

Generative pre-trained transformer based on image content and user personality for caption creation in social media

Ikhsan Ariansyah^{1*}, Shintami Chusnul Hidayati²

^{1,2}Faculty of Electrical Technology and Intelligent Informatics, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia; 6025231045@student.its.ac.id (I.A.) shintami@its.ac.id (S.C.H.)

Abstract: Social media is a platform for sharing information and interactions between users, often involving images with captions that reflect the user's personality. Each user creates distinct captions based on personal traits, making a personalized caption generator beneficial. Currently, existing social media caption generators have limitations, such as requiring payment for full features, lack of support for Bahasa (Indonesian Language), dependency on user input to generate captions, and suboptimal object detection accuracy. To address these issues, a new method is proposed for generating social media captions based on image content and user personality to simplify the caption creation process. This caption generator will be optimized in Bahasa. The content of the image will be explored through image objects and scenery. Image objects are identified using a Graph Convolutional Network (GCN) for personality classification. At the same time, a Convolutional Neural Network (CNN) approach will be employed to detect objects within images, and VGG16 will be used to detect scenery. Then, these three models are combined with a GPT to generate new captions. The model will be trained on public datasets, and subjective evaluation will be used for testing. The outcome of this research is expected to produce relevant captions based on the user's personality, making the captioning process more efficient and relevant to the personality.

Keywords: Caption generation, Generative pre-trained transformer, Object identification, Personality identification, Scene recognition, Social media.

1. Introduction

The 5.0 revolution has accelerated the development of information technology, leading to practical applications that transform services, introduce new service capabilities, and improve existing ones. Artificial intelligence (AI), one of the core technologies in this revolution, has become widely discussed globally (Putra & Wijaya, 2023). AI plays a significant role today, as its advancements enable robots to collaborate with or even replace humans in various tasks (Wang et al., 2020). AI technology encompasses procedures allowing data analysis and implementation, enabling machines to think and act like humans (Purba & Irwansyah, 2022). According to Feng et al. (2022), AI offers several positive impacts, such as enhancing work efficiency and effectiveness, supporting decision-making, improving services and product quality, and performing repetitive tasks traditionally done by humans. However, not everyone views AI development positively. Concerns about privacy, social impact, and potential worker replacement suggest the need for a balanced approach to AI implementation across sectors.

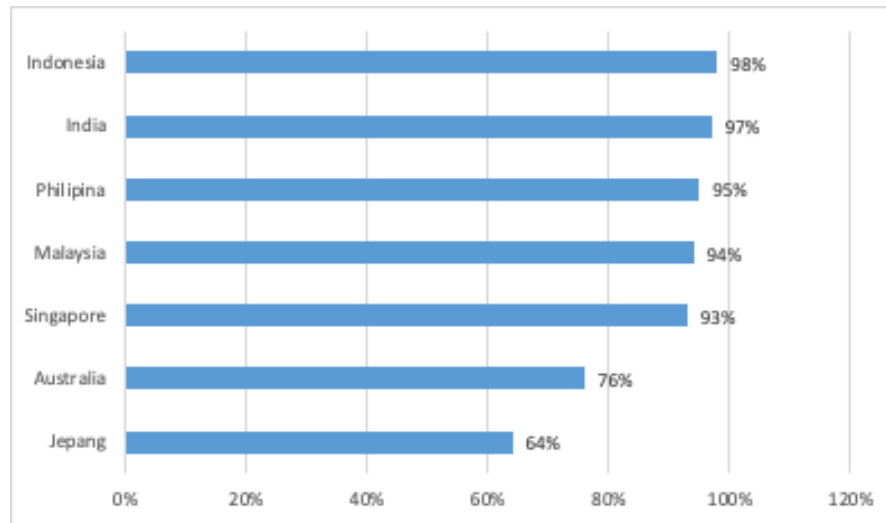


Figure 1.
The use of artificial intelligence in several countries.

Based on data from Figure 1 of research on the use of artificial intelligence among professional workers, Indonesia has the highest adoption rate 98%, followed by India, the Philippines, Malaysia, Singapore, Australia, and Japan. This data shows that artificial intelligence helps professional workers complete work quickly and efficiently. From the data, it can be concluded that most Asian countries prefer using artificial intelligence to assist in their work and most Asian countries prefer artificial intelligence to help complete their work. (Putra & Wijaya, 2023).

Social media has content in the form of images and captions on a post, and the captions on each user are different depending on their personality. The same image does not necessarily produce the same caption. Captions are relevant to images that reflect users' perceptions, emotions, and creativity in expressing themselves. This is what makes each caption produced different. Current technology in the form of a caption generator is one of the latest technologies in creating social media captions quickly and efficiently using artificial intelligence (Sharma et al., 2019). Caption generators have several types of applications that have been made, such as Hootsuite, StoryLab, Copy.ai, and Instagram Caption Generator from Social Pilot supported by chatbots that process language such as Chat GPT (Feng et al., 2022).

Research conducted by (Wang et al., 2020) states that caption generators are created to process millions of texts to produce responses or results that are almost the same as humans. The approach method in caption generators is designed to simplify the process of writing text, helping professionals create exciting content that suits the target audience. Artificial intelligence has been developed as a social media generator, but using this intelligence still depends on the image input process. It is paid to access complete features; a generator in Bahasa still needs to be.

Caption generators can also be used to rearrange text that has been created or posted. The method can be done by adding a new caption to the caption generator tool, which will immediately display a list of new captions to be posted on social media according to the predetermined posting schedule. That way, creating social media captions that often make some people feel complicated and confusing can be done easily and quickly (Sharma et al., 2019). Currently, caption generators still need to consider the user's personality.

Personality is a relatively stable system of internal individual characteristics contributing to consistent thoughts, feelings, and behavior (Mehta et al., 2020). Personality is an essential trait that can distinguish him from others. Personality includes all thoughts, behaviors, feelings, consciousness, and unconsciousness (Carvalho et al., 2020). Therefore, everyone's personality must be different. So, with these differences, some people need help creating captions on social media when making a post. Personality can influence the creation of captions or writings on social media. So, a caption generator

framework is needed that can adjust to personality to facilitate the difficulty of creating captions in a post. The capabilities of the caption generator framework are in great demand by users today. Most social media users need clarification when providing relevant captions to their image uploads (Sazali et al., 2020). Therefore, there is a gap in the current caption generator creation process. These gaps include the absence of support in Bahasa for creating captions, complete features that are still paid, dependence on the image input process, and caption creation that needs to consider each user's personality.

Due to the gaps and backgrounds that have been explained, this study proposes a caption generator framework. The goal is to design a caption generator framework using artificial intelligence so that it can be used to help social media users create captions in Bahasa for free, without having to input images, and by the user's personality in using social media.

2. Related Work

2.1. Artificial Intelligence in Social Media

Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of artificial intelligence include expert systems, natural language processing, speech recognition, and machine vision. Artificial intelligence is often simply a technology component, such as machine learning. Artificial intelligence requires a specialized hardware and software foundation to write and train machine learning algorithms. No single programming language is synonymous with artificial intelligence, but Python, R, and Java, C++ have features that are popular among artificial intelligence developers (Wang et al., 2020). In general, artificial intelligence systems work by using large datasets, analyzing the data for correlations and patterns, and using the patterns to make predictions about future states. In this way, chatbot-fed text samples can learn to produce real-life exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples. New and rapidly developing generative artificial intelligence techniques can generate realistic text, photographs, music, and other media (Seo et al., 2020).

2.2. Caption Generators

Caption generator is an artificial intelligence tool that can help generate creative and exciting text for social media posts. With a caption generator, you only need to enter a short photo description, and the artificial intelligence writer will create excellent text to post (Purba & Irwansyah, 2022). Caption generators can help find ideas for social media photo captions. It uses GPT-3 (Sufi, 2024), a machine-learning algorithm, to generate unique text suggestions. Captions are essential because they can help explain the context of a photo or start a conversation. A good caption will make people want to engage with the content posted on social media. So, it can help increase the reach and engagement of others on social media (Christou, 2023). Caption generators can use deep learning-based techniques, including natural language processing (NLP) and computer vision, to generate text using the Convolutional Neural Network (CNN) algorithm (Kumar et al., 2019). Captioning is a short text given to an image in a book, magazine, or newspaper regarding a description or explanation of the image. This description is given according to the visualization that someone sees within images (Kamal et al., 2020). Text captions are developed to provide a better understanding of the information to be conveyed within images shown. Understanding the image is the main focus of each caption given. Providing text can help you better understand the illustrations of objects in an image. Image caption generation is a computer science problem that involves computer vision (CV), natural language processing (NLP), and Machine Learning (Seo et al., 2020).

2.3. Personality Detection

Personality refers to the enduring characteristics and behaviors that comprise a person's unique adjustment to life, including core traits, interests, drives, values, self-concept, abilities, and emotional patterns. Various theories explain the structure and development of personality in different ways, but all agree that personality helps determine behavior (Sharan & Romano, 2020). The field of personality psychology studies the nature and definition of personality and its development, the structure and

construction of traits, dynamic processes, variations (emphasizing enduring and stable individual differences), and maladaptive forms (Amirhosseini et al., 2024). In psychology, personality is one of the psychological studies born based on experts' thoughts, studies, or findings. The object of personality is everyday human behavior. Everyone certainly has a personality. It should be remembered that everyone's personality is different, so it takes work to understand personality (Schepman & Rodway, 2023).

The Myers-Briggs system (Choong & Varathan, 2021) consists of four preference pairs that reflect different aspects of personality, opposing ways of directing and receiving energy through Extraversion (E) or Introversion (I), obtaining information with Sensing (S) or Intuition (N), drawing conclusions using Thinking (T) or Feeling (F), and approaching the outside world through Judgment (J) or Perception (P). Most people agree that one preference, in a preference pair, best describes their natural way of doing things, where they feel most comfortable being themselves, regardless of whatever role they play in life. When the letters of each of these preferences are combined, 16 different personality types are formed, consisting of different characteristics unique to that type (Amirhosseini et al., 2024). The method of personality classification techniques uses a Convolutional Neural Network (CNN) (Akma et al., 2023). The personality dataset is in the form of text labeled according to each personality's caption. The text is analyzed to identify linguistic patterns that reflect various personalities.

2.4. Image Content Detection

Content on social media often consists of images that display objects or scenery. Objects in images are inanimate objects, humans, and animals, while scenery usually includes the background of an image. Detecting objects and scenery in images is essential to categorize and understand the context of the image, which is then used to generate relevant and informative captions. This study uses the YOLO framework to detect objects in an image because YOLO's ability to detect objects is excellent. VGG16 will be used to detect scenery. VGG16 can detect scenery well and has a fast scenery identification process. YOLO is one of the frameworks that can perform very efficient and accurate object detection, and it is suitable for this study, which requires object detection with high accuracy. From the beginning of its release until now, YOLO has been updated to allow for more complete accuracy and features. The current version of YOLO is YOLOv9 which was released on February 21, 2024 (C.-Y. Wang et al., 2024). YOLOv9 is the latest version of the YOLO detection model. YOLO is known for its object detection; each version has improvements. In YOLOv9, it offers a more effective solution in object detection. With this latest architecture, YOLOv9 is different from previous versions. Details about the architecture in YOLOv9 can be seen in Figure 2.

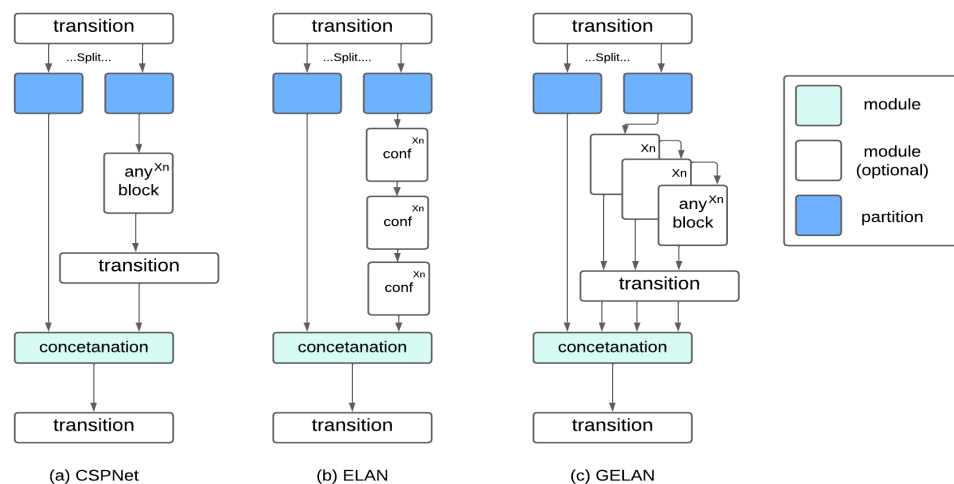


Figure 2.
The Architecture of YOLOv9.

According to research conducted by (C.-Y. Wang et al., 2024), the advantages of YOLOv9 compared to previous versions include a significant reduction in the number of parameters and calculations, Average Precision (AP) in YOLOv9 in recognizing objects is more accurate than the YOLO version below, the speed and accuracy of YOLOv9 can produce better models without sacrificing accuracy and the Generalized ELAN (GELAN) concept in YOLOv9 which makes flexibility higher in architectural design and adaptation.

The monk framework with the pre-trained VGG16 model is a good model for creating scenery. The monk framework is a deep learning framework developed by the Lab of AI. The VGG16 pre-trained model is a Convolutional Neural Network (CNN) model (Arjun & Kumar, 2022) previously trained with a large dataset such as Imagenet. In this case, the VGG16 pre-trained model with the monk framework can be used for image processing to create scenery. This framework also simplifies the process of using pre-trained so that it can be used for complex implementations.

2.5. GPT

Generative Pre-trained Transformer (GPT) is a more widely known language model currently developed by Openai. The GPT model currently developed is GPT-4, which allows the model to learn and produce more natural text (Yenduri et al., 2023). GPT-4 can accept input in document images and other objects and detect them. However, GPT-4 still needs to improve in detecting objects, or there still needs to be complete object detection, and the creation of captions based on image input that considers personality still needs to be improved. The following will discuss the comparison of the current GPT with the research to be carried out.

2.6. Settings

This research used the Python programming language on the Google Colab platform. Google Colab was chosen because it provides an NVIDIA A100 graphics processing unit (GPU), which offers high computational power and enables the acceleration of complex AI model training. This GPU is highly effective for handling models like Graph Convolutional Networks (GCN) requiring intensive computation.

In implementing deep learning models, this study utilized the TensorFlow and Keras libraries for scenery detection on images with the VGG16 model. PyTorch was used for YOLOv9 model inference to detect objects within images. Data processing was performed using Pandas and NumPy to facilitate the manipulation of table- and matrix-shaped data. Meanwhile, PIL (Python Imaging Library) was employed for image processing, and Matplotlib was used for visualizing detection results. Additionally, Deep Translator translated detected object names into Bahasa, adapting them to the local context.

3. Methodology

The research method in this thesis aims to develop an automatic caption generator system for social media using artificial intelligence. This research employs Python to process data. The approach involves a series of systematic stages, ranging from data collection and pre-processing to artificial intelligence model creation and training and model performance evaluation. Each stage is designed to ensure that the developed system can generate relevant and contextual captions that are innovative and engaging, ultimately improving user interaction within social media platforms. The stages of this research are shown in Figure 3.

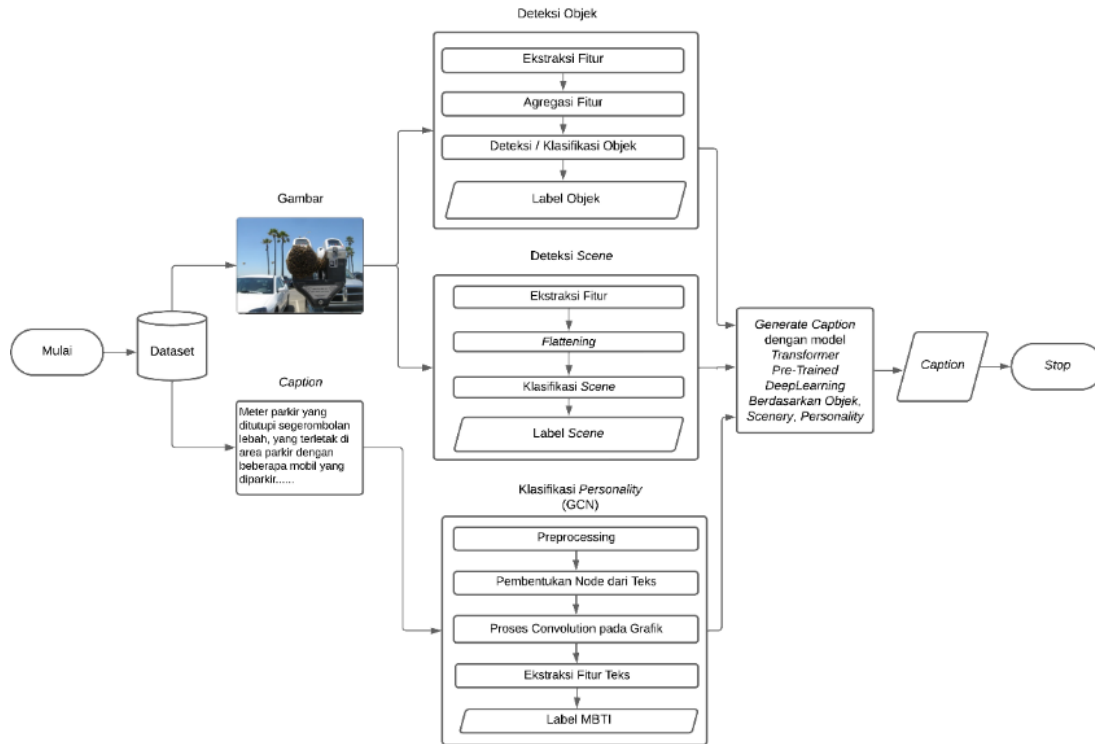


Figure 3.
Stages of researches.

3.1. Data Collection

This study will be sourced from public datasets, a dataset about predictions from the MBTI personality types dataset. This dataset is then used as a reference material to determine captions that match personality. This dataset consists of 16 classes representing personality types in MBTI (Myers-Briggs Type Indicator), and there are at least 210,000 Bahasa captions. Each class represents one of the 16 personality types, such as INTP, ISTP, and others. Data processing will be carried out from the data obtained to produce a text generator framework with artificial intelligence in social media.

3.2. Preprocessing

3.2.1. Text Preprocessing

In developing a caption generator system, pre-processing text data is essential after collecting and pre-processing image data. This step aims to transform text data, namely captions accompanying images, into a more standard format that is easy for AI models to process. Involves techniques designed to clean and prepare text so the model can interpret it effectively. The following stages of text data pre-processing will be used in this study: a) Tokenization, b) Stop Words, c) Text Normalization, d) Stemming and Lemmatization, and e) Word Embedding.

3.3. Image Preprocessing

Pre-processing steps on image data include several stages To prepared it for AI algorithms. In this study, The YOLOv9 is used to classify objects. Meanwhile, to classify scenery, the monk module will be used with the pre-trained model used VGG16. The stages of data pre-processing are as follows:

1. Resize

Resize is an advanced stage of compiling a dataset that changes the image size so that it is manageable and the complexity of classification computation is low. The resize size used in the type classification in this simulation is 224×224 pixels, which is the default input size for all classifiers used (Srinilta & Kanharattanachai, 2019).

2. Data Augmentation

Augmentation techniques are needed to overcome image recognition problems that affect performance when classifying image types, such as blurry images and lighting. Data augmentation is a technique that manipulates data without altering its core essence.

3.4. Model Development

The framework used in this study, namely the YOLO (You Only Look Once) model series, continues to develop, using YOLOv9, which was released on February 21, 2024 (Wang et al., 2024). YOLOv9 is the latest version of YOLO, and it is superior to the previous version. YOLO is one of the frameworks that can perform very efficient and accurate object detection and is suitable for this study, which requires object detection. YOLOv9 focuses on the improvements introduced through the Generalized Efficient Layer Aggregation Network (GELAN) to support any desired computational block. GELAN combines two network architectures, CSPNet and ELAN, and is designed with gradient path planning.

3.5. Object Detection with YOLOv9

The YOLOv9 model was chosen because of its ability to detect objects with high precision and efficient speed. This model analyzes the images in the dataset and then generates a list of detected objects and their confidence values. To explain the flow of the object detection process, here is a Pseudocode describing the main steps in this program.

Table 1.

Pseudocode of object detection process using YOLOv9.

1. INPUT: Image dataset, YOLOv9 model with weights, Inference parameters (image size, confidence threshold, IOU threshold)
OUTPUT: The CSV file contains the object detection results of each image in the dataset.
2. Start
3. Initialize the device to be used (GPU/CPU)
4. Download and load the YOLOv9 model along with the required weights
5. Initialize the parameters for inference (image size, confidence threshold, and IOU threshold)
6.
7. for each image in the dataset
8. Load images from the specified path
9. Preprocess images (resize, normalize, and convert format)
10. Run the YOLOv9 model to perform object prediction
11. Apply Non-Max Suppression (NMS) to reduce redundant predictions
12.
13. for each detected object
14. Get the object label and its confidence value
15. Translate the object label to Bahasa (if necessary)
16. Save the object detection results into a list
17. end for
18.
19. Save the object detection results into a CSV file
20. end for
21. Done

In the above process, each image in the dataset is processed individually. The image is fed into the YOLOv9 model, which outputs the prediction results as object labels and confidence scores. If no object is detected, the program records a particular label "xxxxx." The detection results for each image are then stored in a CSV file for further analysis.

3.6. Scenery Detection with VGG16

The process of detecting scenery in images using VGG16 can be seen in the pseudocode below:

Table 2.

Pseudocode scenery detection process using VGG16.

1. Input: Image dataset, Trained VGG16 model (ImageNet), Image path, Inference parameters (resize and normalization)
Output: CSV file containing scenery prediction results and confidence scores for each image
2. Start
3. Initialize the GPU or CPU to be used
4. Load the trained VGG16 model using the ImageNet dataset
5.
6. for each image in the dataset
7. Load images from the specified path
8. Preprocess images (resize and normalize)
9. Run prediction using the VGG16 model
10.
11. Decode the prediction results to retrieve scenery labels and confidence scores.
12.
13. for each detected scenery label
14. Translate the label to Bahasa
15. Save the scenery prediction results along with the confidence score into a list
16. end for
17.
18. Save the scenery prediction results into a CSV file
19. end for
20. Done

In this process, images from the dataset are taken in stages and processed, and predictions are made using the VGG16 model. The prediction results in the form of scenery labels and confidence scores are then translated into Bahasa using GoogleTranslator and saved into a CSV file.

3.7. Personality Classification with GCN

After the embedding is generated, the GCN model predicts MBTI personality based on the text embedding. This model consists of two convolutional layers that process the nodes in the graph and produce the final prediction in the form of MBTI labels. The flow in the GCN process is shown in the Pseudocode below:

Table 3.

Pseudocode GCN model structure.

1. Input: Graph data (node features and edge index), Initialized GCN model
Output: Final prediction of each node
2. Start
3. Initialize the GCN model with two layers:
4. Layer 1: Extracts features from nodes and generates new representations
5. Layer 2: Generates final predictions for each node
6.
7. In the forward process:
8. Get graph data (node features and edge index)
9. Apply convolution to the nodes
10. Use the ReLU activation function on the output of the first layer
11. Generate final predictions using softmax on the second layer

-
12. Return predictions
 13. Done
-

3.8. Caption Generation with GPT

The GPT is a variant of the GPT model adapted and explicitly trained for the Bahasa language. This model has been enriched with a training layer (attention layer) and Conditional Random Field (CRF) to improve accuracy and relevance in Natural Language Processing (NLP). After completing the classification process, the GPT model will generate automatic captions based on the classification results that were previously trained. The resulting captions describe accurate image content and are adjusted to the user's personality. The illustration of the output results in this study is a caption generated based on objects, scenery, and personality. The pseudocode below illustrates the process flow for understanding how captions are generated.

Table 4.

Pseudocode caption generation process using GPT.

-
1. Input: User uploaded image, MBTI personality, Reference caption (if any), YOLOv9 and VGG16 models
 - Output: Caption generated by GPT, BLEU, and ROUGE values (if reference caption exists)
 2. Start
 3. Set API key for OpenAI
 4. Load the YOLOv9 model for object detection
 5. Load the VGG16 model for scenery detection
 6. Initialize the widget to select MBTI personality and reference caption (if any)
 7. for each image uploaded by the user
 8. Display uploaded images
 9. Object detection in images using YOLOv9
 10. Scenery detection in images using VGG16
 - 11.
 12. if a reference caption is given
 13. Create a prompt for GPT that reflects the reference caption
 14. else
 15. Create a prompt for GPT that reflects the MBTI personality and image elements
 16. endif
 - 17.
 18. Send a prompt to GPT to generate a caption
 19. Display the generated caption to the user
 - 20.
 21. if a reference caption is given
 22. Calculate the BLEU score and ROUGE score between the reference caption and the generated caption
 23. Display BLEU and ROUGE scores to the user
 24. endif
 25. end for
 26. Done
-

The process begins with the user uploading an image. The system then detects objects using YOLOv9 and scenery using VGG16. Captions are generated based on the detected objects and scenery, taking into account the user's personality. If the user provides a reference caption, the generated caption will be tailored to the style and content of the reference.

Table 5.
Results of GCN model training trial.

No	MBTI	GPT		GPT + MBTI	
		Sesuai	Tidak Sesuai	Sesuai	Tidak Sesuai
1	ENFJ	0	3	2	1
2	ENFP	0	3	1	2
3	ENTJ	0	3	3	0
4	ENTP	0	3	1	2
5	ESFJ	0	3	3	0
6	ESFP	1	2	1	2
7	ESTJ	0	3	2	1
8	ESTP	0	3	3	0
9	INFJ	0	3	2	1
10	INFP	0	3	1	2
11	INTJ	0	3	2	1
12	INTP	0	3	2	1
13	ISFJ	0	3	1	2
14	ISFP	1	2	3	0
15	ISTJ	0	3	3	0
16	ISTP	0	3	1	2

3.9. Experiments and Results

Based on the table above, the validation results carried out by psychologists on the MBTI personality types and captions generated by two models, GPT and GPT + MBTI. The GPT model, which relies solely on artificial intelligence without any specific knowledge of MBTI personality types, shows a relatively low level of suitability in generating captions that fit the personality context. Overall, the compliance rate of the GPT model reached around 4.17% of the total validation data. Meanwhile, the GPT + MBTI combination model, which integrates personality information from MBTI into the caption creation process, showed a significant improvement in the suitability of captions to visual and personality contexts. Some personality types achieved a 100% fit, while the average fit rate increased substantially compared to the GPT model alone.

Overall, the GPT + MBTI model has proven more effective in generating captions that reflect Both visuals and the personality traits of the MBTI. The following subchapter will explain the results of the experiments for each model.

3.10. Training and Validation

In this experiment, the GCN model was trained with three training and validation data split configurations (80:20, 70:30, and 75:25). Grid search is used to find the best combination of hyperparameters, namely batch size (128, 256, and 512), learning rate (0.001 and 0.01), hidden dimension size (128, 256, and 512), and number of epochs (100, 200, and 300). The optimal combination of these hyperparameters was selected based on the best performance achieved on each proportion of the data. The best results of this experiment are summarized in Table 6.

Table 6.

Test results of GCN model training.

MBTI	Training data share proportion: Test data											
	80:20				70:30				75:25			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
enfj	0.5	0.39	0.44	145	0.46	0.44	0.45	217	0.46	0.38	0.41	181
enfp	0.49	0.56	0.52	180	0.5	0.54	0.52	270	0.44	0.63	0.52	225
entj	0.58	0.46	0.51	136	0.55	0.5	0.52	203	0.57	0.49	0.53	169
entp	0.48	0.54	0.51	117	0.53	0.49	0.51	176	0.57	0.51	0.54	146
esfj	0.45	0.32	0.38	47	0.58	0.42	0.49	71	0.6	0.42	0.5	59
esfp	0.58	0.47	0.52	40	0.52	0.53	0.53	60	0.46	0.49	0.48	51
estj	0.5	0.43	0.46	44	0.5	0.38	0.43	66	0.51	0.42	0.46	55
estp	0.66	0.79	0.72	29	0.64	0.68	0.66	44	0.76	0.78	0.77	37
infj	0.5	0.58	0.54	184	0.53	0.69	0.6	275	0.56	0.6	0.58	229
infp	0.57	0.49	0.53	180	0.53	0.53	0.53	270	0.49	0.51	0.5	225
intj	0.49	0.62	0.55	181	0.56	0.59	0.57	271	0.49	0.59	0.53	226
intp	0.48	0.5	0.49	143	0.52	0.55	0.53	214	0.56	0.48	0.52	178
isfj	0.51	0.44	0.47	84	0.53	0.4	0.46	126	0.57	0.5	0.53	105
isfp	0.54	0.46	0.49	46	0.5	0.47	0.49	70	0.57	0.47	0.51	58
istj	0.45	0.43	0.44	68	0.47	0.39	0.42	103	0.49	0.4	0.44	86
istp	0.65	0.71	0.68	42	0.61	0.49	0.54	63	0.68	0.52	0.59	52
Accuracy				1666				2499				2082
Macro avg	0.53	0.51	0.52	1666	0.53	0.51	0.52	2499	0.55	0.51	0.52	2082
Weighted avg	0.51	0.51	0.51	1666	0.53	0.53	0.52	2499	0.53	0.52	0.52	2082

Based on Table 6, the GCN model test results show the best performance at the 70:30 data split, with 70% training data and 30% validation data. In this configuration, the model achieved an accuracy of 0.53 and a macro F1-score of 0.52. While the 75:25 data split resulted in an accuracy of 0.52, the performance was still competitive.

Grid search identified the best hyperparameter combination for all three data configurations: batch size 512, hidden dimension 512, learning rate 0.001, and 300 epochs. This combination yielded consistent performance in detecting MBTI types, especially in the 70:30 data configuration. These results emphasize the importance of choosing the optimal data configuration and hyperparameters for the best performance, which forms the basis for the following sections on object detection, scene, and caption generation.

3.11. Object Detection

After finding the best hyperparameter combination using *grid search*, YOLOv9 is used to detect objects in several images from the *dataset*. Table 7 shows the results of image detection trials using YOLOv9 and the confidence score.

Table 7.
Object detection trial results using YOLOv9.

		
Image_2 (ISTJ)	Image_5 (ESTJ)	Image_11 (ENTJ)
tas tangan (0.24), orang (0.68), sepeda motor (0.75), orang (0.77), sepeda motor (0.82), sepeda motor (0.83), sepeda motor (0.88), sepeda motor (0.90), sepeda motor (0.95)	mobil (0.10), orang (0.14), mobil (0.21), meteran parkir (0.42), mobil (0.60), mobil (0.79), meteran parkir (0.87), mobil (0.95)	orang (0.11), ransel (0.12), tas tangan (0.15), orang (0.16), orang (0.16), kursi (0.18), orang (0.23), kursi (0.29), orang (0.36), orang (0.40), ransel (0.42), koper (0.51), payung (0.54), koper (0.59), orang (0.65), orang (0.67), ransel (0.74), koper (0.74), koper (0.74), ransel (0.77)
		
Image_17 (ENTP)	Image_29 (ISTP)	Image_37 (ENFJ)
orang (0.63), orang (0.67), papan luncur (0.75), orang (0.92)	orang (0.11), lampu lalu lintas (0.13), orang (0.22), mobil (0.29), orang (0.34), mouse (0.36), mobil (0.55), sepeda (0.58), mobil (0.61), mobil (0.61), mobil (0.66), mobil (0.82), mobil (0.83), mobil (0.84), hidran	orang (0.44), dasi (0.89), orang (0.95)

	kebakaran (0.90), tanda berhenti (0.92), mobil (0.93)	
--	---	--

From these detection results, the YOLOv9 model shows reliable object detection capabilities with high confidence on various images. For example, in Image_2 (ISTJ), a motorcycle object is detected with a confidence score of 0.95; in Image_37 (ENFJ), a person object is detected with the same confidence score. However, some images show lower confidence scores, such as in Image_29 (ISTP), where the person object only has a confidence score of 0.11, possibly due to image quality or overlapping objects. Overall, YOLOv9 performed efficiently and accurately, especially on good-quality images. This variation in confidence score can be used as a guide to focus on more prominent objects in caption generation, resulting in a more accurate and informative description based on the main detected objects.

3.12. Scenery Detection

After determining the best hyperparameters, the VGG16 model detects *scenery* on various images from the dataset. The scenery detection results for several photos are shown in Table 4.5. In some instances, the detected background objects include multiple items that may be present in the background context, such as motor scooters, parking meters, or limousines. The following are the detection results for select images:

Table 8.
Object detection trial results using VGG16.

		
Image_2 (ISTJ)	Image_5 (ESTJ)	Image_11 (ENTJ)
skuter_motor (0.87), sepeda motor (0.10), helm_tabrakan (0.02)	meteran_parkir (0.59), kubah (0.02), masjid (0.02)	mobil_salju (0.34), usungan (0.13), skuter_motor (0.10)
		
Image_17 (ENTP)	Image_29 (ISTP)	Image_37 (ENFJ)
bikini (0.26), snorkel (0.23), celana renang (0.22)	limousine (0.76), lampu lalu lintas (0.13), taksi (0.07)	pengantin pria (0.85), dasi kupu-kupu (0.05), setelan (0.03)



These results show that the VGG16 model can detect various scenery objects within images. However, the confidence score varies depending on the complexity and visibility of the objects within the images. This scenery detection, information about the object, and MBTI personality will be used in caption generation to create a more detailed image description that fits the visual context.

3.13. Caption Generation

Once the objects and scenery are detected, the GPT model creates captions that reflect the elements within images. This process considers the previously identified information of the object, scenery, and the MBTI personality. If the user provides a reference caption, GPT will customize the caption style. Otherwise, the system automatically generates captions that align with both the image and the user's personality to create personalized, creative, and contextual captions. The results are presented in Table 9, which shows the detected images, objects, scenery, and captions generated by GPT.

Table 9.
Caption creation with GPT.

Image / Personality	Object	Scenery	Caption GPT
 ISTJ	tas tangan (0.24), orang (0.68), sepeda motor (0.75), orang (0.77), sepeda motor (0.82), sepeda motor (0.83), sepeda motor (0.88), sepeda motor (0.90), sepeda motor (0.95)	skuter_motor (0.87), sepeda motor (0.10), helm_tabrakan (0.02)	Dengan tas di tangan dan dikelilingi oleh sepeda motor, kita pastikan setiap detail telah dipertimbangkan; dari keselamatan mengenakan helm hingga to-do list yang harus dilakukan hari ini. Ketertiban dan tanggung jawab adalah matahari yang membuat hari ini bercahaya.
 ESTJ	mobil (0.10), orang (0.14), mobil (0.21), meteran parkir (0.42), mobil (0.60), mobil (0.79), meteran parkir (0.87), mobil (0.95)	meteran_parkir (0.59), kubah (0.02), masjid (0.02)	Ketertiban adalah kunci. Inilah sudut kota kita, diatur dengan disiplin yang luar biasa - setiap mobil memiliki tempatnya, setiap meteran parkir berfungsi dengan baik. Ini bukan hanya pemandangan sehari-hari, tetapi simbol tatanan dan tanggung jawab.
 ENTJ	orang (0.11), ransel (0.12), tas tangan (0.15), orang (0.16), orang (0.16), kursi (0.18), orang (0.23), kursi (0.29), orang (0.36), orang (0.40), ransel (0.42), koper (0.51), payung (0.54), koper (0.59), orang (0.65), orang (0.67), ransel (0.74), koper (0.74), koper (0.74), ransel (0.77)	mobil salju (0.34), usungan (0.13), skuter_motor (0.10)	Di setiap perjalanan, pemimpin selalu siap untuk tantangan berikutnya. Mereka tidak hanya membawa ransel, tetapi juga semua rencana dan antisipasi dalam setiap detail. Hujan atau salju, mereka tetap bergerak maju.
 ENTP	orang (0.63), orang (0.67), papan luncur (0.75), orang (0.92)	bikini (0.26), snorkel (0.23), celana renang (0.22)	Berani mencoba hal baru, mencari petualangan, dibumbui sedikit adrenalin dan keunikan. Karena hidup ini tentang berbagi cerita, jadi kenapa tidak menciptakan yang paling menarik? Ah, dan snorkeling

			dengan celana renang polkadot tentunya!
 ISTP	orang (0.11), lampu lalu lintas (0.13), orang (0.22), mobil (0.29), orang (0.34), mouse (0.36), mobil (0.55), sepeda (0.58), mobil (0.61), mobil (0.61), mobil (0.66), mobil (0.82), mobil (0.83), mobil (0.84), hidran kebakaran (0.90), tanda berhenti (0.92), mobil (0.93)	limousine (0.76), lampu lalu lintas (0.13), taksi (0.07)	Saat melihat kepadatannya, jalanan ini tampak seperti pertandingan catur raksasa, setiap keputusan yang diputuskan oleh setiap individu di sini memiliki dampak langsung pada semua orang. Menikmati tarian dinamis antara mobil dan sepeda, merasa tenang di tengah kaos - inilah sisi dari kota yang sangat saya hargai.
 ENFJ	orang (0.44), dasi (0.89), orang (0.95)	pengantin pria (0.85), dasi kupu-kupu (0.05), setelan (0.03)	Memandu kehidupan dengan kebijaksanaan, memancarkan cinta dalam setiap detail pada setelan terbaiknya. Setia pada tugas, hidup dalam keseimbangan dan harmoni. Layaknya pengantin pria yang siap memimpin pesta pernikahannya, dia selalu hadir untuk memandu dan mengikat semua sekelilingnya dengan kasih sayang dan pemahaman yang mendalam.

3.14. Qualitative and Quantitative Evaluation

3.14.1. Expert Evaluation

The system evaluates the generated captions using two well-known text metrics: BLEU and ROUGE. These metrics assess the quality of automated text, such as in caption generation, machine translation, and summarization. BLEU measures n-gram precision to see how many n-grams in the result text match the reference text. A high BLEU score indicates substantial similarity with the reference text, and the algorithm also applies a brevity penalty to avoid unreasonable scores on short texts. The results of the BLEU score are shown in Table 10.

Table 10.
BLEU SCORE trial results.

Caption	Bleu MBTI	Bleu no MBTI
"Oven microwave dengan pintunya terbuka, mengungkapkan satu CD atau DVD di dalamnya..."	0.0066	0.0055
"Tujuh sepeda motor dari berbagai warna, termasuk perak, merah, dan hitam..."	0.025	0.0035
"Kamar mandi dengan wastafel, toilet dengan tangki, bak mandi, dan tirai kamar mandi."	0.0096	0.0041
"Meter parkir yang ditutupi segerombolan lebah, yang terletak di area parkir dengan beberapa mobil yang diparkir..."	0.0037	0.0052
"Sebuah layang-layang terbang di wilayah tengah atas di atas laut."	0.0113	0.0037

A low BLEU score indicates that the similarity between the generated caption and the reference text still needs to be improved. For example, a score of 0.0018 on a caption signifies very low similarity, indicating that the model still requires optimization in generating captions (Post, 2018).

The ROUGE metric evaluates recall and precision, i.e., how well the generated captions reflect the words or phrases from the reference text. The ROUGE variants used in this study include ROUGE-1, ROUGE-2, and ROUGE-L. The evaluation results for ROUGE-1 show a precision of 0.3, recall of 0.077, and f-measure of 0.123, which indicates that the model successfully captures some words from the reference, but still needs improvement to produce more coherent captions (Lin & Hovy, 2003).

Although the BLEU and ROUGE scores are low, the captions produced could be of better quality. These two metrics only assess text similarity, not aspects of interpretation or creativity. Captions can still be semantically or descriptively relevant, even if their structure differs from the reference text (Ridho Lubis et al., n.d.).

3.15. Quantitative Metrics

In this qualitative test, a psychologist was engaged to validate the MBTI personality predictions and captions generated by the GPT model. This validation ensures the prediction results are technically accurate, semantically relevant, and appropriate to the analyzed visual context and personality characteristics. The psychologist evaluated two main aspects: first, the validation of the MBTI personality predictions generated based on the images; second, the assessment of whether the captions from the GPT model correctly and relevantly described the images. The validation results were grouped into Appropriate (Sesuai) and Inappropriate (Tidak Sesuai) categories to indicate the degree of congruence between the predicted results and the expected interpretation.

This validation provides insights that cannot be achieved with quantitative metrics alone, as it involves subjective interpretations of the meaning and context of images and personalities. Table 11 contains the full results of this validation, showing the evaluation of each caption and the resulting MBTI predictions. It also measures how well the system relates images to personality and the accuracy of the GPT model in captioning.

Table 11.
Expert validation of MBTI predictions and captions generated by GPT¹.

No	Caption	MBTI prediction	Psychologist validation	Caption (GPT)	Psychologist validation
1	Oven microwave dengan pintunya terbuka, mengungkapkan satu CD atau DVD di dalamnya. Microwave dihidupkan, seperti yang ditunjukkan oleh tombol -tombol yang diterangi dan tampilan digital.	ISTJ	Tidak sesuai	Di ambang batas kemajuan digital, namun tetap merangkul keanggunan analog. Mempelajari duniawi dari lensa televisi, sambil menikmati minuman favorit. Kepercayaan dan konsistensi merupakan kunci, dan itu semua ditangkap disini, dalam petikan waktu melalui Polaroid ini.	Sesuai
:	:	:	:	:	:
48	Pengaturan luar ruangan yang hidup dengan banyak orang	ISFJ	Tidak sesuai	Menikmati petualangan menjelajahi piramida batu kasar di tanah	Tidak sesuai

<https://github.com/IkhsanAriansyah1994/mbti/blob/main/Validasi-expert-system.pdf>

yang mengenakan berbagai jenis alas kaki, termasuk sandal jepit, jeans, dan celana pendek. Seekor anjing dengan mantel coklat hadir dalam adegan itu, dikelilingi oleh sandal jepit. Orang - orang tampaknya berada dalam suasana santai, dengan satu orang mengenakan celana jeans di sisi kiri bingkai dan satu lagi di paling kanan, dekat anjing. Seseorang yang mengenakan celana pendek diposisikan lebih terpusat dalam adegan.			Mexico, kita semua tak lupa memperhatikan detail yang sering diabaikan yang ikut menambah kegembiraan. Sama seperti anjing terrier sutra yang menjadi teman setia dalam eksplorasi ini, nexus antara masa lalu dan sekarang.	
--	--	--	--	--

Based on the validation results in Table 11, the system performed well in predicting MBTI personality types and generating relevant captions. The expert rated 29 out of 48 MBTI predictions as 'Sesuai,' achieving an accuracy of about 60.42% in identifying personality types from images. This indicates that the system is quite effective in recognizing the visual elements associated with personality.

In the caption validation, 31 out of 48 captions were deemed 'Sesuai,' achieving an accuracy of 64.58%. This indicates that the GPT model successfully captured the visual context of most images. However, 35.42% of the captions were not 'Sesuai,' especially in images with more abstract or complex visual interpretations, suggesting that there is still room for improvement in deeper semantic understanding.

Furthermore, an experiment was conducted without considering the MBTI, which revealed that only 2 out of 48 objects aligned with the MBTI, representing about 4.17%. These findings only apply to text generated by the GPT model. The results show that the model's capacity to generate relevant text is minimal without referencing personality types. These differences underscore the challenges of producing accurate and meaningful descriptions when personality context is not included. The results of this evaluation can be seen in Table 12.

Table 12.Expert validation of GPT-generated MBTI predictions and captions without considering MBTI².

No	Caption	MBTI prediction	Psychologist validation	Caption (GPT)	Psychologist validation
1	Oven microwave dengan pintunya terbuka, mengungkapkan satu CD atau DVD di dalamnya. Microwave dihidupkan, seperti yang ditunjukkan oleh tombol - tombol yang diterangi dan tampilan digital.	ISTJ	Tidak sesuai	Berada di ruangan aman yang tenang, sebuah televisi dan botol berdiri sebagai saksi bisu era sebelumnya. Di depan osiloskop dan Kamera Polaroid, mereka menceritakan kisah tentang evolusi teknologi.	Tidak Sesuai (INFP)
:	:	:	:	:	:
48	Pengaturan luar ruangan yang hidup dengan banyak orang yang mengenakan berbagai jenis alas kaki, termasuk sandal jepit, jeans, dan celana pendek. Seekor anjing dengan mantel coklat hadir dalam adegan itu, dikelilingi oleh sandal jepit. Orang-orang tampaknya berada dalam suasana santai, dengan satu orang mengenakan celana jeans di sisi kiri bingkai dan satu lagi di paling kanan, dekat anjing. Seseorang yang mengenakan celana pendek diposisikan lebih terpusat dalam adegan.	ISFJ	Tidak sesuai	Sekelompok teman berkumpul di tengah reruntuhan piramida batu kasar, tertawa dan berbagi cerita. Seorang anjing terrier sutra berlarian di antara mereka, menambah keceriaan suasana. Dibalik serunya, tersembunyi keindahan tak tergantikan dari scenery Meksiko kuno ini.	Tidak Sesuai (ESFP)

4. Conclusion and Future Work

Based on the results of the research that has been conducted, several important points can be concluded as follows:

1. This study shows that combining the YOLOv9-based object detection model for objects and the VGG16 model for scenery can accurately identify visual elements in images. The system can recognize various objects and backgrounds in images, which are then used to create relevant descriptions. Applying the grid search technique to determine the optimal hyperparameters also helps improve the accuracy of object and scenery detection in the model.

2. The developed system successfully recognizes the user's personality based on the captions and objects identified within images. Using the MBTI (Myers-Briggs Type Indicator) approach, this model can predict the user's personality type with reasonably high accuracy. Validation by psychology experts shows that the predictions generated by the system are reliable, although some areas still require improvement in handling more complex personality interpretations.

<https://github.com/IkhsanAriansyah1994/mbti/blob/main/Validasi-expert-system-nonMBTI.pdf>²

3. The system successfully combines the identified visual content with the user's personality predictions to produce relevant and appropriate captions. With the help of the GPT model, the resulting captions can capture the essence of the image and the user's personality. Expert validation showed that the generated captions were accurate in most cases, but some captions needed to be more appropriate for images with more abstract visual interpretations.

Copyright:

© 2024 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

References

- [1] Akma, N., Amin, M., Nur, N., Sjarif, A., & Sophiayati Yuhaniz, S. (2023). Convolutional Neural Network Model for Facial Feature Extraction in Personality Traits Recognition Article history. Dalam *Open International Journal of Informatics (OIJI) (RATIPS)* (Vol. 11, Nomor 2).
- [2] Amirhosseini, M. H., Yadav, V., Serpell, J. A., Pettigrew, P., & Kain, P. (2024). An artificial intelligence approach to predicting personality types in dogs. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-52920-9>
- [3] Carvalho, L. de F., Pianowski, G., & Gonçalves, A. P. (2020). Personality differences and COVID-19: Are extroversion and conscientiousness personality traits associated with engagement with containment measures? *Trends in Psychiatry and Psychotherapy*, 42(2), 179–184. <https://doi.org/10.1590/2237-6089-2020-0029>
- [4] Chong, E. J., & Varathan, K. D. (2021). Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in an online social forum. *PeerJ*, 9, e11382. <https://doi.org/10.7717/peerj.11382>
- [5] Christou, P. (2023). How to Use Artificial Intelligence (AI) as a Resource, Methodological, and Analysis Tool in Qualitative Research? *The Qualitative Report*. <https://doi.org/10.46743/2160-3715/2023.6406>
- [6] Feng, S. Y., Lu, K., Tao, Z., Alikhani, M., Mitamura, T., Hovy, E., & Gangal, V. (2022). Retrieve, Caption, and Generate Visual Grounding for Enhancing Common Text Generation Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 10618–10626. <https://doi.org/10.1609/aaai.v36i10.21306>
- [7] Kumar, N. K., Vigneswari, D., Mohan, A., Laxman, K., & Yuvaraj, J. (2019). Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 107–109. <https://doi.org/10.1109/ICACCS.2019.8728516>
- [8] Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, 71–78. <https://doi.org/10.3115/1073445.1073465>
- [9] Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2020). Recent trends in deep learning-based personality detection. *Artificial Intelligence Review*, 53(4), 2313–2339. <https://doi.org/10.1007/s10462-019-09770-z>
- [10] Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191. <https://doi.org/10.18653/v1/W18-6319>
- [11] Purba, & Irwansyah. (2022). User Generated Content dan Pemanfaatan Media Sosial Dalam Perkembangan Industri Pariwisata: Literature Review. *Jurnal Komunikasi Dan Administrasi Publik*, 9(2), 229–238.
- [12] Putra, S. A., & Wijaya, A. (2023). ANALISIS SENTIMEN ARTIFICIAL INTELLIGENCE (AI) PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE LEXICON BASED. *JuSiTik: Jurnal Sistem dan Teknologi Informasi Komunikasi*, 7(1), 21–28. <https://doi.org/10.32524/jusitik.v7i1.1042>
- [13] Ridho Lubis, A., Ramdani Safitri, H., & Lubis, M. (t.t.). IMPROVING TEXT SUMMARIZATION QUALITY BY COMBINING T5-BASED MODELS AND CONVOLUTIONAL SEQ2SEQ MODELS. Dalam *Journal of Applied Engineering and Technological Science* (Vol. 5, Nomor 1).
- [14] Sazali, H., Siregar, Y. D., & Putri, I. A. (2020). The Impact of Instagram Social Media on Religious Behavior of Mosque Youth in Siumbut Baru Village. *Malikussaleh Social and Political Reviews*, 1(1), 25. <https://doi.org/10.29103/mspr.v1i1.2905>
- [15] Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GA AIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction*, 39(13), 2724–2741. <https://doi.org/10.1080/10447318.2022.2085400>
- [16] Seo, E. J., Park, J.-W., & Choi, Y. J. (2020). The Effect of Social Media Usage Characteristics on e-WOM, Trust, and Brand Equity: Focusing on Users of Airline Social Media. *Sustainability*, 12(4), 1691. <https://doi.org/10.3390/su12041691>
- [17] Sharma, G., Kalena, P., Malde, N., Nair, A., & Parkar, S. (2019). Visual Image Caption Generator Using Deep Learning. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3368837>
- [18] Srinilta, C., & Kanharattanachai, S. (2019). Municipal solid waste segregation with CNN. *Proceeding - 5th International Conference on Engineering, Applied Sciences and Technology, ICEAST 2019*, 1–4. <https://doi.org/10.1109/ICEAST.2019.8802522>
- [19] Sufi, F. (2024). Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. *Information*, 15(2), 99. <https://doi.org/10.3390/info15020099>

- [20] Wang, H., Zhang, Y., & Yu, X. (2020). An Overview of Image Caption Generation Methods. *Computational Intelligence and Neuroscience*, 2020, pp. 1–13. <https://doi.org/10.1155/2020/3062706>
- [21] Wayan Sugiartana, I., Studi Administrasi Bisnis Denpasar, P., Sekolah Tinggi Ilmu Sosial dan Politik Wira Bhakti Denpasar Wayan Candra Winetra, I. I., Studi Teknik Elektro Denpasar, P., & Politeknik Negeri Bali Ida Ayu Putu Febri Imawati, I. (2022). *EFEKTIFITAS PENGGUNAAN CODE GENERATOR (UNDAGI CODE CREATOR) DALAM MEMBANGUN SISTEM INFORMASI MODUL ADMIN* (Vol. 12, Nomor 2). Online. <https://ojs.mahadewa.ac.id/index.php/jmti>
- [22] Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. <http://arxiv.org/abs/2305.10435>