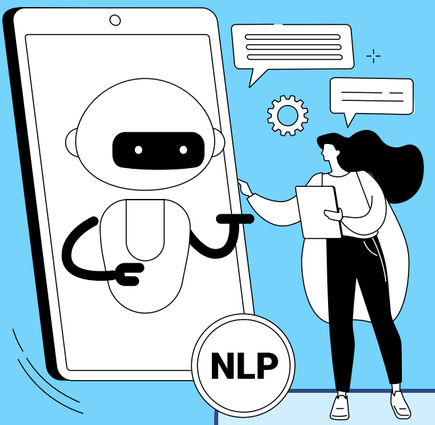


TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN
NHẬP MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG BÀI TOÁN XÁC ĐỊNH TƯƠNG ĐỒNG VĂN BẢN XUYÊN NGỮ ANH-VIỆT

Nhóm Trà đào cam sả

THÀNH VIÊN NHÓM

21120396 - Đào Thị Ngọc Giàu

21120417 - Nguyễn Thị Ngọc Châm

21120446 - Kiên Đinh Mỹ Hạnh

NỘI DUNG

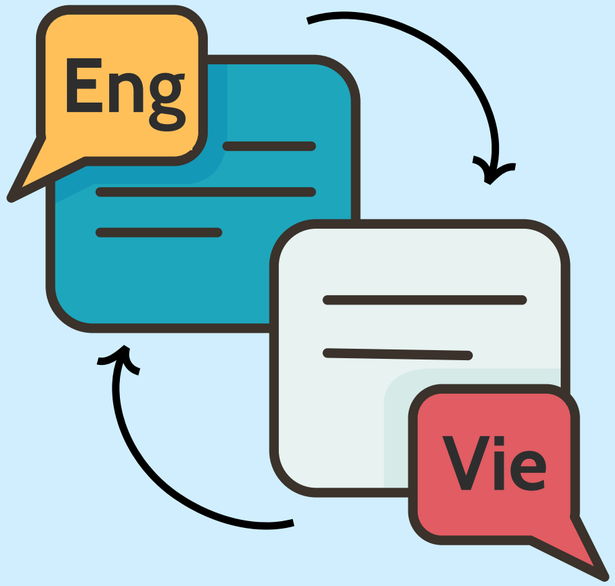
1. Giới thiệu
2. Phương pháp thực hiện
3. Demo

1. Giới thiệu

TƯƠNG ĐỒNG VĂN BẢN XUYÊN NGỮ ANH-VIỆT LÀ GÌ?

Tương đồng văn bản xuyên ngữ Anh-Việt đề cập đến sự tương đồng về mặt ngữ nghĩa của hai đoạn văn bản Anh-Việt dựa trên việc đánh giá thứ tự xuất hiện và ý nghĩa của các từ trong văn bản.

1. Giới thiệu

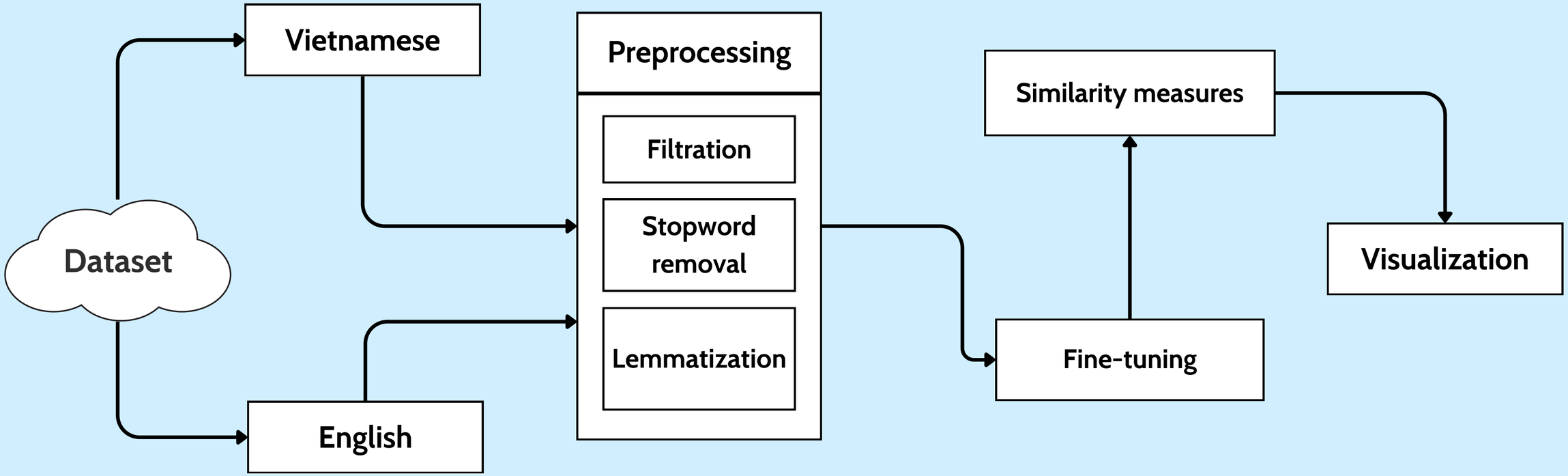


Mục tiêu

- Phát hiện đạo văn xuyên ngữ
- Kiểm tra độ chính xác trong việc dịch ngôn ngữ

2. Phương pháp thực hiện

- **Tổng quan về mô hình**



2. Phương pháp thực hiện

- **Dataset**

	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

2. Phương pháp thực hiện

- **Tiền xử lý văn bản**

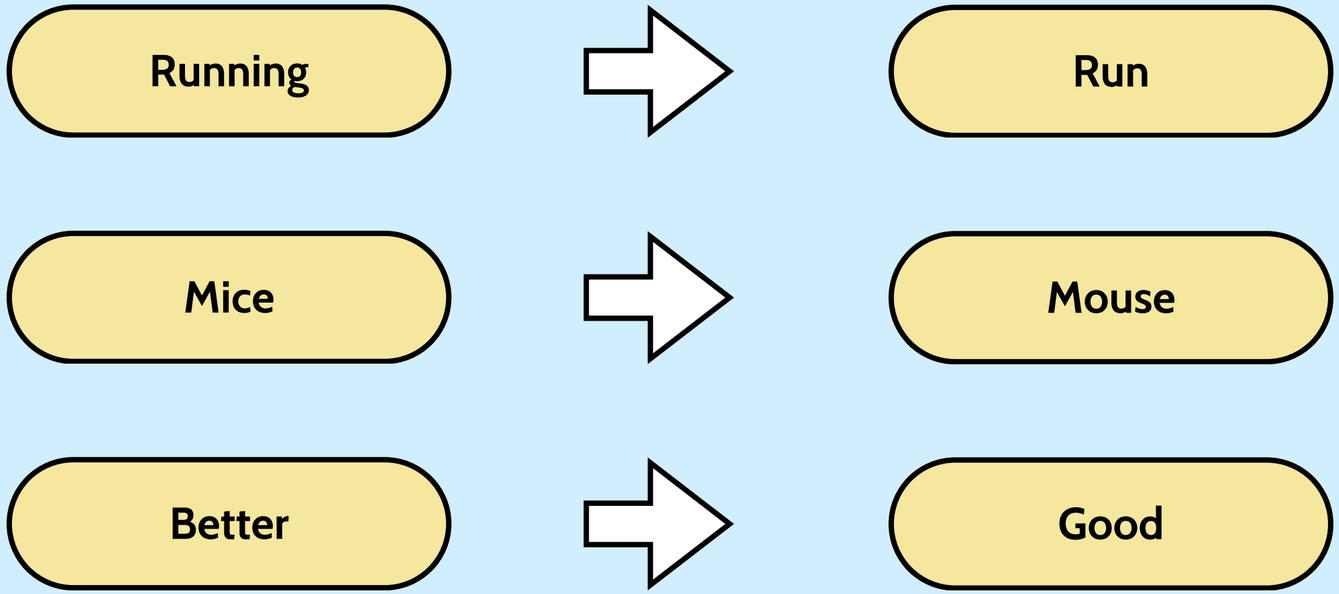
Lọc dữ liệu: Loại bỏ dữ liệu không cần thiết trong văn bản như (, ! & * $ # @ . / ":)

Loại bỏ stopwords: Những stopwords trong tiếng Anh “a”, “the”, “is”, “are”,...

Tách câu thành những từ riêng lẻ: [’I’, ‘am’, ‘a’, ‘student’]

2. Phương pháp thực hiện

- **Lemmatization**



2. Phương pháp thực hiện

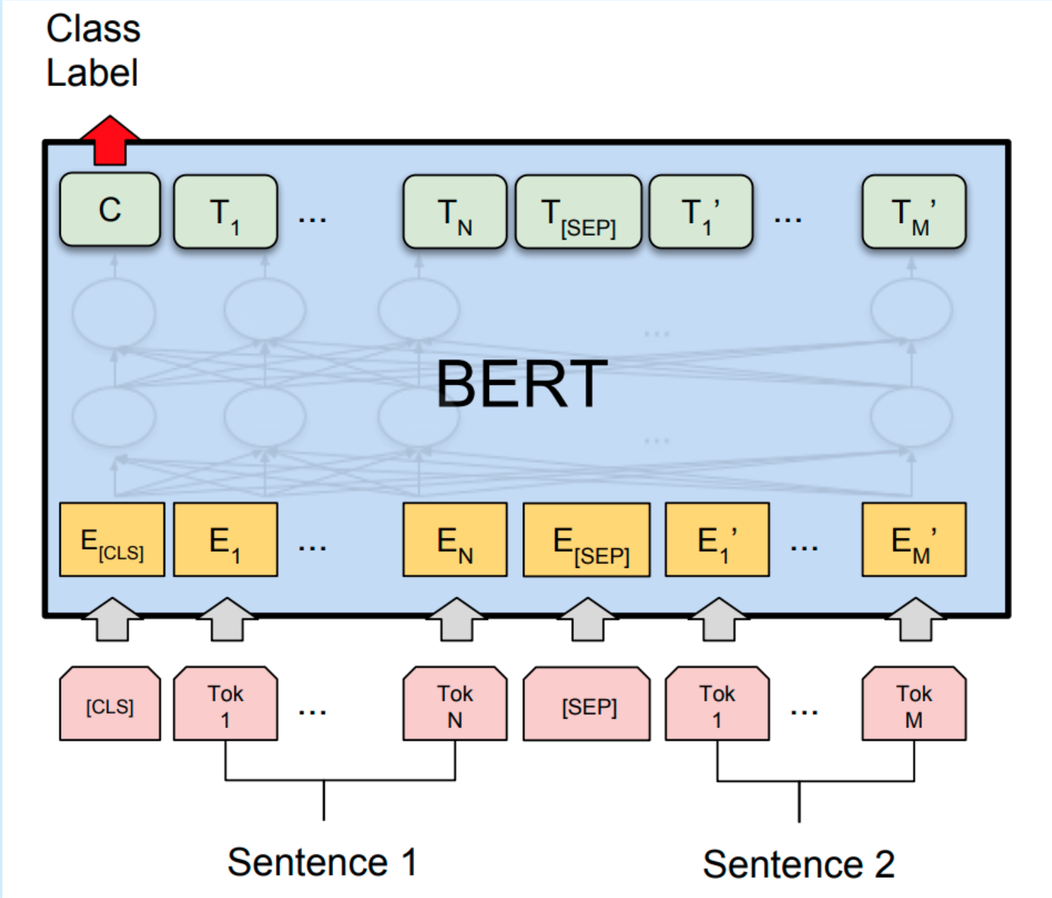
- **mBERT**

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

2. Phương pháp thực hiện

- **mBERT**



2. Phương pháp thực hiện

- **Đo độ tương đồng**

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

Cosine similarity

2. Phương pháp thực hiện

- Các độ đo đánh giá

	precision	recall	f1-score	support
0	0.9404	0.9238	0.9320	512.0000
1	0.9215	0.9385	0.9299	488.0000
accuracy	0.9310	0.9310	0.9310	0.9310
macro avg	0.9309	0.9312	0.9310	1000.0000
weighted avg	0.9312	0.9310	0.9310	1000.0000

3. DEMO

Link video demo

https://drive.google.com/file/d/1_MTJXbbNOuyIcoYScWFCnYlvJg5EWRsx/view?usp=drive_link

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 Đải học Khoa học Tự Nhiên, Đải 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).

Cảm ơn thầy và các bạn đã theo dõi!