University of Information and Technology
AI002 - Artificial Intelligence Thinking

Group 5

Multimodal Sacarsm Detection on Vietnamese Social Media Posts

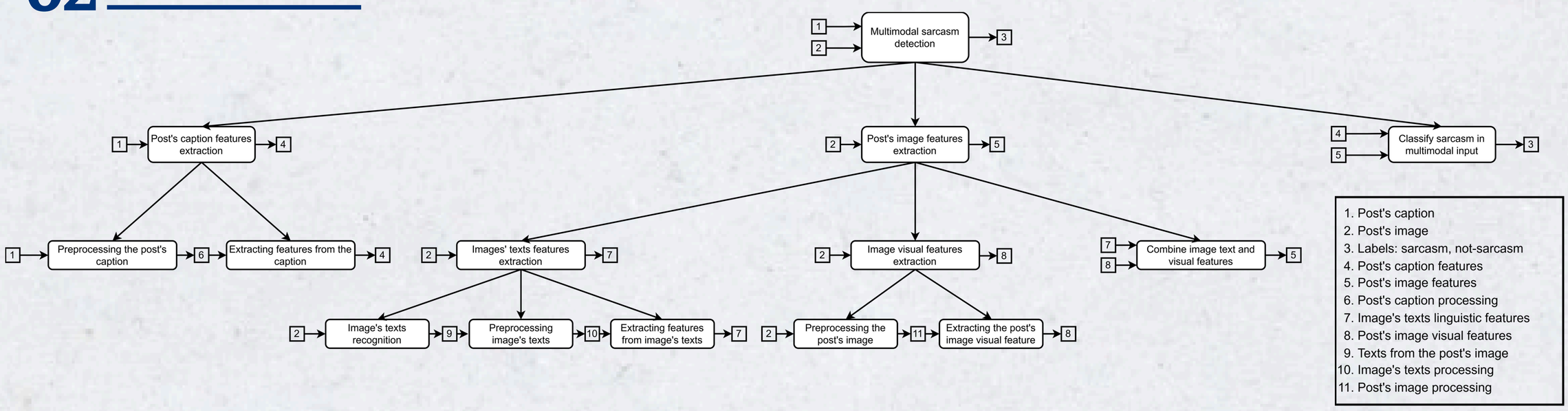
Instructor: *PhD. Ngo Duc Thanh*

Ha Huy Hoang	Dang Vinh Hoi	Nguyen Duy Hoang	Tang Gia Han	Nguyen Xuan Bach
22520460	22520490	22520467	22520394	22520093

01 PROBLEM STATEMENT

- **Input**: Consists of the following two components
  - **Caption**: Text associated with the post.
  - **Image**: A single picture (e.g., photo or graphic) associated with the post.
- **Output**: One of the following two labels
  - **sarcasm**: If the post contains sarcasm based on the analysis of the caption and image.
  - **not-sarcasm**: If the post does not contain sarcasm, indicating a straightforward or literal interpretation of the caption and image.
- **Constraints**:
  - The image must have a minimum height and width of 224x224 pixels (i.e., height and width should be greater than or equal to 224 pixels).
  - If the image contains text, the text within the image must not exceed 256 units (word, character, or emoji).
  - The text (caption) is required to contain at least 1 unit (word, character, or emoji) and no more than 256 units (word, character, or emoji).
  - The text (caption) is recommended to be in standard Vietnamese but may include slang, teencode, or foreign language elements.
- **Requirements**:
  - The model must achieve an accuracy of at least 75% on the test set provided by the customer, as measured under the specified evaluation criteria.
  - The total training and inference time must not exceed 2 hours, with an inference time must not exceed 0.05 seconds per sample, running on Kaggle's Tesla P100-PCIE-16GB GPU (15.89 GB memory, 56 multiprocessors).
  - The inference time, measured from submitting the complete input (image and caption) to receiving the result, must not exceed 5 seconds when deployed via a Streamlit web application on the specified configuration (Model name: AMD EPYC 7B13, CPU frequency: 2449.998 MHz, maximum 2 CPU cores, 2.7GB memory, and up to 50GB storage).

02 DECOMPOSITION



03 EVALUATION

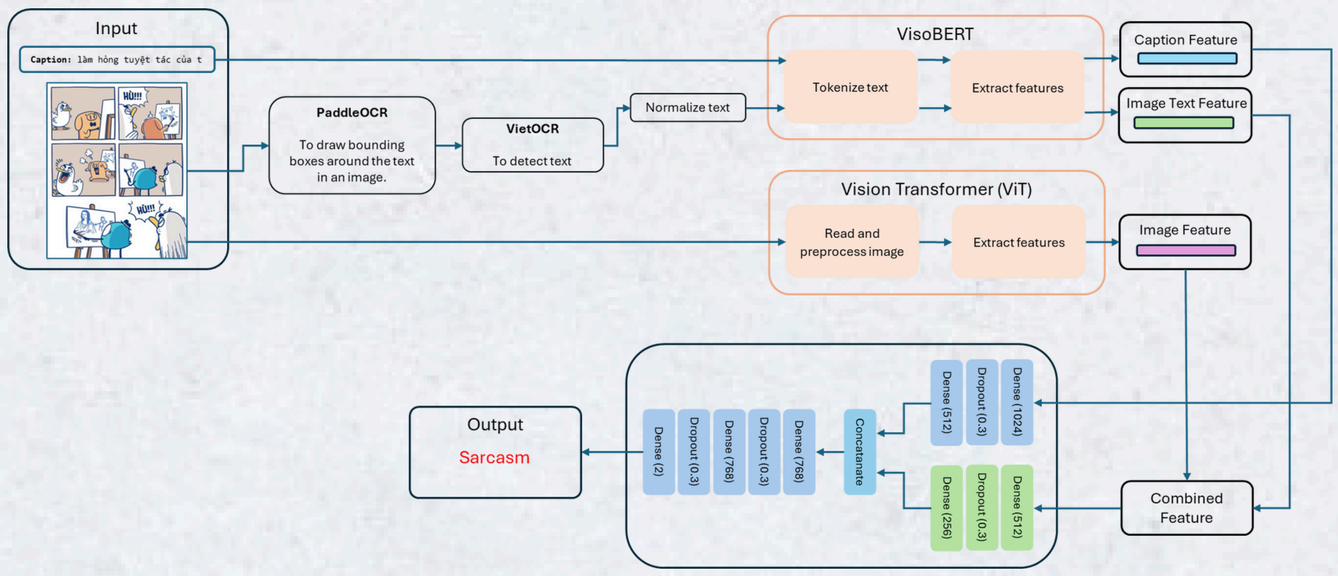
- **Model Performance**:
  - Recall
  - Precision
  - Accuracy
  - F1 Score
- **Efficiency**:
  - Average Inference Time $T = \frac{1}{m}\sum_{i=1}^{m}(t_i)$
- **Performance statistics for processing and prediction times on the test set of the dataset, evaluated using a Tesla P100 GPU in the Kaggle environment**:

Number of test	Processing Time	Prediction Time
2382 (pair image and caption)	109.56s	0.88s

- **Classification report**:

	precision	recall	f1-score	support
not-sarcasm	0.70	0.78	0.74	1071
sarcasm	0.80	0.72	0.76	1311
accuracy			0.75	2382
macro avg	0.75	0.75	0.75	2382
weighted avg	0.76	0.75	0.75	2382

04 ALGORITHM



05 DATA

The dataset for sarcasm detection on Vietnamese social media posts, sourced from the UIT Data Science Challenge 2024 (DSC), has the following label distribution:

- **not-sarcasm**: 6011 samples
- **sarcasm**: 5896 samples

The data is split into an 80% training set (9525 samples) and a 20% test set (2382 samples), with each sample consisting of an image and a caption.

06 CONCLUSION

In this study, we proposed a multimodal approach for detecting sarcasm in Vietnamese social media posts by integrating both image and text data. This approach aimed to leverage visual and textual cues to accurately identify sarcastic intent. Real-world applications of this system, such as automated content moderation or sentiment analysis on Vietnamese social media platforms, would require overcoming challenges like low image quality, inconsistent resolution, and language variability, including the use of slang and non-standard language.