Khazanah Journal of Religion and Technology
Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
https://journal.uinsgd.ac.id/kjrt
Online ISSN: 2987-6060
Volume 2 Issue 2 (2024)
DOI: https://doi.org/10.15575/kjrt.v2i2.1093

IndoT5 (Text-to-Text Transfer Transformer) Algorithm for Paraphrasing Indonesian Language Islamic Sermon Manuscripts

Winda Puspitasari
Department of Informatics
UIN Sunan Gunung Djati Bandung
Jawa Barat, Indonesia
windapuspitasari250303@gmail.com

Dani Ramdani
Department of Informatics
UIN Sunan Gunung Djati Bandung
Jawa Barat, Indonesia
danuramdanu7@gmail.com

Aditya Muhamad Maulana
Department of Informatics
UIN Sunan Gunung Djati Bandung
Jawa Barat, Indonesia
adtyamuhamadmaulana@gmail.com

Received: November 18, 2024; Revised: December 18, 2024; Published: January 7, 2025

Abstract— Often sermons are considered uninteresting because the theme is not varied and repetitive. Although the theme is the same with different delivery can be a variation. The use of Natural Language Processing technology can be used to develop an automatic paraphrasing system in Indonesian that can help create variations in sermon manuscripts. This study focuses on the application of the IndoT5 (Text-to-Text Transfer Transformer) algorithm to build an automatic paraphrasing system in Indonesian. This system is evaluated using BLEU, ROUGE, and METEOR metrics to measure the similarity between the paraphrased text produced by the model and the desired target text. The evaluation results show a BLEU value of 0.28, ROUGE-1 of 0.59, ROUGE-2 of 0.40, ROUGE-L of 0.55, and METEOR of 0.50. This shows that the model is able to produce paraphrases with a moderate level of similarity and maintains fairly good semantic meaning. This model is also able to produce good variations in sermon manuscripts.

Keywords- IndoT5, Paraphrase, Sermon Manuscript, Text-to-Text Transfer Transformer.

I. INTRODUCTION

Language is a dynamic and complex means of communication, playing an important role in supporting human interaction in various fields. In the digital era, natural language processing (NLP) has become a core element of various technological applications, such as machine translation, question-and-answer systems, text summarization, and virtual assistants [1]. One of the fundamental aspects of NLP is the ability to paraphrase, which is the process of rephrasing a text using different words and structures without changing the original meaning [2]. Automatic paraphrasing technology has a significant role in supporting text simplification, increasing content creativity, and large-scale data management [3], especially in an era where accessibility of information is increasingly becoming a priority.

Indonesian, as one of the languages with the largest number of speakers in the world, has more than 275 million speakers spread across various regions [4], and have unique and complex linguistic characteristics. Different grammatical structures, rich vocabulary, and dialect variations present significant challenges in the development of NLP technology

[5]. These challenges include the limited availability of high-quality datasets, the complexity of language structures, and the need for large computational resources [6].

Based on a survey conducted by the Indonesian Internet Service Providers Association (APJII), the internet penetration rate in Indonesia reached 79.05% in early 2024 [7], which reflects the increasing need for local language-based technology to support content digitalization and facilitate technology accessibility. Therefore, innovative solutions are needed that not only improve accuracy and efficiency but are also able to handle the linguistic challenges of the Indonesian language specifically [8].

In recent years, the development of Transformer algorithm-based models has shown outstanding performance in various NLP tasks, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) [9] [10], which has dominated NLP research with its outstanding performance in understanding linguistic context and generating text [11]. One of the latest development innovations is the T5 (Text-to-Text Transfer Transformer) [12], a model designed to handle a variety of text-based tasks in a uniform manner [13]. For the Indonesian context, IndoT5, an adaptation of the T5 model, is presented as a potential solution that utilizes transfer learning on multilingual datasets [14].

However, research that focuses on the development and evaluation of the IndoT5 model for the task of automatic paraphrasing in Indonesian is still limited. There has been no comprehensive study evaluating the ability of this model to produce paraphrases that are accurate, coherent, and maintain the original meaning. This raises a critical question: how can IndoT5-based models be effectively adapted to meet the needs of automatic paraphrasing in Indonesian.

This study aims to develop an automatic paraphrasing system based on IndoT5 for Indonesians. The main focus is to evaluate the performance of the model in producing accurate, coherent paraphrases while maintaining the original meaning. This study uses automatic metrics such as BLEU, ROUGE, and METEOR as well as subjective evaluation to ensure the quality of the paraphrases produced [15]. In addition, this study also explores solutions to overcome challenges in developing automatic paraphrasing systems, such as dataset limitations and grammar complexity. The scope of this study includes evaluating the performance of IndoT5 on automatic paraphrasing tasks for Indonesians using existing datasets, without extensively developing new datasets.

With the implementation of an automatic paraphrasing system based on IndoT5, this study is expected to provide significant contributions to the development of NLP technology for the Indonesian language, especially for variations of sermon scripts. Friday sermons have an important role in providing spiritual, moral, and intellectual guidance for Muslims. Therefore, the quality of sermon scripts must always be considered to ensure that the message delivered is relevant, meaningful, and able to answer the challenges of the times. However, the use of monotonous and less varied sermon scripts can reduce the appeal and effectiveness of the message to be conveyed. Variations in sermon scripts are important to maintain the diversity of themes, increase creativity in delivery, and ensure that religious messages can be well received by various levels of the congregation who have different backgrounds, education, and needs [16]. With the support of technology such as automatic paraphrasing based on IndoT5, developing variations of sermon scripts becomes easier and more efficient. This technology is able to produce variations of texts that maintain the original meaning but with a fresher and more relevant language style. This not only helps preachers in composing more interesting and contextual scripts but also supports the spread of inclusive and adaptive Islamic messages in the digital era. For example, sermon scripts that discuss universal themes such as social justice, education, and the environment can be developed in various delivery styles, so that they are relevant to audiences in both urban and rural areas [17], [18], [19]. The results of this study not only support academic needs, but also have practical impacts, such as text simplification, digital content management, and increasing the accessibility of language-based technology. Furthermore, this study is expected to enrich the multilingual NLP research ecosystem globally and provide strategic insights for the development of local language-based technology.

II. RELATED WORKS

Several studies have been conducted in the application of the Text-to-Text Transfer Transformer (T5) model to various natural language processing (NLP) tasks, including text summarization, key phrase labeling, and automatic text generation. This review discusses research relevant to the development of an automatic paraphrasing system in Indonesian.

Research by I Nyoman Purnama et al. (2023) [20] explores the application of the T5 model for summarizing Indonesian language news documents. The researchers tested three preprocessing scenarios, including stemming and stopwords removal. The best results were obtained by applying stemming without stopwords removal, which resulted in a ROUGE-1 evaluation value of 0.17568. This study shows that proper preprocessing can improve the performance of the T5 model in understanding and summarizing Indonesian language texts.

Another study by Qurrota A'yuna Itsnaini et al. (2023) [21] using a pre-trained T5 model for the abstractive task of summarizing text in Indonesian. The evaluation was conducted using the ROUGE metric with an Indonesian news dataset. The best results obtained were a ROUGE-1 value of 0.68, ROUGE-2 of 0.61, and ROUGE-L of 0.65. Although these results are quite good, this study found several weaknesses in text abstraction, mainly due to the limitations of the dataset used.

In the context of keyphrase labeling, research by Jorge Gabín et al. (2024) [22], propose a docT5keywords model to generate and filter key phrases using T5 architecture. This model uses document title and abstract as input and generates key phrases through classical inference approach and majority voting. The proposed filtering technique also improves the accuracy by eliminating false positives significantly.

Terakhir, penelitian oleh Mohammad Yani et al. (2023) [23], mengembangkan aplikasi peringkas teks bahasa Indonesia yang menerima masukan dalam berbagai format, seperti teks, file dokumen, laman web, dan gambar. Model T5 digunakan untuk menghasilkan ringkasan yang dapat disimpan dalam berbagai format file. Hasil evaluasi menunjukkan bahwa masukan berbentuk dokumen atau teks menghasilkan nilai *ROUGE* yang lebih tinggi dibandingkan masukan non-teks, mencapai rata-rata 0,87.

Berbeda dengan penelitian sebelumnya yang berfokus pada tugas seperti peringkasan teks dan pelabelan frasa kunci, penelitian ini mengadaptasi model IndoT5 untuk pengembangan sistem parafrase otomatis dalam bahasa Indonesia. Dengan demikian, penelitian ini bertujuan untuk mengisi kesenjangan dalam penelitian sebelumnya dengan menghadirkan solusi yang lebih kontekstual dan sesuai untuk tugas parafrase teks dalam bahasa Indonesia.

III. Research Methods

A. *IndoT5 (Text-to-Text Transfer Transformer*)

IndoT5, a variant of the Text-to-Text Transfer Transformer (T5), is designed to enhance various natural language processing tasks, particularly in the Indonesian language. This model excels in text summarization, keyphrase generation, and even drug-target interaction predictions, showcasing its versatility and effectiveness across different applications. IndoT5 has capabilities for text summarization and keyphrase generation.

IndoT5 effectively summarizes lengthy texts, such as news articles and academic papers, by processing inputs in multiple formats (text, documents, web pages) [24]. The model has demonstrated high performance, achieving ROUGE scores of 0.87 for document inputs, indicating its ability to retain essential information while condensing content [24], [25].

In keyphrase generation, the IndoT5 architecture facilitates automatic keyphrase generation, producing relevant phrases that encapsulate document content based on titles and abstracts [26]. This model significantly outperforms traditional extractive methods, achieving over 100% improvement in some evaluations.

IndoT5, a variant of the Text-to-Text Transfer Transformer (T5), is effectively utilized for paraphrasing tasks, demonstrating significant capabilities in generating semantically similar sentences. This model excels in both paraphrase generation and identification, leveraging its transformer architecture to produce fluent outputs. The following sections detail its applications, methodologies, and performance metrics. The model generates multiple paraphrases for a given input, enhancing clarity and variety in expression [27]. IndoT5 is fine-tuned on specific datasets, improving its performance in generating coherent and contextually relevant paraphrases [28]. Hybrid Models: Some studies combine IndoT5 with other architectures, like seq2seq models, to enhance its paraphrasing capabilities by capturing long-term dependencies [29]. Performance Metrics Evaluation Metrics: The effectiveness of IndoT5 is measured using metrics like ROUGE and BLEU, with reported scores indicating strong performance in generating paraphrases [27], [28].

B. *CRISP-DM*

The Cross-Industry Standard Process for Data Mining or CRISP-DM, is a process model that is widely used in data mining and is not dependent on a particular industry [30]. In this study, the author used the stages in the CRISP-DM method.

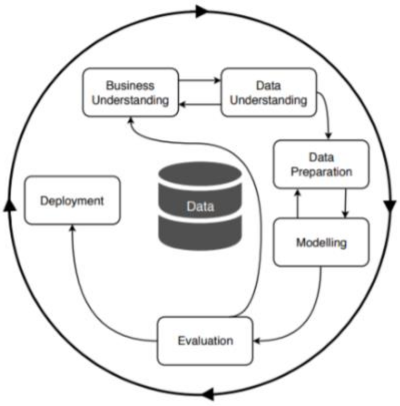


Fig. 1. CRISP-DM Methodology

a. **Business Understanding**, understand the problem topics related to the development of an automatic paraphrasing system in Indonesian. This study aims to apply the IndoT5 (Text-to-Text Transfer Transformer) algorithm to produce a system capable of automatically paraphrasing text, evaluating the quality of the paraphrasing results, and analyzing the effectiveness of the algorithm in understanding and reconstructing the meaning of sentences in Indonesian.
b. **Data Understanding**, takes data sources from the Hugging Face platform using the id-paraphrase-detection dataset provided by Jakarta Research. This dataset contains text pairs in Indonesian marked as paraphrase or not. Each text pair is treated as a data unit for analysis, including two sentences being compared and a classification label (paraphrase or non-paraphrase) as a reference for evaluating system performance.
c. **Data Preparation**, data selection needs to be done by setting inclusion and exclusion criteria to check the quality and relevance of the data. This process includes basic statistical exploration, such as the total number of sentence pairs, label distribution (paraphrase and non-paraphrase), and examination of sentence structure diversity.

d. **Modelling**, using the IndoT5 (Text-to-Text Transfer Transformer) model to build an automatic paraphrasing system based on processed data. This model is trained to understand text input in Indonesian and produce paraphrased text as output while maintaining the same meaning. IndoT5 works with a sequence-to-sequence approach to reconstruct sentences in a different form but still semantically consistent.
e. **Evaluation**, conducted a model evaluation using BLEU, ROUGE, and METEOR metrics to measure the performance of the IndoT5 (Text-to-Text Transfer Transformer) algorithm to evaluate the extent to which the model output can maintain the meaning and quality of the text compared to available references.
f. **Deployment**, carry out the deployment stage by integrating the trained IndoT5 (Text-to-Text Transfer Transformer) model into the application system. After the model is implemented, a web-based interface is also created to make it easier for users to access this paraphrasing system. This website is designed so that users can process text input directly through a form on the web page and get output in the form of text paraphrasing results automatically. Then the application is used to try to paraphrase the sermon manuscript.

IV. Result and Discussion

A. Result of Business Understanding

Business understanding is a study of the research topic being conducted. In this study, the researcher focuses the study on the development of an automatic paraphrasing system for Indonesian by applying the IndoT5 (Text-to-Text Transfer Transformer) algorithm. The development of this paraphrasing system aims to provide a solution in producing accurate and high-quality paraphrased text, especially in natural language processing (NLP) for Indonesian.

Automatic paraphrasing has an important role in various fields, such as education, content writing, and information technology. This system can help users restructure sentences or documents without changing the original meaning, thereby increasing productivity and efficiency in content creation. With the IndoT5 algorithm, which is specifically designed to understand the structure and context of Indonesian, this study focuses on improving the quality of language modeling so that the system can paraphrase more effectively.

This research is important to answer the challenges in processing natural Indonesian language, which often has its own complexities such as synonyms, varying sentence structures, and diverse contexts. The results of this study are expected to provide a significant contribution to the development of NLP-based technology in Indonesia, as well as support the application of paraphrasing systems in various sectors that require automatic text processing.

B. Result of Data Understanding

Collecting data from various sources, checking and characterizing the data, and analyzing its quality are important tasks in this phase [30]. In this study, data was obtained from a dataset available on Hugging Face, namely id-paraphrase-detection developed by Jakarta Research. This dataset is used as a basis for developing and evaluating an automatic paraphrasing system based on the IndoT5 algorithm.

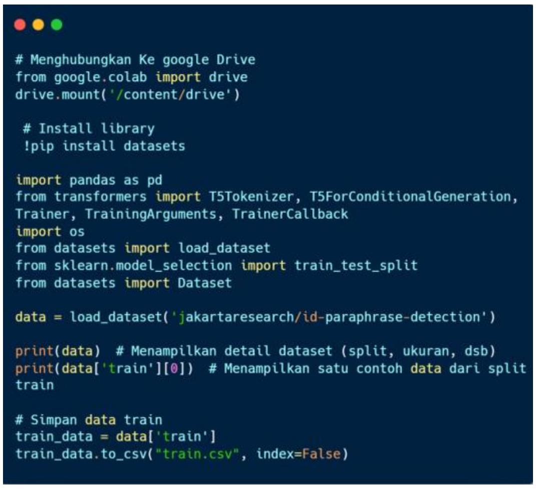
The id-paraphrase-detection dataset contains pairs of sentences in Indonesian that are classified as paraphrases or not. This data covers various variations of sentence structures and contexts of Indonesian usage, so that it can help the model understand the semantic relationship between sentences. Data collection was carried out by accessing the Hugging Face repository using Python and supporting libraries such as datasets and transformers.

After the data has been successfully downloaded, the next step is to check the quality and relevance of the data. This process includes basic statistical exploration, such as the total number of sentence pairs, label distribution (paraphrase and non-paraphrase), and examination of sentence structure diversity. This analysis aims to ensure that the data has adequate quality and can be used effectively in the training and evaluation process of the IndoT5 model for automatic paraphrasing tasks.

C. Proses Pengambilan Data

This study uses the id-paraphrase-detection dataset obtained from Hugging Face, a platform that provides various datasets and natural language processing models. This dataset was developed by Jakarta Research and designed to support the task of paraphrase detection in Indonesian.

The id-paraphrase-detection dataset consists of pairs of sentences in Indonesian that have been categorized into two labels, namely paraphrase and non-paraphrase. Each pair of sentences is equipped with an annotation indicating whether the two sentences have similar or different meanings. This dataset has a variety of sentence structures and contexts, so it greatly supports research related to the development of automatic paraphrasing systems.

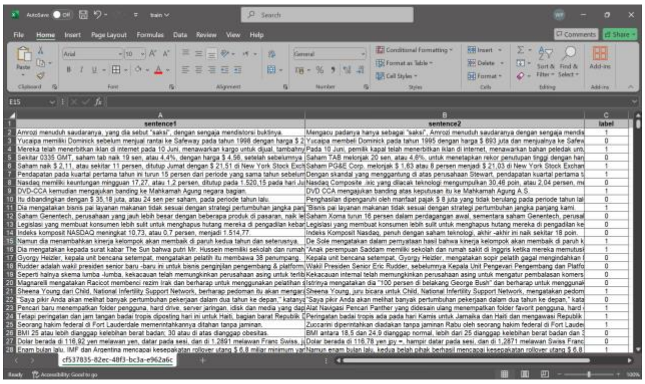


```
# Menghubungkan Ke google Drive
from google.colab import drive
drive.mount('/content/drive')

 # Install library
 !pip install datasets

import pandas as pd
from transformers import T5Tokenizer, T5ForConditionalGeneration,
Trainer, TrainingArguments, TrainerCallback
import os
from datasets import load_dataset
from sklearn.model_selection import train_test_split
from datasets import Dataset

data = load_dataset('jakartaresearch/id-paraphrase-detection')

print(data)  # Menampilkan detail dataset (split, ukuran, dsb)
print(data['train'][0])  # Menampilkan satu contoh data dari split
train

# Simpan data train
train_data = data['train']
train_data.to_csv("train.csv", index=False)
```

Fig. 2. Data Collecting

The dataset obtained will be used in the training, validation, and testing process of the IndoT5 (Text-to-Text Transfer Transformer) model. Before being used, this data will be processed through text pre-processing stages, such as tokenization, normalization, and data cleaning, to ensure optimal input quality for the model.

Figure 2 shows the program code for downloading the id-paraphrase-detection dataset from Hugging Face using the Python programming language. This process involves taking sentence pair data that has been labeled paraphrase and non-paraphrase, then continued with the text pre-processing stage. After getting the id-paraphrase-detection dataset from Hugging Face, the data is saved in CSV format. The next stage is Text Preprocessing to prepare the data before being entered into the model. It can be seen in Figure 3 which is the result of data retrieval with a total of 4,076 data.



	A	B	C
1	sentence1	sentence2	label
2	Amrozi menuduh saudaranya, yang dia sebut "saksi", dengan sengaja mendistorsi buktinya.	Mengacu padanya hanya sebagai "saksi", Amrozi menuduh saudaranya dengan sengaja mendis	1
3	Yucaipa memiliki Dominick sebelum menjual rantai ke Safeway pada tahun 1998 dengan harga $ 2	Yucaipa membeli Dominick pada tahun 1995 dengan harga $ 693 juta dan menjualnya ke Safew	0
4	Mereka telah menerbitkan iklan di internet pada 10 Juni, menawarkan kargo untuk dijual, tambahny	Pada 10 Juni, pemilik kapal telah menerbitkan iklan di internet, menawarkan bahan peledak unt	1
5	Sekitar 0335 GMT, saham tab naik 19 sen, atau 4,4%, dengan harga $ 4,56, setelah sebelumnya	Saham TAB melonjak 20 sen, atau 4,6%, untuk menetapkan rekor penutupan tinggi dengan har	0
6	Saham naik $ 2,11, atau sekitar 11 persen, ditutup Jumat dengan $ 21,51 di New York Stock Exch	Saham PG&E Corp. melonjak $ 1,63 atau 8 persen menjadi $ 21,03 di New York Stock Exchan	1
7	Pendapatan pada kuartal pertama tahun ini turun 15 persen dari periode yang sama tahun sebelum	Dengan skandal yang menggantung di atas perusahaan Stewart, pendapatan kuartal pertama t	1
8	Nasdaq memiliki keuntungan mingguan 17,27, atau 1,2 persen, ditutup pada 1.520,15 pada hari Ju	Nasdaq Composite .ixic yang dilacak teknologi mengumpulkan 30,46 poin, atau 2,04 persen, m	0
9	DVD-CCA kemudian mengajukan banding ke Mahkamah Agung negara bagian.	DVD CCA mengajukan banding atas keputusan itu ke Mahkamah Agung A.S.	1
10	Itu dibandingkan dengan $ 35,18 juta, atau 24 sen per saham, pada periode tahun lalu.	Penghasilan dipengaruhi oleh manfaat pajak $ 8 juta yang tidak berulang pada periode tahun lal	0
11	Dia mengatakan bisnis pai layanan makanan tidak sesuai dengan strategi pertumbuhan jangka pan	"Bisnis pai layanan makanan tidak sesuai dengan strategi pertumbuhan jangka panjang kami.	1
12	Saham Genentech, perusahaan yang jauh lebih besar dengan beberapa produk di pasaran, naik le	Saham Xoma turun 16 persen dalam perdagangan awal, sementara saham Genentech, perusa	0
13	Legislasi yang membuat konsumen lebih sulit untuk menghapus hutang mereka di pengadilan kebar	Legislasi yang membuat konsumen lebih sulit untuk menghapus hutang mereka di pengadilan ke	0
14	Indeks komposit NASDAQ meningkat 10,73, atau 0,7 persen, menjadi 1.514,77.	Indeks Komposit Nasdaq, penuh dengan saham teknologi, akhir -akhir ini naik sekitar 18 poin.	0
15	Namun dia menambahkan kinerja kelompok akan membaik di paruh kedua tahun dan seterusnya.	De Sole mengatakan dalam pernyataan hasil bahwa kinerja kelompok akan membaik di paruh k	1
16	Dia mengatakan kepada surat kabar The Sun bahwa putri Mr. Hussein memiliki sekolah dan rumah	"Anak perempuan Saddam memiliki sekolah dan rumah sakit di Inggris ketika mereka memutusk	1
17	Gyorgy Heizler, kepala unit bencana setempat, mengatakan pelatih itu membawa 38 penumpang.	Kepala unit bencana setempat, Gyorgy Heizler, mengatakan sopir pelatih gagal mengindahkan	0
18	Rudder adalah wakil presiden senior baru -baru ini untuk bisnis penginjilan pengembang & platform	Wakil Presiden Senior Eric Rudder, sebelumnya Kepala Unit Pengevari Pengembang dan Platfo	0
19	Seperti halnya skema lumba -lumba, kekacauan telah memungkinkan perusahaan asing untuk terlib	Kekacauan internal telah memungkinkan perusahaan asing untuk mengatur pembalasan komers	0
20	Magnarelli mengatakan Racicot membenci rezim Irak dan berharap untuk menggunakan pelatihan s	Istrinya mengatakan dia "100 persen di belakang George Bush" dan berharap untuk menggunak	0
21	Sheena Young dari Child, National Infertility Support Network, berharap pedoman itu akan mengar	Sheena Young, juru bicara untuk Child, National Infertility Support Network, mengatakan pedom	1
22	"Saya pikir Anda akan melihat banyak pertumbuhan pekerjaan dalam dua tahun ke depan," katanya	"Saya pikir Anda akan melihat banyak pertumbuhan pekerjaan dalam dua tahun ke depan," kata	0
23	Pencari baru menempatkan folder pengguna, hard drive, server jaringan, idisk dan media yang dap	Alat Navigasi Pencari Panther yang didesain ulang menempatkan folder favorit pengguna, hard	1
24	Tetapi peringatan dan jam tangan badai tropis diposting hari ini untuk Haiti, bagian barat Republik D	Peringatan badai tropis ada pada hari Kamis untuk Jamaika dan Haiti dan mengawasi Republik	1
25	Seorang hakim federal di Fort Lauderdale memerintahkannya ditahan tanpa jaminan.	Zuccarini diperintahkan diadakan tanpa jaminan Rabu oleh seorang hakim federal di Fort Laude	1
26	BMI 25 atau lebih dianggap kelebihan berat badan; 30 atau di atas dianggap obesitas.	BMI antara 18,5 dan 24,9 dianggap normal, lebih dari 25 dianggap kelebihan berat badan dan 3	0
27	Dolar berada di 116,92 yen melawan yen, datar pada sesi, dan di 1,2891 melawan Franc Swiss, j	Dolar berada di 116,78 yen jpy =, hampir datar pada sesi, dan di 1,2871 melawan Swiss Franc	0
28	Enam bulan lalu, IMF dan Argentina mencapai kesepakatan rollover utang $ 6,8 milar minimum yar	Namun enam bulan lalu, kedua belah pihak berhasil mencapai kesepakatan rollover utang $ 6,8	1

Fig. 3. Example of Document Collecting Result

D. Result of Data Preparation

Data selection needs to be done by setting inclusion and exclusion criteria to check the quality and relevance of the data. This process includes basic statistical exploration, such as the total number of sentence pairs, label distribution (paraphrase and non-paraphrase), and examination of sentence structure diversity. Data cleaning steps are carried out with the aim of:

- Displaying sentence pairs labeled 1, indicating that the second sentence is an exact paraphrase of the first sentence;
- Removing the label column to focus the analysis on the text itself;
- Replacing the sentence column to ensure consistent formatting across the data.

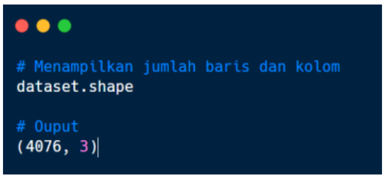
After the cleaning process, weighting is carried out to assess the importance of each word in the sentence, as well as to ensure that the data is ready to be used for further modeling. The data preparation stage in this study was carried out to prepare the dataset so that it can be used in the text analysis process. This process consists of several steps, including loading the dataset, adjusting the dataset, and storing the filtered dataset.

Step 1: Load Dataset

The first step is to load the dataset used in the study. The dataset is downloaded from Google Drive via the link provided. This process involves checking the number of rows and columns, duplicate data, and empty values.

1. Number of Rows and Columns

The loaded dataset has the number of rows and columns displayed to ensure the data is complete and according to research needs. The results of checking the number of rows and columns can be seen in Figure 4 below.



```
# Menampilkan jumlah baris dan kolom
dataset.shape

# Ouput
(4076, 3)
```

Fig. 4. Number of Row and Column

From Figure 4, it can be seen that the dataset has 4,076 rows and 3 columns. This information shows that the dataset has a large enough amount of data to be analyzed.

2. Duplication Check

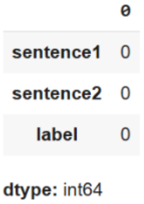
This process is done to ensure that there is no identical or duplicate data in the dataset. Based on the results of the check, a number of duplicate rows were found which are displayed in Table 1. In Table 1, the same rows in the sentence1 and sentence2 columns are identified as duplicate data.

Table 1. Duplication Checking

No	Sentence1	Sentence2	Label
2464	Werdegar juga mengatakan server Intel tidak dirugikan oleh pesan komputer dan ribuan penerima dapat meminta agar email berhenti, yang dihormati Hamidi.	Dia juga mencatat bahwa Tenet telah memberikan komite 19 volume dokumen tentang intelijen sebelum perang yang telah tersedia bagi anggota DPR.	1
3836	Jutaan barel lagi, yang dibeli oleh kilang Spanyol Cepsa SA, akan dimuat ke kapal tanker Spanyol, Sandra tapias.	Sebuah kapal tanker Spanyol, Sandra Tabias, akan sarat dengan jutaan barel, dibeli oleh kilang Spanyol Cepsa SA, pada sore hari.	1

3. Checking Empty Values

The dataset is also checked for empty values. If there are empty values, the data will be processed further in the cleaning stage. The results of checking for empty values can be seen in Figure 5.



```
           0
sentence1  0
sentence2  0
    label  0

dtype: int64
```

Fig. 5. Empty Value Checking

Based on Figure 5, it can be seen that all columns in the dataset, namely sentence1, sentence2, and label, have an empty value of 0. This indicates that the dataset is complete and does not require further imputation or handling of empty values.

Stage 2: Dataset Adjustment

Once the dataset is loaded and checked, the next step is to customize the dataset according to the research needs. This process includes:

1. Counting the Number of Labels

The dataset is divided based on labels, namely label 1 for paraphrase and label 0 for non-paraphrase. The results of calculating the number of data based on labels are shown as follows:

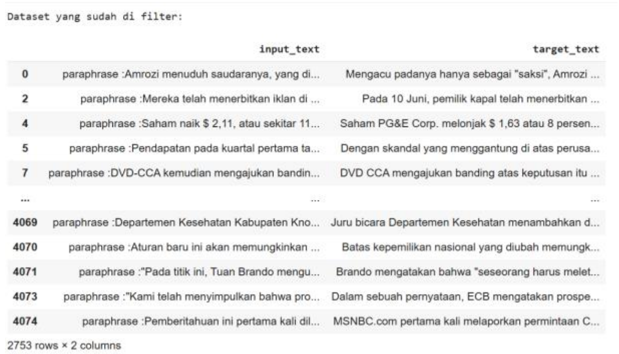
- Number of data with label = 1: 2,753 data
- Number of data with label = 0: 1,323 data

2. Filter Dataset

Data with label 1 is filtered for further research. This dataset is then modified by adding two new columns, namely input_text and target_text.

- The input_text column contains the text of the first sentence (sentence1) with the label "paraphrase:" added.
- The target_text column contains the text of the second sentence (sentence2).

The results of the dataset filter are shown in Figure 6.



Dataset yang sudah di filter:

	input_text	target_text
0	paraphrase :Amrozi menuduh saudaranya, yang di...	Mengacu padanya hanya sebagai "saksi", Amrozi ...
2	paraphrase :Mereka telah menerbitkan iklan di ...	Pada 10 Juni, pemilik kapal telah menerbitkan ...
4	paraphrase :Saham naik $ 2,11, atau sekitar 11...	Saham PG&E Corp. melonjak $ 1,63 atau 8 persen...
5	paraphrase :Pendapatan pada kuartal pertama ta...	Dengan skandal yang menggantung di atas perusa...
7	paraphrase :DVD-CCA kemudian mengajukan bandin...	DVD CCA mengajukan banding atas keputusan itu ...
...	...	...
4069	paraphrase :Departemen Kesehatan Kabupaten Kno...	Juru bicara Departemen Kesehatan menambahkan d...
4070	paraphrase :Aturan baru ini akan memungkinkan ...	Batas kepemilikan nasional yang diubah memungk...
4071	paraphrase :"Pada titik ini, Tuan Brando mengu...	Brando mengatakan bahwa "seseorang harus melet...
4073	paraphrase :"Kami telah menyimpulkan bahwa pro...	Dalam sebuah pernyataan, ECB mengatakan prospe...
4074	paraphrase :Pemberitahuan ini pertama kali dil...	MSNBC.com pertama kali melaporkan permintaan C...

2753 rows × 2 columns

Fig. 6. Data Filtering

Based on Figure 6, the dataset has been successfully filtered by only including data labeled 1. The input_text column combines the label "paraphrase" with the values from the sentence1 column, while the target_text column contains the values from the sentence2 column.

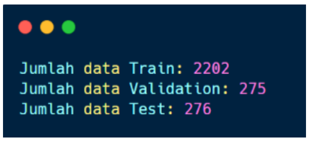
3. Dataset Storage

The filtered dataset is then saved in CSV format with the file name filtered_dataset_paraphrase.csv. This dataset storage is done to ensure the data is ready to be used in the next stage.

E. *Modelling Result*

Data modeling involves selecting a modeling technique, constructing test cases, and creating a model. explaining the reasons for the selection. Specific parameters must be set to build the model [30]. In modeling with the IndoT5 (Text-to-Text Transfer Transformer) algorithm, it involves the process of dataset processing, tokenization, and model training. The final model results along with the tokenizer are stored in a specific directory to facilitate the use of the model in the implementation of the Indonesian language automatic paraphrasing system.

This stage is carried out to divide the filtered dataset into three main subsets, namely Train, Validation, and Test. This division aims to ensure that the model can be trained, validated, and tested separately, thereby reducing the risk of overfitting. The dataset division process is carried out with a ratio of 80:10:10 using the train_test_split function. The results of the amount of data for each subset can be seen in Figure 7.



```
Jumlah data Train: 2202
Jumlah data Validation: 275
Jumlah data Test: 276
```

Fig. 7. Number of Training, Validation, and Testing Data

Based on Figure 7, the dataset is divided into:

- Train: 2,202 data, used to train the model.
- Validation: 275 data, were used to evaluate the model's performance during the training process.
- Test: 276 data, used to test the model after the training process is complete.

This division ensures that each subset has non-overlapping data, making model evaluation more accurate.

- **Tokenization**

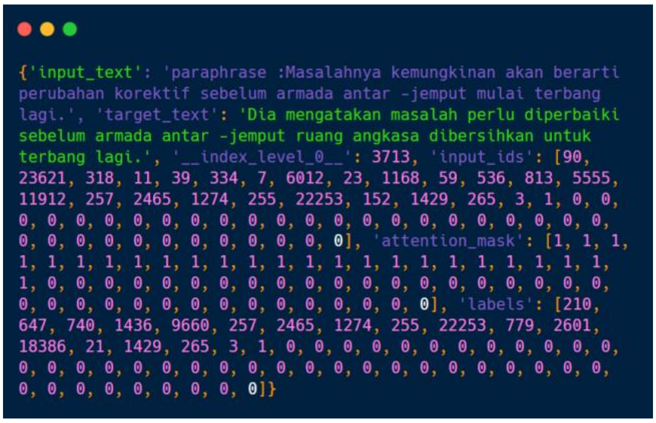
At this stage, the tokenization process is carried out to convert text into a numeric representation using the IndoT5-base-paraphrase tokenizer model. This tokenizer is used to process text in the input_text and target_text columns, with the maximum token length parameter (max_length) adjusted based on the longest tokenization length in the training data. The tokenization process also involves truncation (truncation if the text length exceeds the maximum limit) and padding (addition of padding so that all texts have the same length). In addition, tokenization produces a numeric token representation for the label by adjusting the input in the input_ids column. The tokenization results can be seen in Figure 8.

Based on Figure 8, the input text is successfully converted into a numeric token representation. The result of this tokenization ensures that the data is ready to be used by the IndoT5-base-paraphrase model for the training and validation process. The token representation includes:

- Input IDs: Token representation of the input text.

- Labels: Token representation of the target text, used for the training process.
- Attention Mask: Used to mark the active token position in the input sequence.

This process is an important stage in data preprocessing to ensure that the data is compatible with the transformer-based model architecture.



```
{'input_text': 'paraphrase :Masalahnya kemungkinan akan berarti
perubahan korektif sebelum armada antar -jemput mulai terbang
lagi.', 'target_text': 'Dia mengatakan masalah perlu diperbaiki
sebelum armada antar -jemput ruang angkasa dibersihkan untuk
terbang lagi.', '__index_level_0__': 3713, 'input_ids': [90,
23621, 318, 11, 39, 334, 7, 6012, 23, 1168, 59, 536, 813, 5555,
11912, 257, 2465, 1274, 255, 22253, 152, 1429, 265, 3, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'labels': [210,
647, 740, 1436, 9660, 257, 2465, 1274, 255, 22253, 779, 2601,
18386, 21, 1429, 265, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0]}
```

Fig. 8. Data Filtering

- **Fine Tuning IndoT5 Model**

This stage aims to train the IndoT5-base-paraphrase model using the tokenized dataset. The fine-tuning process involves configuring training parameters through the TrainingArguments class, which includes setting the output directory, batch size, number of epochs, learning rate, and mixed precision (fp16) to improve training efficiency.

The model is trained using the Trainer API, which manages the training and evaluation process based on the training and validation datasets. Here is a summary of the fine-tuning process and results:

- Model Used: IndoT5-base-paraphrase
- Main Training Parameters:
  - Batch size for training and evaluation: 8
  - Number of epochs: 5
  - Learning rate: 3e-5
  - Evaluation strategy: Every end of epoch

The fine-tuning process results in a model that is saved in the destination directory, along with the adapted tokenizer. The training results are shown through the training loss and validation loss metrics for each epoch, as shown in Table 2.

Table 2. IndoT5 in Dataset Fine Tuning

No	Training Loss	Validation Loss
1	0.448100	0.855902
2	0.414000	0.873149
3	0.442700	0.867400
4	0.481600	0.852144
5	0.518400	0.842879

Based on Table 2, the training loss and validation loss values tend to be stable after several epochs, indicating that the model has learned optimally from the dataset without any indication of overfitting. The final model has an average training loss of 0.4493, with a validation loss value at the last epoch of 0.8429. This trained model is ready to be used for text paraphrasing tasks, with better capabilities because it is adjusted to the given dataset.

F. *Evaluation*

At the evaluation stage, the evaluation results are based on the research objectives that have been set. Therefore, the results must be interpreted and further actions need to be determined [30]. The purpose of evaluation is to compare the output of the automatic paraphrasing system with the provided reference. Metrics such as BLEU, ROUGE, and METEOR can be used to assess the quality of the model in preserving the meaning and structure of sentences. BLEU measures the similarity between the model's output text and the reference, ROUGE evaluates the information coverage, and METEOR measures the semantic closeness between the model's text and the reference.

After the training process is complete, the fine-tuned model is evaluated using a test dataset to measure its performance in the text paraphrasing task. The evaluation is done by calculating several key metrics, namely BLEU, ROUGE, and METEOR, which measure the similarity between the paraphrased text generated by the model and the desired target text.

- **Evaluation Process**
  1. Text Paraphrasing
     The model processes the input text from the test dataset using the paraphrase function, producing paraphrased text that is compared to the target text (ground truth).
  2. Prediction and Reference Extraction
     The paraphrased text (prediction) is extracted to be compared to the reference (target text) from the test dataset.
  3. Evaluation Metric Calculation
     - BLEU, measures the degree of similarity based on word alignment and sequence.
     - ROUGE, assesses similarity using n-gram overlap calculation, including ROUGE-1, ROUGE-2, and ROUGE-L.
     - METEOR, combines word alignment with synonymy and grammatical flexibility.

- **Evaluation Result**
  The following are the evaluation results obtained:

1. Example of paraphrasing results on a dataset

Table 3 presents examples of original text, targets, and paraphrase results produced by the model.

Table 3. Paraphrase Result with IndoT5 Model

Original Text	Target Text	Paraphrased Text
Dia memproyeksikan Vanderpool akan	Produk yang menampilkan	Dia mengatakan bahwa Vanderpool

Original Text	Target Text	Paraphrased Text
tersedia dalam lima tahun ke depan.	Vanderpool akan dirilis dalam waktu lima tahun, katanya.	akan tersedia dalam lima tahun ke depan.
Terlepas dari perbedaan mereka, para pemimpin AS dan Uni Eropa mengatakan ada bidang kesepakatan.	Terlepas dari perbedaan ini dan lainnya, para pemimpin AS dan Uni Eropa bersikeras mereka bekerja sama dengan baik.	Terlepas dari perbedaan mereka, para pemimpin AS dan Uni Eropa mengatakan ada bidang kesepakatan di antara mereka.
John Logsdon, seorang anggota dewan, mengatakan bahwa "Spaceflight manusia telah menjadi tempat di mana perbedaan pendapat tidak diterima.	Faktanya, kata John Logsdon, seorang anggota dewan dan pakar kebijakan ruang angkasa Universitas George Washington, "Spaceflight manusia telah menjadi tempat di mana perbedaan pendapat tidak diterima.	Spacefight manusia telah menjadi tempat di mana perbedaan pendapat tidak diterima.
Itu berarti Demokrat dapat memblokir calon Mahkamah Agung melalui filibuster jika mereka bisa mendapatkan 40 anggota mereka untuk menyetujui.	Tetapi Demokrat dapat memblokir calon potensial melalui filibuster jika mereka bisa mendapatkan 41 suara.	Ini berarti bahwa Demokrat dapat memblokir calon Mahkamah Agung melalui filibuster jika mereka bisa mendapatkan 40 anggota
Charles Howell III mengambil beberapa pengetahuan lokal setahun yang lalu yang memberikan wawasan yang sangat dibutuhkan di babak pembukaan turnamen peringatan di Dublin, Ohio.	Charles Howell III mengambil beberapa pengetahuan lokal setahun yang lalu yang memberikan wawasan yang sangat dibutuhkan di babak pembukaan turnamen peringatan hari Kamis.	Charles Howell mengambil beberapa pengetahuan lokal setahun yang lalu, memberikan wawasan yang sangat dibutuhkan di babak pembukaan

Based on Table 3 above, it can be seen that the model can produce paraphrased text that is close to the target text while maintaining the main meaning and sentence structure. However, in some cases, there are minor differences in word choice or sentence order, which are still within reasonable limits for the paraphrasing task.

2. Evaluation Metrics

To evaluate the performance of the paraphrasing model that has been generated, three main evaluation metrics are used, namely BLEU, ROUGE, and METEOR. Table 4 shows the results of the quantitative evaluation scores based on the test dataset used. Figure 9 visualizes the results of the evaluation of the paraphrasing model metrics that have been analyzed.

Table 4. Evaluation Result

No	Metrik Evaluasi	Hasil
1	*BLEU*	0.28
2	*ROUGE-1*	0.59
3	*ROUGE-2*	0.40
4	*ROUGE-L*	0.55
5	*ROUGE-Lsum*	0.55
6	*METEOR*	0.50

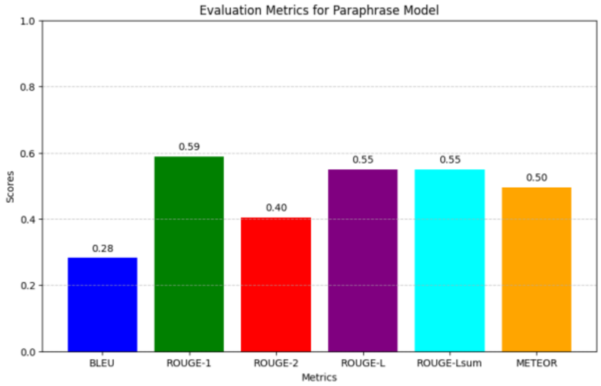


Fig. 9. Visualization of Evaluation Result

From the evaluation results above, it can be seen that the model shows good performance with quite high scores on the BLEU, ROUGE, and METEOR metrics. This indicates the model's ability to produce relevant and reference-based paraphrased texts, although there is room for further improvement in the future.

G. Deployment Result

In the deployment stage, the trained model is implemented into the system to be used directly by users. Therefore, the deployment process must be designed by considering the production environment and the research objectives that have been set [30]. The purpose of deployment is to ensure that the model can work optimally in real-world scenarios. This stage involves integrating the model with the application, creating a web-based interface for user interaction, testing performance in a production environment, and monitoring model performance to maintain consistency of results. In addition, post-deployment evaluation is carried out to ensure that the model can provide relevant, quality output that is in accordance with user needs. The deployment stage is carried out to integrate the paraphrase model into a system that can be used by users. The trained IndoT5 model is stored in a specific directory so that it can be reloaded without the need for retraining. The deployment process involves several main steps as follows:

1. Loading Model and Tokenizer

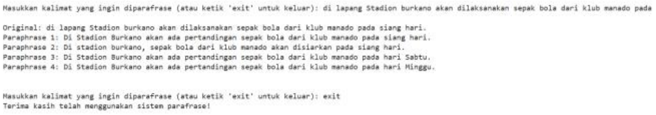
The IndoT5 model and tokenizer are loaded from a specified directory. The model is then moved to an available device, i.e. GPU if available or CPU as an alternative.

2. Functions for Inference

The generate_paraphrase function is designed to generate paraphrased sentences. This function uses sampling techniques such as top-k, top-p, and parameter settings such as num_beams and temperature to generate various paraphrase variations. Sentence input is processed by adding the prompt "paraphrase:" to match the model training format.

3. Paraphrase Results

The system is able to accept sentence input from the user, process it, and generate four versions of the paraphrased sentence. An example of the output from the paraphrasing process can be seen in Figure 10.



```
Masukkan kalimat yang ingin diparafrase (atau ketik 'exit' untuk keluar): di lapang Stadion burkano akan dilaksanakan sepak bola dari klub manado pada

Original: di lapang Stadion burkano akan dilaksanakan sepak bola dari klub manado pada siang hari.
Paraphrase 1: Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada siang hari.
Paraphrase 2: Di stadion burkano, sepak bola dari klub manado akan disiarkan pada siang hari.
Paraphrase 3: Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada hari Sabtu.
Paraphrase 4: Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada hari Minggu.


Masukkan kalimat yang ingin diparafrase (atau ketik 'exit' untuk keluar): exit
Terima kasih telah menggunakan sistem parafrase!
```

Fig. 10. Example of Paraphrase Result

In Figure 10, is an example of the results of paraphrasing using the IndoT5 model with the following explanation:

- In the example of the input sentence: "*di lapang Stadion burkano akan dilaksanakan sepak bola dari klub manado pada siang hari* (in the Buro Stadium field, a soccer match from the Manado club will be held in the afternoon.)"
- The model produces four versions of the paraphrase sentences as follows:
  - Paraphrase 1: *Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada siang hari.*
  - Paraphrase 2: *Di stadion burkano, sepak bola dari klub manado akan disiarkan pada siang hari.*
  - Paraphrase 3: *Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada hari Sabtu.*
  - Paraphrase 4: *Di Stadion Burkano akan ada pertandingan sepak bola dari klub manado pada hari Minggu.*

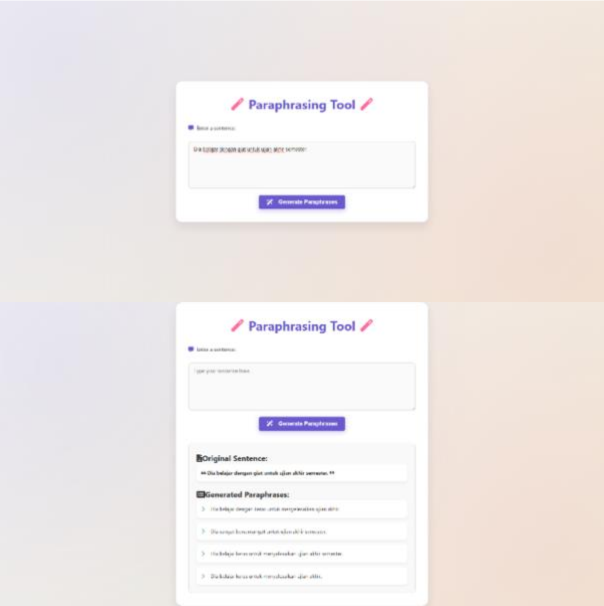
4. User Interface

This paraphrasing system is equipped with a web-based user interface to increase the ease of access for end users. This interface is designed using a modern web framework, where users can:

- Enter the sentence to be paraphrased through the input form.
- Generate up to four versions of the paraphrased sentence instantly.
- View the paraphrased results directly on the web page in an easy-to-read format.

This interface allows users to interact with the model intuitively without requiring technical knowledge. Figure 6 below shows the implementation result of the developed web interface.

Figure 11 shows the user interface of the paraphrasing system that has been deployed in web form. Users can easily enter the sentence to be paraphrased into the input field, and the system will automatically generate several paraphrasing options. The paraphrasing results are displayed directly on the interface for easy comparison and selection of the appropriate option by the user. With a simple and user-friendly interface, the system provides an intuitive experience, making it easy for users to access paraphrasing technology without technical difficulties. This shows the readiness of the system for use in real-world contexts, allowing users to obtain paraphrasing results quickly and efficiently. This study also tested the model on several sermon manuscripts to be paraphrased so that the manuscripts are more varied. Figures 12 and 13 show examples of paraphrasing results of sermon manuscripts.

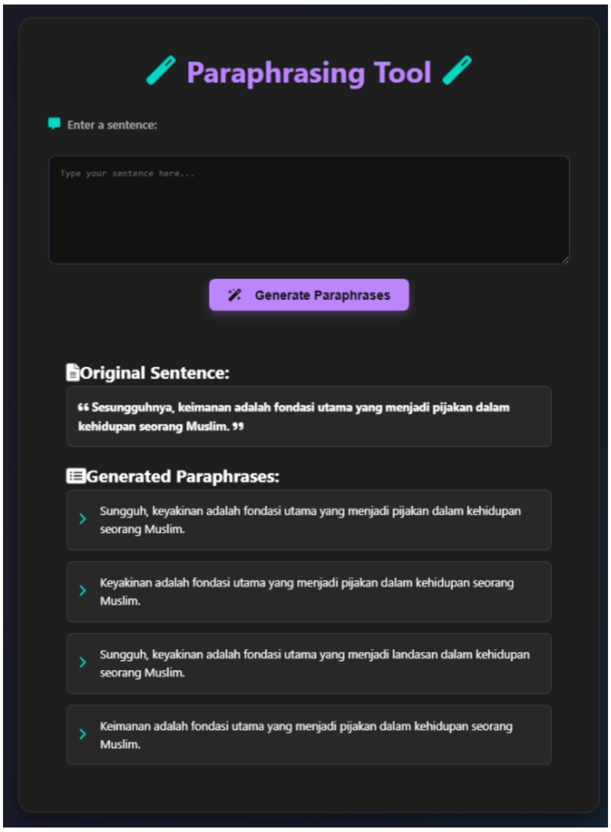


Original Sentence:

“ Dia belajar dengan giat untuk ujian akhir semester. ”

Generated Paraphrases:

Fig. 11. User Interface of Paraphrase Application



Original Sentence:

“ Sesungguhnya, keimanan adalah fondasi utama yang menjadi pijakan dalam kehidupan seorang Muslim. ”

Generated Paraphrases:

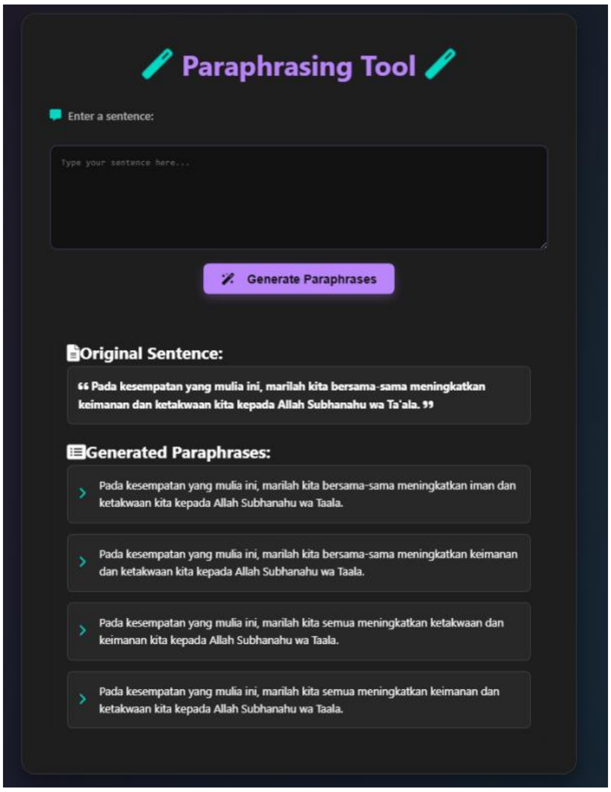
› Sungguh, keyakinan adalah fondasi utama yang menjadi pijakan dalam kehidupan seorang Muslim.

› Keyakinan adalah fondasi utama yang menjadi pijakan dalam kehidupan seorang Muslim.

› Sungguh, keyakinan adalah fondasi utama yang menjadi landasan dalam kehidupan seorang Muslim.

› Keimanan adalah fondasi utama yang menjadi pijakan dalam kehidupan seorang Muslim.

Fig. 12. First Example of Sermon Paraphrase



“Pada kesempatan yang mulia ini, marilah kita bersama-sama meningkatkan keimanan dan ketakwaan kita kepada Allah Subhanahu wa Ta'ala.”

- Pada kesempatan yang mulia ini, marilah kita bersama-sama meningkatkan iman dan ketakwaan kita kepada Allah Subhanahu wa Taala.
- Pada kesempatan yang mulia ini, marilah kita bersama-sama meningkatkan keimanan dan ketakwaan kita kepada Allah Subhanahu wa Taala.
- Pada kesempatan yang mulia ini, marilah kita semua meningkatkan ketakwaan dan keimanan kita kepada Allah Subhanahu wa Taala.
- Pada kesempatan yang mulia ini, marilah kita semua meningkatkan keimanan dan ketakwaan kita kepada Allah Subhanahu wa Taala.

Fig. 13. Second Example of Sermon Paraphrase

V. CONCLUSION

This study develops an automatic Indonesian paraphrasing system using the IndoT5 (Text-to-Text Transfer Transformer) algorithm. Evaluation is carried out by calculating key metrics such as BLEU, ROUGE, and METEOR, which measure the similarity between the paraphrased text generated by the model and the desired target text. Based on the evaluation results, the model achieved a BLEU value of 0.28, ROUGE-1 of 0.59, ROUGE-2 of 0.40, ROUGE-L of 0.55, and METEOR of 0.50. These results indicate that the model has a fairly good ability to produce paraphrases that match the target text, although there is still room for improvement.

For example, for the input sentence "In the Burkano Stadium field, soccer from the Manado club will be held during the day," the model successfully produced four versions of the paraphrased sentence with relevant and meaningful variations, reflecting the model's ability to produce diverse paraphrased texts. In addition, this system is equipped with a web-based user interface that makes it easy for users to enter sentences, generate up to four paraphrased versions, and view the results directly in an easy-to-read format.

However, the relatively low BLEU value indicates that the model can still be improved in producing more accurate and coherent paraphrases. The relatively high ROUGE-L and METEOR values indicate that the model is able to maintain most of the meaning and structure of the original text. Based on these results, the IndoT5-based automatic paraphrasing system has shown good performance in producing paraphrased text, but further research is needed to improve the quality of the paraphrasing results, especially in terms of deeper semantic similarity. Additional evaluation with larger and more diverse datasets and the use of additional models can strengthen the system's performance in the future. Further research is recommended to use larger and more diverse datasets to improve the quality of the paraphrasing results. In addition, it is recommended to compare the performance of the IndoT5 algorithm with other models such as BART, T5 Multilingual, Pegasus, and mBERT in order to evaluate the effectiveness of each model in developing an automatic paraphrasing system. In addition to Indonesian, research can also be expanded by supporting other languages such as English, Spanish, or regional languages in Indonesia to create a more comprehensive and globally useful multilingual paraphrasing system.

REFERENCES

[1] U. Saokani, M. Irfan, D. S. Maylawati, R. J. Abidin, I. Taufik, and R. N. Hay's, “Comparison of the Fisher-Yates Shuffle and the Linear Congruent Algorithm for Randomizing Questions in Nahwu Learning Multimedia,” *Khazanah Journal of Religion and Technology*, vol. 1, no. 1, pp. 10–14, Jun. 2023, doi: 10.15575/kjrt.v1i1.159.

[2] M. Urva Madani and R. Ardianti, “Prosiding Seminar Nasional PBSI-III Tahun 2020 Tema: Inovasi Pembelajaran Bahasa dan Sastra Indonesia Guna Mendukung Merdeka Belajar pada Era Revolusi Industry 4.0 dan Society TEKNIK PARAFRASE DALAM KETRAMPILAN MENULIS UNTUK MENGHINDARI PLAGIARISME.”

[3] A. F. Sberdevices and R. Sberbank, “Russian Paraphrasers: Paraphrase with Transformers,” 2021. [Online]. Available: https://huggingface.co/

[4] admin, “Bahasa Indonesia Disetujui Menjadi Bahasa Resmi Sidang Umum UNESCO,” badanbahasa.kemdikbud.go.id. Accessed: Dec. 15, 2024. [Online]. Available: https://badanbahasa.kemdikbud.go.id/berita-detail/4081/bahasa-indonesia-disetujui-menjadi-bahasa-resmi-sidang-umum-unesco

[5] L. Xue *et al.*, “mT5: A massively multilingual pre-trained text-to-text transformer,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11934

[6] M. Fuadi, A. Dharma Wibawa, and S. Sumpeno, “idT5: Indonesian Version of Multilingual T5 Transformer.” [Online]. Available: https://huggingface.co/muchad/idt5-base

[7] Admin, “APJII Jumlah Pengguna Internet Indonesia Tembus 221 Juta Orang,” apjii.or.id. Accessed: Dec. 15, 2024. [Online]. Available: https://apjii.or.id/berita/d/apjii-jumlah-pengguna-internet-indonesia-tembus-221-juta-orang

[8] A. Wulandari, “PEMBELAJARAN BAHASA INDONESIA YANG INOVATIF DAN KREATIF DI SMP PADA ERA KURIKULUM MERDEKA.”

[9] H. Judul, “Halaman Depan (cover).”

[10] Bunga Dea Laraswati, “Transformer dalam Machine Learning: Model Dibalik GPT, BERT, dan T5,” blog.algorit.ma. Accessed: Dec. 18, 2024. [Online]. Available: https://blog.algorit.ma/transformer-machine-learning/

[11] M. Yani, N. Siti Khodijah, and M. Mustamiin, “Aplikasi Peringkas Teks Bahasa Indonesia Menggunakan Model Text-to-Text Transfer Transformer (T5)”, doi: 10.37817/ikraith-informatika.v9i2.

[12] Belajar Data Science di Rumah, “Tren Terbaru dalam NLP Machine Learning,” dqlab.id. Accessed: Dec. 15, 2024. [Online]. Available: https://dqlab.id/tren-terbaru-dalam-nlp-machine-learning

[13] L. Xue *et al.*, “mT5: A massively multilingual pre-trained text-to-text transformer,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11934

[14] M. Fuadi, A. Dharma Wibawa, and S. Sumpeno, "idT5: Indonesian Version of Multilingual T5 Transformer." [Online]. Available: https://huggingface.co/muchad/idt5-base

[15] A. F. Sberdevices and R. Sberbank, "Russian Paraphrasers: Paraphrase with Transformers," 2021. [Online]. Available: https://huggingface.co/

[16] E. Sugiri, S. Sodiq, and A. Yusuf, "Penggunaan Variasi Bahasa Dalam Khotbah Salat Jumat Berdasarkan Stratifikasi Sosial Jamaah Di Masjid-Masjid Wilayah Provinsi Jawa Timur: Suatu Kajian Sosiolinguistik," 2018.

[17] Z. Nuryana, "Pemanfaatan Teknologi Informasi dalam Pendidikan Agama Islam," *TAMADDUN*, vol. 19, no. 1, p. 75, Mar. 2019, doi: 10.30587/tamaddun.v0i0.818.

[18] M. Alamsyah, "Pemanfaatan Perkembangan Teknologi Informasi Dan Komunikasi Untuk Meningkatkan Mutu Dakwah," *Jurnal An-nasyr: Jurnal Dakwah Dalam Mata Tinta*, vol. 10, no. 1, pp. 48–62, Apr. 2023, doi: 10.54621/jn.v10i1.605.

[19] P. Yualita, "Menulis Paragraf dengan Teknik Parafrasa Menggunakan Software Paraphraser," *Jurnal Kajian Bahasa, Sastra dan Pengajaran (KIBASP)*, vol. 7, no. 1, pp. 12–26, Aug. 2023, doi: 10.31539/kibasp.v7i1.6409.

[20] U. Primakara, "IMPLEMENTASI PERINGKAS DOKUMEN BERBAHASA INDONESIA MENGGUNAKAN METODE TEXT TO TEXT TRANSFER TRANSFORMER (T5) I Nyoman Purnama 1) , Ni Nengah Widya Utami 2) Program Studi Sistem Informasi 1) , Sistem Informasi Akutansi 2)."

[21] Q. A. Itsnaini, M. Hayaty, A. D. Putra, and N. A. M. Jabari, "Abstractive Text Summarization using Pre-Trained Language Model 'Text-to-Text Transfer Transformer (T5),'" *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 124–131, Apr. 2023, doi: 10.33096/ilkom.v15i1.1532.124-131.

[22] J. Gabín, M. E. Ares, and J. Parapar, "Enhancing Automatic Keyphrase Labelling with Text-to-Text Transfer Transformer (T5) Architecture: A Framework for Keyphrase Generation and Filtering," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.16760

[23] M. Yani, N. Siti Khodijah, and M. Mustamiin, "Aplikasi Peringkas Teks Bahasa Indonesia Menggunakan Model Text-to-Text Transfer Transformer (T5)", doi: 10.37817/ikraith-informatika.v9i2.

[24] M. Yani, N. S. Khodijah, R. Rendi, and M. Mustamiin, "Aplikasi Peringkas Teks Bahasa Indonesia Menggunakan Model Text-to-Text Transfer Transformer (T5)," *IKRA-ITH Informatika : Jurnal Komputer dan Informatika*, vol. 9, no. 2, pp. 78–86, Oct. 2024, doi: 10.37817/ikraith-informatika.v9i2.4392.

[25] I Nyoman Purnama and Ni Nengah Widya Utami, "Implementasi Peringkas Dokumen Berbahasa Indonesia Menggunakan Metode Text-to-Text Transfer Transformer (T5)," *Jurnal Teknologi Informasi dan Komputer*, vol. 9, no. 4, Aug. 2023, doi: 10.36002/jutik.v9i4.2531.

[26] J. Gabín, M. E. Ares, and J. Parapar, "Enhancing Automatic Keyphrase Labelling with Text-to-Text Transfer Transformer (T5) Architecture: A Framework for Keyphrase Generation and Filtering," *Preprint submitted to Elsevier*, p. 1, Sep. 2024.

[27] D. Kubal and H. Palivela, "Unified Model for Paraphrase Generation and Paraphrase Identification," Apr. 23, 2021. doi: 10.20944/preprints202104.0630.v1.

[28] Q. A. Itsnaini, M. Hayaty, A. D. Putra, and N. A. M. Jabari, "Abstractive Text Summarization using Pre-Trained Language Model 'Text-to-Text Transfer Transformer (T5),'" *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 124–131, Apr. 2023, doi: 10.33096/ilkom.v15i1.1532.124-131.

[29] E. Egonmwan and Y. Chali, "Transformer and seq2seq model for Paraphrase Generation," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 249–255. doi: 10.18653/v1/D19-5627.

[30] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[31] M. Muhammad and M. R. Muhammad, "Building trust in e-commerce: A proposed shari'ah compliant model," *Journal of Internet Banking and Commerce*, vol. 18, Dec. 2013.

[32] M. Akram, N. Khan, and M. N. Anjum, "Perceived Financial Transparency and Loyalty in the Islamic Banking Sector of Pakistan: Exploring the Role of Trust and Age," *Contemporary Issues in Social Sciences and Management Practices*, vol. 3, no. 2, pp. 14–25, Jun. 2024, doi: 10.61503/cissmp.v3i2.160.

[33] M. A. Khan, "Justice in economics: an Islamic perspective," in *Islamic Economics and Human Well-being*, Edward Elgar Publishing, 2024, pp. 66–108. doi: 10.4337/9781035333691.00012.

[34] M. Roberts-Lombard and D. J. Petzer, "Do you want my loyalty? Then understand what drives my trust – a conventional and Islamic banking perspective," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 17, no. 3, pp. 532–551, Jul. 2024, doi: 10.1108/IMEFM-10-2023-0412.

[35] E. Elmahjub, "Artificial Intelligence (AI) in Islamic Ethics: Towards Pluralist Ethical Benchmarking for AI," *Philos Technol*, vol. 36, no. 4, p. 73, Dec. 2023, doi: 10.1007/s13347-023-00668-x.

[36] K. Albar, A. Abubakar, and A. Arsyad, "Islamic Business Ethics in Online Commerce: A Perspective from Maqashid Shariah by Imam Haramain," vol. 07, no. 2, pp. 273–289, 2023, doi: 10.33852/jurnalin.v7i2.501.