KIMage: Exploring Aging Perceptions Through User-Generated Photos Using a Mixed-Method Approach Combining Computer Vision and Natural Language Processing

Thomas Secco
IDACUS
Furtwangen University
Furtwangen, Deutschland
thomas.secco@mines-albi.fr

Liane-Marina Meßmer
IDACUS
Furtwangen University
Furtwangen, Deutschland
Liane-Marina.Messmer@hs-furtwangen.de

Christoph Reich
IDACUS
Furtwangen University
Furtwangen, Deutschland
christoph.reich@hs-furtwangen.de

***Abstract*—The study of aging perceptions has traditionally relied on self-report questionnaires, often limited by predefined categories and explicit cognitions. With the advent of modern technology, particularly the ubiquity of smartphones, new methodologies for capturing views on aging in real-time have emerged. This project explores a novel approach to assessing views on aging through the analysis of user-generated photographs and descriptions. We developed a mixed-method framework combining computer vision and natural language processing to label and analyze these images and their associated texts. By engaging participants from diverse age groups, our study identifies the varying content and relative importance of aging-related themes across the lifespan. Our findings reveal age-specific foci and offer insights into multidimensional and multidirectional views on aging. This research not only contributes to the methodological advancement in aging studies but also informs strategies to promote healthy aging and intergenerational dialogue**

***Index Terms*—Aging perceptions, Computer vision, Natural language processing**

CONTENTS

I. Introduction

A. Background

Views on aging significantly influence individual behaviors, societal norms, and public policy. Traditional methods of capturing these views predominantly involve self-report quantitative questionnaires, which, while valuable, often constrain participants to predefined categories and do not fully capture the complexity of aging perceptions. Recent advancements in technology provide an opportunity to explore more nuanced and dynamic methods of assessment. Smartphones, always at hand, enable the ecological momentary capture of everyday experiences, offering a rich, real-time perspective on how individuals perceive aging.

B. Aims

This project aims to systematically investigate the differences in the content and relative importance of views on aging across various age groups. By utilizing a photo-based assessment method, we seek to capture both qualitative and quantitative data, providing a comprehensive understanding of aging perceptions. Specifically, we aim to:

1) Develop a robust framework for labeling and analyzing photographs and descriptions related to aging.
2) Identify age-specific themes and the relative emphasis placed on different aspects of aging.
3) Explore the multidimensionality and multidirectionality of views on aging across the lifespan.

C. Methodology

To achieve these aims, we engaged participants from three distinct age groups: young adults (20-30 years), young-old adults (50-69 years), and old-old adults (70+ years). Participants were asked to take photographs representing what aging means to them and provide brief descriptions for each image. Using a combination of computer vision techniques for image analysis and natural language processing for text analysis, we labeled and categorized the collected data. This mixed-method approach allows for a detailed examination of the content and sentiment expressed in the photographs and descriptions. [1] [2]

D. Significance

Our research addresses several gaps in the current understanding of aging perceptions. By moving beyond traditional self-report methods, we capture a more authentic and immediate view of how aging is perceived in everyday life. The insights gained from this study have the potential to inform interventions aimed at promoting positive aging and enhancing intergenerational communication. Moreover, our methodological advancements contribute to the broader field of psychological and gerontological research, providing a template for future studies exploring complex social and psychological phenomena.

E. Workflow Diagram

The handling of the datas in summarized in the worflow diagram below:

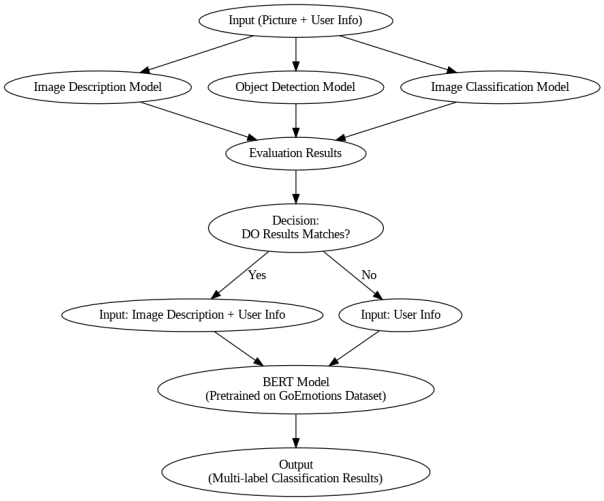

Fig. 1: Workflow Diagram.

F. Structure of the Paper

The paper is structured as follows: We begin with a detailed description of our data collection process, followed by an overview of the analytical techniques employed. We then present our findings, highlighting key themes and variations across age groups. Finally, we discuss the implications of our results for theory, practice, and future research, and conclude with a reflection on the strengths and limitations of our study.

II. Data Collection Process

A. Participants and Recruitment

The study targeted three distinct age groups from a city in North-Western Germany: younger adults (20 to 30 years), young-old adults (50 to 69 years), and old-old adults (70 years and above). Participants were approached in various public spaces and briefed about the study's goal to gather images reflecting personal perceptions of aging. Those interested were invited to a preliminary meeting where they provided written informed consent in accordance with ethical standards. [3]

B. Data Collection Methodology

At the initial meeting, participants received instructions to capture images over a five to eight-day period using their own digital devices or a single-use camera provided by the research team. They were

advised to avoid photographing recognizable faces to ensure privacy. Permissible images included partial views, back shots, and public scenes where individual identities were obscured. Participants could also use images of people from publicly available media that did not allow for individual identification. They were encouraged to freely choose subjects that they believed best represented their perception of aging.

C. Submission and Initial Processing

Participants returned their cameras or submitted digital storage devices containing the images. All photographic content was transferred to secure, pseudonymized electronic storage to maintain confidentiality. Any textual descriptions provided by the participants, explaining the context or intent behind each image, were also digitized and stored anonymously.

D. Post-Collection Review

Following the submission, participants were asked to attend a second meeting where they reviewed their photographs along with their accompanying text. This step was crucial for clarifying any ambiguous or incomplete descriptions and ensured that the research team accurately understood the intent behind each image.

E. Ethical Considerations

This study was conducted in full compliance with the ethical principles of the German Psychological Society [4] as well as the Declaration of Helsinki [5]. The research design was reviewed and approved by an institutional review board. Written informed consent was obtained from all participants prior to any data collection, and strict measures were implemented to protect the privacy and anonymity of the participants throughout the study.

F. Data Distribution Analysis

To ensure a comprehensive understanding of aging perceptions across different age groups, we categorized the collected data into six main categories: Physical Changes, Living Environment, Competencies, Resilience, Social Embeddedness, and Lifestyle & Engagement. Each main category was further divided into subcategories, capturing various dimensions of aging as perceived by participants. The following analysis delves into the significance of the data distribution across these Categories. See Annex for the diagrams.

a) Physical Changes: The category "Physical Changes" covers several critical aspects of aging. Notably, Functional Change (29.6%) and Illness (18.2%) are the most significant subcategories. This highlights that physical functionality and health-related concerns are predominant themes in how participants perceive aging. Appearance Change (10.8%) also represents a considerable focus, indicating that changes in physical appearance are a notable concern. The combined focus on Functional Change and Illness suggests that maintaining physical health is a priority in aging perceptions.

b) Living Environment: In the "Living Environment" category, a significant portion of the data (53.3%) is not specifically related to this category, indicating a potential overlap with other categories or a lack of explicit emphasis on living environment aspects. Among the specified subcategories, Infrastructure (29.1%) is the most prominent, followed by Aids in Everyday Life (10.8%). This suggests that participants are particularly concerned with the physical and infrastructural aspects that support their daily living as they age.

c) Competencies: The "Competencies" category shows a dominant focus on areas not explicitly related to the main category (77.2%). Within the specified subcategories, Education and Mental Stimulation (15.4%) is more emphasized than Craft and Technology (7.4%). This indicates that mental engagement and lifelong learning are seen as important aspects of aging, though they are less frequently highlighted compared to other categories.

d) Resilience: In the "Resilience" category, Control (36.5%) emerges as a significant theme, reflecting the importance of maintaining autonomy and control over one's life during aging. The next substantial subcategory, Experience of Time (11.4%), indicates that how individuals perceive and value their time is also a critical aspect. A large portion (45.3%) of data is not explicitly related to resilience, suggesting that this category may overlap with broader themes of aging.

e) Social Embeddedness: "Social Embeddedness" reveals that Family and Friends (25.1%) are crucial to participants' perceptions of aging, underscoring the importance of social connections. Loneliness (5.7%) and Love (10.3%) are also notable, reflecting both the positive and negative aspects of social relationships in aging. A significant portion of the data (49.9%) is not specifically categorized here, indicating that social embeddedness is a broad and multifaceted theme.

f) Lifestyle & Engagement: The "Lifestyle & Engagement" category shows that Being Active and Healthy Behavior (28.5%) is a primary concern, indicating a strong emphasis on maintaining an active and healthy lifestyle. Cultural Experience and Enjoyment (9.1%) and Pleasure and Recreation (15.4%) also highlight the importance of engaging in enjoyable activities. A considerable amount of data (24.2%) falls outside the specified subcategories, suggesting diverse interpretations of lifestyle and engagement in aging.

g) Overall Insights: The data distribution reveals several key insights. Physical health, functional ability, and social connections are prominent themes in aging perceptions. There is a strong emphasis on maintaining control and autonomy in aging. Mental

engagement and lifelong learning are valued, though less frequently highlighted. Social relationships, both positive and negative, play a significant role in how aging is perceived. A substantial portion of data across categories is not explicitly categorized, indicating overlapping themes and the multifaceted nature of aging perceptions.

These findings underscore the complexity and diversity of aging perceptions, highlighting the need for targeted interventions and strategies that address both the physical and social dimensions of aging. By understanding these distributions, we can better tailor efforts to promote healthy aging and enhance intergenerational dialogue.

III. Image Processing

For each submitted photograph, an initial descriptive caption was generated using an image description model [6]. However, this model did not provide a confidence score for its outputs. To establish confidence and enhance the reliability of the image interpretations, additional image classification and object detection models were employed.

A. Integration of Multiple Models

The image classification and object detection models included a CLIP (Contrastive Language–Image Pretraining) image classification model developed by OpenAI [7] , a pre-trained Faster R-CNN model with a ResNet50 backbone and a Feature Pyramid Network (FPN) object detection model [8], and the YOLO (You Only Look Once) model for real-time object detection [9]. The output of each model was compared to determine the confidence level of the descriptive caption. The process involved the following steps:

- Generate an initial caption using the captioning model.
- Obtain classification output from the image classification model.
- Detect objects in the image with the two object detection models.
- Standardize the output (removal of underscores and conversion to lowercase) for uniformity.

B. Semantic Analysis

Semantic similarity between the outputs of the models and the initial descriptive caption was calculated using a custom similarity function. This function is based on the medium-sized English language model provided by spaCy (`en_core_web_md`), which is utilized due to its balance between size and accuracy. This model enables effective semantic analysis of terms by considering their contextual likeness. The function assesses whether elements recognized by the models were mentioned either directly or indirectly (through semantically similar words) in the initial caption, allowing for a nuanced analysis of data congruence.

C. Categorization of Model Outputs

The integration of multiple model outputs involves categorizing detected elements into three groups based on the frequency of their recurrence across models:

- **High Confidence Group:** Elements that are identified by all three models are placed in this category. This group signifies the highest level of confidence in the relevance of the detected elements to the concept of aging as depicted in the images.
- **Medium Confidence Group:** Elements that are identified by two of the three models are categorized here. These elements have moderate confidence and support the themes identified, though with less certainty than the high confidence group.
- **Low Confidence Group:** Elements detected by only one model fall into this category, indicating the lowest confidence level. These may represent more subjective or less universally recognized aspects of aging.

D. Weighted Semantic Scoring Mechanism

The confidence score for each image caption is calculated using a weighted sum of similarity scores between terms detected in the image and terms in the caption. This is performed by categorizing detected terms based on the agreement across models and then measuring their semantic similarity to the caption. The formula for computing the weighted score is as follows:

$$S = \left(\frac{\sum_{i \in C_{\text{all}}} \max_{w \in \text{Words}}(\text{simi}(i,w))}{|C_{\text{all}}|} \times 0.6 \right) + \left(\frac{\sum_{i \in C_{\text{two}}} \max_{w \in \text{Words}}(\text{simi}(i,w))}{|C_{\text{two}}|} \times 0.3 \right) + \left(\frac{\sum_{i \in C_{\text{one}}} \max_{w \in \text{Words}}(\text{simi}(i,w))}{|C_{\text{one}}|} \times 0.1 \right) \quad (1)$$

Where:

- S is the total confidence score as a percentage.
- C_{all}, C_{two}, and C_{one} are the sets of class names found in all three, two, and one models' outputs, respectively.
- Words is the set of words in the caption.
- $\text{simi}(i, w)$ is the similarity function returning a score between 0 (no similarity) and 1 (identical) for the semantic similarity between class name i and caption word w.
- The factors 0.6, 0.3, and 0.1 are the weights assigned to the scores from C_{all}, C_{two}, and C_{one} respectively, reflecting their relative importance in the final score.

This scoring mechanism quantifies the relevance of detected image elements to the described aging concepts within the captions, reflecting the accuracy and depth of the models' interpretations.

The weights 0.6, 0.3, and 0.1 are arbitrary and may be subject to modifications in the future, with AI or other methods, to improve the accuracy and reliability of the scoring mechanism.

IV. MULTI-LABEL CLASSIFICATION PROCESS

A. Integration and Data Preparation

In the initial phase of our multi-label classification process, we focus on selecting textual data that integrates user inputs with image descriptions, ensuring a minimum confidence threshold of 50% for the image descriptions. If the confidence score is below 50%, then the multi-label classification process will only take as input the text given by the user. This careful selection guarantees the accuracy and relevance of the data fed into our system. The data is then divided into subsets: 70% for training (245 datas) and 30% for testing, with the latter equally split into validation and test datasets (53 and 53). This structure facilitates thorough validation of the model's predictive capabilities across varied data segments. [10]

B. Handling Valence and Productivity Columns

In our multi-label classification process, we carefully integrated the Valence and Productivity columns to streamline the model's input and ensure accurate label prediction. The Valence column originally categorized data into four distinct labels: positive (1), negative (2), neutral (3), and ambivalent (4). Similarly, the Productivity column classified data into three labels: productive/constructive/functional (1), unproductive/destructive/afunctional (2), and not assessable (3).

To utilize the same multi-label classification model for both valence and productivity, we transformed and merged these columns into new binary labels. This transformation process involved the following steps:

1) Creating New Labels: From the Valence column, we extracted and created two new binary labels: *positive* and *negative*. From the Productivity column, we generated two new binary labels: *productive* and *unproductive*.

2) Mapping Original Labels to New Labels: For the Valence column:

- *positive* was assigned a value of 1 if the original Valence value was 1 (positive) or 4 (ambivalent).
- *negative* was assigned a value of 1 if the original Valence value was 2 (negative) or 4 (ambivalent).

For the Productivity column:

- *productive* was assigned a value of 1 if the original Productivity value was 1 (productive/constructive/functional).
- *unproductive* was assigned a value of 1 if the original Productivity value was 2 (unproductive/destructive/afunctional).

We combined the newly created binary labels (*positive*, *negative*, *productive*, *unproductive*) to generate a multi-label representation for each data point. This combination allowed our multi-label classification model to handle both valence and productivity predictions simultaneously.

3) Reconstructing Original Columns: After the model's predictions, we reconstructed the original Valence and Productivity columns by mapping the predicted labels back to their respective categories. For instance:

- If both *positive* and *negative* labels were predicted as 1, the Valence was classified as ambivalent.
- If only *positive* was predicted as 1, the Valence was classified as positive.
- If only *negative* was predicted as 1, the Valence was classified as negative.
- If neither *positive* nor *negative* was predicted as 1, the Valence was classified as neutral.
- Similarly, if *productive* was predicted as 1, the Productivity was classified as productive.
- If *unproductive* was predicted as 1, the Productivity was classified as unproductive.
- If neither *productive* nor *unproductive* was predicted as 1, the Productivity was classified as not assessable.

By transforming and merging the Valence and Productivity columns into binary labels for the multi-label classification model, we ensured a more efficient and unified approach to predicting emotional and functional aspects of the data. This method facilitated accurate and comprehensive predictions, maintaining the integrity of the original data attributes while leveraging the power of multi-label classification.

C. BERT Model Configuration

After preparing the data, we configure the BERT model for our task [11]. We start by using the `BertTokenizer` to standardize and optimize text processing. This tokenizer splits text into tokens, converts them to numerical IDs, and adds special tokens required by BERT. It also handles padding and truncation, ensuring all input sequences have the same length for efficient processing.

We then create a custom dataset class, which manages tokenization, batching, and data preparation. This class loads data, tokenizes it using the `BertTokenizer`, generates attention masks, and organizes the data into batches. These steps ensure the data is in the optimal format for the BERT model, facilitating effective and accurate processing.

By carefully configuring the BERT model with these tools, we ensure that it can process the input data efficiently and accurately, leading to better performance on our specific task.

D. Pre-training on the GoEmotion Dataset

To enhance our model's performance for our specific task, we pre-train it on the GoEmotion dataset [12], which closely aligns with our classification goals. This dataset is rich in emotional content, making it ideal for

helping the model grasp the subtleties of emotional expression in text.

Pre-training involves adapting the model to understand the nuances of emotional content. This process involves training the model on the GoEmotion dataset before fine-tuning it on our specific task. By doing so, we provide the model with a strong foundational understanding of emotional text, improving its ability to accurately classify emotions in subsequent tasks. [13]

This pre-training step ensures that the model is well-prepared for the fine-tuning phase, leading to better performance and more accurate results in our specific application.

E. Model Training, Fine-Tuning, and Validation

After pre-training, the BERT model undergoes a critical training and fine-tuning phase. We employ the `BCEWithLogitsLoss` function suitable for our multi-label tasks and the `AdamW` optimizer, known for effectively managing sparse gradients. Throughout this phase, we test and adjust the model's settings to fine-tune its performance.

F. Threshold and Hyperparameter Optimization and Model Evaluation

After the training and fine-tuning phases, we finalize the optimal settings for our model. The following bullet points explain the role and importance of each hyperparameter and setting:

- **Classification Threshold (0.4):** This threshold is set to balance precision and recall optimally, crucial for accurately distinguishing between different labels in our multi-label classification task. It determines the cut-off point at which a probability prediction is considered positive.
- **Batch Size (32):** Chosen to provide a balance between processing efficiency and memory usage, allowing for stable gradient estimation. It affects how many data samples the model sees before making updates to its weights, impacting both training speed and model convergence.
- **Learning Rate (1e-04):** This rate ensures smooth and effective convergence during the training process. It controls the size of the updates to the model's weights with each batch of data, playing a critical role in how quickly the model learns and stabilizes its accuracy.
- **Dropout Rate (0.3):** Utilized to prevent overfitting by randomly dropping units (along with their connections) during the training phase. This rate helps the model to generalize better, making it robust to slight variations in input data.

These parameters are continuously monitored and adjusted if necessary, using the Hamming loss metric to ensure the accuracy of predictions across all labels.

V. Results

In this section, we present a comprehensive evaluation of our predictive model's performance.The following subsections will delve into the specifics of each evaluation phase, providing quantitative metrics and visual aids to elucidate the model's capabilities and limitations. This structured approach ensures a thorough understanding of the predictive model's reliability and accuracy, crucial for determining the optimal parameters for its application in our study.

A. First Results

To assess the accuracy of our predictive model, we employed an evaluation function designed to compare predicted labels against true labels across various thresholds. The `eval` function defines thresholds from -0.1 to 2.0, in steps of 0.1, and calculates the number of accurate predictions at each threshold level. A threshold of -0.1 indicates that the models did not give any prediction.

```
def eval(y_pred, y_true):
    bench = [-0.1 + i / 10 for i in range(22)
        ]
    good = []
    for j in range(len(bench)):
        t = 0
        for i in range(y_pred.shape[0]):
            pred_row = y_pred.iloc[i]
            true_row = y_true.iloc[i]
            # Exclude valence and
                productivity classes
            pred_labels = [col for col in
                pred_row.index[:-2] if
                pred_row[col] == 1]
            true_labels = [col for col in
                true_row.index[:-2] if
                true_row[col] == 1]
            c = sum([1 for y in pred_labels
                if y in true_labels])
            recall = c / len(true_labels) if
                len(true_labels) > 0 else
                -0.05
            precision = c / len(pred_labels)
                if len(pred_labels) > 0 else
                -0.05
            if len(pred_labels) == 0:
                recall = -0.05
            test = recall + precision
            if (j != len(bench) - 1 and bench
                [j] <= test < bench[j + 1])
                or (j == len(bench) - 1 and
                test >= bench[j]):
                t += 1
        good.append(t)
    return pd.DataFrame(good, index=bench,
        columns=['Count'])
```

The function follows these key steps:

1) **Threshold Definition:** Establishes benchmarks (`bench`) at intervals of 0.1, ranging from -0.1 to 2.0. A threshold of -0.1 indicates no predictions were made by the model.
2) **Prediction and True Labels Comparison:** Iterates through each prediction and true label pair, identifying the presence of predicted labels within the true labels. We exclude the last two

indices in 'pred_row.index[:-2]' to avoid including the valence and productivity classes in the comparison.

3) **Precision and Recall Calculation:** Computes precision as the ratio of correctly predicted labels to the total predicted labels and recall as the ratio of correctly predicted labels to the total true labels.
4) **Threshold Assignment:** Assigns counts to bins based on the threshold criteria, determining the number of predictions meeting each threshold level.

The outcome of this evaluation provides a detailed count of accurate predictions at each threshold level, allowing us to understand the performance and reliability of the predictive model under varying conditions. This comprehensive assessment is crucial for determining the optimal threshold for precision and recall in the context of our study.

a) Results on the Training Dataset: The evaluation function tested the model's performance across a range of thresholds from -0.1 to 2.0. The aim was to determine the number of accurate predictions at each threshold level, helping identify the optimal threshold for balancing precision and recall in our multi-label classification task.

The results of this evaluation are tabulated below:

Threshold	Count
-0.1	0
0.0	0
0.1	0
⋮	⋮
1.9	0
2.0	245

TABLE I: Model predictions at various thresholds

These results indicate that no pictures had the sum of precision and recall ranging from 0 to 1.9 and the model gave a prediction for each picture. However, at a threshold of 2.0, the model correctly predicted all 245 instances in the training dataset. This finding suggests that the model perfectly on the training dataset, with no errors and all the true labels have been found.

b) Results on the Validation Dataset: The evaluation function tested the model's performance on the validation dataset across a range of thresholds from -0.1 to 2.0. This analysis helps to verify the model's generalizability and performance on unseen data.

Counts of Accurate Predictions: The sum of precision and recall at each threshold level is detailed below:

For more detail results, the precision and recall are also calculated separately.

Recall: Recall measures the model's ability to identify all relevant instances correctly.

Precision: Precision assesses the accuracy of the predictions, measuring the proportion of correct predictions among the predicted categories.

TABLE II: Counts of accurate predictions at various thresholds

Threshold	Count
-0.1	15
0.0	14
0.1	0
0.2	0
0.3	0
0.4	0
0.5	0
0.6	0
0.7	0
0.8	0
0.9	0
1.0	2
1.1	0
1.2	0
1.3	1
1.4	0
1.5	8
1.6	1
1.7	0
1.8	0
1.9	0
2.0	12

TABLE III: Recall at various thresholds

Threshold	Recall
-0.1	15
0.0	14
0.1	0
0.2	0
0.3	1
0.4	0
0.5	7
0.6	1
0.7	0
0.8	0
0.9	0
1.0	15

TABLE IV: Precision at various thresholds

Threshold	Precision
-0.1	15
0.0	14
0.1	0
0.2	0
0.3	0
0.4	0
0.5	5
0.6	0
0.7	0
0.8	0
0.9	0
1.0	19

Analysis of Results: The analysis of the results reveals several key points:

- The model shows a significant number of correct predictions at extremely low (-0.1) and high (2.0) thresholds, suggesting a dichotomy in its confidence level; it either predicts with high certainty or remains uncertain.
- The recall and precision values indicate that the model is capable of capturing a significant num-

ber of true positives at a threshold of 1.0, but this comes with a trade-off in precision.

B. Box Plot

To visually represent the performance of the multi-label classification model, we generated box plots for the key metrics—precision, recall, F1-score, and accuracy—across all labels except for Valence and Productivity. The box plots provide a clear visualization of the distribution and variability of these metrics, calculated only on the validation dataset.

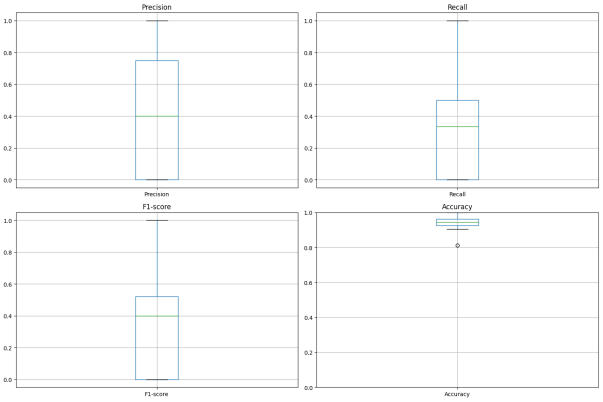

Fig. 2: Box plot across all the metrics.

Looking into the specifics of the diagram:

- **Precision**: The box plot shows a median precision around 0.5, with a range extending from near 0 to 1. This indicates that while the model achieves high precision for some labels, it struggles with others.
- **Recall**: The recall distribution is similar to precision, with a median around 0.4. This suggests that the model's ability to identify all relevant instances varies significantly across labels.
- **F1-Score**: The F1-score, which balances precision and recall, also shows a wide range, reflecting the variability in the model's performance.
- **Accuracy**: The accuracy plot shows a higher median value, around 0.95, with most values clustered between 0.9 and 1.0. This indicates that the model is generally accurate, but the presence of outliers suggests occasional misclassifications.

C. Result per label

The detailed performance metrics for each category, highlighting both the best and worst performing labels, are summarized in Table V.

Category	Precision	Recall	F1-Score
Best Performing Labels			
Appearance Change	0.50	1.00	0.67
Functional Change	0.83	0.71	0.77
Family and Friends	1.00	0.60	0.75
Pet	1.00	1.00	1.00
Cultural Experience and Enjoyment	1.00	0.50	0.67

TABLE V: Precision, Recall, and F1-Score for each Category

The results presented in Table V indicate that the model performs well in certain categories, such as *Pet*, *Functional Change*, and *Family and Friends*, achieving high precision, recall, and F1-scores. However, the model fails to identify other categories, such as *Dying or Death*, *Mobility*, *Education and Mental Stimulation*, *Openness or Wisdom*, *Festivities and Traditions*, *Craft and Technology* and *Loneliness*, where precision, recall, and F1-scores are all zero.

This discrepancy suggests that the model struggles with categories that may be less visually distinct or more abstract in nature. The poor performance in these areas could be due to several factors, including a lack of sufficient training data for these specific labels, inherent difficulty in visually distinguishing these themes, or limitations in the model's ability to interpret complex and nuanced aspects of aging. Further refinement of the model, including the incorporation of additional training data and more sophisticated feature extraction techniques, may be necessary to improve performance across all categories.

D. Valence and Productivity

a) Productivity Classification: The classification results for productivity, as illustrated in the confusion matrix (Figure 3), demonstrate the model's ability to differentiate between Productive/Constructive, Unproductive/Destructive, and Not Assessable classes.

The overall accuracy of the productivity classification indicates a strong performance in identifying Productive/Constructive instances, while the model shows some confusion between Unproductive/Destructive and Not Assessable categories.

b) Valence Classification: The valence classification results, as depicted in the confusion matrix (Figure 4), provide insights into the model's performance in identifying Positive, Negative, Neutral, and Ambivalent sentiments.

The valence classification results suggest that the model is effective in identifying Positive sentiments but shows some confusion when distinguishing between Negative, Neutral, and Ambivalent sentiments.

c) Analysis and Implications: The confusion matrices for both productivity and valence classifications reveal strengths and areas for improvement in the model's performance. The high accuracy in identifying Productive/Constructive and Positive classes highlights the model's robustness in these areas. However, the misclassifications in Unproductive/Destructive and Ambivalent categories suggest a need for further refinement.

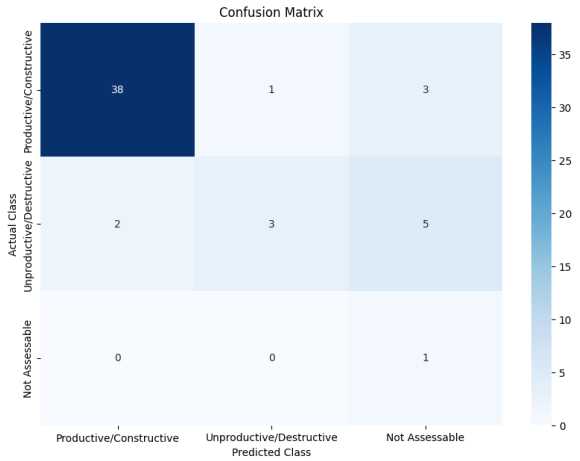

Fig. 3: Confusion Matrix for Productivity Classification

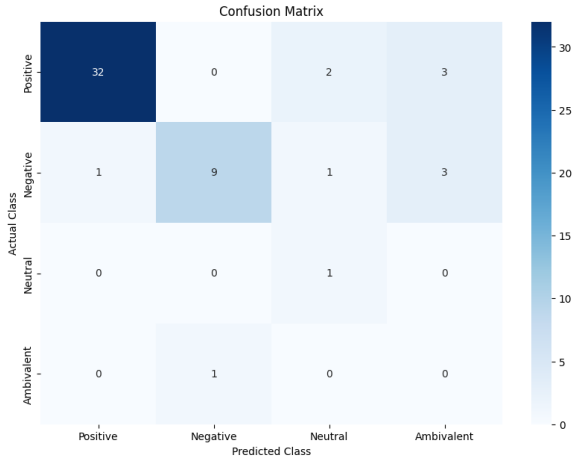

Fig. 4: Confusion Matrix for Valence Classification

VI. Future Work

The preceding results demonstrate the effectiveness of our approach. In this section, we explore potential strategies for enhancing these outcomes in future work.

A. Data Augmentation Techniques

To enhance the performance and robustness of machine learning models, it is essential to employ data augmentation techniques. These techniques help create more diverse and extensive training datasets, allowing models to generalize better and perform well on unseen data. This section delves into various data augmentation strategies for both image and text data. [14] [15]

The image augmentation can be used on the pictures before going through the image description model and combined with augmenting the text given by the user. However, it is also possible to only use text augmentation in our case: using it on both the image description and the user input, without changing the pictures.

a) Image Augmentation: Image augmentation involves applying a variety of transformations to the original images to create new, altered versions while maintaining the original class labels [16]. Below, we discuss some common image augmentation techniques and provide examples based on the provided image:

- **Rotation**: Images are rotated by a certain degree, typically between -30 to +30 degrees. This helps the model become invariant to the orientation of the object.
- **Flipping**: Images are flipped horizontally or vertically. This is useful when the object's orientation does not affect its class.
- **Scaling**: Images are resized to different scales, either zooming in or out. This ensures the model can handle objects of varying sizes.
- **Translation**: Images are shifted horizontally or vertically. This helps the model to be invariant to the position of the object within the frame.
- **Brightness Adjustment**: The brightness of the images is adjusted to create variations in lighting conditions.
- **Color Jitter**: Random changes are made to the hue, saturation, and contrast of the images. This helps the model generalize better under different color conditions.
- **Noise Addition**: Random noise is added to the images to make the model more robust to noisy data.
- **Cutout**: Random patches are cut out from the images to make the model focus on less obvious parts of the image.

Those techniques are shown on this picture example:

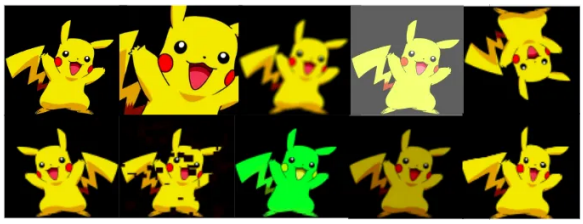

Fig. 5: Image Augmentation Techniques [17]

b) Text Augmentation: Text augmentation is equally important in natural language processing (NLP) to improve model robustness and performance by increasing the variety and amount of textual data [18] [19]. Some common techniques include:

- **Synonym Replacement**: Randomly replacing words with their synonyms to create different versions of the same sentence.
  - *Example*: "The quick brown fox jumps over the lazy dog" can become "The fast brown fox leaps over the lazy dog".
- **Random Insertion**: Adding new words at random positions in the sentence.

  - *Example*: "The quick brown fox jumps over the lazy dog" can become "The quick and brown fox jumps over the lazy dog".
- **Random Deletion**: Randomly removing words from the sentence.
  - *Example*: "The quick brown fox jumps over the lazy dog" can become "The quick fox jumps over the lazy dog".
- **Random Swap**: Swapping the positions of two words in the sentence.
  - *Example*: "The quick brown fox jumps over the lazy dog" can become "The brown quick fox jumps over the lazy dog".
- **Back Translation**: Translating the text to another language and then back to the original language to create a paraphrased version.
  - *Example*: Translating "The quick brown fox jumps over the lazy dog" to French and back to English might result in "The fast brown fox leaps over the lazy dog".
- **Sentence Shuffling**: Shuffling the order of sentences in a paragraph to create new versions.
  - *Example*: Given two sentences, "The quick brown fox jumps over the lazy dog. It is a sunny day.", shuffling can create "It is a sunny day. The quick brown fox jumps over the lazy dog."
- **Token Insertion and Deletion**: Adding or removing characters or words in tokens to introduce variety.
  - *Example*: "The quick brown fox jumps over the lazy dog" can become "The quick brown fox jumps ovre the lazy dog".

These augmentation techniques are crucial for creating a diverse and robust dataset, enabling models to perform better on unseen data by learning from a wider range of variations in the input data. By applying these techniques, we can significantly enhance the model's ability to generalize and perform well in real-world scenarios.

By using these data augmentation techniques, both image and text datasets can be enriched, leading to more effective and reliable models capable of handling diverse real-world data.

B. Better Labeling Strategies

As mentioned in part V.B, it seems that the that the model struggles with categories that may be less visually distinct or more abstract in nature. As such, changing or improving the data labeling may allow us to have better results

a) Improved Data Labeling: To enhance the accuracy and consistency of data labeling, implementing detailed annotation guidelines and conducting training sessions for annotators may be a solution. Providing annotators with comprehensive guidelines and examples for each category, particularly abstract concepts like Dying or Death and Loneliness, can ensure more consistent labeling. [20]

b) Data Augmentation and Balancing: Employing data augmentation and balancing techniques can address the issue of insufficient training data for certain categories. Synthetic data generation using Generative Adversarial Networks (GANs) can create realistic images representing underrepresented categories, thus balancing the dataset. For text data, advanced augmentation techniques, such as back-translation and contextual synonym replacement, can increase the variety of text examples. Additionally, utilizing the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for underrepresented categories and reducing samples in overrepresented categories can prevent the model from becoming biased. [21]

VII. Conclusion

This study has successfully demonstrated the potential of combining computer vision and natural language processing techniques to analyze user-generated photographs and descriptions for assessing views on aging. By developing a robust mixed-method framework, we were able to effectively label and categorize images and their associated texts, providing a comprehensive analysis of aging perceptions across different age groups.

Our approach involved the integration of multiple machine learning models, including CLIP, Faster R-CNN, and YOLO, to generate high-confidence descriptive captions for the images. The semantic similarity between model outputs and initial captions was quantified using a custom similarity function, ensuring accurate and nuanced interpretations of the data. This multi-model integration and the weighted semantic scoring mechanism provided a reliable framework for analyzing complex and abstract themes related to aging.

The multi-label classification process, which included pre-training on the GoEmotion dataset and fine-tuning with our specific data, further enhanced the model's ability to predict valence and productivity labels accurately.

Our findings reveal that physical health, functional ability, and social connections are prominent themes in aging perceptions, with a strong emphasis on maintaining control and autonomy. The model's performance varied across different categories, highlighting the need for more refined labeling strategies and data augmentation techniques to address the challenges posed by less visually distinct or more abstract themes.

In summary, this research underscores the importance of leveraging modern technology to enhance the accuracy and depth of data interpretation in aging studies. The technical innovations introduced in this study provide a foundation for future research to build upon, ensuring continued progress in the application of computer science methodologies to psychological and social research domains.

VIII. Annex

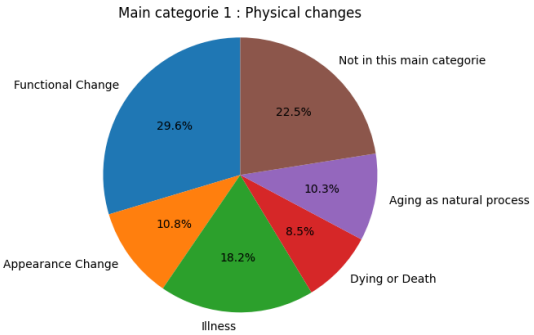

Fig. 6: Physical Changes

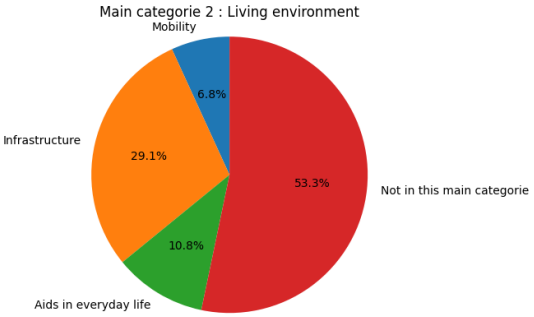

Fig. 7: Living Environment

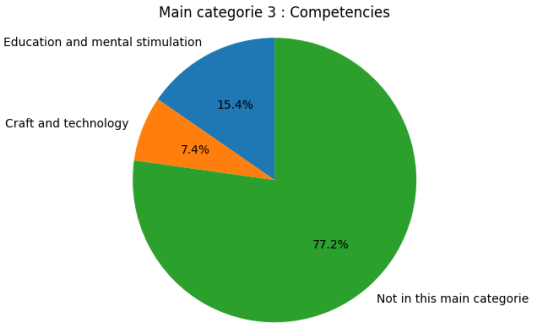

Fig. 8: Competencies

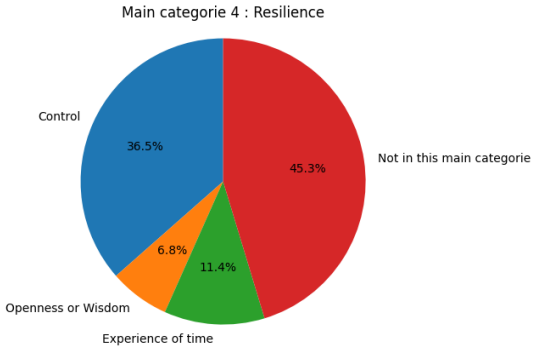

Fig. 9: Resilience

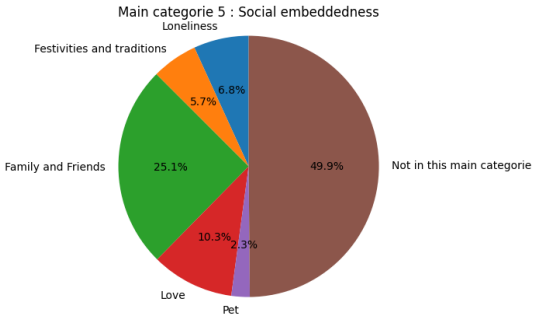

Fig. 10: Social Embeddedness
)

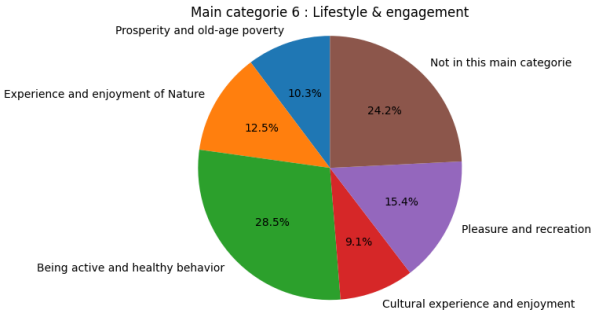

Fig. 11: Lifestyle & Engagement

REFERENCES

[1] V. Klusmann, "Ageing is in the eye of the beholder: Capturing images of ageing with photographs," 08 2018.

[2] L. G. Aziliz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, and C. Lemey, "Machine learning and natural language processing in mental health: a systematic review (preprint)," 07 2019.

[3] M. McCormack, A. Adams, and E. Anderson, "Taking to the streets: the benefits of spontaneous methodological innovation in participant recruitment," pp. 228–241, 2013. [Online]. Available: https://doi.org/10.1177/1468794112451038

[4] G. P. Society, "Ethical principles of the german psychological society (dgp) and the association of german professional psychologists (bdp)." [Online]. Available: https://www.reseaupsychologues.eu/attachment/48103/

[5] W. M. Association, "Wma declaration of helsinki – ethical principles for medical research involving human subjects – wma – the world medical association." [Online]. Available: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

[6] M. Franke, "Image captioning — neural pragmatic natural language generation," https://michael-franke.github.io/npNLG/08-grounded-LMs/08c-NIC-pretrained.html.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016. [Online]. Available: https://arxiv.org/abs/1506.01497

[9] Ultralytics, "Home - ultralytics yolo docs," https://docs.ultralytics.com/.

[10] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, vol. 2, no. 3, p. 249–262, Jul. 2018. [Online]. Available: http://dx.doi.org/10.1007/s41664-018-0068-2

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[12] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," 2020. [Online]. Available: https://arxiv.org/abs/2005.00547

[13] P. Su and K. Vijay-Shanker, "Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction," *BMC Bioinformatics*, vol. 23, no. 1, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1186/s12859-022-04642-w

[14] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1186/s40537-019-0197-0

[15] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1712.04621

[16] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, vol. 137, p. 109347, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323000481

[17] Jyotsana, "Image augmentation techniques. table of contents — by jyotsana — medium," https://medium.com/@jyotsana.cg/image-augmentation-techniques-798243f6afdf.

[18] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," 2019. [Online]. Available: https://arxiv.org/abs/1901.11196

[19] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of Big Data*, vol. 8, no. 1, Jul. 2021. [Online]. Available: http://dx.doi.org/10.1186/s40537-021-00492-0

[20] R. Snow, "Cheap and fast — but is it good? evaluating non-expert annotations for natural language tasks." [Online]. Available: https://aclanthology.org/D08-1027.pdf

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, Jun. 2002. [Online]. Available: http://dx.doi.org/10.1613/jair.953