ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Đào Thị Ngọc Giàu – 21120396
Nguyễn Thị Ngọc Châm – 21120417
Kiên Đinh Mỹ Hạnh – 21120446

Báo cáo cuối kỳ – Ứng dụng mô hình ngôn ngữ lớn trong bài toán xác định tương đồng văn bản xuyên ngữ Anh-Việt

Môn học: Nhập môn xử lý ngôn ngữ tự nhiên

GIÁO VIÊN HƯỚNG DẪN
Thầy Nguyễn Hồng Bửu Long
Thầy Lê Thanh Tùng
Thầy Lương An Vinh

Tp. Hồ Chí Minh, tháng 01/2024

Lời cảm ơn

Chúng em xin gửi lời cảm ơn chân thành đến thầy ***Nguyễn Hồng Bửu Long***– giảng viên lý thuyết và thầy ***Lê Thanh Tùng,*** thầy ***Lương An Vinh*** – giảng viên thực hành môn học *Nhập môn xử lý ngôn ngữ tự nhiên* đã trang bị cho nhóm chúng em tài liệu hướng dẫn cùng những kiến thức và kỹ năng cơ bản cần có để hoàn thành bài thực hành này.

Tuy nhiên trong quá trình thực hiện đồ án, vì kiến thức chuyên ngành còn hạn chế nên nhóm chúng em vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá và trình bày báo cáo. Rất mong nhận được sự quan tâm, góp ý của các thầy để bài báo cáo thực hành của nhóm chúng em được đầy đủ và hoàn chỉnh hơn.

Chúng em xin chân thành cảm ơn!

Mục lục

I. Đánh giá mức độ hoàn thành

STT	MSSV	Họ và tên	Phân công	Đóng góp
1	21120396	Đào Thị Ngọc Giàu	- Xây dựng ứng dụng tích hợp mô hình ngôn ngữ lớn. - Demo ứng dụng. - Viết báo cáo	100%
2	21120417	Nguyễn Thị Ngọc Châm	- Tìm hiểu và sử dụng mô hình ngôn ngữ lớn trong bài toán xác định tương đồng văn bản xuyên ngữ Anh-Việt. - Thiết kế slide báo cáo. - Viết báo cáo	100%
3	21120446	Kiên Đinh Mỹ Hạnh	- Thu thập thập dữ liệu, tạo dataset. - Báo cáo thuyết trình. - Viết báo cáo	100%

II. Nội dung đồ án

1. Giới thiệu đề tài

1.1. Ứng dụng mô hình ngôn ngữ lớn trong bài toán xác định tương đồng văn bản xuyên ngữ Anh-Việt.

- Bài toán đánh giá tương đồng văn bản xuyên ngữ Anh-Việt là một thách thức trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Mục tiêu chính của bài toán này là đo lường mức độ tương đồng giữa hai đoạn văn bản, một bằng tiếng Anh và một bằng tiếng Việt. Đánh giá tương đồng văn bản có thể áp dụng trong nhiều ứng dụng, bao gồm:

+ **Dịch máy**: Đánh giá mức độ tương đồng giữa văn bản nguồn và văn bản dịch để đảm bảo chất lượng của quá trình dịch.

+ **Tìm kiếm thông tin**: Xác định mức độ tương đồng giữa các đoạn văn bản để cải thiện hiệu suất tìm kiếm và truy xuất thông tin.

+ **Phân loại văn bản**: Đánh giá sự tương đồng giữa các bài báo, bài viết hoặc văn bản khác nhau để hỗ trợ quá trình phân loại.

+ **Kiểm tra đồng nghĩa và chuyển đổi ngôn ngữ**: Đánh giá mức độ tương đồng giữa các biểu diễn ngôn ngữ Anh và Việt để hỗ trợ các ứng dụng như chuyển đổi ngôn ngữ và kiểm tra đồng nghĩa.

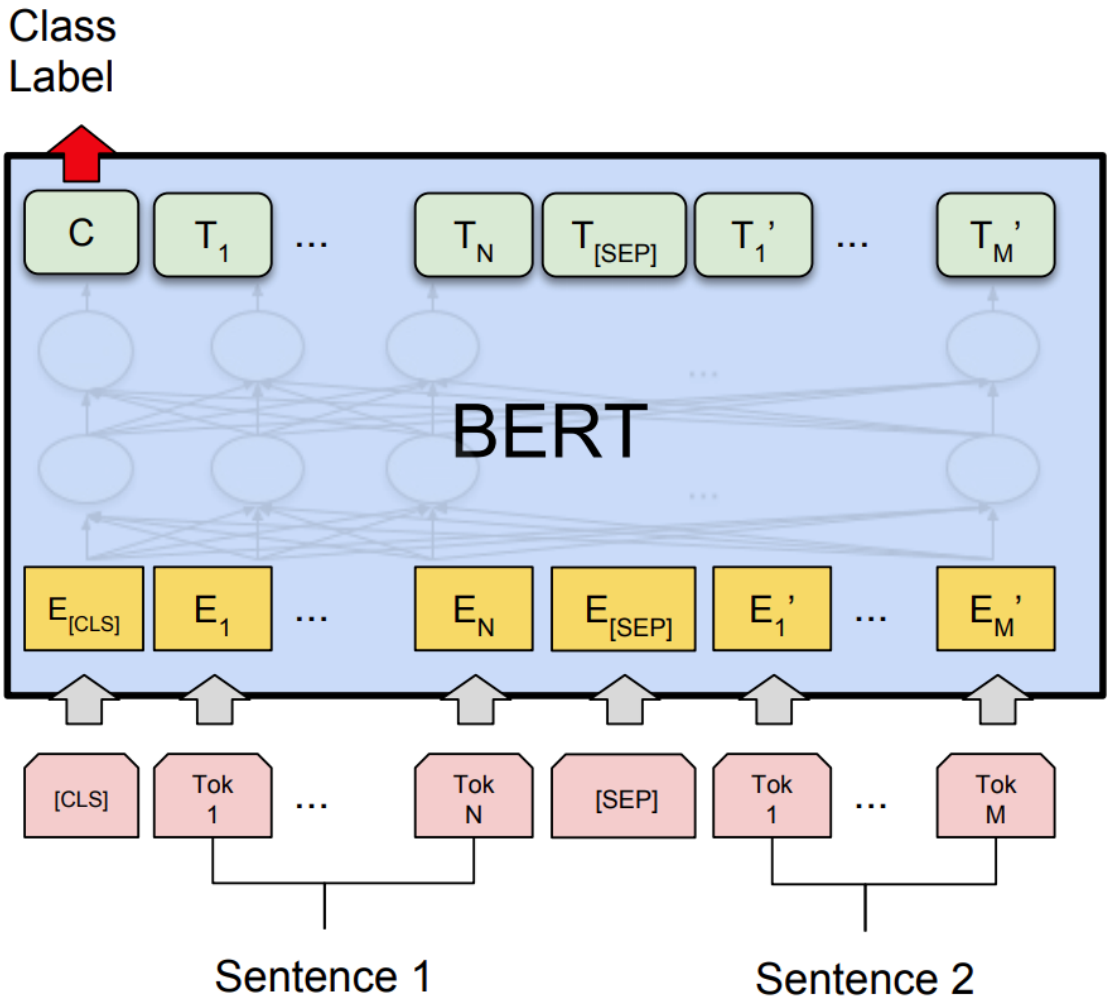
- Cách tiếp cận bài toán này thường sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên, biểu diễn từ (word embeddings), và mô hình học máy như mô hình BERT, GPT, hoặc các mô hình siamese network. Đối với tiếng Anh và tiếng Việt, có thể sử dụng các bộ dữ liệu song ngữ để huấn luyện mô hình với cặp câu tương đồng và không tương đồng.

- Đánh giá hiệu suất của mô hình thường được thực hiện thông qua các độ đo như precision, recall, F1-score, hoặc cosine similarity giữa các biểu diễn văn bản. Bài toán này đặt ra nhiều thách thức do sự đa dạng của ngôn ngữ và khác biệt về cấu trúc câu trong cả hai ngôn ngữ.

1.2. Mô hình mBERT

- mBERT là một BERT đa ngôn ngữ được đào tạo trước trên 104 ngôn ngữ, được phát hành bởi tác giả của bài báo gốc về kho lưu trữ GitHub chính thức của Google Research: **google-research / bert on** Tháng 11/2018. mBERT tuân theo cấu trúc tương tự của BERT.

- Sự khác biệt duy nhất là rằng mBERT được đào tạo trước về dữ liệu Wikipedia được ghép nối cho 104 ngôn ngữ và nó hoạt động tốt một cách đáng ngạc nhiên so với việc nhúng từ đa ngôn ngữ trên zero-shot chuyển giao đa ngôn ngữ trong tập dữ liệu XNLI, tập dữ liệu **Suy luận ngôn ngữ tự nhiên xuyên ngữ (XNLI)** đã trở thành một tập dữ liệu tiêu chuẩn cho mục đích này. mBERT được sử dụng rộng rãi cho các nhiệm vụ đa ngôn ngữ.



1.3. Các công trình liên quan

- Nghiên cứu của Potthast [1] đã phân loại các phương so sánh độ tương đồng ngữ nghĩa xuyên ngữ theo năm mô hình như trong Bảng 1. Nhóm mô hình dựa trên cấu trúc với phương pháp có phương pháp CL-CNG [2] làm đại diện. Ý tưởng chính của phương pháp này là so sánh các cặp câu sử dụng các n-gram được trích xuất từ các từ liên tiếp nhau trong câu. Phương pháp này không đạt hiệu quả cao trên các cặp ngôn ngữ khác cấu trúc cú pháp hoặc không cùng nhóm ngôn ngữ, nên không áp dụng được hiệu quả cho cặp ngôn ngữ Anh - Việt. Bên cạnh đó, phương pháp CL-CTS [3] đại diện cho nhóm mô hình dựa trên tự

điển, có ý tưởng chính là biểu diễn văn bản dưới dạng véctơ khái niệm và tiến hành so sánh độ tương đồng hai văn bản dựa trên hai véctơ của chúng.

- Với nhóm mô hình dựa trên kho ngữ liệu song song, phương pháp CL-ASA [4] được phát triển dựa trên công nghệ dịch máy thống kê. Với hai văn bản d và d' thuộc hai ngôn ngữ khác nhau L và L', phương pháp tính toán xác suất mỗi từ ở d là bản dịch của mỗi từ ở d' dựa trên cặp kho ngữ liệu song song thuộc hai ngôn ngữ L và L'. Từ xác suất các cặp từ là bản dịch của nhau, tính toán xác suất hai văn bản d và d' là bản dịch của nhau. Mô hình này phụ thuộc nhiều vào chất lượng kho ngữ liệu và mô hình Length và chỉ hiệu quả cao với cặp câu được dịch bởi các chuyên gia hay dịch tự động.

Bảng 1. Các mô hình so sánh độ tương đồng ngữ nghĩa xuyên ngữ

Tên nhóm mô hình	Phương pháp đại diện
Mô hình dựa trên cấu trúc (Syntax-based model)	Phương pháp CL-CNG (McNamee và Mayfield, 2004)
Mô hình dựa trên tự điển (Dictionary-based model)	Phương pháp CL-CTS (Gupta, 2012)
Mô hình dựa trên kho ngữ liệu song song (Parallel corpus- based model)	Phương pháp CL-ASA (Pinto, 2009)
Mô hình dựa trên kho ngữ liệu có thể so sánh (Comparable corpus- based model)	Phương pháp CL-KGA (M. Franco-Salvador, 2015)
Mô hình dựa trên dịch tự động (Machine translation- based model)	Phương pháp dịch và phân tích đơn ngữ (Barrón-Cedeno, 2012)

2. Phương pháp cài đặt

2.1. Sử mô hình ngôn ngữ lớn cho bài toán xác định tương đồng ngữ nghĩa văn bản xuyên ngữ Anh-Việt.

- Ý tưởng thực hiện: Sử dụng mô hình ngôn ngữ lớn huấn luyện trước và thực hiện fine-tuning cho bài toán xác định độ tương đồng giữa 2 văn bản Anh -Việt.

+ **Bước 1:** Chuẩn bị dữ liệu:

- Thu thập và chuẩn bị dữ liệu văn bản gồm những cặp câu tiếng Anh và tiếng Việt, đánh giá độ tương đồng của từng cặp câu.
- Tiền xử lý dữ liệu.
- Phân chia dữ liệu thành các tập huấn luyện (train) và thử nghiệm (test).

+ **Bước 2:** Fine-tuning (Tinh chỉnh): Thực hiện tinh chỉnh trên mô hình ngôn ngữ lớn bằng cách sử dụng dữ liệu huấn luyện đã chuẩn bị.

+ **Bước 3:** Biểu diễn văn bản: Sử dụng mô hình ngôn ngữ lớn để biểu diễn các đoạn văn bản trong tập dữ liệu.

+ **Bước 4:** Đo độ tương đồng: Sử dụng độ đo tương đồng cosine similarity để đo lường sự tương đồng giữa các biểu diễn văn bản.

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

+ **Bước 5:** Đánh giá hiệu suất: Sử dụng tập dữ liệu thử nghiệm (test) để đánh giá hiệu suất của mô hình.

2.1.1. Các bước tiến hành tạo tập dữ liệu

- Tập dữ liệu ban đầu được tải về từ hugging face với độ lớn hơn 11.000 cặp câu Anh-Việt tương đồng nhau về mặt ngữ nghĩa. Các cặp câu Anh-Việt được thu thập từ nhiều tập dữ liệu song ngữ như TED2020 v1, wikimedia v20210402, WikiMatrix v1, OpenSubtitiles v2018.

- Cột **"en"** chứa câu tiếng Anh, cột **"vi"** chứa câu tiếng Việt.

	en	vi	source
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	TED2020 v1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	wikimedia v20210402
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	WikiMatrix v1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	OpenSubtitles v2018
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	OpenSubtitles v2018
...	...	...	...
11220	It's important.	Đây là điều quan trọng.	TED2020 v1
11221	I looked at him and I saw myself.	Mình nhìn vào nó, và nhìn thấy chính con người...	OpenSubtitles v2018
11222	In India it is distributed mainly on the plain...	Tại Ấn Độ, nó phân bố chủ yếu trên các vùng đồ...	WikiMatrix v1
11223	He moved to MIO Biwako Shiga in 2015.	Anh chuyển đến MIO Biwako Shiga năm 2015.	WikiMatrix v1
11224	The words genius and musical are used in the s...	Những từ ngữ như thiên tài âm nhạc đã được sử ...	wikimedia v20210402

11225 rows × 3 columns

- Cắt tập dữ liệu còn 10.000 dòng để tiện cho các bước xử lý

	en	vi	source
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	TED2020 v1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	wikimedia v20210402
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	WikiMatrix v1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	OpenSubtitles v2018
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	OpenSubtitles v2018
...	...	...	...
9995	She is the most disgusting of journalists.	Cô ta là ký giả đáng ghét nhất trong đám ký giả.	OpenSubtitles v2018
9996	Legolas then tells Thranduil he must leave, an...	Legolas sau đó nói với Thranduil rằng anh ta p...	WikiMatrix v1
9997	British America gained large amounts of new te...	Mỹ thuộc Anh đã giành được một lượng lớn lãnh ...	WikiMatrix v1
9998	A statement from the Ministry of Information a...	Một tuyên bố của Bộ Thông tin và Nghệ thuật nó...	WikiMatrix v1
9999	Shaobing contains a variety of stuffings that ...	Shaobing chứa nhiều loại độn có thể được chia ...	WikiMatrix v1

10000 rows × 3 columns

- Thay đổi nội dung tập dữ liệu

+ **Bước 1:** Thay cột **"source"** bằng cột **"label"**, đánh số 1 (tương đồng) cho toàn bộ cột **"label"**.

	en	vi	label
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	1
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	1
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	1
...	...	...	...
9995	She is the most disgusting of journalists.	Cô ta là ký giả đáng ghét nhất trong đám ký giả.	1
9996	Legolas then tells Thranduil he must leave, an...	Legolas sau đó nói với Thranduil rằng anh ta p...	1
9997	British America gained large amounts of new te...	Mỹ thuộc Anh đã giành được một lượng lớn lãnh ...	1
9998	A statement from the Ministry of Information a...	Một tuyên bố của Bộ Thông tin và Nghệ thuật nó...	1
9999	Shaobing contains a variety of stuffings that ...	Shaobing chứa nhiều loại độn có thể được chia ...	1

10000 rows × 3 columns

- **Bước 2:** Tách tập dữ liệu thành 2 tập dữ liệu con có độ lớn bằng nhau.

+ Tập dữ liệu 1:

	en	vi	label
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	1
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	1
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	1
...	...	...	...
4995	The states, however, invested their money into...	Các tiểu bang, tuy nhiên, đầu tư tiền của họ v...	1
4996	DID patients have changed their body chemistry...	Bệnh nhân đa nhân cách (DID) thay đổi cấu trúc...	1
4997	Since 1668 the church has been part of the ben...	Từ năm 1668 nhà thờ thánh St Martin cùng thuộc...	1
4998	Pitt became Prime Minister in December 1783, w...	Tháng 12 năm 1783, Pitt trở thành Thủ tướng, W...	1
4999	So she went in search of her hedgehog.	Vì vậy, cô đã đi tìm kiếm của hedgehog cô.	1

5000 rows × 3 columns

+ Tập dữ liệu 2:

	en	vi	label
5000	There's...something different about you.	Có cái gì đó rất khác về anh.	1
5001	"The true sweetness of wine... is one flavor."	"Vị ngọt thực sự của rượu... chính là một hươn...	1
5002	There are environmental concerns.	Có những mối nguy hiểm về môi trường.	1
5003	While the original Lombardi's closed its doors...	Trong khi các Lombardi ban đầu đóng cửa vào nă...	1
5004	The Guiroue, a tributary of the Osse, flows no...	Sông Guiroue, một nhánh của sông Osse, chảy th...	1
...	...	...	...
9995	She is the most disgusting of journalists.	Cô ta là ký giả đáng ghét nhất trong đám ký giả.	1
9996	Legolas then tells Thranduil he must leave, an...	Legolas sau đó nói với Thranduil rằng anh ta p...	1
9997	British America gained large amounts of new te...	Mỹ thuộc Anh đã giành được một lượng lớn lãnh ...	1
9998	A statement from the Ministry of Information a...	Một tuyên bố của Bộ Thông tin và Nghệ thuật nó...	1
9999	Shaobing contains a variety of stuffings that ...	Shaobing chứa nhiều loại độn có thể được chia ...	1

5000 rows × 3 columns

- **Bước 3:** Xáo trộn các câu ở cột **"vi"** và đánh **"label"** bằng **0** ở tập dữ liệu 1.

+ Tập dữ liệu 1 trước khi xáo trộn:

	en	vi	label
0	And what I think the world needs now is more c...	Và tôi nghĩ điều thế giới đang cần bây giờ là ...	1
1	The group is named after Bangkok, the capital ...	Nhóm được đặt theo tên của Bangkok, thủ đô của...	1
2	It is surrounded by rivers (Simpson and Coyhai...	Nó được bao quanh bởi các con sông (Simpson và...	1
3	Four years, and you never talked to her.	Bốn năm, và cậu chưa nói gì với cô ấy.	1
4	The man that's giving the test has serious dou...	Người cung cấp bài kiểm tra đó đang nghi ngờ v...	1
...	...	...	...
4995	The states, however, invested their money into...	Các tiểu bang, tuy nhiên, đầu tư tiền của họ v...	1
4996	DID patients have changed their body chemistry...	Bệnh nhân đa nhân cách (DID) thay đổi cấu trúc...	1
4997	Since 1668 the church has been part of the ben...	Từ năm 1668 nhà thờ thánh St Martin cùng thuộc...	1
4998	Pitt became Prime Minister in December 1783, w...	Tháng 12 năm 1783, Pitt trở thành Thủ tướng, W...	1
4999	So she went in search of her hedgehog.	Vì vậy, cô đã đi tìm kiếm của hedgehog cô.	1

5000 rows × 3 columns

+ Tập dữ liệu 1 sau khi xáo trộn:

	en	vi	label
0	And what I think the world needs now is more c...	Chào mừng các bạn đến với Los Angeles và triển...	0
1	The group is named after Bangkok, the capital ...	Rivera là một họa sĩ hàng đầu của trường phái ...	0
2	It is surrounded by rivers (Simpson and Coyhai...	Trong cộng đồng Do Thái Chính Thống và Bảo Thủ...	0
3	Four years, and you never talked to her.	- Thế còn những cuốn này?	0
4	The man that's giving the test has serious dou...	Nó được cấp bằng sáng chế ở Mỹ vào năm 1961 (B...	0
...	...	...	...
4995	The states, however, invested their money into...	Ông ấy nhìn tôi rồi nhìn Donald, và ông nói, k...	0
4996	DID patients have changed their body chemistry...	em sẽ nói tất cả những gì các anh cần biết.	0
4997	Since 1668 the church has been part of the ben...	Nó chẳng chữa được gì cả, chỉ làm chậm lại.	0
4998	Pitt became Prime Minister in December 1783, w...	Trong khi sự tác động giữa thuế suất và thu nh...	0
4999	So she went in search of her hedgehog.	Thật là điều kỳ diệu cho phía Mỹ khi chỉ có ch...	0

5000 rows × 3 columns

- **Bước 4:** Gộp 2 tập dữ liệu đã phân tách trước đó lại thành 1. Ta được tập dữ liệu dùng cho bài toán.

	en	vi	label
0	And what I think the world needs now is more c...	Chào mừng các bạn đến với Los Angeles và triển...	0
1	The group is named after Bangkok, the capital ...	Rivera là một họa sĩ hàng đầu của trường phái ...	0
2	It is surrounded by rivers (Simpson and Coyhai...	Trong cộng đồng Do Thái Chính Thống và Bảo Thủ...	0
3	Four years, and you never talked to her.	- Thế còn những cuốn này?	0
4	The man that's giving the test has serious dou...	Nó được cấp bằng sáng chế ở Mỹ vào năm 1961 (B...	0
...	...	...	...
9995	She is the most disgusting of journalists.	Cô ta là ký giả đáng ghét nhất trong đám ký giả.	1
9996	Legolas then tells Thranduil he must leave, an...	Legolas sau đó nói với Thranduil rằng anh ta p...	1
9997	British America gained large amounts of new te...	Mỹ thuộc Anh đã giành được một lượng lớn lãnh ...	1
9998	A statement from the Ministry of Information a...	Một tuyên bố của Bộ Thông tin và Nghệ thuật nó...	1
9999	Shaobing contains a variety of stuffings that ...	Shaobing chứa nhiều loại độn có thể được chia ...	1

10000 rows × 3 columns

- **Bước 5:** Xáo trộn ngẫu nhiên các dòng dữ liệu

	en	vi	label
0	Is it serious, you two?	Nghiêm túc chứ, hai người ấy?	1
1	Well, it's average size.	Phiên bản PC gặp khó khăn hơn, với một số nhận...	0
2	Depends on what you're gonna do to me once you...	Hình như suốt ngày không có gì để làm.	0
3	I hear they got nice beaches, too.	Giờ tất cả đều là kẻ thù của chúng ta.	0
4	The following year she led the program Movete.	- Thằng đó xài trứng thiệt!	0
...	...	...	...
9995	Hope Pete's getting a shot of this.	Hy vọng Pete chụp được cảnh này.	1
9996	The replica took several months to build and c...	Phải mất vài tháng để xây dựng và chi phí khoả...	1
9997	But when our story begins, he was better known...	Nhưng khi tôi biết câu chuyện về anh ấy. anh ấ...	1
9998	No, sir, but that's what they did.	Năm 1989, chúng tôi đi đến phía bắc.	0
9999	But what excites me is that since I first put ...	Nhưng điều làm tôi hào hứng nhất là từ khi thú...	1

10000 rows × 3 columns

- **Tiền xử lý dữ liệu văn bản tiếng Anh và tiếng Việt:** Tiền xử lý văn bản làm cho văn bản dễ xử lý hơn và giảm độ phức tạp của dữ liệu. Dưới đây là mô tả chi tiết các bước tiền xử lý:

+ **Bước 1:** Chuyển đổi văn bản thành chữ thường (**`lowercase`**):

→ Mục tiêu: Chuyển đổi tất cả các ký tự thành chữ thường để đồng nhất hóa dữ liệu.

+ **Bước 2:** Loại bỏ ký tự không cần thiết (**`[^a-zA-Z0-9\s]`** hoặc **`[^\w\s]`**):

→ Mục tiêu: Loại bỏ các ký tự không phải là chữ cái, chữ số hoặc khoảng trắng.

+ **Bước 3:** Tách từ (**`split`** hoặc **`word_tokenize`**):

→ Mục tiêu: Chia văn bản thành các từ riêng lẻ để tiếp tục xử lý từng từ.

+ **Bước** 4: Loại bỏ từ dừng (**`stop words`**):

→ Mục tiêu: Loại bỏ các từ không mang ý nghĩa như **"and"**, **"the"**, **"is"**,... để giảm kích thước của dữ liệu và tăng độ chính xác.

+ **Bước** 5: Stemming hoặc Lemmatization (**`PorterStemmer`** hoặc **`WordNetLemmatizer`**):

→ Mục tiêu: Chuyển đổi từ về dạng gốc để giảm số lượng biến thể và làm cho từ ngữ đồng nhất hóa hơn.

+ **Stemming (`PorterStemmer`)**: Cắt bớt các hậu tố của từ để đạt được hình thức gốc.

+ **Lemmatization (`WordNetLemmatizer`)**: Chuyển đổi từ về dạng từ điển (định danh) của từ.

+ **Bước 6:** Ghép từ sau khi xử lý lại thành câu **(`join`):**

→ Mục tiêu: Tạo lại câu từ các từ đã được xử lý.

- Dưới đây là mô tả chi tiết mỗi bước áp dụng cho tiếng Anh **(`language='en'`)** và tiếng Việt **(`language='vi'`):**

+ Tiếng Anh **(`language='en'`)**:

- Chuyển đổi văn bản thành chữ thường **(`lower()`).**
- Loại bỏ ký tự không cần thiết **(`[^a-zA-Z0-9\s]`).**
- Tách từ **(`split`).**
- Loại bỏ từ dừng **(`stop words`).**
- Stemming **(`PorterStemmer`).**
- Lemmatization **(`WordNetLemmatizer`).**

+ Tiếng Việt **(`language='vi'`):**

- Chuyển đổi văn bản thành chữ thường **(`lower()`).**
- Loại bỏ ký tự không cần thiết **(`[^\w\s]`).**
- Tách từ **(`word_tokenize`).**
- Loại bỏ từ dừng **(`stop words`).**
- Nhận diện thực thể **(`ner(word)[0][0]`).**

Cuối cùng, áp dụng các bước tiền xử lý này cho cả cột **“en”** (tiếng Anh) và **“vi”** (tiếng Việt) trong DataFrame **“df”**. Kết quả cuối cùng là DataFrame được hiển thị với văn bản đã được tiền xử lý.

	en	vi	label
0	seriou two	nghiêm túc hai	1
1	well averag size	phiên bản pc nhận xét rút so sánh bất lợi tựa ...	0
2	depend gonna get insid	hình như suốt	0
3	hear got nice beach	kẻ thù	0
4	follow year led program movet	thằng xài trứng thiệt	0
...	...	...	...
9995	hope pete get shot	hy vọng pete chụp cảnh	1
9996	replica took sever month build cost estim 1 mi...	xây dựng chi phí 1 triệu usd	1
9997	stori begin better known johnni bull walker	câu chuyện johnny bull walker	1
9998	sir	1989 đi bắc	0
9999	excit sinc first put forward particular idea f...	hào hứng thúc đẩy ý tưởng khoa học bước tiến	1

10000 rows × 3 columns

2.1.2. Fine-tuning mô hình mBERT

- Quá trình fine-tuning mô hình sử dụng PyTorch và Hugging Face Transformers

+ Token hóa dữ liệu:

- Dữ liệu đào tạo **(train_texts)** và dữ liệu phát triển **(dev_texts)** được token hóa bằng **tokenizer** được tạo từ mô hình **mBERT** đã chọn.
- Thực hiện việc cắt (**truncation**) và thêm đệm (**padding**) để đảm bảo mọi đầu vào có cùng độ dài.

+ Tạo Dataset cho PyTorch:

- Sử dụng các mã hóa từ bước trước để tạo ra các đối tượng **CustomDataset** cho cả tập dữ liệu đào tạo và tập dữ liệu phát triển.
- Mỗi mục trong dataset bao gồm các đặc trưng đầu vào (**encodings**) và nhãn tương ứng (**labels**).

+ Tạo DataLoader cho PyTorch:

- Sử dụng **DataLoader** để tạo các **mini-batch** từ dataset.
- **Batch size** được chọn là **16**, và dữ liệu được xáo trộn trong quá trình huấn luyện.

+ Chuẩn bị GPU:

- Kiểm tra xem GPU có sẵn hay không và đặt mô hình lên GPU.
- Ở đây nhóm sử dụng **T4 GPU 16GB VRAM** của Google Colab.

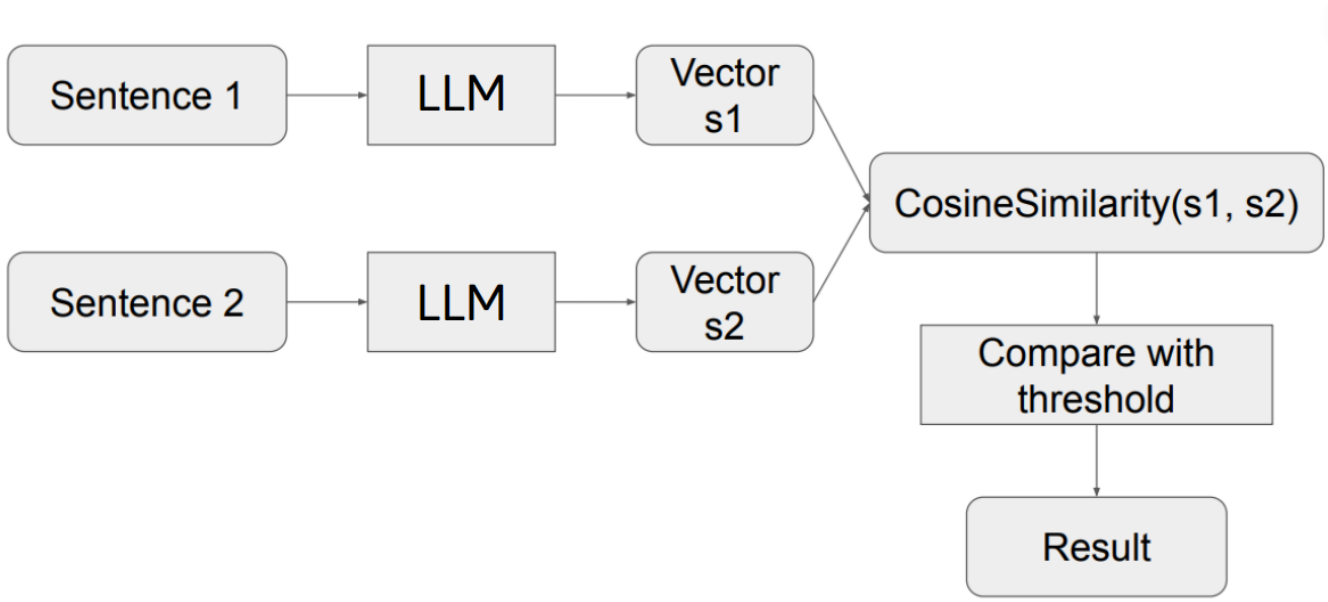
+ Tiến hành Fine-tuning:

- Sử dụng tối ưu hóa **AdamW** với **learning rate** là **2e-5**.
- Sử dụng giảm **learning rate** theo chu kỳ **(OneCycleLR)** để cập nhật **learning rate** trong suốt quá trình huấn luyện.
- Huấn luyện mô hình qua một số **epochs** được chọn (ở đây **epochs = 3**).
- Với mỗi **epoch**:
  - Đặt mô hình vào chế độ huấn luyện.

- Với mỗi **batch** trong tập dữ liệu đào tạo, thực hiện các bước sau:
  - Chuyển dữ liệu lên GPU nếu cần.
  - Đặt **gradients** bằng **0 (optimizer.zero_grad()).**
  - Chạy dữ liệu qua mô hình (**outputs = model(**inputs)).**
  - Tính **loss** và thực hiện **backpropagation (loss.backward()).**
  - Cập nhật trọng số mô hình **(optimizer.step()).**
  - Cập nhật **learning rate (scheduler.step()).**

Bằng cách thực hiện những bước này qua mỗi **epoch**, quá trình **fine-tuning** giúp mô hình tối ưu hóa các trọng số dựa trên dữ liệu đào tạo, với mong muốn là mô hình sẽ học được biểu diễn tốt cho nhiệm vụ đánh giá độ tương đồng văn bản xuyên ngữ Anh-Việt.

2.1.3. Sử dụng mô hình fine-tuned để biểu diễn các đoạn văn bản và đánh giá sự tương đồng bằng độ đo cosine similarity



- Sau khi tính độ đo tương đồng cosine similarity, ta gán nhãn độ tương đồng ứng với từng mức độ:

+ Không tương đồng: CosineSimilarity = [0;0.5).

+ Khá tương đồng: CosineSimilarity = [0.5;0.8).

+ Rất tương đồng: CosineSimilarity = [0.8;1.0).

2.2. Xây dựng ứng dụng cho phép nhập văn bản để so sánh độ tương đồng ngữ nghĩa văn bản xuyên ngữ Anh-Việt

- Ý tưởng thực hiện: Tạo web server, dùng HTML/CSS để tạo giao diện người dùng sau đó tích hợp mô hình đánh giá tương đồng văn bản vào ứng dụng web.

- Mô hình được xây dựng bằng framework FastAPI và được lưu trữ trong file model.py

Các endpoint API

/: Trả về trang web chính sử dụng template Jinja2.

/measure_dis: Nhận 2 câu văn bản, tính toán độ tương đồng theo cosine, manhattan, euclidean, trả về kết quả dưới dạng JSON.

/cluster: Nhận một tập văn bản và số lượng cụm, thực hiện phân cụm KMeans, trả về kết quả HTML hiển thị biểu đồ phân cụm.

- Tương tác với mô hình

+ Hàm get_model được sử dụng để khởi tạo và cung cấp mô hình cho các endpoint API

+ Các endpoint API gọi các phương thức của mô hình để thực hiện các tác vụ xử lý văn bản và trả về kết quả.

3. Mô tả ngữ liệu

- Tập ngữ liệu gồm 10.000 cặp câu Anh-Việt được gán nhãn:

+ Cột **"en"** chứa văn bản tiếng Anh.

+ Cột **"vi"** chứa văn bản tiếng Việt.

+ Cột **"label"** đánh giá sự tương đồng của 2 câu Anh-Việt (1: Tương đồng, 0: Không tương đồng).

	en	vi	label
0	Is it serious, you two?	Nghiêm túc chứ, hai người ấy?	1
1	Well, it's average size.	Phiên bản PC gặp khó khăn hơn, với một số nhận...	0
2	Depends on what you're gonna do to me once you...	Hình như suốt ngày không có gì để làm.	0
3	I hear they got nice beaches, too.	Giờ tất cả đều là kẻ thù của chúng ta.	0
4	The following year she led the program Movete.	- Thằng đó xài trứng thiệt!	0
...	...	...	...
9995	Hope Pete's getting a shot of this.	Hy vọng Pete chụp được cảnh này.	1
9996	The replica took several months to build and c...	Phải mất vài tháng để xây dựng và chi phí khoả...	1
9997	But when our story begins, he was better known...	Nhưng khi tôi biết câu chuyện về anh ấy. anh ấ...	1
9998	No, sir, but that's what they did.	Năm 1989, chúng tôi đi đến phía bắc.	0
9999	But what excites me is that since I first put ...	Nhưng điều làm tôi hào hứng nhất là từ khi thú...	1

10000 rows × 3 columns

- Tập ngữ liệu qua tiền xử lý:

	en	vi	label
0	seriou two	nghiêm túc hai	1
1	well averag size	phiên bản pc nhận xét rút so sánh bất lợi tựa ...	0
2	depend gonna get insid	hình như suốt	0
3	hear got nice beach	kẻ thù	0
4	follow year led program movet	thằng xài trứng thiệt	0
...	...	...	...
9995	hope pete get shot	hy vọng pete chụp cảnh	1
9996	replica took sever month build cost estim 1 mi...	xây dựng chi phí 1 triệu usd	1
9997	stori begin better known johnni bull walker	câu chuyện johnny bull walker	1
9998	sir	1989 đi bắc	0
9999	excit sinc first put forward particular idea f...	hào hứng thúc đẩy ý tưởng khoa học bước tiến	1

10000 rows × 3 columns

- Chia tập huấn luyện:

+ Số lượng samples:

	Dataset	Size
0	Training data size	8000
1	Testing data size	1000
2	Development data size	1000

- Tỷ lệ phân chia nhãn:

	Training data label	Testing data label	Development data label
0	3987	501	512
1	4013	499	488

4. Thực nghiệm

- Kết quả các độ đo đánh giá:

	Metric	Value
0	Accuracy	0.954000
1	Precision	0.954134
2	Recall	0.954000
3	F1 Score	0.954006
4	Validation Loss	0.139785

	precision	recall	f1-score	support
0	0.962302	0.947266	0.954724	512.000
1	0.945565	0.961066	0.953252	488.000
accuracy	0.954000	0.954000	0.954000	0.954
macro avg	0.953933	0.954166	0.953988	1000.000
weighted avg	0.954134	0.954000	0.954006	1000.000

- Kết quả so sánh sự tương đồng:

+ Test 1:

```
Sentence 1: I love study NLP
Sentence 2: Tôi thích học NLP
Cosine Similarity Score: 0.99995416
```

+ Test 2:

```
Sentence 1:
Twice rose to domestic fame in 2016 with their single Cheer Up, which charted at number one on the Gaon Digital Chart,
became the best-performing single of the year, and won Song of the Year at the Melon Music Awards and Mnet Asian Music Awards.
Their next single, TT, from their third EP Twicecoaster: Lane 1, topped the Gaon charts for four consecutive weeks.
The EP was the highest selling Korean girl group album of 2016. Within 19 months after debut, Twice had already sold over 1.2 million units of their four EPs and special album.
As of 2022, the group has sold over 15 million albums cumulatively in South Korea and Japan.
Sentence 2:
Twice trở nên nổi tiếng trong nước vào năm 2016 với đĩa đơn Cheer Up, đứng ở vị trí số một trên Gaon Digital Chart,
trở thành đĩa đơn có thành tích tốt nhất trong năm và giành giải Bài hát của năm tại Melon Music Awards và Mnet Asian Music Awards.
Đĩa đơn tiếp theo của họ, TT, trích từ EP thứ ba Twicecoaster: Lane 1, đứng đầu bảng xếp hạng Gaon trong bốn tuần liên tiếp.
EP này là album của một nhóm nhạc nữ Hàn Quốc bán chạy nhất năm 2016. Trong vòng 19 tháng sau khi ra mắt, Twice đã bán được hơn 1,2 triệu bản trong số 4 EP và album của họ.
Tính đến cuối năm 2021, Twice đã bán được hơn 7,29 triệu bản album tại Hàn Quốc vào năm 2019.
 Tính đến tháng 12 năm 2020, nhóm đã bán được hơn 10 triệu bản album tại Hàn Quốc và Nhật Bản.
Cosine Similarity: 0.9179484844207764
```

+ Test 3:

```
Sentence 1: I like hang out with my friends because they are so kind to me.
Sentence 2: Tôi thích đi chơi với bạn tôi.
Cosine Similarity: 0.7148616909980774
```

5. Kết luận

Qua đề tài nhóm đã được hiểu thêm về các mô hình ngôn ngữ lớn và ứng dụng chúng trong các nhiệm vụ xử lý ngôn ngữ tự nhiên. Tuy nhiên do tài nguyên còn hạn chế nên nhóm chưa thực sự hoàn thiện được mô hình, cần tinh chỉnh thêm với bộ dữ liệu lớn hơn và điều chỉnh các tham số để đạt hiệu quả cao nhất. Ngoài ra ứng dụng tích hợp mô hình ngôn ngữ lớn còn khá đơn giản và chưa ổn định, chỉ mới được chạy trên máy ảo Google colab, chưa sử dụng được như một ứng dụng độc lập trên nhiều thiết bị khác nhau và có thể sẽ xảy ra lỗi trong quá trình thử nghiệm.

6. Tài liệu tham khảo

Emerging Cross-lingual Structure in Pretrained Language Models: Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, Veselin Stoyanov - Department of Computer Science, Johns Hopkins University (2020).

XÁC ĐỊNH TƯƠNG ĐỒNG XUYÊN NGỮ ANH - VIỆT SỬ DỤNG MÔ HÌNH ĐỒ THỊ: Lê Thành Nguyên, Trần Gia Trọng Nhân, Trần Công Hậu, Đinh Điền - Trường Đại học Khoa học Tự Nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh (2019).

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019)