

A dark blue background with a network of light blue circles connected by thin lines, creating a geometric pattern.

# IAI-UET

**VIỆN TRÍ TUỆ NHÂN TẠO**

ĐẠI HỌC CÔNG NGHỆ - ĐẠI HỌC QUỐC GIA HÀ NỘI

YOUR ONLY LIMIT IS YOU



**VIỆN TRÍ TUỆ NHÂN TẠO**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ  
ĐẠI HỌC QUỐC GIA HÀ NỘI

Chủ đề:

# Sử dụng mô hình BERT trong các ứng dụng xử lý NNTN

Nhóm 10

22022602 Bùi Đức Mạnh

22022666 Lê Việt Hùng

22022522 Đàm Thái Ninh

YOUR ONLY LIMIT IS YOU



# NỘI DUNG TRÌNH BÀY

1 Giới thiệu

2 Mục tiêu của báo cáo

3 Sentiment Analysis

4 NER

5 Question Answering

6 Multitask: Dịch máy + SA

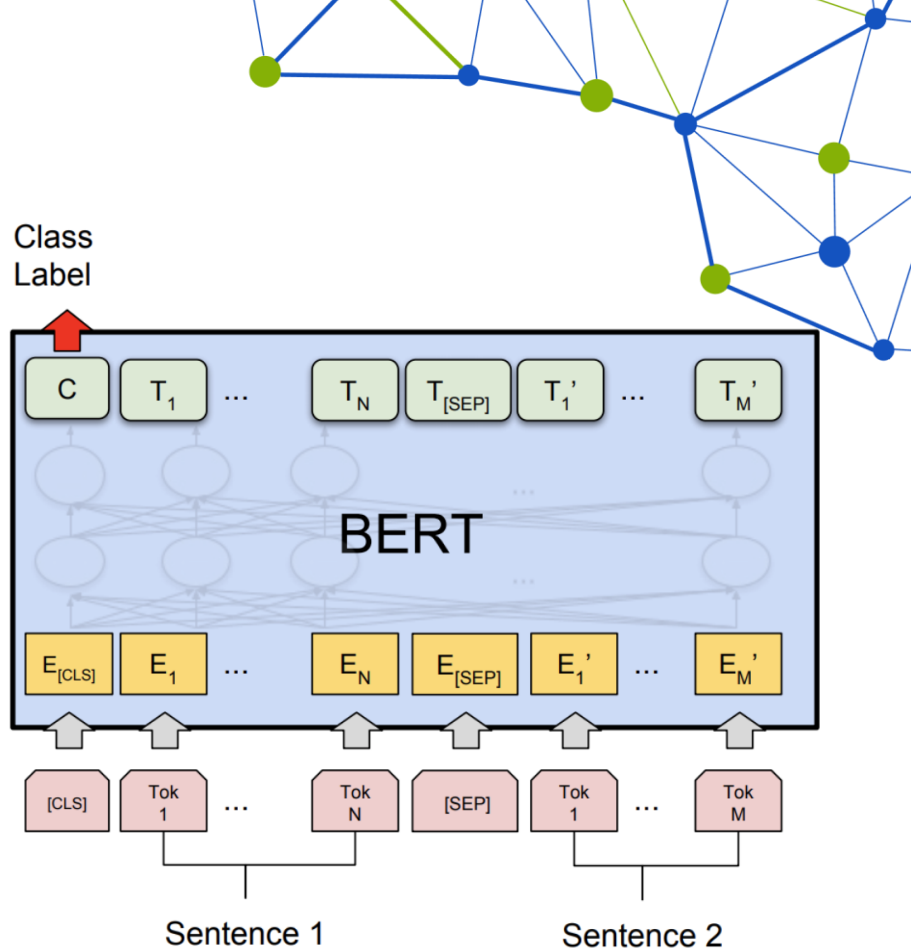
7 Phân công công việc

8 Kết luận

# 1 Giới thiệu

## 1.1 BERT

- BERT (Bidirectional Encoder Representations from Transformers) một kiến trúc mới cho lớp bài toán Language Representation.
- BERT được thiết kế để đào tạo ra các vector đại diện cho ngôn ngữ văn bản thông qua ngữ cảnh 2 chiều (trái và phải).



## 1.1 BERT

BERT là một khái niệm đơn giản nhưng lại mang lại hiệu quả cực lớn trong thực tế. Nó đã thu được kết quả tối ưu mới nhất cho 11 nhiệm vụ xử lý ngôn ngữ tự nhiên, bao gồm việc đẩy kết quả của nhiệm vụ GLUE benchmark lên 80.4% (cải tiến thêm 7.6%) (Hình 1.1) và SQuAD v.1.1 với F1 score trên tập test đạt 93.2% (cải tiến thêm 1.5%), tốt hơn con người 2% (Hình 1.2).

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language <a href="https://arxiv.org/abs/1810.04805">https://arxiv.org/abs/1810.04805</a>	87.433	93.160
2 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) Google Brain & CMU	84.454	90.490

Hình 1.1: Kết quả của BERT trên GLUE benchmark

Rank	Model	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	QNLI	RTE
1	BERT: 24-layers, 1024-hidden, 16-heads	80.4	60.5	94.9	85.4/89.3	87.6/86.5	89.3/72.1	86.7	91.1	70.1
2	Singletask Pretrain Transformer	72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	88.1	56.0
3	BiLSTM+ELMo+Attn	70.5	36.0	90.4	77.9/84.9	75.1/73.3	84.7/64.8	76.4	79.9	56.8

Hình 1.2: Kết quả của BERT trên SQuAD v.1.1

# 1 Giới thiệu

## 1.2 BART

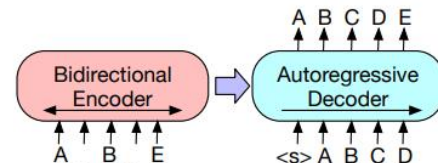
BART (Bidirectional and Auto-Regressive Transformers) là một mô hình kết hợp kiến trúc của bộ mã hóa hai chiều (như BERT) và bộ giải mã tự hồi quy (như GPT). Cụ thể:

- **Bộ mã hóa (Encoder):** Sử dụng kiến trúc Transformer hai chiều, cho phép mô hình nắm bắt ngữ cảnh toàn diện từ cả hai phía của mỗi token.
- **Bộ giải mã (Decoder):** Áp dụng cơ chế tự hồi quy, giúp mô hình có khả năng sinh văn bản tuần tự.



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitrary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

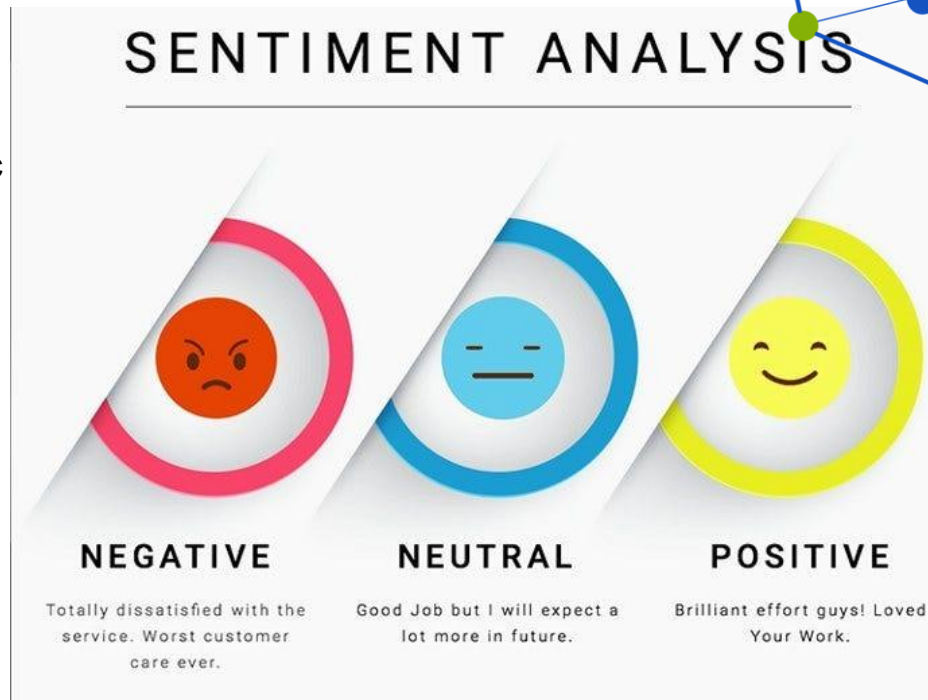
## 2 Mục tiêu của báo cáo

- Tìm hiểu các ứng dụng của mô hình BERT trong xử lý ngôn ngữ tự nhiên (NLP).
- Ứng dụng mô hình BERT cho ba tác vụ cụ thể: phân tích cảm xúc (Sentiment Analysis), trả lời câu hỏi (Q&A), và nhận dạng thực thể có tên (Named Entity Recognition - NER). • Nghiên cứu và triển khai mô hình BART cho đa nhiệm vụ (multi-task learning) gồm dịch máy và Sentiment Analysis, nhằm đánh giá khả năng của BART trong việc xử lý đồng thời nhiều tác vụ.
- Trình bày và so sánh kết quả giữa các mô hình ngôn ngữ xây dựng theo BERT trên các tác vụ đã chọn, nhằm đánh giá hiệu năng và ưu nhược điểm của từng mô hình khi áp dụng cho ngữ cảnh tiếng Việt.
- Với BART, so sánh hiệu suất giữa mô hình BARTPho Word Base (150M params) và BARTPho Word (420M params) trên tác vụ dịch máy Anh-Việt, Việt-Anh và Sentiment Analysis, nhằm đánh giá khả năng cải thiện hiệu suất của mô hình khi tăng kích thước mô hình

## 3 Sentiment Analysis

### 3.1. Giới thiệu

- **Bài toán:** phân loại các đánh giá/tài liệu tiếng Việt thành một trong ba loại: tích cực, tiêu cực và trung tính.
- **Input:** một đoạn văn bản tiếng Việt.
- **Output:** một trong ba nhãn: tích cực, tiêu cực hoặc trung tính.





## 3 Sentiment Analysis

### 3.2. Dữ liệu

#### VLSP 2016 Shared Task

- Data characteristics:** Bộ dữ liệu này bao gồm các đánh giá của người dùng về các thiết bị công nghệ theo ba loại: "tiêu cực", "tích cực" và "trung tính".

Table 2.1: Thống kê dữ liệu

Dataset	Train			Test			Totally
	Positive	Negative	Neutral	Positive	Negative	Neutral	
VLSP 2016	23,952	23,756	22,292	5988	5939	5573	87,500
AIVIVN 2019	7424	5446	0	1856	1361	0	16,087

#### AIVIVN 2019

- Data characteristics:** Bộ dữ liệu này bao gồm nhận xét của người dùng về đánh giá sản phẩm trên các trang thương mại điện tử. Dữ liệu chứa các đánh giá của người dùng theo hai loại bao gồm "tích cực" và "tiêu cực"

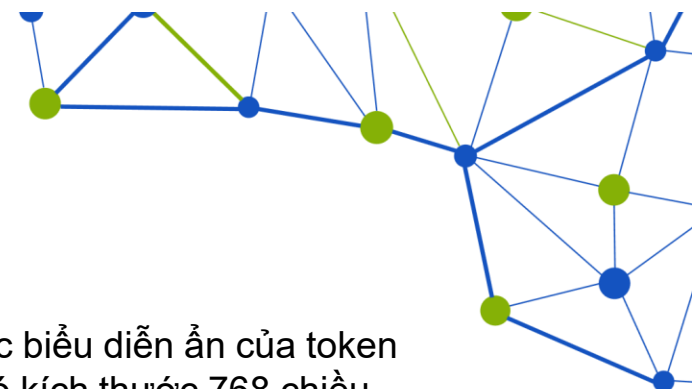
Table 2.2: Thống kê số lượng từ sử dụng trong một câu

	VLSP 2016	AIVIVN 2019
Mean	143.45	92.66
Std	204.17	99.19
Min	1	1
25%	35	39
50%	72	68
75%	165	114
Max	4113	2788

## 3 Sentiment Analysis

### 3.3. Hướng tiếp cận

- **Multilingual BERT-base, PhoBERT, PhoBERTv2:** Ghép nối các biểu diễn ẩn của token [CLS] từ 5 lớp cuối cùng. Mỗi vector biểu diễn của token [CLS] có kích thước 768 chiều. Vector kết quả sau khi ghép nối sẽ có kích thước 3840 chiều ( $768 * 5$ ). Vector này sau đó được đưa vào một mô-đun MLP để tạo ra kết quả phân loại cuối cùng.
- **BARTPho-word-base, BARTPho-word:** áp dụng phương pháp "text-to-text". Thêm một câu prompt vào đầu vào để chỉ định tác vụ cụ thể mà mô hình cần thực hiện. Mô hình sau đó sẽ xử lý toàn bộ đầu vào, bao gồm cả câu prompt, để tạo ra kết quả đầu ra tương ứng với tác vụ đó.



## 3 Sentiment Analysis

### 3.4. Kết quả

- Trên cả hai bộ dữ liệu, các mô hình đều thể hiện hiệu suất cao, với độ chính xác trên 98% đối với VLSP 2016 và trên 88% đối với AIVIVN 2019.
- VLSP 2016 có vẻ là một bộ dữ liệu "dễ" hơn, trong khi AIVIVN 2019 thách thức hơn và phản ánh tốt hơn khả năng tổng quát hóa của các mô hình.

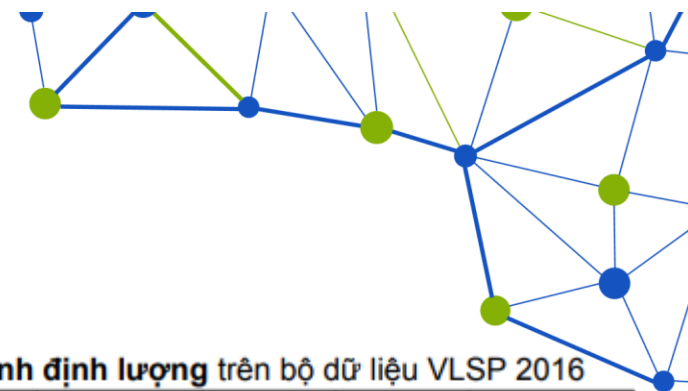


Table 2.3: So sánh định lượng trên bộ dữ liệu VLSP 2016

Model	Precision	Recall	F1	Accuracy
Multilingual BERT	0.9994	0.9994	0.9994	0.9994
phoBERT	<b>0.9996</b>	<b>0.9996</b>	<b>0.9996</b>	<b>0.9996</b>
phoBERTv2	<u>0.9995</u>	<u>0.9995</u>	<u>0.9995</u>	<u>0.9995</u>
BARTPho-word-base	0.9899	0.9899	0.9899	0.9899
BARTPho-word	0.988	0.988	0.988	0.988

Table 2.4: So sánh định lượng trên bộ dữ liệu AIVIVN 2019

Model	Precision	Recall	F1	Accuracy
Multilingual BERT	0.6007	0.5881	0.5942	0.8831
phoBERT	0.6177	0.5948	0.606	0.8946
phoBERTv2	0.6188	0.6024	0.6102	0.9027
BARTPho-word-base	<u>0.8960</u>	<u>0.9044</u>	<u>0.8979</u>	<u>0.8990</u>
BARTPho-word	<b>0.9018</b>	<b>0.9081</b>	<b>0.9041</b>	<b>0.9055</b>

Tô đậm biểu thị cho kết quả tốt nhất, gạch chân biểu thị cho kết quả tốt thứ hai.

# 4 NER

## 4.1 Giới thiệu

- **Bài toán:** xác định và phân loại các thực thể như tên người, địa điểm, tổ chức trong văn bản.
- **Input:** một đoạn văn bản tiếng Việt.
- **Output:** Các thực thể trong văn bản được đánh dấu và phân loại theo các nhãn như "người", "địa điểm", "tổ chức", v.v.



## 4 NER

### 4.2 Dữ liệu

#### VLSP 2016 NER Shared Task

- Data characteristics:** Bộ dữ liệu này bao gồm các nhãn gồm POS, Chunk và NER đã được gán sẵn. Tuy nhiên trong task này, ta chỉ quan tâm đến nhãn NER.

Table 3.2: Phân phối các nhãn NER trong tập dữ liệu

Nhãn	Số lượng
O	62,395
B-PER	1,371
B-LOC	1,314
I-PER	976
I-LOC	552
I-ORG	545
B-ORG	343
I-MISC	51

Table 3.1: Dữ liệu huấn luyện mô hình NER

Word	POS	Chunk	NER
Dương	Np	B-NP	B-PER
là	V	B-VP	O
một	M	B-NP	O
chủ	N	B-NP	O
cửa	N	B-NP	O
hàng	N	I-NP	O
lâu	A	B-AP	O
năm	N	B-NP	O
ở	E	B-PP	O
Hà	Np	B-NP	B-LOC
Nội	Np	I-NP	I-LOC
.	CH	O	O

## 4 NER

### 4.2 Dữ liệu

#### VLSP 2016 NER Shared Task

- Data characteristics:** Bộ dữ liệu này bao gồm các nhãn gồm POS, Chunk và NER đã được gán sẵn. Tuy nhiên trong task này, ta chỉ quan tâm đến nhãn NER.

Table 3.1: Dữ liệu huấn luyện mô hình NER

Word	POS	Chunk	NER
Dương	Np	B-NP	B-PER
là	V	B-VP	O
một	M	B-NP	O

Table 3.2: Phân phối các nhãn NER trong t

Nhãn	Số lượng
O	62,395
B-PER	1,371
B-LOC	1,314
I-PER	976
I-LOC	552
I-ORG	545
B-ORG	343
I-MISC	51

Phân phối dữ liệu cho thấy tập dữ liệu NER của bạn có sự mất cân bằng mạnh giữa các nhãn, với nhãn O chiếm phần lớn. Các thực thể về người và địa điểm là phổ biến nhất, trong khi tổ chức và các thực thể khác ít phổ biến hơn. Điều này phản ánh đúng thực tế rằng phần lớn văn bản không liên quan đến các thực thể cụ thể, nhưng cũng đặt ra thách thức trong việc huấn luyện mô hình cần phải phân biệt được các thực thể từ phần lớn dữ liệu không liên quan.

## 4 NER

### 4.3 Hướng tiếp cận

Huấn luyện các mô hình Named Entity Recognition (NER) dựa trên ba mô hình ngôn ngữ khác nhau: Multilingual BERT-base , PhoBERT, và PhoBERTv2. Các mô hình được huấn luyện với cùng một thiết lập tham số để đảm bảo tính nhất quán và công bằng trong so sánh.

Các tham số được thiết lập như sau:

- max\_len: 128
- batch\_size: 16
- epochs: 10
- learning\_rate: 2e-5

## 4 NER

### 4.4 Kết quả

Table 3.3: Kết quả kiểm tra của PhoBERT

Nhãn	Precision	Recall	F1-Score	Số lượng
B-LOC	0.94	0.91	0.92	179
B-MISC	0.73	1.00	0.85	11
B-ORG	0.61	0.92	0.73	12
B-PER	0.97	0.97	0.97	188
I-LOC	0.85	0.85	0.85	61
I-MISC	0.55	1.00	0.71	11
I-ORG	0.89	0.85	0.87	40
I-PER	0.99	0.99	0.99	154
O	1.00	1.00	1.00	7445

Table 3.5: Kết quả kiểm tra của Multilingual BERT

Nhãn	Precision	Recall	F1-Score	Số lượng
B-LOC	0.93	0.96	0.94	364
B-MISC	1.00	0.91	0.95	11
B-ORG	0.76	0.76	0.76	29
B-PER	0.99	0.95	0.97	224
I-LOC	0.91	0.90	0.90	109
I-MISC	0.85	0.85	0.85	13
I-ORG	0.90	0.83	0.86	63
I-PER	0.98	0.97	0.98	200
O	1.00	1.00	1.00	10988

Table 3.4: Kết quả kiểm tra của PhoBERTv2

Nhãn	Precision	Recall	F1-Score	Số lượng
B-LOC	0.89	0.87	0.88	179
B-MISC	0.73	1.00	0.85	11
B-ORG	0.59	0.83	0.69	12
B-PER	0.97	0.98	0.98	188
I-LOC	0.75	0.87	0.80	61
I-MISC	0.45	0.82	0.58	11
I-ORG	0.75	0.82	0.79	40
I-PER	0.99	0.97	0.98	154
O	1.00	1.00	1.00	7445

Table 3.6: So sánh kết quả giữa ba mô hình NER

Nhãn	PhoBERT			PhoBERTv2			Multilingual BERT		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
B-LOC	0.94	0.91	0.92	0.89	0.87	0.88	0.93	0.96	0.94
B-MISC	0.73	1.00	0.85	0.73	1.00	0.85	1.00	0.91	0.95
B-ORG	0.61	0.92	0.73	0.59	0.83	0.69	0.76	0.76	0.76
B-PER	0.97	0.97	0.97	0.97	0.98	0.98	0.99	0.95	0.97
I-LOC	0.85	0.85	0.85	0.75	0.87	0.80	0.91	0.90	0.90
I-MISC	0.55	1.00	0.71	0.45	0.82	0.58	0.85	0.85	0.85
I-ORG	0.89	0.85	0.87	0.75	0.82	0.79	0.90	0.83	0.86
I-PER	0.99	0.99	0.99	0.99	0.97	0.98	0.98	0.97	0.98
O	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00



# 5 Q&A

## 5.1 Giới thiệu

- **Bài toán:** phát triển một hệ thống có khả năng tự động trả lời các câu hỏi do người dùng đưa ra.
- **Input:** bao gồm hai thành phần chính:
  - Câu hỏi (Question): Đây là câu hỏi mà người dùng đặt ra, có thể là một câu hỏi đơn giản hoặc phức tạp, bao gồm nhiều dạng câu hỏi như câu hỏi yes/no, câu hỏi dạng liệt kê, hay câu hỏi đòi hỏi phân tích sâu.
  - Văn bản chứa câu trả lời (Context/Passage): Đây là đoạn văn bản chứa thông tin cần thiết để trả lời câu hỏi. Văn bản này có thể là một tài liệu, một đoạn văn, hoặc thậm chí là toàn bộ cơ sở dữ liệu thông tin.
- **Output:** phân loại thành 2 thành phần chính
  - Extractive QA (Câu trả lời trích xuất): câu trả lời được trích xuất trực tiếp từ đoạn văn bản (context) đã cho.
  - Abstractive QA (Câu trả lời tổng hợp): hệ thống không chỉ đơn thuần là trích xuất thông tin từ đoạn văn bản mà còn có khả năng tổng hợp, diễn đạt lại thông tin theo cách mới.

## 5.2 Dữ liệu

## BERT Vietnamese Question Answering Dataset (VQA)

- Data characteristics:** bộ dữ liệu được thiết kế cho bài toán trả lời câu hỏi (Question Answering) trong ngôn ngữ tiếng Việt

Train Dataset					
Số câu hỏi	2976				
Số câu hỏi có câu trả lời	985				
	answer_start	answer_end	Độ dài câu hỏi	Độ dài ngữ cảnh	Độ dài câu trả lời
Giá trị trung bình	25.14	29.31	31.70	249.61	4.16
Độ lệch chuẩn	64.60	68.18	11.14	168.01	6.65
Giá trị nhỏ nhất	0.00	0.00	12.00	20.00	0.00
Phân vị 25%	0.00	0.00	24.00	113.75	0.00
Phân vị 50%	0.00	0.00	29.00	219.00	0.00
Phân vị 75%	13.00	26.00	38.00	338.00	9.00
Giá trị lớn nhất	757.00	764.00	72.00	1224.00	51.00

Table 4.1: Thống kê về Train Dataset sử dụng cho bài toán Question Answering

- `answer_start`: Vị trí bắt đầu của câu trả lời (answer) trong ngữ cảnh (context)
- `answer_end`: Vị trí kết thúc của câu trả lời (answer) trong ngữ cảnh (context)

Dev Dataset					
Số câu hỏi	478				
Số câu hỏi có câu trả lời	141				
	answer_start	answer_end	Độ dài câu hỏi	Độ dài ngữ cảnh	Độ dài câu trả lời
Giá trị trung bình	25.77	29.52	35.36	269.86	3.75
Độ lệch chuẩn	67.99	71.69	11.40	164.89	6.30
Giá trị nhỏ nhất	0.00	0.00	14.00	41.00	0.00
Phân vị 25%	0.00	0.00	27.00	136.50	0.00
Phân vị 50%	0.00	0.00	33.00	251.00	0.00
Phân vị 75%	0.00	18.00	43.00	369.75	8.00
Giá trị lớn nhất	572.00	588.00	74.00	1231.00	32.00

Table 4.2: Thống kê về Dev Dataset sử dụng cho bài toán Question Answering

- `answer_start`: Vị trí bắt đầu của câu trả lời (answer) trong ngữ cảnh (context)
- `answer_end`: Vị trí kết thúc của câu trả lời (answer) trong ngữ cảnh (context)

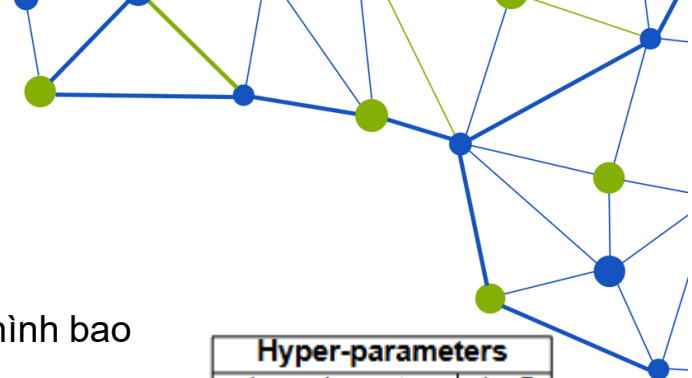
## 5 Q&A

### 5.3 Hướng tiếp cận

Trong nghiên cứu này, chúng tôi tiến hành thử nghiệm với 5 mô hình bao gồm:

- BERT Base Multilingual Cased
- XLM RoBERTa Large
- Multilingual Bert Base Cased Vietnamese
- XLM RoBERTa Large SQuAD2
- Vi MRC Large

Chúng tôi không sử dụng PhoBERT và PhoBERTv2 do hai mô hình này chưa hỗ trợ offset mapping, một tính năng quan trọng cho các bài toán Extractive QA. Thay vào đó, PhoBERT và PhoBERTv2 thích hợp hơn với các bài toán Abstractive QA, hoặc đòi hỏi kiến thức chuyên sâu để áp dụng chúng một cách hiệu quả trong Extractive QA.



Hyper-parameters	
Learning rate	1e-5
Train Batch size	4
Dev Batch size	4
Epochs	10
Weight decay	0.01
Warmup steps	100

Các siêu tham số được sử dụng trong các mô hình

## 5 Q&A

### 5.4 Kết quả

- Tất cả các mô hình đều cho kết quả khá tốt, với F1 score từ 0.78 trở lên cho cả vị trí bắt đầu và kết thúc.
- Vi MRC Large đạt hiệu suất cao nhất với F1 score 0.97 cho cả hai chỉ số.
- Hầu hết các mô hình đều có sự chênh lệch nhỏ nhưng không đáng kể giữa start\_f1 và end\_f1.

Model	start_f1	end_f1
BERT Base Multilingual Cased (baseline)	0.79	0.78
XLNet Large	0.87	0.84
Multilingual Bert Base Cased Vietnamese	0.81	0.80
XLNet Large SQuAD2	<u>0.88</u>	<u>0.86</u>
Vi MRC Large	<b>0.97</b>	<b>0.97</b>

Table 4.4: Kết quả các mô hình trên bộ dữ liệu tiếng Việt cho bài toán Question Answering

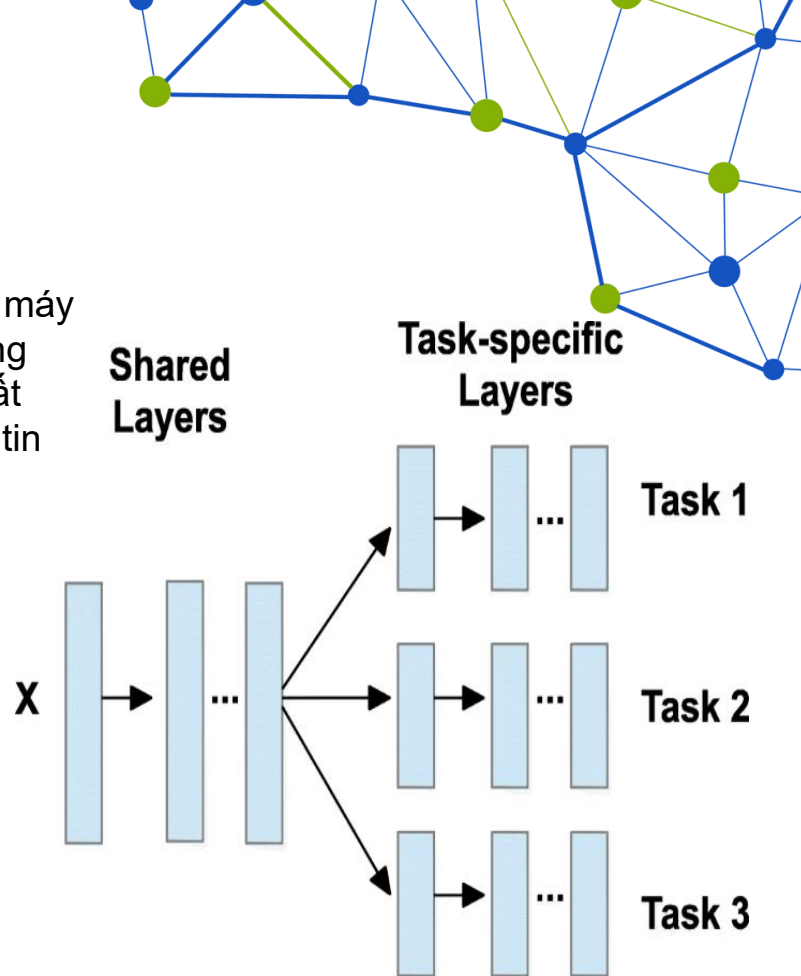
- Tô đậm** biểu thị cho kết quả tốt nhất, gạch chân biểu thị cho kết quả tốt thứ hai.
- start\_f1: F1 score của vị trí bắt đầu của câu trả lời trong ngữ cảnh
- end\_f1: F1 score của vị trí kết thúc của câu trả lời trong ngữ cảnh

## 6 Multitask

### 6.1 Giới thiệu

Multitask learning là một phương pháp huấn luyện mô hình máy học mà một mô hình được huấn luyện trên nhiều tác vụ cùng một lúc. Mục tiêu của multitask learning là cải thiện hiệu suất của mô hình trên tất cả các tác vụ bằng cách chia sẻ thông tin giữa các tác vụ.

Triển khai mô hình BART cho đa nhiệm vụ (multi-task learning) gồm dịch máy và Sentiment Analysis



# 6 Multitask

## 6.1 Giới thiệu

### Dịch máy

- **Bài toán:** chuyển đổi văn bản từ một ngôn ngữ nguồn sang một ngôn ngữ đích, trong đó bảo toàn ý nghĩa và cấu trúc ngữ pháp của văn bản gốc.
- **Input:** một đoạn văn bản trong ngôn ngữ nguồn (tiếng Anh hoặc tiếng Việt).
- **Output:** một đoạn văn bản trong ngôn ngữ đích, tương ứng với nội dung của văn bản nguồn.

### Sentiment Analysis

- **Bài toán:** phân loại các đánh giá/tài liệu tiếng Việt thành một trong hai loại: tích cực, tiêu cực.
- **Input:** một đoạn văn bản tiếng Việt.
- **Output:** một trong hai nhãn: tích cực, tiêu cực.

## 6 Multitask

### 6.2 Dữ liệu

#### IWSLT English-Vietnamese 2015

- **Data characteristics:** Bộ dữ liệu này bao gồm các cặp câu song ngữ Anh-Việt được sử dụng trong bài toán dịch máy

Table 5.1: Thống kê bộ dữ liệu IWSLT English-Vietnamese 2015

	Training Set	Validate Set	Test Set
Number of Sentence Pairs	133,166	1,553	1,268
Average Sentence Length (English)	20.3	18	21.1
Average Sentence Length (Vietnamese)	24.8	22	26.5

Dữ liệu cho bài toán Sentiment Analysis đã được trình bày ở phần 3.

## 6 Multitask

### 6.3 Hướng tiếp cận

- Chuyển đổi tất cả các nhiệm vụ thành định dạng "text-to-text". Điều này có nghĩa là mô hình nhận đầu vào là một văn bản mô tả hoặc điều kiện và phải sinh ra một văn bản đầu ra.
- Mô hình được huấn luyện với mục tiêu maximum likelihood không phụ thuộc vào loại tác vụ nào.
- Để chỉ ra tác vụ mà mô hình cần thực hiện, chúng tôi thêm một tiền tố văn bản cụ thể cho chuỗi đầu vào gốc trước khi đưa vào mô hình.



## 6 Multitask

### 6.3 Hướng tiếp cận

- Để chỉ ra tác vụ mà mô hình cần thực hiện, chúng tôi thêm một tiền tố văn bản cụ thể cho chuỗi đầu vào gốc trước khi đưa vào mô hình.

source	target	task
Hơi thất vọng chỉ tiết máy bay không đạt yêu ...	0	sa
In fact , while we drove 140,000 miles , peopl...	Trên thực tế nó đã lái cả 140,000 dặm , mà thậ...	mt-en-vi
Và nếu phải chọn 1 nhóm kiểu như Trường đại họ...	And if I had to pick a group that I think is o...	mt-vi-en
Nên các bạn có thể -- nhảy trong một không gia...	So you can have -- and you can do it in a very...	mt-vi-en
Cậu trông có vẻ bận . Cậu có khoẻ không ?	You look busy . How are you ?	mt-vi-en
...	...	...
Nhưng những bong bóng đó có màng rất giống với...	But those bubbles have membranes very similar ...	mt-vi-en
micro thu âm tiếng nhỏ	0	sa
Shop phục vụ rất tốt.\nCó phản hồi hỗ trợ. Lỗi...	1	sa
On the left there is the PackBot from iRobot .	Bên trái đây là PackBot từ iRobot .	mt-en-vi
Ví như Andrew Wilder , được sinh ra ở khu vực ...	Andrew Wilder , for example , born on the Paki...	mt-vi-en

Ví dụ đầu vào trước khi tiền xử lý.

source	target	task
Classify the sentiment: Hơi thất vọng chỉ tiế...	0	sa
Translate English to Vietnamese: In fact , whi...	Trên thực tế nó đã lái cả 140,000 dặm , mà thậ...	mt-en-vi
Translate Vietnamese to English: Và nếu phải c...	And if I had to pick a group that I think is o...	mt-vi-en
Translate Vietnamese to English: Nên các bạn c...	So you can have -- and you can do it in a very...	mt-vi-en
Translate Vietnamese to English: Cậu trông có ...	You look busy . How are you ?	mt-vi-en
...	...	...
Translate Vietnamese to English: Nhưng những b...	But those bubbles have membranes very similar ...	mt-vi-en
Classify the sentiment: micro thu âm tiếng nhỏ	0	sa
Classify the sentiment: Shop phục vụ rất tốt...	1	sa
Translate English to Vietnamese: On the left t...	Bên trái đây là PackBot từ iRobot .	mt-en-vi
Translate Vietnamese to English: Ví như Andrew...	Andrew Wilder , for example , born on the Paki...	mt-vi-en

Ví dụ đầu vào sau khi tiền xử lý.

## 6 Multitask

### 6.4 Kết quả

Table 5.4: So sánh định lượng tất cả mô hình trên 3 nhiệm vụ

Model	Task 1: Dịch Anh-Việt	Task 2: Dịch Việt-Anh	Task 3: Sentiment Analysis	
			VLSP 2016	AIVIVN 2019
BARTPho-base	20.99	16.69	<b>0.9899</b>	0.899
BARTPho	<b>22.05</b>	<b>18.23</b>	0.988	<b>0.9055</b>

Mô hình BARTPho thể hiện hiệu suất vượt trội so với BARTPho-base trong hầu hết các nhiệm vụ:

- Trong dịch máy, BARTPho đạt điểm BLEU-4 cao hơn cho cả hai hướng dịch (22.05 so với 20.99 cho Anh-Việt và 18.23 so với 16.69 cho Việt-Anh).
- Trong SA, BARTPho cho kết quả tốt hơn trên tập dữ liệu AIVIVN 2019 (0.9055 so với 0.899).

=> Kết quả này chứng minh lợi thế của mô hình có kích thước lớn hơn trong việc nắm bắt và xử lý thông tin ngôn ngữ phức tạp

## 6 Multitask

### 6.4 Kết quả

#### So sánh định tính

- **Input:** Thank you .
- **Predict:**
  - BARTPho-base: Cảm ơn.
  - BARTPho: Xin cảm ơn.
- **Target:** Cảm ơn .

*Nhận xét:* Cả hai mô hình đều cho kết quả chính xác. BARTPho thêm từ "Xin" làm câu lịch sự hơn, trong khi BARTPho-base gần với câu đích hơn.

- **Input:** Đây là chân dung gia đình .
- **Predict:**
  - BARTPho-base: This is the family picture.
  - BARTPho: This is a family portrait.
- **Target:** This is a family portrait .

*Nhận xét:* BARTPho cho kết quả chính xác hoàn toàn, trong khi BARTPho-base dùng từ "picture" thay vì "portrait".

## 6 Multitask

### 6.4 Kết quả

#### So sánh định tính

- **Input:** Và họ thả tôi ra . Đó quả là một phép màu .
- **Predict:**
  - BARTPho-base: And they pulled me out. It is a color.
  - BARTPho: And they pull me out. That is a magic.
- **Target:** And they let me go . It was a miracle .

*Nhận xét:* Cả hai mô hình đều gặp khó khăn với câu này. BARTPho-base dịch sai nghĩa của "phép màu", trong khi BARTPho dịch gần đúng hơn nhưng vẫn chưa chính xác.

## 6 Multitask

### 6.4 Kết quả

#### So sánh định tính

Các ví dụ này tập trung vào nhiệm vụ phân loại cảm xúc, với 1 đại diện cho cảm xúc tích cực và 0 đại diện cho cảm xúc tiêu cực.

Input	BARTPho-base	BARTPho	Target
Đồng hồ yêu lắm ạ thích lắm luôn 🥰🥰🥰	1	1	1
quạt không chạy êm có tiếng ồn	1	0	0
tì vi_dep qua	0	1	1
Nếu các bạn đi theo con đường âm_thanh thì ko nên bỏ_qua con này đâu	0	1	1
Oke	0	1	1

Table 5.5: So sánh kết quả phân loại cảm xúc

*Nhận xét:* BARTPho thể hiện hiệu suất tốt hơn trong nhiệm vụ phân loại cảm xúc, với 4/5 trường hợp chính xác so với 2/5 của BARTPho-base.

Qua các ví dụ trên, có thể thấy BARTPho thường cho kết quả tốt hơn BARTPho-base, đặc biệt trong các nhiệm vụ phức tạp và phân loại cảm xúc. Tuy nhiên, cả hai mô hình đều còn có những hạn chế nhất định trong việc xử lý ngôn ngữ tự nhiên.

## 6 Multitask

### 6.5 Thử nghiệm trên dữ liệu chưa biết

Mô hình của chúng tôi được huấn luyện đa nhiệm vụ, bao gồm phân tích cảm xúc (sentiment analysis) và dịch Anh-Việt. **Nhưng phần dữ liệu huấn luyện cho nhiệm vụ phân tích cảm xúc chỉ chứa văn bản tiếng Việt.**

Để đánh giá khả năng tổng quát hóa của mô hình, chúng tôi tiến hành kiểm thử trên **tập dữ liệu Twitter Sentiment Analysis với dữ liệu hoàn toàn bằng tiếng Anh**. Tập dữ liệu này chứa các tweet với nhãn cảm xúc tích cực, tiêu cực hoặc trung lập.

Table 5.8: Kết quả trên tập dữ liệu Twitter Sentiment Analysis

Model	Precision	Recall	F1	Accuracy
Baseline (Majority Class)	0.5087	0.5087	0.5087	0.5087
BARTPho-word-base	0.4932	0.4872	0.4872	0.7345
BARTPho-word	<b>0.8236</b>	<b>0.8227</b>	<b>0.823</b>	<b>0.8235</b>

## 7 Phân công công việc

Sinh viên	Nhiệm vụ	Đóng góp
<ul style="list-style-type: none"><li>Bùi Đức Mạnh (NT)</li></ul>	<ul style="list-style-type: none"><li>Đảm nhận task Sentiment Analysis</li><li>Đảm nhận task Multitask: Dịch máy + SA</li><li>Viết báo cáo chương 2, 5</li><li>Xây dựng demo</li><li>Làm slide báo cáo</li></ul>	<ul style="list-style-type: none"><li>33.3%</li></ul>
<ul style="list-style-type: none"><li>Đàm Thái Ninh</li></ul>	<ul style="list-style-type: none"><li>Đảm nhận task Named Entity Recognition</li><li>Viết báo cáo chương 3</li></ul>	<ul style="list-style-type: none"><li>33.3%</li></ul>
<ul style="list-style-type: none"><li>Lê Việt Hùng</li></ul>	<ul style="list-style-type: none"><li>Đảm nhận task Question Answering</li><li>Viết báo cáo chương 4</li></ul>	<ul style="list-style-type: none"><li>33.3%</li></ul>

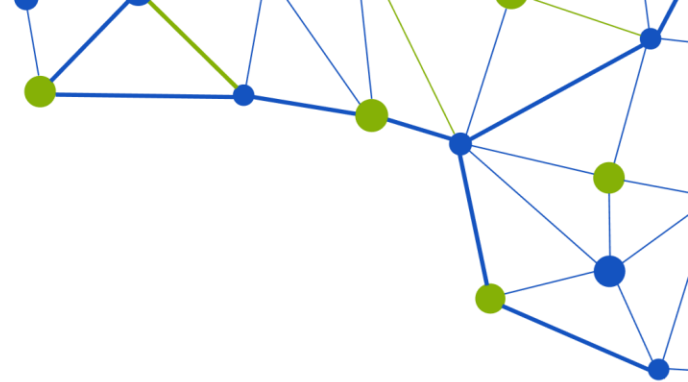
## Hiệu suất tổng thể

- Mô hình BARTPho thể hiện hiệu suất vượt trội so với BARTPho-base trong hầu hết các nhiệm vụ:
- Trong dịch máy: BARTPho đạt điểm BLEU-4 cao hơn cho cả hai hướng dịch (22.05 so với 20.99 cho Anh-Việt và 18.23 so với 16.69 cho Việt-Anh).
- Trong phân tích cảm xúc, BARTPho cho kết quả tốt hơn trên tập dữ liệu AIVIVN 2019 (0.9055 so với 0.899).
- Kết quả này chứng minh lợi thế của mô hình có kích thước lớn hơn trong việc nắm bắt và xử lý thông tin ngôn ngữ phức tạp.

## Ưu điểm của multi-task learning

- Phương pháp này cho phép một mô hình duy nhất thực hiện nhiều tác vụ khác nhau, tiết kiệm tài nguyên và thời gian huấn luyện.
- Kết quả cho thấy việc học đồng thời nhiều nhiệm vụ có thể cải thiện hiệu suất tổng thể của mô hình.





## Hạn chế

- Khó khăn trong việc nắm bắt ngữ cảnh và cảm xúc phức tạp
- Hạn chế trong xử lý các cụm từ đặc thù văn hóa
- Đôi khi gặp khó khăn với các biểu hiện mỉa mai hoặc gián tiếp trong phân tích cảm xúc

## Tầm quan trọng của dữ liệu

- Nghiên cứu nhấn mạnh vai trò quan trọng của chất lượng dữ liệu huấn luyện, đặc biệt trong việc gán nhãn chính xác cho các tác vụ phân loại

## Sản phẩm của nhóm

- Link github: [https://github.com/NinhDT22022522/BTL\\_NLP\\_2024](https://github.com/NinhDT22022522/BTL_NLP_2024)
- Link demo: <https://huggingface.co/spaces/mc0c0z/btl-nlp>



# IAI-UET

**VIỆN TRÍ TUỆ NHÂN TẠO**

ĐẠI HỌC CÔNG NGHỆ - ĐẠI HỌC QUỐC GIA HÀ NỘI

**THANKS YOU!**



[facebook.com/iaiu](https://facebook.com/iaiu)