Multimodal Idiomaticity Representation

In Task 2 of the coursework, a system needs to be developed to address the challenges associated with idiomatic language. The primary challenge is that some phrases can have multiple meanings, one that aligns with the literal interpretation of the words (e.g., *a real rotten apple*) and another that conveys a figurative meaning.

To tackle this, the algorithm's input will consist of a sentence containing a phrase that can be interpreted either literally or figuratively, along with three images and their corresponding captions. The algorithm should accurately understand the context and predict which image best represents the intended meaning of the phrase in the given sentence

Architecture Overview

The model leverages pre-trained models,BERT for text processing and EfficientNetB0 for image processing , and using transfer learning and fine-tuning for the image-sentence matching challenge.

- **Image Preprocessing**
    - The function loads an image from a file and preprocesses it to make it compatible with the EfficientNetB0 model, which is used later in the algorithm for image feature extraction
- **Clean DataFrame Function**
    - This function cleans the input DataFrame to ensure data consistency and quality, which is essential for training and evaluating the model
- **Text Encoding (BERT)**:
    - Sentences and captions are tokenized, producing input_ids and attention_mask because BERT requires tokenized inputs with attention masks to process text efficiently and focus only on meaningful tokens, ignoring padding.
    - A custom BertLayer built on TFBertModel generates sequence embeddings, which are reduced to 768-dimensional vectors via global average pooling.
    - **Output**: Embeddings for sentences and captions .
- **Image Encoding (EfficientNetB0)**:
    - Images preprocessed to 224x224x3 resolution.
    - Preprocessing images to match EfficientNetB0's input requirements enables effective feature extraction, aligning with the model's pre-training on ImageNet.
    - EfficientNetB0 extracts features, followed by global average pooling to produce a 1280-dimensional vector, which is then projected to 128 dimensions using a dense layer with ReLU activation.
    - A 128-dimensional image embedding.

- **Data Augmentation**:
    - **LLM-based Augmentation**:The goal of this section is to augment the training dataset by generating paraphrased versions of the 'sentence' and 'image_caption' columns.
    - This method uses an LLM (T5) to generate paraphrases and a sentence embedding model (Sentence-BERT) to filter them, ensuring they remain semantically close to the originals.
- **Zero-shot Prediction**:
    - A pre-trained BART model performs zero-shot classification to score caption-sentence relevance, serving as an ensemble component.
    - Produce a set of probability scores (zero_shot_probs) that can be combined with fine-tuned model predictions.
    - BART-large-MNLI excels at zero-shot tasks because it can generalize entailment relationships to new, unseen data using a hypothesis template.
- **Multimodal Fusion**:
    - **Combination**: The image and caption embeddings are concatenated and projected to 128 dimensions. The sentence embedding is similarly projected to 128 dimensions.
    - **Final Combination**: These two 128-dimensional vectors are concatenated into a 256-dimensional vector.
    - **Prediction**: A dense layer with sigmoid activation outputs a probability (0 to 1) indicating the likelihood of an image-sentence match.
- **Prediction and Evaluation**:
    - The fine-tuned model predicts probabilities for test data.
    - **Ensemble**: Combines model predictions (weighted 0.7) with zero-shot predictions (weighted 0.3).
    - **Evaluation**: Measures performance using accuracy , F1 score,Precision and Recall ,supplemented by visualizations of predicted versus true image-sentence pairs.

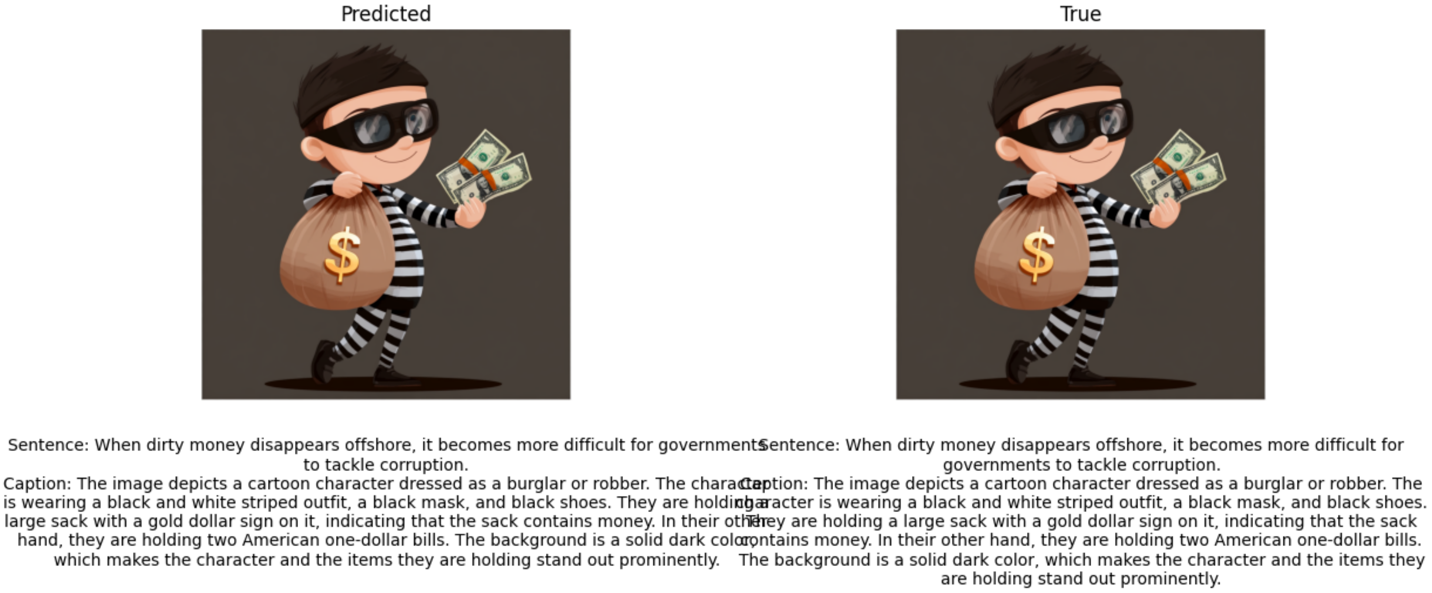
Evaluation:

The table below presents the model’s performance based on the F1 score.

Model’s Performace	
Test Accuracy	88%
Test F1 Score	87%
Test Precision	87%

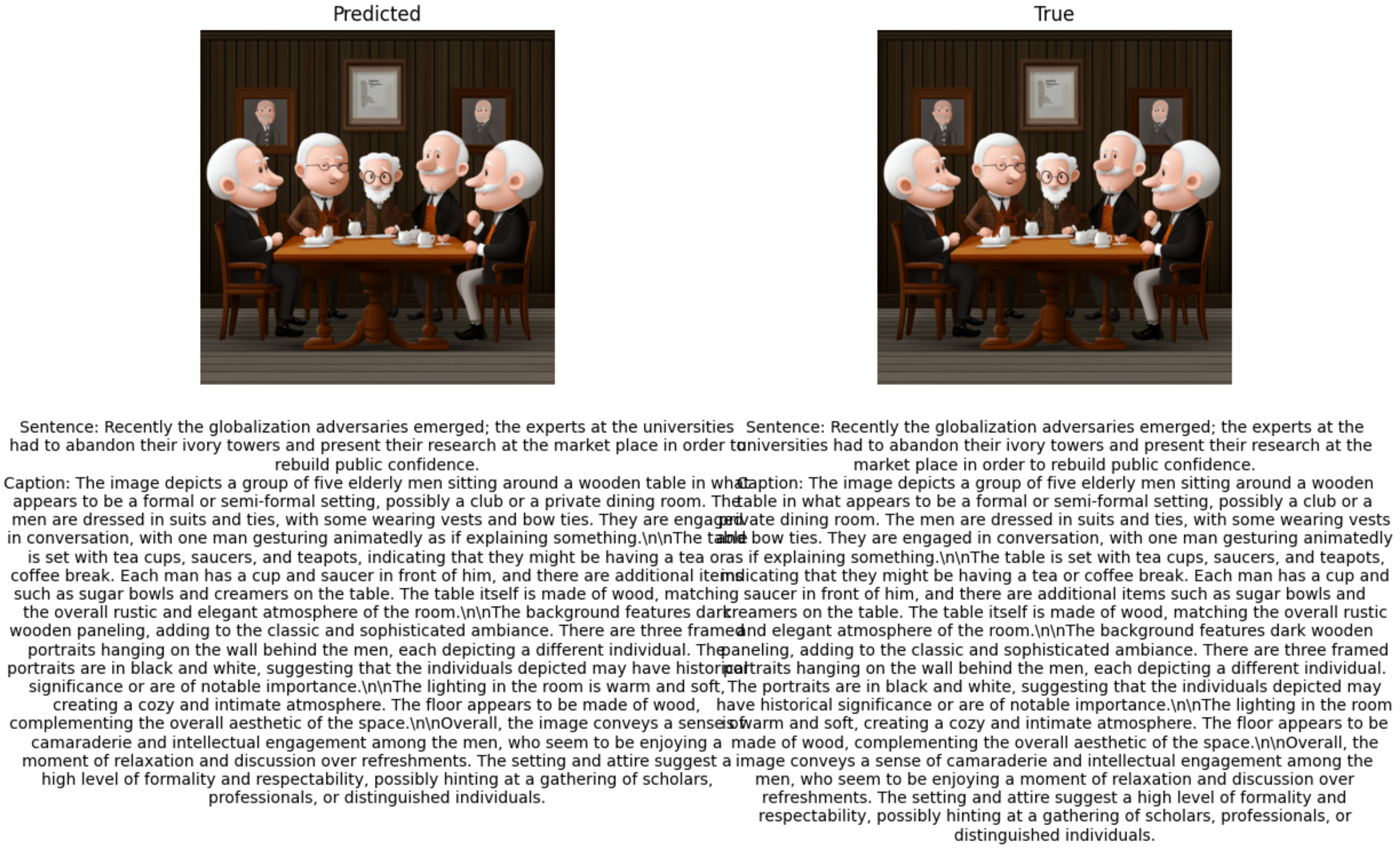
Model's Performace	
Test Recall Score	88%

The image below displays the model's output, showing the actual image corresponding to the given sentence alongside the image predicted by the model.



Sentence: When dirty money disappears offshore, it becomes more difficult for governments to tackle corruption.
Caption: The image depicts a cartoon character dressed as a burglar or robber. The character is wearing a black and white striped outfit, a black mask, and black shoes. They are holding a large sack with a gold dollar sign on it, indicating that the sack contains money. In their other hand, they are holding two American one-dollar bills. The background is a solid dark color, which makes the character and the items they are holding stand out prominently.

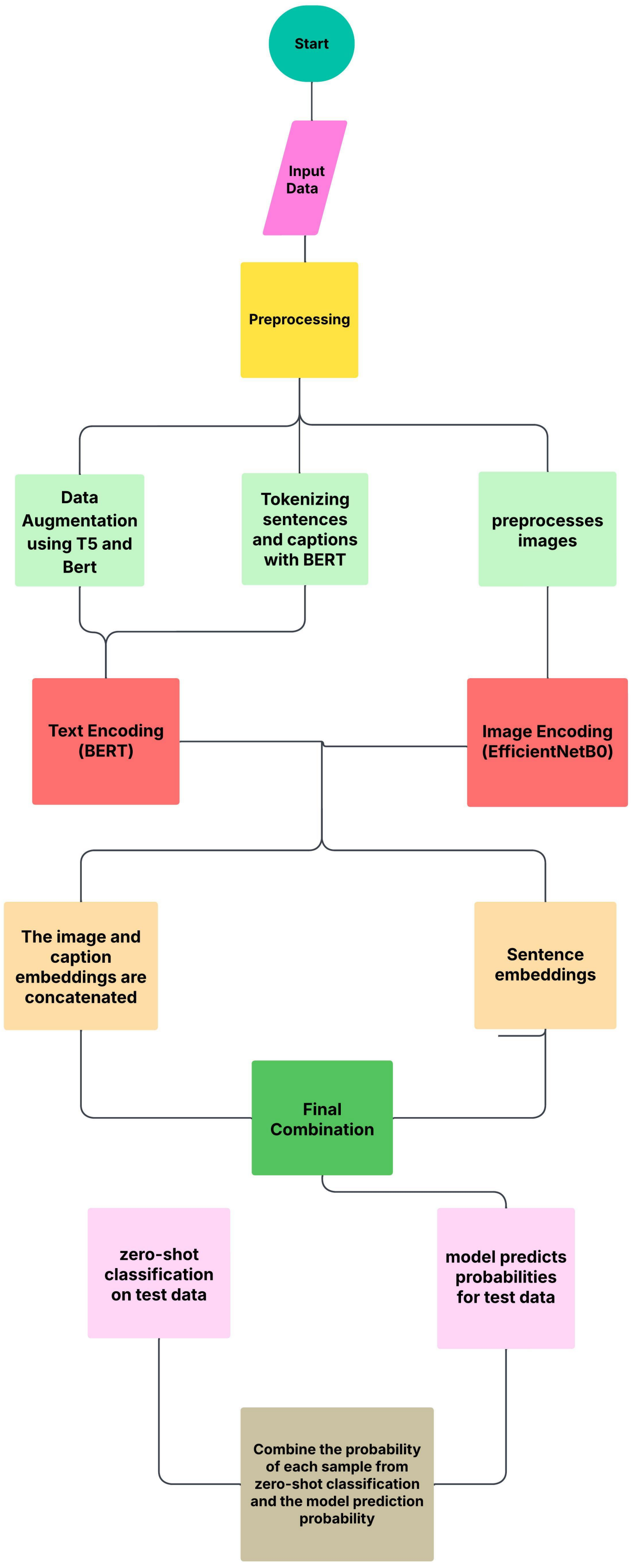
Sentence: When dirty money disappears offshore, it becomes more difficult for governments to tackle corruption.
Caption: The image depicts a cartoon character dressed as a burglar or robber. The character is wearing a black and white striped outfit, a black mask, and black shoes. They are holding a large sack with a gold dollar sign on it, indicating that the sack contains money. In their other hand, they are holding two American one-dollar bills. The background is a solid dark color, which makes the character and the items they are holding stand out prominently.



Sentence: Recently the globalization adversaries emerged; the experts at the universities had to abandon their ivory towers and present their research at the market place in order to rebuild public confidence.
Caption: The image depicts a group of five elderly men sitting around a wooden table in what appears to be a formal or semi-formal setting, possibly a club or a private dining room. The men are dressed in suits and ties, with some wearing vests and bow ties. They are engaged in conversation, with one man gesturing animatedly as if explaining something.\n\nThe table is set with tea cups, saucers, and teapots, indicating that they might be having a tea or coffee break. Each man has a cup and saucer in front of him, and there are additional items such as sugar bowls and creamers on the table. The table itself is made of wood, matching the overall rustic and elegant atmosphere of the room.\n\nThe background features dark wooden paneling, adding to the classic and sophisticated ambiance. There are three framed portraits hanging on the wall behind the men, each depicting a different individual. The portraits are in black and white, suggesting that the individuals depicted may have historical significance or are of notable importance.\n\nThe lighting in the room is warm and soft, creating a cozy and intimate atmosphere. The floor appears to be made of wood, complementing the overall aesthetic of the space.\n\nOverall, the image conveys a sense of camaraderie and intellectual engagement among the men, who seem to be enjoying a moment of relaxation and discussion over refreshments. The setting and attire suggest a high level of formality and respectability, possibly hinting at a gathering of scholars, professionals, or distinguished individuals.

Sentence: Recently the globalization adversaries emerged; the experts at the universities had to abandon their ivory towers and present their research at the market place in order to rebuild public confidence.
Caption: The image depicts a group of five elderly men sitting around a wooden table in what appears to be a formal or semi-formal setting, possibly a club or a private dining room. The men are dressed in suits and ties, with some wearing vests and bow ties. They are engaged in conversation, with one man gesturing animatedly as if explaining something.\n\nThe table is set with tea cups, saucers, and teapots, indicating that they might be having a tea or coffee break. Each man has a cup and saucer in front of him, and there are additional items such as sugar bowls and creamers on the table. The table itself is made of wood, matching the overall rustic and elegant atmosphere of the room.\n\nThe background features dark wooden paneling, adding to the classic and sophisticated ambiance. There are three framed portraits hanging on the wall behind the men, each depicting a different individual. The portraits are in black and white, suggesting that the individuals depicted may have historical significance or are of notable importance.\n\nThe lighting in the room is warm and soft, creating a cozy and intimate atmosphere. The floor appears to be made of wood, complementing the overall aesthetic of the space.\n\nOverall, the image conveys a sense of camaraderie and intellectual engagement among the men, who seem to be enjoying a moment of relaxation and discussion over refreshments. The setting and attire suggest a high level of formality and respectability, possibly hinting at a gathering of scholars, professionals, or distinguished individuals.

The Flowchart below demonstrates the workflow of the designed architecture:


Start
Input Data
Preprocessing
Data Augmentation using T5 and Bert
Tokenizing sentences and captions with BERT
preprocesses images
Text Encoding (BERT)
Image Encoding (EfficientNetB0)
The image and caption embeddings are concatenated
Sentence embeddings
Final Combination
zero-shot classification on test data
model predicts probabilities for test data
Combine the probability of each sample from zero-shot classification and the model prediction probability