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 1 (2024)
DOI: https://doi.org/10.15575/kjrt.v2i1.822

SoulScripture: Chatbot using Bidirectional Encoder Representations from Transformers as a Medium of Spiritual Guidance

Andhika Malik
Department of Informatics
UIN Sunan Gunung Djati Indonesia
Bandung, Indonesia
tugassdhika@gmail.com

Elman Sidik
Department of Informatics
UIN Sunan Gunung Djati Indonesia
Bandung, Indonesia
elmansidiq4@gmail.com

Andhika Putra Gefadri
Department of Informatics
UIN Sunan Gunung Djati Indonesia
Bandung, Indonesia
andhikagefadri@gmail.com

Alika Putie Syadrina
Department of Informatics
UIN Sunan Gunung Djati Indonesia
Bandung, Indonesia
syadrinalika@gmail.com

Abstract— Mental health is an important aspect of human life. Many people face stress, anxiety, and distress daily without adequate support to manage these conditions. Islamic teachings from the Quran and Hadith provide wisdom as a source of inspiration and inner peace. However, accessing and understanding these teachings requires specialized knowledge and often the help of experts. With the advancement of machine learning, these teachings can be made more accessible and accurate. The SoulScripture app offers an innovative solution to support mental health by combining the wisdom of the Quran and Hadith through AI technology. Using the Bidirectional Encoder Representations from Transformers (BERT) model and Transformer architecture, the app can understand and provide relevant advice that anyone can access anytime. This research is significant because it offers a new approach to leveraging technology to support mental well-being, especially for communities underserved by conventional mental health services. The app was developed using self-supervised learning to understand the text of the Hadith without external annotation. This process involves several stages, such as user input, data preprocessing, and text analysis, to generate relevant answers. It is hoped that the SoulScripture application can serve as a source of inspiration and support for individuals in controlling stress and maintaining peace of mind, as well as supporting the achievement of the Sustainable Development Goals (SDGs) related to mental health.

Keywords- AI, BERT, mental health, self-supervised learning, transformer, Quran, Hadith

I. INTRODUCTION

Mental health is one of the important aspects of human life. Many people face stress, anxiety, and pressures of daily life without having adequate sources of support to manage these conditions. Meanwhile, the Qur'an and Hadith teachings contain a lot of wisdom that can be a source of inspiration and inner peace for individuals. However, accessing and understanding these teachings requires special knowledge and often help from experts. Along with the development of the era, the field of machine learning is increasingly booming. Machine learning can be used to build comprehensive religious applications, such as translations of the Qur'an and Hadith, which allow easier and more accurate access to these teachings [1]. Based on the problems above, the implementation of machine learning is self-supervised learning, which allows AI systems to learn human language patterns from very large textual datasets without the need for labeled data [2].

The SoulScripture application is an innovative solution to support mental health by combining the wisdom of the Qur'an and Hadith through AI technology. Using the BERT base model and Transformer Architecture, this application can understand and provide relevant advice that anyone can access anytime. This research is important because it offers a new approach to leveraging technology to support mental well-being, especially for communities that conventional mental health services may underserve. The general public is not yet fully aware of the potential of AI technology in accessing and understanding religious teachings to support mental health. In addition, there is still a gap between the need for mental support and the availability of resources that can be easily and accurately accessed. Researchers will develop the SoulScripture application using self-supervised learning technology to understand the text of the Hadith without the need for external annotation. This process involves user input, pre-processing, and text analysis to produce relevant answers. With these steps, it is hoped that the SoulScripture application can serve as a source of inspiration and support for individuals in managing stress and maintaining peace of mind, as well as supporting the achievement of the Sustainable Development Goals (SDGs) related to mental health .

II. Related Works

Self-supervised learning (SSL) is a learning method that uses labels obtained "for free" from input data (x) through various transformations and conventional supervision objectives to predict the labels (SSL). The representation obtained in this way will be meaningful for advanced tasks with limited labeled data and linear classification. SSL is known as predictive learning because supervised learning is taken directly from the data [3].

Self-supervised learning has helped a lot in its application in the world of health, one example being a study entitled "A Mobile Deep Learning Model on COVID-19 CT-Scan Classification." This study used multi-task and self-supervised learning in the data collection process, which then got an F1 score of 0.90 and an AUC of 0.98 [4].

Artificial intelligence has focused on natural language processing. Traditional techniques such as rule-based rules and statistical modeling have produced satisfactory results, but their weaknesses appear when faced with increasingly complex language processing tasks. self-supervised learning allows models to learn from unlabeled data through self-learning. In its application, although deep learning has made great progress in various language processing tasks, current challenges indicate that there is still much work to be done to improve its performance and relevance in a broader context. Solutions such as transfer learning and integration with traditional approaches have been proposed to address this issue, but a broader approach is needed. A previous study provided a comprehensive analysis of the current state and future prospects in the use of Deep Learning in natural language processing (NLP) [5], [6], [7], especially in chatbot task [8], [9].

The causes of environmental damage are anthropocentric views, lack of understanding of religious texts, and lack of human awareness of the universe. One way to address this issue is to understand the prophetic messages of the environment in the hadith. One of the prophetic messages in preserving the environment is the purpose of maintenance, inclusive ownership, positive contribution, use based on utility, sustainable programs, restrictions on use, and self-supervised [10]. Several kinds of research discuss NLP with Hadith as objects, such as text categorization [11], search engine [12].

Based on previous studies, Self-supervised learning (SSL) has shown significant potential in various fields. SSL is not only effective in improving model performance on natural language processing tasks but has also been successfully applied in the health world, as shown in a study on COVID-19 CT-Scan classification with an SSL-based deep learning model, which produced an F1 score of 0.90 and an AUC of 0.98 . In addition, understanding and applying religious texts through a self-supervised approach also offers a solution to increase environmental awareness and understanding of prophetic messages in the hadith. Thus, this study provides a strong foundation for further development in using SSL in mental health, language processing, and environmental maintenance.

III. Research Methods

This section explains the steps and methods implemented to build the SoulScripture Application, starting with data collection, Application development, and finally, data analysis and interpretation.

A. *Data Collecting and Interpreting*

This research process begins by collecting hadith text data that covers various topics relevant to mental health and well-being from trusted sources. In this data analysis and interpretation process, the data collected from the algorithm evaluation and user experience is analyzed to evaluate whether the SoulScripture application has succeeded in achieving the stated goals and answering the research questions.

B. *SoulScripture Application Development*

The next process is to develop the SoulScripture application by integrating self-supervised learning into the system. In addition, this application will utilize the Transformer architecture, specifically BERT, to enhance natural language processing capabilities. BERT is one of the most advanced language models used in natural language processing (NLP). The advantage of BERT is its ability to understand the context of words in a sentence bidirectionally, which allows this model to capture deeper and more relevant meaning from the processed text. By implementing BERT, the SoulScripture application will be able to understand user questions and problems more accurately and provide more relevant and precise answers. This process involves developing an intuitive and functional user interface to allow users to enter questions or searches and receive the correct answers. Figure 1 shows a flowchart of how the application works.

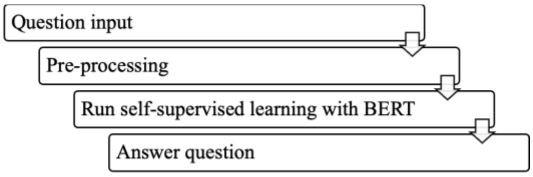

Fig 1. Application Flow Activities

Starting from user input regarding mental health problems or disorders experienced. From the user input in the form of a narrative, the system will perform data preprocessing. It will enter self-supervised learning, then matching and analysis by the system, and the system will issue relevant answers from the user input.

C. *Bidirectional Encoder Representations from Transformers (BERT)*

Bidirectional Encoder Representations from Transformers (BERT) is a transformative model in natural language processing (NLP), first introduced by Google researchers in 2018 [13]. Bidirectional Encoder Representations from Transformers (BERT) is a powerful neural network architecture primarily designed for natural language processing (NLP) tasks [14]. Its bidirectional training approach allows for a nuanced understanding of context, making it particularly effective in sentiment analysis, conversational comprehension, and recommendation systems. For instance, BERT has demonstrated superior performance in sentiment classification compared to traditional models like logistic regression and LSTM, showcasing its advanced capabilities in processing text data [15]. Additionally, BERT's architecture has been successfully adapted for various domains, including stock price modeling and music recommendation systems, where it leverages contextual relationships to enhance predictive accuracy [16], [17]. Furthermore, BERT has been utilized in cybersecurity, specifically in intrusion detection systems, where it outperforms conventional methods in classifying network activities [18]. Overall, BERT's versatility and effectiveness across diverse applications underscore its significance in the field of machine learning and NLP.

IV. Result and Discussion

In this study, researchers built a system to return relevant hadith and Quranic verses based on the questions given. This system uses various stages ranging from data processing, feature extraction, to text matching. By using the BERT Base Model with the Cosine Similarity algorithm and developed with the Transformer architectural style, a web-based AI Chatbot named Soulscripture can provide recommendations for hadith and Quranic verses according to the problems inputted by the user.

A. *Dataset*

The dataset used is a collection of hadiths consisting of 9 narrators[1] and the dataset is a collection of Al-Quran verses[2]. The dataset consists of hadiths based on the books Musnad Ahmad (26,363 hadiths), Musnad Syafi'i (1,800 hadiths), Riyadhus Shalihin (372), Riyadhus Shalihin Arabic (850), Sahih Bukhari (7,008 hadiths), Sahih Muslim (5,362 hadiths), Sunan Abu Daud (4,590 hadiths), Sunan Ibnu Majah (4,332 hadiths), Sunan Nasa'i (5,662 hadiths), Sunan Tirmidhi (3,891 hadiths), Muwatho' Malik (1,594 hadiths), and Musnad Darimi (3367 hadiths).

B. *Data Processing*

Importing some required libraries or modules, one of which is 'SQLite3', connects the prepared dataset to the SQLite database. Before that, the dataset's contents in the form of hadith and verses of the Qur'an undergo a data cleaning process. The function of this text cleaning is to clean text from special characters and numbers; we create a 'clean_text' function that uses regular expressions to retain only letters and spaces and change all text to lowercase.

```
import  sqlite3  #konek  ke  database  sqlite3  unntuk
ngebaca isi
import pandas as pd
import re

# Fungsi untuk membersihkan teks
def clean_text(text):
    # Menghapus karakter khusus dan angka
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    text = text.lower().strip()
    return text

# Path ke file SQL
hadith_db_path = './datasets/hadist_database.db'

# Membuat koneksi ke database
conn_hadith = sqlite3.connect(hadith_db_path)

# Mendapatkan daftar tabel hadith
hadith_tables  =  ['musnad_darimi',  'musnad_syafii',
'muwatho_malik', 'shahih_bukhari',
                   'shahih_muslim', 'sunan_abu_daud',
'sunan_ibnu_majah', 'sunan_nasai', 'sunan_tirmidzi']

# Menggabungkan semua hadith menjadi satu dataframe
all_hadiths = pd.concat([pd.read_sql_query(f"SELECT *
FROM    {table}",    conn_hadith)    for    table    in
hadith_tables])
all_hadiths['cleaned_text']                          =
all_hadiths['terjemah'].apply(clean_text)            #
ng3bersihin tanda baca yang ada di kolom terjemah

print(all_hadiths.head())

# Path ke file SQL
quran_db_path = './datasets/quran.db'

# Membuat koneksi ke database
conn_quran = sqlite3.connect(quran_db_path)

# Mendapatkan data Quran
ayat_df    =    pd.read_sql_query("SELECT    *    FROM
table_ayat", conn_quran)
ayat_df['cleaned_text']                              =
ayat_df['Terjemahan'].apply(clean_text)
```

[1] https://github.com/irsyadulibad/hadits-database.git

[2] https://github.com/alfianyusufabdullah/DatabaseAlquranQ.git

```
print(ayat_df.head())
```

Then, the dataset will be filled into the database using SQLite and combined into one dataframe for data cleaning using the previously created ‘clean_text’ function.

```
# Menggunakan tokenizer BERT
tokenizer = BertTokenizer.from_pretrained('bert-base-
uncased')

# Menggunakan model BERT untuk representasi teks
model     =     BertModel.from_pretrained('bert-base-
uncased')
```

C. Feature Extraction with BERT

We employ the BERT (Bidirectional Encoder Representations from Transformers) model to extract high-quality embeddings from the cleaned text, capturing intricate semantic relationships and contextual nuances. These embeddings serve as robust representations, facilitating downstream tasks such as classification and clustering.

D. Tokenization with BERT

The purpose of tokenization is to break down the text into manageable units, which are then transformed into embeddings or vector representations. These embeddings encapsulate the semantic and syntactic properties of the original text, enabling the model to effectively process and analyze the information. This step is crucial for tasks such as text classification, sentiment analysis, and natural language understanding.

E. Embedding Function

Using the get_embeddings function to get the embedding from the text. This function processes text in batches for efficiency and uses the representation from the last layer of the BERT model. The resulting embedding is then saved in the .pkl file format using the pickle module.

```
def get_embeddings(texts, batch_size=32):
    embeddings = [] # ngebuat list kosong buat ngisi
hasil embedding dari setiap batch
    # looping untuk memproses teks dalam batch
    for i in tqdm(range(0, len(texts), batch_size),
desc="Processing Batches"):
        batch_texts = texts[i:i+batch_size]
            inputs    =    tokenizer(batch_texts,
return_tensors='pt', padding=True, truncation=True)
        with torch.no_grad():
            outputs = model(**inputs)
                batch_embeddings       =
outputs.last_hidden_state.mean(dim=1)
        embeddings.append(batch_embeddings)
    return torch.cat(embeddings)
```

F. Relevant Text Search

By creating a function find_relevant_texts that uses cosine similarity to find the text that is most similar to a given question. This function returns the five most relevant texts.

```
def find_relevant_texts(question, embeddings, texts):
    question_embedding = get_embeddings([question])
                  similarities                     =
cosine_similarity(question_embedding, embeddings)
     most_similar_idx  =  similarities.argsort()[0,  -
5:][::-1]
     valid_indices = [idx for idx in most_similar_idx
if idx < len(texts)]
    return [texts.iloc[idx] for idx in valid_indices]
```

In this implementation, we use BERT to obtain a text representation that can be used for relevant text search. The algorithm used is cosine similarity to calculate the similarity between the question text and the text in the database. The architecture of this system involves several stages, including data processing, feature extraction using the BERT model, and text matching with cosine similarity. With this approach, we can build an efficient system to search for relevant hadith and Quranic verses based on the given question. This implementation can be extended and refined by using more sophisticated models or by adding additional features to improve search accuracy.

V. Conclusion

In this study, we have successfully developed a system that uses self-supervised learning technology and the BERT model to help individuals access and understand the teachings of the Qur'an and Hadith that are relevant to mental health. Through the SoulScripture application, users can obtain relevant advice based on the questions they ask, generated through a text-matching process using cosine similarity.

The study's results show that this approach is effective in providing relevant text recommendations. The system can clean data, tokenize, and produce accurate text embeddings. Self-supervised learning allows the model to learn human language patterns without the need for labeled data, thereby increasing the system's efficiency and flexibility in processing various types of questions.

In addition, the Transformer architecture used in the BERT model provides a powerful ability to understand the context of words in sentences bidirectionally, allowing the application to capture deeper meaning from the text being processed. This is especially important in natural language processing and text matching, where understanding context is key to producing accurate and relevant answers.

The implementation of SoulScripture makes a significant contribution to supporting mental health, especially for communities that conventional mental health services may underserve. By providing easy and accurate access to the wisdom of the Quran and Hadith, this application can be a source of inspiration and inner peace for many individuals. In the future, this research can be expanded by using more sophisticated models or adding additional features to improve search accuracy. In addition, further evaluation of the user experience and the real impact of this application on users' mental well-being would be very useful in refining and optimizing the functionality of the SoulScripture application.

REFERENCES

[1] D. Maulidiya, I. Musarofah, R. S. Hasnani, and A. N. Aeni, “Aplikasi HOT-Day: Hadits of the Day Sebagai Media Edukasi Pengamalan Hadits,” *AL-HIKMAH (Jurnal Pendidikan dan Pendidikan Agama Islam)*, vol. 4, no. 1, pp. 57–65, 2022.

[2] M. A. Bora, Ansarullah Lawi, I Made Sondra Wijaya, and Tia Andini Salsabilla, “Mengoptimalkan Kenyamanan Kognitif: Analisis Ergonomis terhadap Interaksi Pengguna dengan AI Chatbots,” *Ranah Research : Journal of Multidisciplinary Research and Development*, vol. 6, no. 4, pp. 710–723, Jun. 2024, doi: 10.38035/rrj.v6i4.869.

[3] D. Spathis, I. Perez-Pozuelo, L. Marques-Fernandez, and C. Mascolo, “Breaking away from labels: The promise of self-supervised machine learning in intelligent health,” *Patterns*, vol. 3, no. 2, p. 100410, Feb. 2022, doi: 10.1016/j.patter.2021.100410.

[4] P. E. Susanto, A. Kurniawardhan, D. H. Fudholi, and R. Rahmadi, “A Mobile Deep Learning Model on Covid-19 CT-Scan Classification,” *International Journal of Artificial Intelligence Research*, vol. 6, no. 2, Jul. 2022, doi: 10.29099/ijair.v6i1.257.

[5] Z. Kaddari, Y. Mellah, J. Berrich, M. G. Belkasmi, and T. Bouchentouf, “Natural Language Processing: Challenges and Future Directions,” 2021, pp. 236–246. doi: 10.1007/978-3-030-53970-2_22.

[6] J. Liu, X. Han, C. Deng, and J. Feng, “Robust Self-Supervised Learning with Contrast Samples for Natural Language Understanding,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2024, pp. 10076–10080. doi: 10.1109/ICASSP48485.2024.10448238.

[7] P. E. Susanto, A. Kurniawardhan, D. H. Fudholi, and R. Rahmadi, “A Mobile Deep Learning Model on Covid-19 CT-Scan Classification,” *International Journal of Artificial Intelligence Research*, vol. 6, no. 2, Jul. 2022, doi: 10.29099/ijair.v6i1.257.

[8] H. K. K., A. K. Palakurthi, V. Putnala, and A. Kumar K., “Smart College Chatbot using ML and Python,” in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, Jul. 2020, pp. 1–5. doi: 10.1109/ICSCAN49426.2020.9262426.

[9] M. Muliyono and S. Sumijan, “Identifikasi Chatbot dalam Meningkatkan Pelayanan Online Menggunakan Metode Natural Language Processing,” *Jurnal Informatika Ekonomi Bisnis*, pp. 142–147, Sep. 2021, doi: 10.37034/infeb.v3i4.102.

[10] M. Akmaluddin, “Diskursus Penelitian Al-quran dan Hadis dengan Ilmu Pengetahuan Modern,” in *Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat UNIMUS 2017*, Muhammadiyah University Semarang, 2020.

[11] M. F. Afianto, Adiwijaya, and S. Al-Faraby, “Text Categorization on Hadith Sahih Al-Bukhari using Random Forest,” *J Phys Conf Ser*, vol. 971, p. 012037, Mar. 2018, doi: 10.1088/1742-6596/971/1/012037.

[12] I. Taufik, M. Jaenudin, F. U. Badriyah, B. Subaeki, and O. T. Kurahman, “The search for science and technology verses in Qur’an and hadith,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 1008–1014, Apr. 2021, doi: 10.11591/eei.v10i2.2629.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[14] E. Gogoulou, “Using Bidirectional Encoder Representations from Transformers for Conversational Machine Comprehension,” 2019.

[15] S. Alaparthi and M. Mishra, “Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey,” Jul. 2020.

[16] P. Chaudhry, “Bidirectional Encoder Representations from Transformers for Modelling Stock Prices,” *Int J Res Appl Sci Eng Technol*, vol. 10, no. 2, pp. 896–901, Feb. 2022, doi: 10.22214/ijraset.2022.40406.

[17] N. Yadav and A. K. Singh, “Bi-directional Encoder Representation of Transformer model for Sequential Music Recommender System,” in *Forum for Information Retrieval Evaluation*, New York, NY, USA: ACM, Dec. 2020, pp. 49–53. doi: 10.1145/3441501.3441503.

[18] M. Vubangsi, T. R. Mangai, A. Olukayode, A. S. Mubarak, and F. Al-Turjman, “BERT-IDS: an intrusion detection system based on bidirectional encoder representations from transformers,” in *Computational Intelligence and Blockchain in Complex Systems*, Elsevier, 2024, pp. 147–155. doi: 10.1016/B978-0-443-13268-1.00021-2.