RoBERTa, as depicted in Table 2 through a graph illustrating consistent and parallel amelioration between training and validation data. An in-depth scrutiny of the loss reduction graph unveils IndoBERT's substantial reduction in loss, indicative of its stable convergence over the temporal course.

D. Model Evaluation

The comprehensive evaluation of the models transpired through a meticulously orchestrated two-stage process, employing the parameters n_best_size 100 and n_best_size 500, each corresponding to the max_length parameter [35]. This deliberate configuration sought to systematically scrutinize the influence of distinct maximum capacity constraints on each model, thereby discerning potential disparities contingent upon the imposed limitations. The overarching objective of this evaluative framework was to ascertain whether the models manifested significant performance variations under the applied constraints. The evaluative protocol was rigorously executed for each model, involving a nuanced variation of parameter values. A detailed exposition of the empirical findings resulting from this evaluative undertaking is systematically presented and elucidated in Tables 4 and 5, thereby enriching the academic discourse on the discernible nuances in model performance across diverse parameter configurations.

The evaluation of model performance unfolded in a systematic bifurcation, employing the parameters n_best_size 100 and n_best_size 500. This methodological stratagem sought to elucidate the nuanced impact of diverse maximum capacity constraints on each model, thereby discerning potential disparities contingent upon the imposed limitations. The evaluative process was meticulously executed for each model through a deliberate variation of parameter values, affording a comprehensive exploration of their respective performances under distinct configurations. The detailed and granular findings stemming from this evaluative endeavor are meticulously cataloged and expounded upon in Tables 4 and 5, thereby providing a comprehensive empirical foundation for discerning the performance differentials between the models

Table 4. Evaluate Model with N_Best_Size 100

No.	**Model Architecture**	***Exact Match***	***F1-Score***
1	RoBERTa	69,50	78,62
2	IndoBERT	69,79	80,34

Table 4 presents a meticulous exposition of the Evaluation Metrics (EM) and F1-Score values delineating the performance of each model, as assessed under the n_best_size parameter set at 100. The empirical findings elucidate that the IndoBERT model architecture exhibits superior performance relative to its counterpart, RoBERTa, within the transformer paradigm. Specifically, the IndoBERT model attains an Exact Match value of 69.79, surpassing RoBERTa by 0.29. Furthermore, the F1-Score of 80.34 achieved by the IndoBERT model outstrips RoBERTa by a margin of 1.72. This discerning evaluation not only provides a granular insight into the nuanced differentials in performance between the transformer architectures but also enriches the academic dialogue on the intricacies of model evaluation metrics in the domain of question answering systems.

Table 5. Evaluate Model with N_Best_Size 500

No.	**Model Architecture**	***Exact Match***	***F1-Score***
1	RoBERTa	79,53	90,18
2	IndoBERT	81,17	91,32

Table 5 provides a comprehensive exposition of Evaluation Metrics (EM) and F1-Score values, wherein the parameters n_best_size and max_length_answer are systematically configured to 500. The empirical findings underscore the discernible impact of parameter adjustment on performance metrics, indicating a noteworthy enhancement in both EM and F1-Score values. Notably, the IndoBERT architecture continues to manifest superior performance in contrast to RoBERTa, attaining an EM value of 81.17, a notable increment of 1.64, and an F1-

Score of 91.32, surpassing RoBERTa by 1.14. This empirical scrutiny not only illuminates the sensitivity of model performance to parameter configurations but also substantiates the sustained proficiency of the IndoBERT architecture in question answering tasks. The comprehensive comparative evaluation results, further illustrated in Figure 6, contribute substantively to the nuanced understanding of transformer model dynamics in the context of question answering systems.

Tables 4 and 5 serve as visual representations elucidating the comparative analysis of F1-score values between the RoBERTa and IndoBERT models. The discerned findings unequivocally underscore the superior performance exhibited by the IndoBERT model across scenarios characterized by n values set at 100 and 500. Noteworthy is the substantive impact of variations in the parameter n on the overall model evaluation results, thereby substantively augmenting the performance scores for both models. Additionally, the tables present detailed bar graphs juxtaposing the Exact Match (EM) values between RoBERTa and IndoBERT. These graphical depictions distinctly reveal the consistent outperformance of IndoBERT over RoBERTa in both scenarios characterized by n values of 100 and 500. It is pivotal to recognize that the observed alterations in the parameter n also exert a discernible influence on the augmentation of scores for both models, thereby contributing to a nuanced understanding of the model dynamics in question answering tasks

Henceforth, it is deducible that IndoBERT demonstrates superiority over RoBERTa in the establishment of a resilient and effective Question Answering System (QAS) predicated on the utilization of a dataset in the Indonesian language. This discerning inference is drawn from a comprehensive analysis encompassing diverse evaluation metrics, including but not limited to F1-Score and Exact Match values, thereby substantiating the pragmatic viability of IndoBERT as a more proficient model for question answering tasks in the specified linguistic context. The pronounced disparities in performance underscore the nuanced intricacies influencing the efficacy of transformer models in the realm of natural language understanding and QAS development.

IV. CONCLUSION

The present investigation embarks upon a comprehensive juxtaposition of transformer methodologies, employing a dataset formatted in alignment with the Stanford Question Answering Dataset (SQuAD) and featuring information pertinent to Indonesian campuses. The ensuing comparative analysis yields discernible outcomes, unequivocally demonstrating the superior performance of the IndoBERT model over its counterpart, RoBERTa, specifically in the realm of question and answer processing. Additionally, based on the empirical findings, a strategic recommendation is posited, advocating for the augmentation of dataset size through the enrichment of individual data points. This strategic augmentation is proposed with the aim of potentially amplifying the prediction accuracy of the model, thus contributing to the ongoing refinement and optimization of question answering systems within the Indonesian language context.

Author Contributions: *Fedryanto Dartiko*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Mochammad Yusa*: Software, Investigation, Data Curation, Writing - Original Draft. *Aan Erlansari*: Investigation, Data Curation. *Shaikh Ameer Basha*: Review.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data cannot be openly shared for the protection of study participant privacy.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Fedryanto Dartiko: http://orcid.org/0000-0002-4126-9247
Mochammad Yusa: http://orcid.org/0000-0002-5550-5597
Aan Erlansari: http://orcid.org/0000-0002-4387-2226
Shaikh Ameer Basha: http://orcid.org/0000-0003-1641-4614

REFERENCES

[1] O. Sanllorente, R. Ríos-Guisado, L. Izquierdo, J. L. Molina, E. Mourocq, and J. D. Ibáñez-Álamo, “The importance of university campuses for the avian diversity of cities,” Urban For. Urban Green., vol. 86, p. 128038, 2023, doi: https://doi.org/10.1016/j.ufug.2023.128038.

[2] S. Colby, “HCRC 2022: A Novel Conference Approach for Disseminating Information on Assessing the Healthfulness of College Campuses,” J. Nutr. Educ. Behav., vol. 55, no. 7, Supplement, p. 108, 2023, doi: https://doi.org/10.1016/j.jneb.2023.05.230.

[3] A. Ansari, M. Maknojia, and A. Shaikh, “Intelligent question answering system based on artificial neural network,” in 2016 IEEE International Conference on Engineering and Technology (ICETECH), 2016, pp. 758–763. doi: 10.1109/ICETECH.2016.7569350.

[4] Y. Tan et al., “Research on Knowledge Driven Intelligent Question Answering System for Electric Power Customer Service,” Procedia Comput. Sci., vol. 187, pp. 347–352, 2021, doi: https://doi.org/10.1016/j.procs.2021.04.072.

[5] T. Shao, Y. Guo, H. Chen, and Z. Hao, “Transformer-Based Neural Network for Answer Selection in Question Answering,” IEEE Access, vol. 7, pp. 26146–26156, 2019, doi: 10.1109/ACCESS.2019.2900753.

[6] X. Hu, S. Zhu, and T. Peng, "Hierarchical attention vision transformer for fine-grained visual classification," J. Vis. Commun. Image Represent., vol. 91, p. 103755, 2023, doi: https://doi.org/10.1016/j.jvcir.2023.103755.

[7] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," COLING 2020 - 28th Int. Conf. Comput. Linguist. Proc. Conf., pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[9] H. Ngai, Y. Park, J. Chen, and M. Parsapoor, "Transformer-Based Models for Question Answering on COVID19," pp. 1–7, 2021, [Online]. Available: http://arxiv.org/abs/2101.11432

[10] Y. Wijaya, Rahmaddeni, and F. Zoromi, "Chatbot Designing Information Service for New Student Registration Based on AIML and Machine Learning," JAIA - J. Artif. Intell. Appl., vol. 1, no. 1, pp. 01–10, 2020, doi: 10.33372/jaia.v1i1.638.

[11] A. F. Hanifah and R. Kusumaningrum, "Non-Factoid Answer Selection in Indonesian Science Question Answering System using Long Short-Term Memory (LSTM)," Procedia Comput. Sci., vol. 179, pp. 736–746, 2021, doi: https://doi.org/10.1016/j.procs.2021.01.062.

[12] T. Noraset, L. Lowphansirikul, and S. Tuarob, "WabiQA: A Wikipedia-Based Thai Question-Answering System," Inf. Process. Manag., vol. 58, no. 1, p. 102431, 2021, doi: https://doi.org/10.1016/j.ipm.2020.102431.

[13] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, M. M. Anwar, and A. B. M. S. Ali, "Reading comprehension based question answering system in Bangla language with transformer-based learning," Heliyon, vol. 8, no. 10, p. e11052, 2022, doi: https://doi.org/10.1016/j.heliyon.2022.e11052.

[14] N. Volkmann, A. Riedel, N. Kemper, and B. Spindler, "Applied Research Note: Comparison of different methods for beak measurements in laying hens," J. Appl. Poult. Res., vol. 32, no. 4, p. 100373, 2023, doi: https://doi.org/10.1016/j.japr.2023.100373.

[15] Z. A. Guven and M. O. Unalir, "Natural language based analysis of SQuAD: An analytical approach for BERT," Expert Syst. Appl., vol. 195, p. 116592, 2022, doi: https://doi.org/10.1016/j.eswa.2022.116592.

[16] X. Zhou, D. Nurkowski, A. Menon, J. Akroyd, S. Mosbach, and M. Kraft, "Question answering system for chemistry—A semantic agent extension," Digit. Chem. Eng., vol. 3, p. 100032, 2022, doi: https://doi.org/10.1016/j.dche.2022.100032.

[17] Z. H. Syed, A. Trabelsi, E. Helbert, V. Bailleau, and C. Muths, "Question Answering Chatbot for Troubleshooting Queries based on Transfer Learning," Procedia Comput. Sci., vol. 192, pp. 941–950, 2021, doi: https://doi.org/10.1016/j.procs.2021.08.097.

[18] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," Molecules, vol. 26, no. 4, 2021, doi: 10.3390/molecules26041111.

[19] A. Apicella, F. Isgrò, A. Pollastro, and R. Prevete, "On the effects of data normalization for domain adaptation on EEG data," Eng. Appl. Artif. Intell., vol. 123, p. 106205, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106205.

[20] S. Ullah et al., "TNN-IDS: Transformer neural network-based intrusion detection system for MQTT-enabled IoT Networks," Comput. Networks, vol. 237, p. 110072, 2023, doi: https://doi.org/10.1016/j.comnet.2023.110072.

[21] S. Zhang, C. Lian, B. Xu, J. Zang, and Z. Zeng, "A token selection-based multi-scale dual-branch CNN-transformer network for 12-lead ECG signal classification," Knowledge-Based Syst., vol. 280, p. 111006, 2023, doi: https://doi.org/10.1016/j.knosys.2023.111006.

[22] Z. Zhang, J. David, and J. Liu, "Batch sizing control of a flow shop based on the entropy-function theorems," Expert Syst. Appl., vol. 213, p. 118958, 2023, doi: https://doi.org/10.1016/j.eswa.2022.118958.

[23] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, "Automated ICD coding using extreme multi-label long text transformer-based models," Artif. Intell. Med., vol. 144, p. 102662, 2023, doi: https://doi.org/10.1016/j.artmed.2023.102662.

[24] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in NeurIPS EMC^2 Workshop, 2019.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," CoRR, vol. abs/1810.0, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[26] H. Sharma and A. S. Jalal, "A survey of methods, datasets and evaluation metrics for visual question answering," Image Vis. Comput., vol. 116, p. 104327, 2021, doi: https://doi.org/10.1016/j.imavis.2021.104327.

[27] Y. Kim, S. Bang, J. Sohn, and H. Kim, "Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers," Autom. Constr., vol. 134, p. 104061, 2022, doi: https://doi.org/10.1016/j.autcon.2021.104061.

[28] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online: Association for Computational Linguistics, 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[29] N. B. Kurniawan and others, "A systematic literature review on survey data collection system," in 2018 International Conference on Information Technology Systems and Innovation (ICITSI), 2018, pp. 177–181. doi: 10.1109/ICITSI.2018.8696036.

[30] J. Lorenz, F. Barthel, D. Kienzle, and R. Lienhart, "Haystack: A Panoptic Scene Graph Dataset to Evaluate Rare Predicate Classes," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 62–70. doi: 10.1109/ICCVW60793.2023.00013.

[31] N. M. Kebonye, "Exploring the novel support points-based split method on a soil dataset," Measurement, vol. 186, p. 110131, 2021, doi: https://doi.org/10.1016/j.measurement.2021.110131.

[32] S. Wang et al., "Advances in Data Preprocessing for Biomedical Data Fusion: An Overview of the Methods, Challenges, and Prospects," Inf. Fusion, vol. 76, pp. 376–421, 2021, doi: https://doi.org/10.1016/j.inffus.2021.07.001.

[33] H. Zhu, H. Peng, Z. Lyu, L. Hou, J. Li, and J. Xiao, "Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation," Expert Syst. Appl., vol. 215, p. 119369, 2023, doi: https://doi.org/10.1016/j.eswa.2022.119369.

[34] G.-K. Wu, J. Xu, Y.-D. Zhang, B.-Y. Wen, and B.-P. Zhang, "Weighted feature fusion of dual attention convolutional neural network and transformer encoder module for ocean HABs classification," Expert Syst. Appl., vol. 243, p. 122879, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122879.

[35] C. M. Greco, A. Simeri, A. Tagarelli, and E. Zumpano, "Transformer-based language models for mental health issues: A survey," Pattern Recognit. Lett., vol. 167, pp. 204–211, 2023, doi: https://doi.org/10.1016/j.patrec.2023.02.016.