**ĐẠI HỌC ĐÀ NẴNG**
**TRƯỜNG ĐẠI HỌC CNTT &TT VIỆT - HÀN**

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

# PHÂN TÍCH CẢM XÚC CỦA KHÁCH HÀNG VỀ CƠ SỞ Y TẾ

**Thành viên** : Lê Hồng Anh –21AD002
**GVHD** : TS. Nguyễn Văn Bình

**Đà Nẵng – 4/2025**

## 1 Giới thiệu đề tài

Mục tiêu của đề tài là áp dụng các phương pháp xử lý ngôn ngữ tự nhiên (NLP) và học máy (Machine Learning) để phân tích cảm xúc (sentiment analysis) trong các đánh giá của khách hàng về các cơ sở y tế trên Google Maps. Cụ thể, đề tài sẽ xây dựng một hệ thống tự động có khả năng phân loại các đánh giá của khách hàng thành ba nhóm cảm xúc chính: tích cực, tiêu cực và trung tính. Qua đó, hệ thống sẽ cung cấp một cái nhìn tổng thể và rõ ràng về mức độ hài lòng của khách hàng đối với các dịch vụ y tế mà họ đã trải nghiệm.

Đề tài sẽ sử dụng các thuật toán học máy và kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến để xử lý và phân tích dữ liệu đánh giá, từ đó xác định được các yếu tố chính ảnh hưởng đến cảm xúc của khách hàng. Việc phân loại này không chỉ giúp cung cấp thông tin về mức độ hài lòng mà còn giúp các cơ sở y tế nhận diện được những vấn đề tồn đọng trong dịch vụ của mình, từ đó đưa ra các quyết định cải thiện chất lượng dịch vụ một cách chính xác và hiệu quả hơn.

## 2 Beautiful Soup 4

Beautiful Soup 4 là một thư viện Python mạnh mẽ và phổ biến được sử dụng để phân tích cú pháp HTML và XML, giúp dễ dàng trích xuất dữ liệu từ các trang web. Thư viện này rất hữu ích trong việc thu thập dữ liệu từ web (web scraping), đặc biệt là khi làm việc với các trang web có cấu trúc không hoàn hảo hoặc không chuẩn.

## 3 Selenium

Selenium là một công cụ tự động hóa mạnh mẽ dành cho việc kiểm thử ứng dụng web trên nhiều trình duyệt như Chrome, Firefox, và Safari. Nó cho phép mô phỏng các thao tác người dùng như nhấp chuột, nhập liệu, và điều hướng giữa các trang. Selenium hỗ trợ nhiều ngôn ngữ lập trình như Python, Java, C#, Ruby, và JavaScript, đồng thời hoạt động trên các nền tảng như Windows, macOS, và Linux. Với khả năng tương tác linh hoạt và hỗ trợ kiểm thử song song qua Selenium Grid, đây là công cụ lý tưởng cho tự động hóa web và thu thập dữ liệu.

## 4 BERT-base-multilingual-cased

### 4.1 Giới thiệu về BERT-base-multilingual-cased

***BERT-base-multilingual-cased*** là một mô hình ngôn ngữ đa ngữ (multilingual) được phát triển bởi Google, dựa trên kiến trúc BERT (Bidirectional Encoder Representations from Transformers). Mô hình này được huấn luyện trên dữ liệu từ 104 ngôn ngữ khác nhau, bao gồm các ngôn ngữ phổ biến như tiếng Anh, tiếng Việt, tiếng Tây Ban Nha, tiếng Trung, và nhiều ngôn ngữ khác. Mô hình này có thể áp dụng cho nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP), như phân loại văn bản, phân tích cảm xúc, nhận diện thực thể tên (NER), và câu trả lời câu hỏi (question answering). Mô hình BERT-base-

multilingual-cased  có 12 lớp, 12 đầu chú ý, kích thước đầu ra 768, với ~110 triệu tham số, được huấn luyện trên dữ liệu Wikipedia và nguồn công khai.

BERT-base-multilingual-cased được huấn luyện với hai nhiệm vụ chính để học biểu diễn ngôn ngữ:

Mô hình ngôn ngữ bị che giấu (MLM): lấy một câu, mô hình che giấu ngẫu nhiên 15% các từ trong đầu vào sau đó chạy toàn bộ câu bị che giấu qua mô hình và phải dự đoán các từ bị che giấu. Điều này khác với các mạng nơ-ron hồi quy truyền thống (RNN) thường thấy các từ theo thứ tự, hoặc từ các mô hình tự hồi quy như GPT che giấu các mã thông báo tương lai. Nó cho phép mô hình học cách biểu diễn hai chiều của câu.

Dự đoán câu tiếp theo (NSP): các mô hình nối hai câu bị che giấu làm đầu vào trong quá trình tiền huấn luyện. Đôi khi chúng tương ứng với các câu nằm cạnh nhau trong văn bản gốc, đôi khi thì không. Sau đó, mô hình phải dự đoán xem hai câu có theo sau nhau hay không.

Theo cách này, mô hình sẽ học được cách biểu diễn bên trong của các ngôn ngữ trong tập huấn luyện, sau đó có thể được sử dụng để trích xuất các đặc điểm hữu ích cho các tác vụ tiếp theo: ví dụ, nếu bạn có một tập dữ liệu các câu được gắn nhãn, bạn có thể huấn luyện một bộ phân loại chuẩn bằng cách sử dụng các đặc điểm do mô hình BERT tạo ra làm đầu vào.

### 4.2 Tính năng nổi bật của BERT-base-multilingual-cased

- Đa ngữ: Mô hình có khả năng xử lý văn bản từ nhiều ngôn ngữ khác nhau, giúp nó hữu ích trong các tình huống cần phân tích dữ liệu ngôn ngữ đa dạng.

- Cased: Mô hình này "cased", có nghĩa là nó phân biệt chữ hoa và chữ thường trong quá trình huấn luyện và dự đoán, điều này giúp xử lý chính xác hơn các ngôn ngữ có sự phân biệt chữ hoa và chữ thường như tiếng Đức, tiếng Pháp, và tiếng Việt.

- Chuyển đổi hai chiều: BERT sử dụng cơ chế Transformers để hiểu ngữ cảnh của từ trong cả hai hướng (trái sang phải và phải sang trái), giúp cải thiện khả năng nắm bắt ngữ nghĩa chính xác trong câu.

- Ứng dụng đa dạng: Mô hình này có thể được áp dụng cho nhiều bài toán NLP, đặc biệt là trong việc xử lý các ngôn ngữ không phải tiếng Anh, giúp mở rộng khả năng của BERT cho các ngôn ngữ khác.

### 4.3 Cách thức BERT-base-multilingual-cased xử lý văn bản

1. Tiếp nhận văn bản đầu vào: Văn bản thô được đưa vào hệ thống để xử lý.

2. Tiền xử lý bằng tokenizer: Văn bản được tách thành các đơn vị con (subword), thêm các token đặc biệt như [CLS] và [SEP], sau đó chuyển thành chỉ số (ID) để mô hình hiểu được.

3. Đưa dữ liệu vào mô hình: Các chỉ số token cùng với thông tin mặt nạ chú ý (attention mask) được đưa vào mô hình BERT để xử lý.

4. Sinh đầu ra dạng vector: Mô hình trả về các vector đặc trưng cho từng token trong câu. Mỗi token được biểu diễn bằng một vector nhiều chiều mang thông tin ngữ nghĩa theo ngữ cảnh.

5. Trích xuất thông tin cần thiết: Tùy vào mục đích sử dụng, có thể lấy vector của từng token hoặc chỉ lấy vector tại vị trí [CLS] để đại diện cho toàn bộ câu, đặc biệt trong các bài toán phân loại văn bản.

### 4.4 Tại sao BERT-base-multilingual-cased phù hợp với đề tài này

BERT-base-multilingual-cased là lựa chọn lý tưởng cho phân tích cảm xúc của khách hàng về cơ sở y tế từ đánh giá trên Google Maps nhờ các tính năng nổi bật của nó. Mô hình hỗ trợ 104 ngôn ngữ, bao gồm cả tiếng Việt và tiếng Anh, giúp phân tích đánh giá từ khách hàng ở nhiều quốc gia khác nhau. Với khả năng phân biệt chữ hoa và chữ thường (cased), BERT giúp xử lý chính xác các ngôn ngữ như tiếng Việt. Cơ chế Transformer hai chiều của BERT cho phép hiểu ngữ cảnh của các câu đánh giá, giúp xác định chính xác cảm xúc tích cực, tiêu cực hay trung lập. Mô hình cũng có thể phân loại cảm xúc và nhận diện thực thể tên, giúp thu thập thông tin chi tiết từ các đánh giá dài và phức tạp. Từ đó, BERT-base-multilingual-cased giúp cung cấp cái nhìn sâu sắc về cảm nhận của khách hàng đối với cơ sở y tế.

## TÀI LIỆU THAM KHẢO


[1] https://huggingface.co/google-bert/bert-base-multilingual-cased

[2] https://pypi.org/project/beautifulsoup4/

[3] https://pypi.org/project/selenium/

[4] https://viblo.asia/p/tien-xu-li-du-lieu-van-ban-voi-nltk-Az45b0LgZxY

[5] https://www.nltk.org/howto/wordnet.html

[6] https://omwn.org/omw1.html