ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

ĐỒ ÁN MÔN HỌC
XỬ LÝ NGÔN NGỮ TỰ NHIÊN

ĐÁNH GIÁ HIỆU SUẤT MÔ HÌNH TÓM TẮT VĂN BẢN DU LỊCH

Giảng viên hướng dẫn: TS. Nguyễn Trọng Chỉnh
Lớp: CS221.Q11
Nhóm sinh viên thực hiện:

1. **Trần Ngọc Danh – 22520200**
2. **Phạm Huỳnh Tấn Khang – 22520624**
3. **Huỳnh Ngọc Trang – 22521510**

TP. HCM, tháng 1 năm 2026

LỜI CẢM ƠN

Nhóm xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đến Quý Thầy Cô Trường Đại học Công nghệ Thông tin – Đại học Quốc Gia TP.HCM, đặc biệt là Quý Thầy Cô khoa Khoa học máy tính, những người đã truyền đạt cho chúng em những kiến thức nền tảng quý báu và tạo điều kiện thuận lợi để nhóm có thể hoàn thành tốt đồ án môn học Xử lý ngôn ngữ tự nhiên.

Đặc biệt, nhóm xin bày tỏ lòng biết ơn sâu sắc đến Thầy Nguyễn Trọng Chỉnh – giảng viên môn Xử lý ngôn ngữ tự nhiên, người đã tận tình hướng dẫn, hỗ trợ và định hướng cho nhóm trong suốt quá trình thực hiện đề tài "Đánh giá hiệu suất Mô hình tóm tắt văn bản du lịch". Nhờ vào sự chỉ bảo và những kiến thức chuyên môn quý giá từ Thầy, nhóm đã tiếp thu được nhiều kiến thức bổ ích và có thể vận dụng một cách hiệu quả trong quá trình thực hiện đồ án.

Dựa trên những kiến thức được học trên lớp cùng với sự chủ động trong việc tìm hiểu thêm các công cụ và phương pháp hiện đại, nhóm đã nỗ lực hoàn thành đồ án với tinh thần trách nhiệm cao. Trong thời gian thực hiện, nhóm không ngừng học hỏi, nghiên cứu và ứng dụng những kiến thức mới để đạt được kết quả tốt nhất.

Tuy nhiên, do thời gian có hạn và kinh nghiệm thực tiễn còn hạn chế, nhóm không tránh khỏi những thiếu sót trong quá trình triển khai và viết báo cáo. Nhóm rất mong nhận được những ý kiến đóng góp từ Thầy để hoàn thiện hơn về mặt kiến thức, kỹ năng và chuẩn bị tốt hơn cho các đề tài nghiên cứu trong tương lai.

Một lần nữa, nhóm xin chân thành cảm ơn Quý Thầy Cô!

NHẬN XÉT CỦA GIẢNG VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Giải nghĩa
CLS	Classification Token (Token đặc biệt đầu câu trong BERT dùng cho bài toán phân loại)
CPU	Central Processing Unit (Bộ xử lý trung tâm).
GPU	Graphics Processing Unit (Bộ xử lý đồ họa - dùng để tăng tốc huấn luyện AI).
HTML	HyperText Markup Language (Ngôn ngữ đánh dấu siêu văn bản - thường là nhiễu cần lọc bỏ)
LLMs	Large Language Models (Mô hình ngôn ngữ lớn)
LR	Logistic Regression (Hồi quy Logistic)
ML	Machine Learning (Học máy)
NER	Named Entity Recognition (Nhận diện thực thể định danh)
NLP	Natural Language Processing (Xử lý ngôn ngữ tự nhiên)
ROUGE	Recall-Oriented Understudy for Gisting Evaluation (Độ đo đánh giá chất lượng tóm tắt văn bản).
TF-IDF	Term Frequency - Inverse Document Frequency (Kỹ thuật thống kê đánh giá trọng số của từ)

TÓM TẮT ĐỒ ÁN

Trong bối cảnh bùng nổ thông tin du lịch trực tuyến, nhu cầu tự động tóm tắt các bài viết để nắm bắt nhanh thông tin cốt lõi (địa điểm, chi phí, thời gian) trở nên cấp thiết. Tuy nhiên, các mô hình tóm tắt tạo sinh (Abstractive) hiện đại thường gặp khó khăn khi triển khai trên tập dữ liệu nhỏ, dẫn đến hiện tượng "ảo giác" – sinh ra thông tin sai lệch, điều tối kỵ trong miền dữ liệu du lịch. Đồ án này tập trung giải quyết bài toán tóm tắt tin tức du lịch tiếng Việt với ràng buộc kép: dữ liệu huấn luyện hạn chế (100 mẫu) và yêu cầu độ trung thực tuyệt đối.

Nghiên cứu tiến hành xây dựng một bộ dữ liệu chuẩn hóa, qua đó phân tích và xác nhận đặc thù cấu trúc "Tháp ngược" (Inverted Pyramid) của báo chí du lịch. Dựa trên cơ sở này, chúng tôi đề xuất kiến trúc tóm tắt trích xuất lai (Hybrid Extractive) kết hợp giữa sức mạnh hiểu ngữ nghĩa của PhoBERT (sử dụng kỹ thuật Mean Pooling) và sự ổn định của thuật toán Hồi quy Logistic (Logistic Regression) được cài đặt thủ công. Hệ thống sử dụng các đặc trưng ngữ nghĩa kết hợp với đặc trưng cấu trúc (vị trí câu) để xếp hạng và chọn lọc thông tin.

Các thực nghiệm được triển khai so sánh giữa mô hình đề xuất với các phương pháp Thống kê (TextRank, TF-IDF) và Học sâu (ViT5, BARTpho). Kết quả cho thấy, mặc dù mô hình tạo sinh ViT5 đạt điểm số định lượng cao nhất (ROUGE-1 ≈77.87%), nhưng lại mắc lỗi ảo giác nghiêm trọng về số liệu. Ngược lại, mô hình đề xuất PhoBERT + Logistic Regression đạt hiệu suất cân bằng (ROUGE-1 ≈71.71%) nhưng đảm bảo độ trung thực 100% . Nghiên cứu cũng chứng minh vai trò tiên quyết của đặc trưng "Vị trí câu" (Lead Bias) trong việc nhận diện thông tin quan trọng. Kết quả này khẳng định rằng với dữ liệu nhỏ và yêu cầu chính xác cao, phương pháp trích xuất kết hợp tri thức ngôn ngữ bản địa là giải pháp tối ưu hơn so với các mô hình sinh văn bản phức tạp.

Chương 1. GIỚI THIỆU

1.1. Đặt vấn đề

Trong kỷ nguyên số, du lịch trực tuyến đang phát triển bùng nổ, kéo theo sự gia tăng chóng mặt của lượng thông tin trên các nền tảng báo chí và mạng xã hội. Du khách ngày nay đối mặt với tình trạng "quá tải thông tin" khi phải xử lý hàng trăm bài viết, bài đánh giá và gợi ý lịch trình để tìm kiếm các dữ liệu cốt lõi như địa điểm, chi phí và thời gian. Nhu cầu về một công cụ có khả năng tự động tóm tắt nội dung, chắt lọc ý chính một cách nhanh chóng và chính xác là vô cùng cấp thiết.

Tuy nhiên, các mô hình ngôn ngữ lớn (LLMs) hiện đại dù có khả năng viết văn mượt mà nhưng thường mắc phải lỗi "ảo giác" (Hallucination) – tự sinh ra các thông tin sai lệch không có trong văn bản gốc. Trong miền dữ liệu du lịch, việc sai lệch một con số (ví dụ: giá vé 50.000đ thành 500.000đ) hay sai tên địa danh là điều tối kỵ, làm mất hoàn toàn giá trị sử dụng của bản tóm tắt.

Xuất phát từ thực tế trên, đề tài "Đánh giá hiệu suất mô hình tóm tắt văn bản du lịch" được lựa chọn thực hiện nhằm giải quyết bài toán tóm tắt tự động trên miền dữ liệu tiếng Việt đặc thù. Nghiên cứu tập trung vào việc tìm kiếm một giải pháp cân bằng giữa hiệu năng và tài nguyên, đặc biệt trong bối cảnh dữ liệu huấn luyện hạn chế nhưng yêu cầu độ trung thực tuyệt đối. Đề tài không chỉ so sánh các phương pháp học máy truyền thống và học sâu hiện đại, mà còn đề xuất một hướng tiếp cận lai (Hybrid) để tối ưu hóa bài toán này.

1.2. Mục tiêu nghiên cứu

Mục tiêu chung hướng đến xây dựng, triển khai và đánh giá hiệu quả của các kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP) từ cơ bản đến nâng cao cho bài toán tóm tắt tin tức du lịch tiếng Việt. Từ đó, xác định được mô hình tối ưu nhất phù hợp với điều kiện dữ liệu nhỏ và yêu cầu cao về bảo toàn thông tin định lượng.

Để đạt được mục tiêu đề ra, đề tài tập trung đi tìm lời giải cho 3 câu hỏi cốt lõi, tương ứng với các quyết định kỹ thuật trong quá trình thực hiện:

1. Với tập dữ liệu nhỏ, phương pháp *Trích xuất (Extractive)* hay *Tóm lược (Abstractive)* mang lại độ chính xác thông tin cao hơn?
2. Các đặc trưng ngôn ngữ học (cụ thể là *vị trí câu - lead bias*) đóng vai trò như thế nào trong việc nhận diện thông tin quan trọng của tin tức du lịch?
3. Việc kết hợp *embedding ngữ nghĩa tiền huấn luyện (PhoBERT)* với *mô hình phân loại tuyến tính (Logistic Regression)* ảnh hưởng như thế nào đến hiệu suất tóm tắt trích xuất trên dữ liệu du lịch tiếng Việt?

1.3. Phạm vi nghiên cứu

- *Dữ liệu:* Tập trung vào các bài báo tin tức du lịch tiếng Việt với các thể loại chính: tin tức sự kiện, bài trải nghiệm, bài gợi ý.
- *Phương pháp:* Giới hạn so sánh trong 5 mô hình đại diện: TextRank (Unsupervised), TF-IDF + Logistic Regression (Classic ML), PhoBERT + Logistic Regression (Hybrid), ViT5 và BARTpho (Deep Learning).
- *Đầu ra:* Bản tóm tắt văn bản thuần túy, không bao gồm tóm tắt đa phương tiện.

1.4. Yêu cầu kĩ thuật

Hệ thống tóm tắt cần đáp ứng các tiêu chuẩn kỹ thuật sau:

- *Độ trung thực:* Phải bảo toàn 100% các thông tin định lượng (Giá tiền, Ngày tháng) và định danh (Tên địa điểm). Tuyệt đối không chấp nhận hiện tượng sinh tin sai lệch.
- *Khả năng xử lý tiếng Việt:* Mô hình phải hiểu được ngữ nghĩa của từ ghép, từ đồng nghĩa và cấu trúc câu phức trong tiếng Việt (sử dụng các bộ Pre-trained model bản địa như PhoBERT).
- *Hiệu quả tài nguyên:* Mô hình cần có kiến trúc gọn nhẹ, khả năng huấn luyện ổn định trên tập dữ liệu nhỏ, tránh hiện tượng quá khớp (Overfitting) và có tốc độ phản hồi nhanh.

1.5. Thách thức

Trong quá trình thực hiện, nghiên cứu đối mặt với các thách thức chính:

- *Ràng buộc về dữ liệu:* Số lượng mẫu huấn luyện ít khiến các mô hình Deep Learning khó hội tụ và dễ học vẹt. Việc làm sạch dữ liệu (loại bỏ chú thích ảnh, nhiễu HTML) đòi hỏi quy trình xử lý kỹ lưỡng.
- *Đặc thù ngôn ngữ:* Tiếng Việt có tính đa nghĩa và cấu trúc linh hoạt. Các bài báo du lịch thường sử dụng văn phong miêu tả cảm xúc, gây khó khăn cho việc xác định câu chứa thông tin cốt lõi.
- *Vấn đề đánh giá:* Các độ đo tự động (như ROUGE) chủ yếu dựa trên sự trùng lặp từ vựng, chưa phản ánh hết được tính đúng đắn về mặt ngữ nghĩa và thông tin. Do đó, cần kết hợp đánh giá định tính bởi con người.

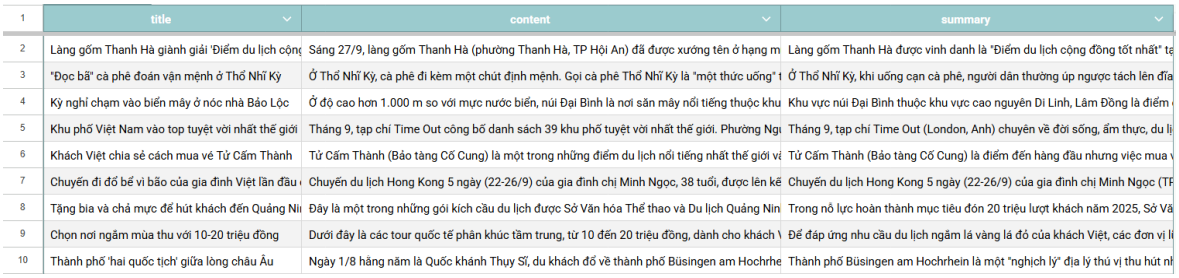
Chương 2. DỮ LIỆU

2.1. Tổng quan bộ dữ liệu

2.1.1. Nguồn gốc

Bộ dữ liệu được xây dựng dựa trên các bài báo du lịch tiếng Việt được thu thập từ chuyên mục Du lịch của báo điện tử VnExpress. Việc lựa chọn nguồn dữ liệu này đảm bảo ba tiêu chí:

- *Chất lượng ngôn ngữ:* Văn phong chuẩn mực, đúng ngữ pháp tiếng Việt.
- *Độ tin cậy:* Thông tin về địa danh, giá cả, lịch trình có độ chính xác cao.
- *Tính đa dạng:* Bao gồm nhiều thể loại như tin tức sự kiện, bài trải nghiệm, bài gợi ý lịch trình.



	title	content	summary
2	Làng gốm Thanh Hà giành giải 'Điểm du lịch cộn	Sáng 27/9, làng gốm Thanh Hà (phường Thanh Hà, TP Hội An) đã được xướng tên ở hạng m	Làng gốm Thanh Hà được vinh danh là "Điểm du lịch cộng đồng tốt nhất" tạ
3	"Đọc bã" cà phê đoán vận mệnh ở Thổ Nhĩ Kỳ	Ở Thổ Nhĩ Kỳ, cà phê đi kèm một chút định mệnh. Gọi cà phê Thổ Nhĩ Kỳ là "một thức uống" t	Ở Thổ Nhĩ Kỳ, khi uống cạn cà phê, người dân thường úp ngược tách lên đĩa
4	Kỳ nghỉ chạm vào biển mây ở nóc nhà Bảo Lộc	Ở độ cao hơn 1.000 m so với mực nước biển, núi Đại Bình là nơi săn mây nổi tiếng thuộc khu	Khu vực núi Đại Bình thuộc khu vực cao nguyên Di Linh, Lâm Đồng là điểm
5	Khu phố Việt Nam vào top tuyệt vời nhất thế giới	Tháng 9, tạp chí Time Out công bố danh sách 39 khu phố tuyệt vời nhất thế giới. Phường Ngu	Tháng 9, tạp chí Time Out (London, Anh) chuyên về đời sống, ẩm thực, du lị
6	Khách Việt chia sẻ cách mua vé Tử Cấm Thành	Tử Cấm Thành (Bảo tàng Cố Cung) là một trong những điểm du lịch nổi tiếng nhất thế giới và	Tử Cấm Thành (Bảo tàng Cố Cung) là điểm đến hàng đầu nhưng việc mua v
7	Chuyến đi đổ bể vì bão của gia đình Việt lần đầu	Chuyến du lịch Hong Kong 5 ngày (22-26/9) của gia đình chị Minh Ngọc, 38 tuổi, được lên kế	Chuyến du lịch Hong Kong 5 ngày (22-26/9) của gia đình chị Minh Ngọc (TP
8	Tặng bia và chả mực để hút khách đến Quảng Nì	Đây là một trong những gói kích cầu du lịch được Sở Văn hóa Thể thao và Du lịch Quảng Nin	Trong nỗ lực hoàn thành mục tiêu đón 20 triệu lượt khách năm 2025, Sở Vă
9	Chọn nơi ngắm mùa thu với 10-20 triệu đồng	Dưới đây là các tour quốc tế phân khúc tầm trung, từ 10 đến 20 triệu đồng, dành cho khách V	Để đáp ứng nhu cầu du lịch ngắm lá vàng lá đỏ của khách Việt, các đơn vị lữ
10	Thành phố 'hai quốc tịch' giữa lòng châu Âu	Ngày 1/8 hằng năm là Quốc khánh Thụy Sĩ, du khách đổ về thành phố Büsingen am Hochrhe	Thành phố Büsingen am Hochrhein là một "nghịch lý" địa lý thú vị thu hút nh

Hình 2.1: Dữ liệu các bài báo

2.1.2. Quy mô và cấu trúc

Trong bối cảnh nguồn lực gán nhãn thủ công hạn chế, đồ án tập trung giải quyết bài toán trên quy mô dữ liệu nhỏ chất lượng cao. Bộ dữ liệu bao gồm 100 cặp mẫu, mỗi mẫu gồm hai thành phần chính:

- *Văn bản gốc (Input):* Nội dung toàn văn của bài báo.
- *Tóm tắt mẫu (Target):* Bản tóm tắt do biên tập viên thực hiện (thường là phần Sapo), đảm bảo cô đọng đầy đủ ý chính.

Bảng 2.1: Ví dụ mẫu dữ liệu trong tập dataset

title	content	summary
Hà Nội vào top những	Hà Nội xếp thứ 5 trong danh sách các điểm đến ngắm lá mùa thu đẹp nhất châu Á. Tạp chí Anh miêu tả thủ đô của Việt Nam "có một mùa thu rất riêng" so với những điểm đến khác trong châu lục. Hà Nội chưa phải điểm đến quen thuộc với khách quốc tế để ngắm mùa thu. Tạp chí này gợi ý du khách nên dạo	Thủ đô Hà Nội vừa được tạp chí Time Out (Anh) xếp vị trí thứ 5 trong danh sách các điểm đến ngắm lá mùa thu đẹp nhất châu Á. Khác với vẻ

điểm đến mùa thu đẹp nhất châu Á	quanh hồ Hoàn Kiếm để chiêm ngưỡng sắc lá đỏ, vàng đặc trưng. Đây cũng là mùa thu hoạch ở miền Bắc, trùng dịp Tết Trung thu, mang đến nhiều trải nghiệm văn hóa đặc sắc. Ngoài ngắm cảnh, du khách có thể hòa mình vào lễ hội với âm nhạc truyền thống, múa lân và thưởng thức bánh trung thu nhân hạt sen, trứng muối, đậu xanh. "Khi màn đêm buông xuống, phố phường sáng bừng bởi muôn kiểu đèn lồng, tạo nên khung cảnh lung linh, cuốn hút", tạp chí của Anh mô tả. Ấn phẩm này cũng gợi ý du khách tháng 10 là thời điểm đẹp nhất để trải nghiệm mùa thu Hà Nội. Ngoài Thủ đô của Việt Nam, 6 cái tên còn lại xuất hiện trong danh sách còn có Ibaraki và Kyoto (Nhật Bản); đảo Nami (Hàn Quốc); khu thắng cảnh quốc gia Alishan (Đài Loan, Trung Quốc); công viên rừng quốc gia Trương Gia Giới và Vườn quốc gia Thung lũng Cửu Trại Câu (Trung Quốc). Time Out là tạp chí chuyên về đời sống, ẩm thực, du lịch, nghệ thuật và giải trí. Được thành lập năm 1968 ở London (Anh), tạp chí hiện hoạt động như một chuyên trang số, cung cấp gợi ý về nhà hàng, sự kiện và trải nghiệm độc đáo, trở thành nguồn tham khảo đáng tin cậy cho du khách lẫn người dân địa phương.	đẹp quen thuộc của các nước ôn đới, mùa thu Hà Nội được mô tả là "rất riêng" và chưa được nhiều khách quốc tế biết tới. Tạp chí gợi ý du khách nên đến vào tháng 10 để dạo bước quanh hồ Hoàn Kiếm ngắm sắc lá vàng đỏ và hòa mình vào không khí Tết Trung thu. Đây là dịp để trải nghiệm văn hóa độc đáo với múa lân, đèn lồng rực rỡ và thưởng thức các loại bánh trung thu truyền thống. Ngoài Hà Nội, danh sách còn có những cái tên lừng danh như Kyoto (Nhật Bản), đảo Nami (Hàn Quốc) và Cửu Trại Câu (Trung Quốc).

2.2. Tiền xử lý dữ liệu

Để đảm bảo mô hình học được các đặc trưng ngôn ngữ thực sự thay vì "học mẹo" từ nhiễu, quy trình làm sạch dữ liệu được thực hiện qua các bước sau.

2.2.1. Loại bỏ nhiễu

- *Thẻ HTML & Metadata:* Loại bỏ toàn bộ các thẻ định dạng, icon, đường dẫn quảng cáo.
- *Thông tin rác:* Xóa bỏ ngày đăng, tên tác giả, nguồn bài để tránh mô hình nhầm lẫn đây là nội dung chính.

2.2.2. Xử lý chú thích ảnh

Trong báo chí du lịch, các dòng chú thích dưới ảnh thường chứa nội dung tóm tắt ngắn gọn của đoạn văn đó. Nếu giữ lại, mô hình trích xuất sẽ có xu hướng "gian lận" bằng cách chỉ chọn các câu caption này mà không cần hiểu ngữ cảnh bài báo (hiện tượng Data Leakage).

Giải pháp: *Sử dụng biểu thức chính quy (Regex) và luật vị trí để nhận diện và xóa bỏ hoàn toàn các dòng chú thích ảnh.*

2.2.3. Chuẩn hóa ngôn ngữ

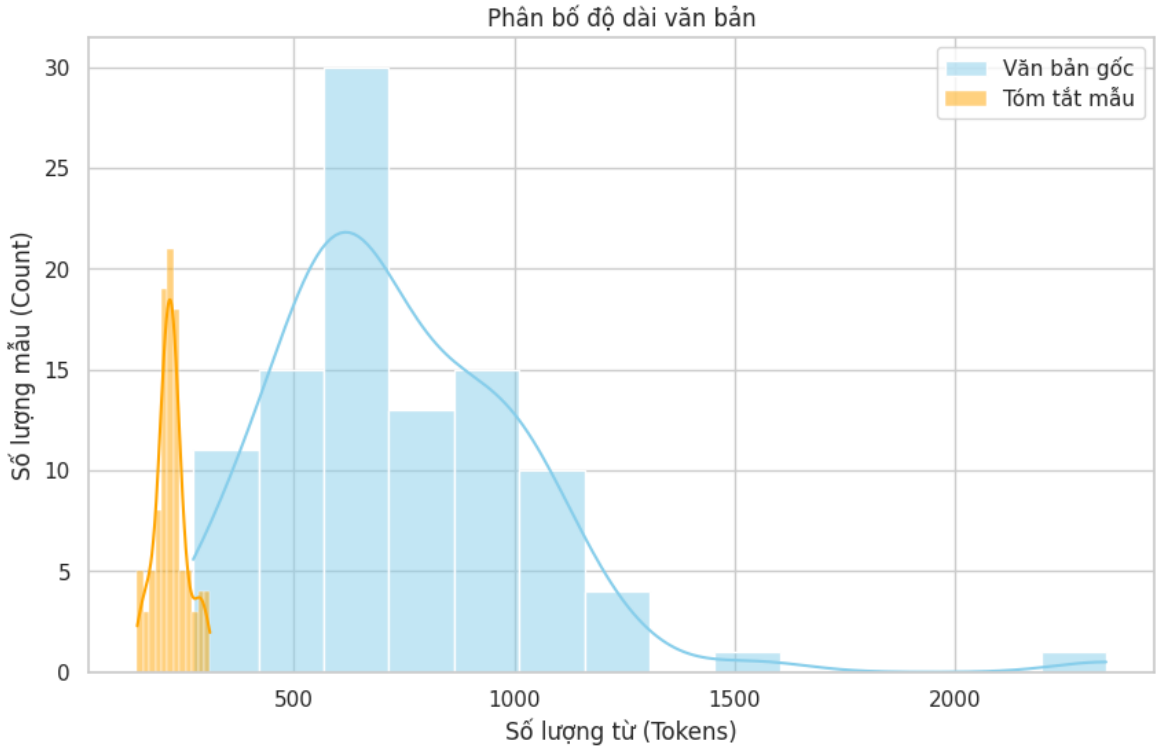
- *Tách câu (Sentence Segmentation):* Sử dụng thư viện Underthesea để tách đoạn văn thành danh sách các câu đơn. Đây là đơn vị xử lý cơ bản cho nhóm phương pháp Trích xuất.
- *Tách từ (Word Segmentation):* Áp dụng cho nhóm mô hình Deep Learning (BARTpho) để mô hình hiểu được các từ ghép tiếng Việt.
  Ví dụ: "đất_nước", "khách_sạn" thay vì "đất", "nước", "khách", "sạn".

2.3. Phân tích đặc điểm dữ liệu

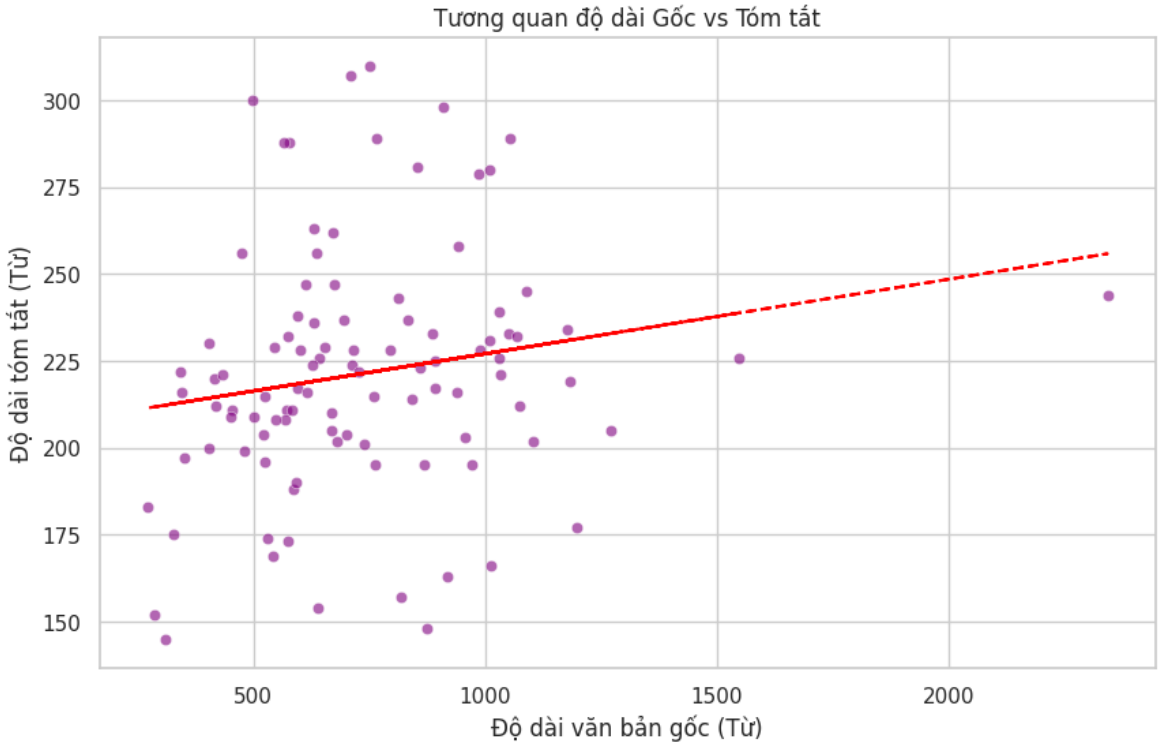
2.3.1. Thống kê độ dài

Kết quả thống kê mô tả trên 100 mẫu dữ liệu cho thấy:

- *Độ dài trung bình bài gốc:* ≈ 500 - 800 từ/bài.
- *Độ dài trung bình tóm tắt:* ≈ 50 - 80 từ/bài.
- *Tỷ lệ nén:* Trung bình 1/10. Tức là bản tóm tắt chỉ giữ lại khoảng 10% dung lượng thông tin cô đọng nhất.



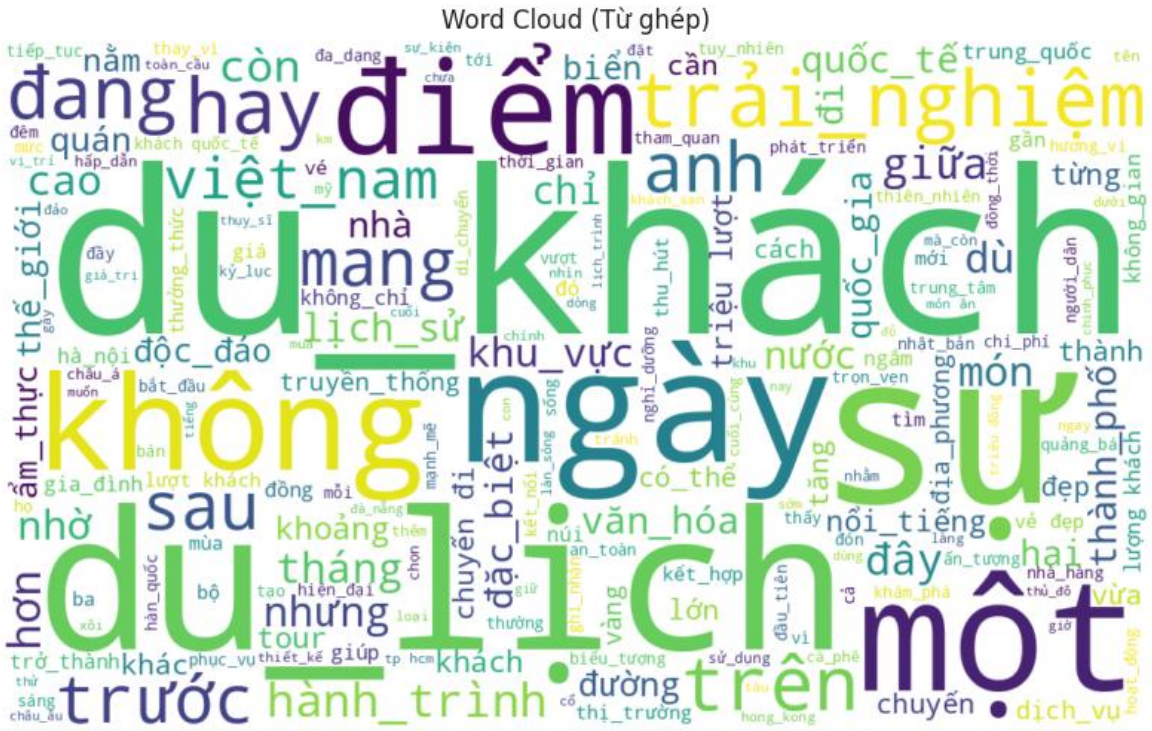
Hình 2.2: Biểu đồ phân bố độ dài văn bản



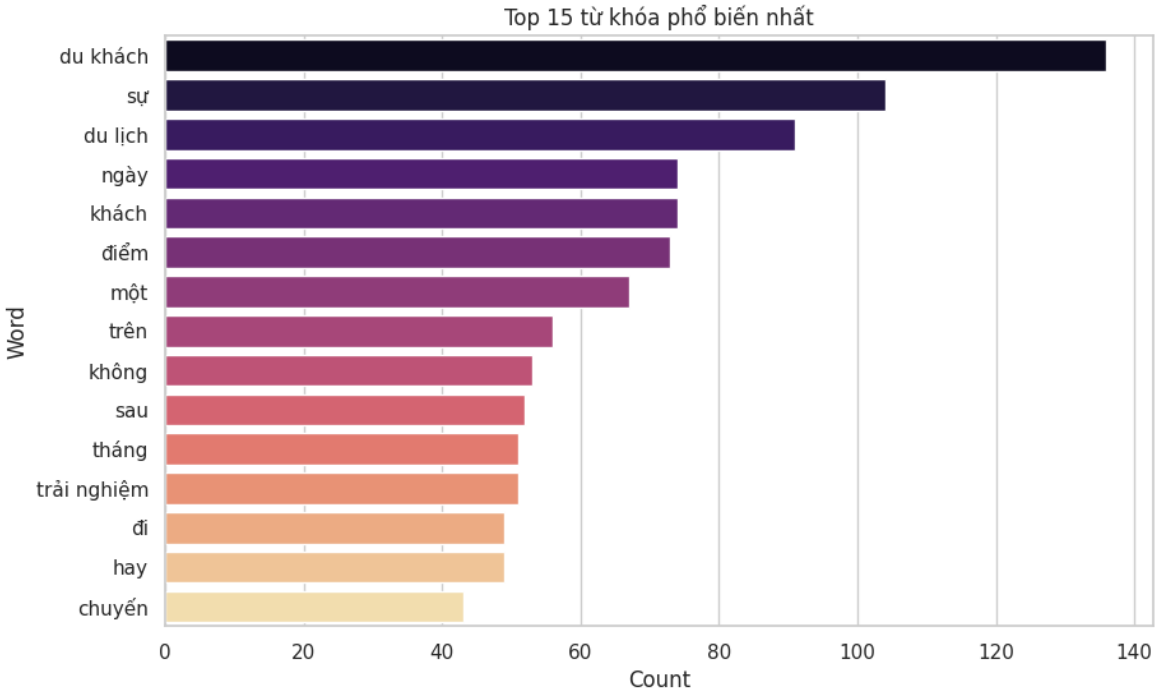
Hình 2.3: Tương quan độ dài văn bản gốc và tóm tắt

2.3.2. Phân bố từ vựng và chủ đề

Để hiểu rõ ngữ cảnh của tập dữ liệu, nghiên cứu đã thực hiện thống kê tần suất từ vựng sau khi loại bỏ các hư từ (stopwords) và chuẩn hóa từ ghép bằng thư viện Underthesea.



Hình 2.4: Tần suất xuất hiện của các từ ghép du lịch



Hình 2.5: Top các từ xuất hiện nhiều nhất

Kết quả trực quan hóa bằng đám mây từ (Word Cloud) và biểu đồ tần suất (Bar Chart) cho thấy:

- ***Các từ khóa nổi bật:*** Tập trung dày đặc vào lĩnh vực du lịch như *"du khách", "trải nghiệm", "hành trình", "check-in", "địa điểm"*.
- ***Tính nhất quán:*** Sự xuất hiện lặp lại của các từ khóa này xác nhận bộ dữ liệu đã được thu thập đúng miền, không bị lẫn tạp các tin tức xã hội hay pháp luật khác.

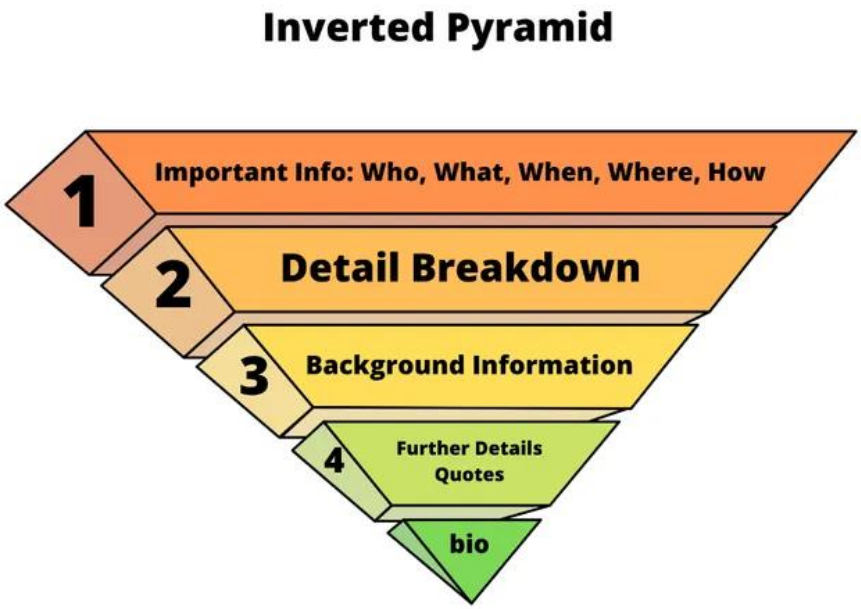
2.3.3. Cấu trúc tháp ngược và hiện tượng lead bias

Phân tích sự trùng khớp thông tin giữa tóm tắt mẫu và bài gốc cho thấy:

- Hơn **80%** thông tin trong bản tóm tắt mẫu được lấy từ **20%** nội dung đầu tiên của bài báo (thường là đoạn Sapo và đoạn thứ nhất).
- Các đoạn sau thường đi sâu vào chi tiết bổ trợ hoặc trải nghiệm cá nhân.

Sự phân bố này được minh họa cụ thể qua hình sau, trong đó tầng quan trọng nhất (5W1H) luôn nằm ở đỉnh tháp.

*Từ phân tích trên, có thể nhận định rằng dữ liệu tuân thủ chặt chẽ cấu trúc **tháp ngược**. Điều này gợi ý rằng đặc trưng **vị trí câu** sẽ đóng vai trò cực kỳ quan trọng trong việc định hướng cho mô hình tóm tắt.*



Hình 2.6:Sơ đồ phân bố thông tin theo cấu trúc tháp ngược

2.3.4. Đặc điểm về thực thể

Tin tức du lịch có mật độ thực thể định danh rất cao, đặc biệt là:

- *Địa danh:* Tên thành phố, điểm tham quan.
- *Định lượng:* Giá tiền, ngày tháng, thời gian mở cửa.

Yêu cầu đặt ra cho bài toán này là độ trung thực tuyệt đối. Việc sai lệch một con số (ví dụ: giá tour 5 triệu thành 500k) là không thể chấp nhận. Đây là thách thức lớn đối với các mô hình Tạo sinh (Abstractive) vốn dễ mắc lỗi ảo giác.

2.4. Phân chia dữ liệu

Để đảm bảo tính khách quan trong đánh giá thực nghiệm, bộ dữ liệu được chia theo tỷ lệ **80/20**:

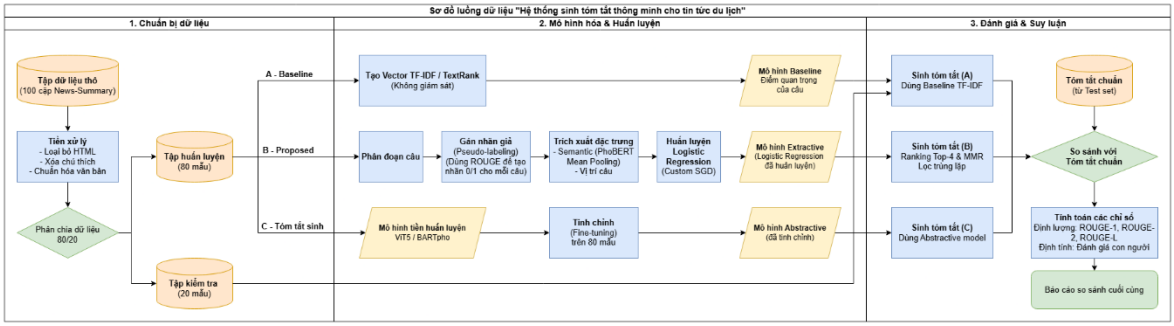
- *Tập huấn luyện (Training Set - 80 bài):* Dùng để tính toán thống kê TF-IDF, huấn luyện Logistic Regression và Fine-tune Transformers.
- *Tập kiểm thử (Test Set - 20 bài):* Dữ liệu chưa từng thấy, được cách ly hoàn toàn để đảm bảo tính khách quan khi đánh giá kết quả cuối cùng.

Chương 3. PHƯƠNG PHÁP SỬ DỤNG

3.1. Tổng quan kiến trúc hệ thống

Để đánh giá toàn diện hiệu suất tóm tắt trên tập dữ liệu nhỏ, nghiên cứu xây dựng một quy trình xử lý thống nhất gồm ba giai đoạn chính:

- Tiền xử lý & Gán nhãn: Chuẩn hóa dữ liệu đầu vào và tạo nhãn giả cho bài toán học máy có giám sát.
- Mô hình hóa: Triển khai song song hai nhánh tiếp cận:
    - Nhánh trích xuất: Sử dụng các mô hình từ cơ bản (TextRank, TF-IDF) đến nâng cao (PhoBERT + Logistic Regression).
    - Nhánh tạo sinh: Sử dụng các mô hình Deep Learning tiên tiến (ViT5, BARTpho).
- Suy luận & Đánh giá: Sinh bản tóm tắt và đánh giá dựa trên các độ đo định lượng (ROUGE) và định tính (Human Evaluation).



Hình 3.1: Sơ đồ luồng dữ liệu của hệ thống

3.2. Cơ sở lý thuyết

3.2.1. Tóm tắt trích xuất và bài toán xếp hạng câu

Tóm tắt trích xuất hoạt động dựa trên nguyên lý chọn lọc các câu quan trọng nhất từ văn bản gốc và ghép lại thành bản tóm tắt mà không thay đổi nội dung câu.

Về mặt toán học, bài toán này được mô hình hóa thành bài toán *Phân loại nhị phân (Binary Classification)* hoặc *Xếp hạng câu (Sentence Ranking)*. Với một văn bản D gồm n câu $S = \{s_1, s_2, \dots, s_n\}$, mục tiêu là gán cho mỗi câu s_i một nhãn $y_i \in \{0,1\}$ (1 là chọn, 0 là bỏ) hoặc một điểm số $score(s_i)$ để chọn ra k câu có điểm cao nhất.

3.2.2. Biểu diễn văn bản và PhoBERT

Để máy tính xử lý được ngôn ngữ, văn bản được chuyển đổi thành các vector số học.

- *Hạn chế của phương pháp truyền thống (TF-IDF):* TF-IDF (Term Frequency - Inverse Document Frequency) chỉ dựa trên tần suất xuất hiện của từ. Nó coi các từ là độc lập và không nắm bắt được ngữ nghĩa (ví dụ: không hiểu mối liên hệ giữa "khách sạn" và "nơi lưu trú").
- *Mô hình PhoBERT:* PhoBERT là mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer (RoBERTa), được huấn luyện trên 20GB dữ liệu văn bản tiếng Việt. Nhờ cơ chế Attention, PhoBERT tạo ra các vector ngữ cảnh hóa, giúp máy hiểu được nghĩa của từ dựa trên ngữ cảnh xung quanh và giải quyết tốt các vấn đề từ đồng nghĩa, đa nghĩa trong tiếng Việt.

3.2.3. Hồi quy Logistic (Logistic Regression)

Logistic Regression là một thuật toán học máy tuyến tính dùng cho bài toán phân loại. Dù đơn giản, nó đặc biệt hiệu quả với các tập dữ liệu nhỏ nhờ khả năng kiểm soát tốt hiện tượng quá khớp (Overfitting). Hàm giả thuyết của Logistic Regression sử dụng hàm Sigmoid để đưa đầu ra về khoảng (0, 1), biểu thị xác suất:

$$P(y = 1 \mid x) = \sigma(W^T x + b) = \frac{1}{1 + e^{-(W^T x + b)}}$$

Trong đó: x là vector đặc trưng đầu vào, W là bộ trọng số cần học, b là hệ số bias.

3.3. Phương pháp đề xuất: PhoBERT kết hợp Logistic Regression

Đây là mô hình trọng tâm của đồ án, được thiết kế theo kiến trúc lai nhằm tận dụng sức mạnh hiểu ngữ nghĩa của Deep Learning (PhoBERT) và sự ổn định của Machine Learning (Logistic Regression).

3.3.1. Kỹ thuật gán nhãn giả

Do tập dữ liệu gốc chỉ có cặp {Văn bản, Tóm tắt mẫu} mà không có nhãn chi tiết cho từng câu, nghiên cứu sử dụng thuật toán tham lam (Greedy Algorithm) dựa trên độ đo ROUGE để tự động sinh nhãn huấn luyện:

- *Input:* Danh sách câu trong văn bản gốc và văn bản tóm tắt mẫu.

- *Quy trình:* Tính điểm ROUGE-L của từng câu gốc so với tóm tắt mẫu. Câu nào có độ trùng khớp cao nhất (vượt ngưỡng quy định) sẽ được gán nhãn **1** (Quan trọng), các câu còn lại gán nhãn **0**.
- *Output:* Tập dữ liệu huấn luyện có nhãn từng câu để đưa vào mô hình Logistic Regression.

3.3.2. Trích xuất đặc trưng

Mỗi câu văn s_i được chuyển đổi thành một vector đặc trưng tổng hợp v_{input} kết hợp giữa ba thành phần:

Đặc trưng ngữ nghĩa:

- Sử dụng PhoBERT-base để mã hóa câu văn.
- *Kỹ thuật Mean Pooling:* Thay vì lấy vector của token đầu tiên [CLS] (vốn hoạt động kém hiệu quả trên các tác vụ similarity khi chưa fine-tune), nghiên cứu sử dụng kỹ thuật Mean Pooling – tính trung bình cộng vector của tất cả các token trong câu ở lớp ẩn cuối cùng.
- *Kết quả:* Vector $v_{sem} \in \mathbb{R}^{768}$.

Đặc trưng cấu trúc:

- Dựa trên phân tích "Tháp ngược" tại Chương 2, vị trí câu là yếu tố dự báo quan trọng.
- *Position Embedding:* Tạo một giá trị vô hướng biểu thị vị trí tương đối của câu trong đoạn văn (được chuẩn hóa về khoảng [0, 1]). Các câu đầu bài (Sapo) sẽ có giá trị vị trí thấp, giúp mô hình học được hiện tượng Lead Bias.
- *Kết quả:* Giá trị $v_{pos} \in \mathbb{R}^1$.

Đặc trưng độ dài:

- Giá trị chuẩn hóa biểu thị độ dài câu (số từ), giúp mô hình ưu tiên các câu có lượng thông tin vừa đủ.
- Kết quả: $v_{len} \in R^1$

➢ ***Vector tổng hợp:*** Vector input cuối cùng cho mô hình phân loại là sự ghép nối của các đặc trưng trên: $v_{input} = [v_{sem}; v_{pos}; v_{len}] \in \mathbb{R}^{770}$

3.3.3. Huấn luyện mô hình

Mô hình Logistic Regression được cài đặt thủ công với cấu hình tối ưu hóa như sau:

- *Thuật toán:* Gradient Descent (Cập nhật trọng số ngược chiều đạo hàm).
- *Learning Rate*: $\eta = 0.1$
- *Số vòng lặp*: 3000 vòng.
- *Xử lý mất cân bằng*: Sử dụng tham số *class_weight = "balance"*, tự động điều chỉnh trọng số phạt trong hàm Loss để mô hình chú trọng vào lớp thiểu số (câu quan trọng).

3.3.4. Chiến lược suy luận

Trong giai đoạn kiểm thử, quy trình sinh tóm tắt được thực hiện như sau:

- *Dự đoán:* Tính xác suất $P(y = 1 \mid x)$ cho tất cả các câu trong bài báo mới.
- *Xếp hạng:* Sắp xếp các câu theo thứ tự xác suất giảm dần và chọn ra Top-4 câu có điểm số cao nhất.
- *Lọc trùng lặp*: Duyệt qua danh sách Top-4. Nếu một câu có độ tương đồng Cosine với các câu đã chọn lớn hơn ngưỡng 0.8, câu đó sẽ bị loại bỏ để tránh dư thừa thông tin.
- *Sắp xếp lại:* Các câu được chọn sẽ được sắp xếp lại theo thứ tự xuất hiện trong văn bản gốc để đảm bảo tính mạch lạc.

3.4. Các phương pháp đối chứng

Để đánh giá khách quan hiệu quả của mô hình đề xuất, nghiên cứu triển khai thêm hai nhóm mô hình so sánh:

3.4.1. Nhóm Baseline (Thống kê)

- *TF-IDF Ranking:* Xếp hạng câu dựa trên tổng điểm TF-IDF của các từ trong câu.
- *TextRank:* Sử dụng lý thuyết đồ thị, coi mỗi câu là một đỉnh và độ tương đồng giữa các câu là cạnh. Áp dụng thuật toán PageRank để tìm ra các câu trung tâm quan trọng nhất. Đây là phương pháp không giám sát (Unsupervised).

3.4.2. Nhóm Tạo sinh (Abstractive - Deep Learning)

- *Mô hình:* Sử dụng ViT5 (VietAI) và BARTpho (VinAI). Đây là các mô hình Transformer kiến trúc Seq2Seq (Encoder-Decoder) tiên tiến nhất hiện nay cho tiếng Việt.
- *Cơ chế:* Fine-tuning (Tinh chỉnh) mô hình trên tập dữ liệu huấn luyện. Đầu vào là bài báo gốc, đầu ra là chuỗi văn bản tóm tắt mới.
- *Mục đích:* Kiểm chứng giả thuyết rằng với dữ liệu nhỏ, mô hình tạo sinh hiện đại dễ gặp hiện tượng "học vẹt" và "ảo giác" so với phương pháp trích xuất.

3.5. Phương pháp đánh giá

3.5.1. Đánh giá định lượng

Sử dụng bộ chỉ số ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – tiêu chuẩn vàng trong đánh giá tóm tắt văn bản:

- *ROUGE-1:* Đo độ trùng khớp của từ đơn (unigram) → Đánh giá độ bao phủ nội dung.
- *ROUGE-2:* Đo độ trùng khớp của cụm từ đôi (bigram) → Đánh giá độ trôi chảy cục bộ.
- *ROUGE-L:* Đo chuỗi con chung dài nhất → Đánh giá cấu trúc câu.

3.5.2. Đánh giá định tính

Do ROUGE chỉ dựa trên sự trùng lặp từ ngữ bề mặt, nghiên cứu thực hiện đánh giá thủ công trên 20 bài test dựa trên các tiêu chí:

- *Độ trung thực:* Bản tóm tắt có chứa thông tin sai lệch (sai giá tiền, sai địa danh) so với bài gốc hay không *(Đây là tiêu chí quan trọng nhất).*
- *Độ dư thừa:* Bản tóm tắt có chứa các câu lặp ý hay không.

Chương 4. THỰC NGHIỆM VÀ KẾT QUẢ

4.1. Cài đặt thực nghiệm

4.1.1. Môi trường

Để đảm bảo tính khả thi và khả năng tái lập kết quả, các thực nghiệm được triển khai trên nền tảng đám mây Google Colab với cấu hình:

- *Phần cứng:* CPU Intel Xeon (2 core @ 2.20GHz), GPU NVIDIA Tesla T4 (16GB VRAM), RAM 12GB.
- *Thư viện lập trình:* Python 3.10, PyTorch, Transformers (HuggingFace), VnCoreNLP, Scikit-learn, ROUGE-Score.

4.1.2. Tham số huấn luyện chi tiết

Các tham số được thiết lập tối ưu cho từng nhóm mô hình như sau:

Nhóm Baseline (TextRank & TF-IDF + LR)

- *TextRank:* Cửa sổ trượt (Window size) = 2 câu; Hàm tương đồng: Cosine Similarity.
- *TF-IDF:* N-gram range = (1, 2); Loại bỏ từ hiếm (min_df = 2).

Nhóm Abstractive (ViT5 & BARTpho)

- *Backbone:* vietai/vit5-base và vinai/bartpho-word-base.
- *Training:* Learning Rate = 2e-5; Batch Size = 4; Epochs = 5 (Dừng sớm - Early Stopping để tránh Overfitting trên dữ liệu nhỏ).
- *Generation:* Beam Search với num_beams = 2.

Mô hình đề xuất (PhoBERT + Custom LR)

- *Input Vector:* 770 chiều (768 ngữ nghĩa + 1 vị trí + 1 độ dài).
- *Logistic Regression:*
  - Learning Rate (η): 0.1
  - Số vòng lặp: 3000
  - Class Weight: “Balanced”
- *Tham số suy luận:* Top-K = 4 câu; Ngưỡng lọc trùng lặp = 0.8

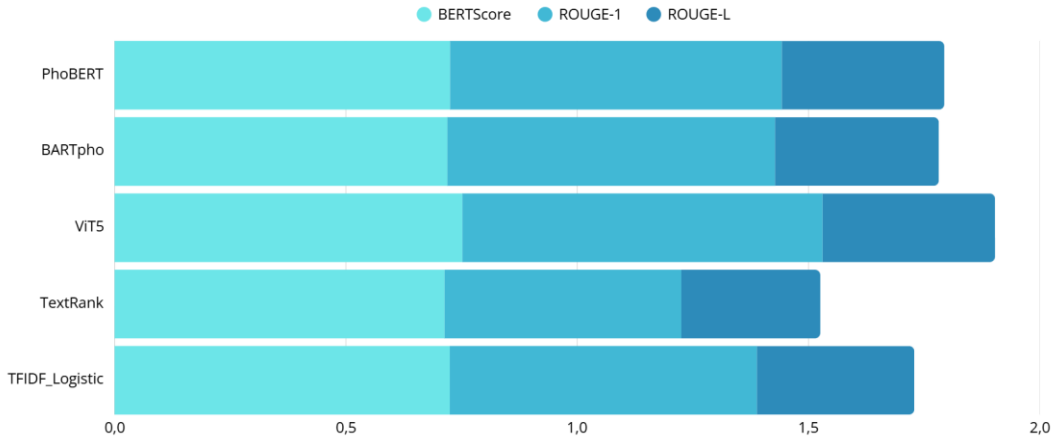
4.2. Kết quả định lượng

Kết quả đánh giá trên tập kiểm thử, sử dụng độ đo ROUGE được tổng hợp chi tiết dưới đây.

Bảng 4.1: Bảng so sánh hiệu năng ROUGE

Mô hình	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)
TextRank (Unsupervised)	51.22	28.95	30.08
TF-IDF + LR (Supervised)	66.40	35.97	34.04
PhoBERT + LR (Đề xuất)	**71.71**	**37.73**	**35.08**
BARTpho (Abstractive)	70.84	36.96	35.42
ViT5 (Abstractive)	**77.87**	**44.93**	**37.28**

Để có cái nhìn trực quan hơn về sự chênh lệch hiệu suất giữa các nhóm phương pháp, Hình 4.1 dưới đây biểu diễn so sánh điểm ROUGE-1 và ROUGE-L.



Hình 4.1: Biểu đồ so sánh điểm ROUGE giữa 5 mô hình thực nghiệm

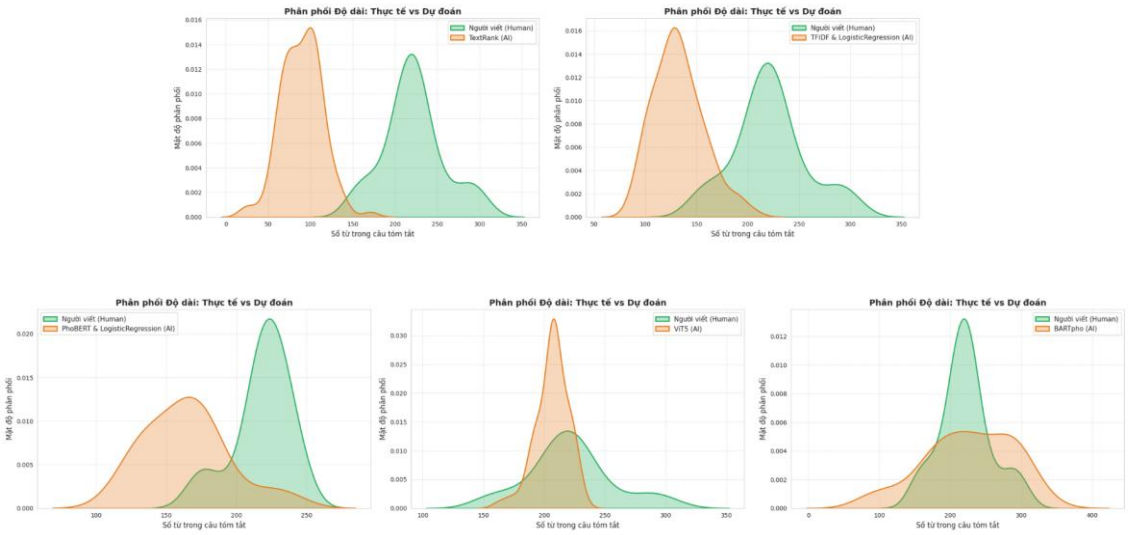
Nhận xét: *Quan sát Hình 4.1, có thể thấy rõ sự vượt trội của nhóm mô hình Deep Learning (ViT5, BARTpho, PhoBERT - các thanh màu sáng) so với nhóm thống kê truyền thống (TextRank, TF-IDF - các thanh màu tối). Trong đó, ViT5 đạt đỉnh cao nhất về độ khớp từ vựng.*

4.2.1. Kết quả phân tích đặc trưng đầu ra

Bên cạnh điểm ROUGE, nghiên cứu đi sâu phân tích đặc điểm hành vi của mô hình thông qua độ dài và vị trí thông tin.

- **Phân phối độ dài**

  Hình 4.2 so sánh mật độ phân phối độ dài bản tóm tắt do máy sinh ra (màu cam) so với con người viết (màu xanh).



Hình 4.2: So sánh phân phối độ dài tóm tắt giữa máy và người

Nhận xét: *Đường phân phối của AI (màu cam) thường hẹp và nhọn hơn, tập trung quanh mức 200 từ, cho thấy mô hình có xu hướng sinh ra các bản tóm tắt có độ dài ổn định, trong khi con người có độ biến thiên lớn hơn tùy theo cảm xúc.*

- **Vị trí thông tin**

Để kiểm chứng giả thuyết tháp ngược, Hình 4.3 thể hiện tần suất vị trí các câu được mô hình lựa chọn.

Hình 4.3: Phân bố vị trí câu được chọn bởi các mô hình

Nhận xét: *Quan sát Hình 4.3, các cột màu đỏ dồn trọng số rất cao vào khoảng vị trí đầu tiên (0.0 - 0.2), chứng minh các mô hình học máy hiện đại đều tự động phát hiện và khai thác triệt để cấu trúc Sapo của bài báo du lịch.*

Chương 5. PHÂN TÍCH VÀ BÀN LUẬN KẾT QUẢ

5.1. So sánh hiệu năng

5.1.1. Sự vượt trội về điểm số của Deep Learning (ViT5)

Dựa trên kết quả *Bảng 4.1*, ViT5 là mô hình đạt điểm số cao nhất trên cả 3 chỉ số *(ROUGE-1 ≈77.87%).* Điều này phản ánh khả năng vượt trội của kiến trúc Transformer trong việc học mô hình ngôn ngữ và sinh ra các từ ngữ khớp với văn bản mẫu. Với bài toán tóm tắt thông thường, ViT5 sẽ là người chiến thắng.

5.1.2. Vấn đề ảo giác trên dữ liệu nhỏ

Tuy nhiên, khi kiểm tra định tính trên nội dung du lịch, nhóm mô hình tạo sinh (ViT5/BARTpho) cho thấy nhược điểm rõ ràng: *lỗi ảo giác*

- *Nguyên nhân:* Do tập dữ liệu huấn luyện quá nhỏ (80 mẫu), mô hình Deep Learning bị overfitting. Nó học thuộc lòng cách viết câu văn trôi chảy nhưng không học được cách liên kết các con số thực tế.
- *Hậu quả:* Mô hình tự sinh ra các thông tin sai lệch.
  Ví dụ: Bài gốc nói "Vé 50.000đ", ViT5 có thể sinh ra "Vé 500.000đ" hoặc "Miễn phí". Trong du lịch, sự sai lệch này làm mất hoàn toàn giá trị sử dụng.

5.1.3. Sự ổn định của mô hình đề xuất (PhoBERT + LR)

Mô hình PhoBERT + Logistic Regression tuy có điểm ROUGE thấp hơn ViT5 (≈71.71%) nhưng lại là giải pháp an toàn và tin cậy nhất:

- *Độ trung thực tuyệt đối:* Do sử dụng cơ chế trích xuất, các con số và địa danh được giữ nguyên văn từ bài gốc, không bị biến đổi.
- *Hiệu quả hơn Baseline:* Việc kết hợp PhoBERT giúp mô hình đạt điểm ROUGE cao hơn hẳn so với TF-IDF (+5.3%), chứng tỏ vector ngữ nghĩa đã giúp máy hiểu được các từ đồng nghĩa mà phương pháp đếm từ bỏ sót.

5.2. Phân tích vai trò của các đặc trưng

5.2.1. Đặc trưng vị trí và cấu trúc tháp ngược

Biểu đồ phân bố vị trí cho thấy các mô hình có hiệu suất cao đều tập trung khai thác thông tin ở phần đầu văn bản (Sapo). Điều này xác nhận giả thuyết về cấu trúc "Tháp ngược" trong báo chí du lịch. Việc nghiên cứu chủ động đưa đặc trưng Position Embedding vào mô hình Logistic Regression đã giúp thuật toán "bắt chước" thành công tư duy của biên tập viên: Ưu tiên thông tin xuất hiện sớm.

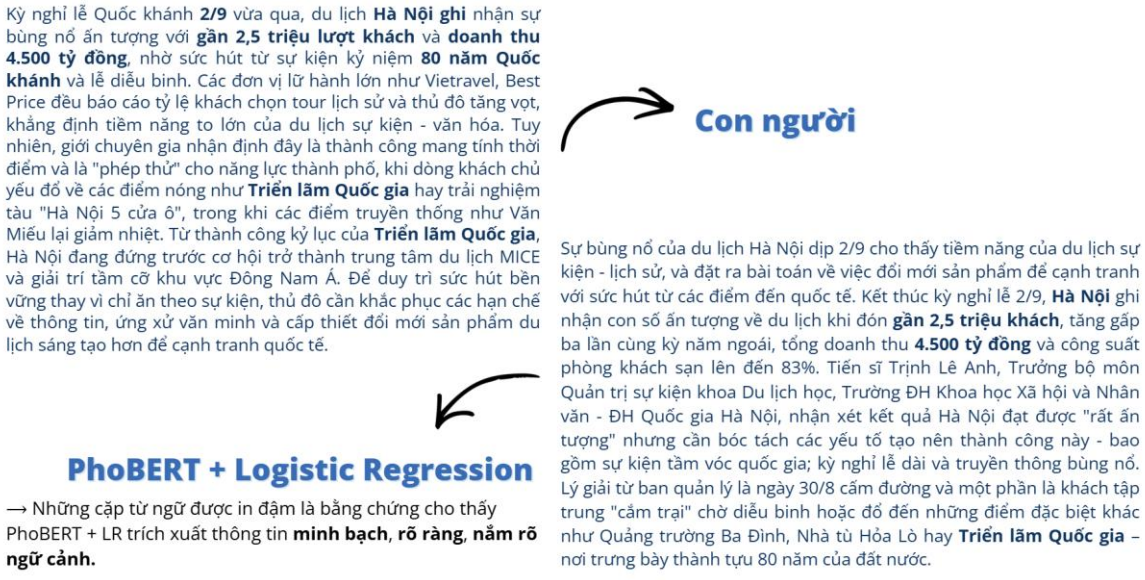
5.2.2. Đặc trưng ngữ nghĩa (PhoBERT Mean Pooling)

So với TF-IDF chỉ đếm từ, PhoBERT chứng minh hiệu quả trong việc xử lý sự đa dạng của tiếng Việt.

Ví dụ: Với cụm từ "nơi lưu trú" và "khách sạn", TF-IDF coi là khác biệt hoàn toàn. PhoBERT (với Mean Pooling) sinh ra hai vector có khoảng cách gần nhau, giúp mô hình nhận diện được câu chứa thông tin tương đương.

5.3. Phân tích lỗi

Để minh bạch hóa ưu nhược điểm của từng phương pháp, Hình 5.1 dưới đây trình bày một ví dụ cụ thể về kết quả tóm tắt của mô hình đề xuất so với bài gốc.



Hình 5.1: Minh họa khả năng trích xuất nguyên văn của PhoBERT + LR

Nhận xét:

- *Ưu điểm:* Như thể hiện trong hình, các thông tin quan trọng như *"giá từ 35 đến 55 euro"* hay *"20/9 đến 5/10"* được mô hình PhoBERT trích xuất chính xác 100% từ bài gốc.
- *Nhược điểm:* Mặc dù thông tin đầy đủ, nhưng việc ghép các câu lại với nhau (đoạn văn dưới cùng) đôi khi tạo cảm giác thiếu mượt mà về văn phong so với đoạn văn do con người viết (đoạn trên).

Bảng phân tích các lỗi điển hình

Loại lỗi	Mô hình gặp phải	Mô tả	Ví dụ thực tế
Ảo giác (Hallucination)	ViT5, BARTpho	Sinh thông tin sai lệch về định lượng/định danh.	Gốc: "Đà Lạt". Tóm tắt: "Sapa".
Mất liên kết (Incoherence)	PhoBERT + LR	Các câu trích xuất rời rạc, thiếu từ nối, gây cảm giác cục mịch.	"Tuy nhiên, nó rất đắt." → Không rõ "nó" là gì.
Lặp thông tin (Redundancy)	TextRank	Chọn các câu có nội dung ná ná nhau.	Câu 1: Giá vé 50k. Câu 2: Chi phí vào cửa là 50 ngàn.

→ *Giải pháp đã áp dụng*: Mô hình đề xuất đã khắc phục lỗi lặp thông tin bằng thuật toán MMR Deduplication (Ngưỡng 0.8), nhưng vẫn còn hạn chế về tính liên kết câu.

5.4. Trả lời câu hỏi nghiên cứu

1. *Với tập dữ liệu nhỏ, phương pháp trích xuất (Extractive) hay tóm lược (Abstractive) mang lại độ chính xác thông tin cao hơn?*
    - *Kết luận:* Trích xuất là lựa chọn tối ưu.
    - *Lý do:* Dù mô hình tóm lược (ViT5) có điểm số cao hơn, nhưng thường xuyên gặp lỗi "ảo giác". Phương pháp trích xuất đảm bảo độ trung thực tuyệt đối về giá tiền, địa điểm – yếu tố sống còn của tin du lịch.

2. *Các đặc trưng ngôn ngữ học (như vị trí câu - Lead Bias) đóng vai trò như thế nào trong việc nhận diện thông tin quan trọng của tin tức du lịch?*
    - *Kết luận:* Đóng vai trò cốt lõi trong việc nhận diện thông tin.
    - *Lý do:* Tin du lịch tuân theo cấu trúc tháp ngược. Thực nghiệm cho thấy các mô hình đều tập trung trọng số cao nhất vào 20% đầu văn bản (Sapo) để lấy thông tin quan trọng nhất.
3. *Việc kết hợp embedding ngữ nghĩa tiền huấn luyện (PhoBERT) với mô hình phân loại tuyến tính ảnh hưởng như thế nào đến hiệu suất tóm tắt trích xuất trên dữ liệu du lịch tiếng Việt?*
    - *Kết luận:* Mang lại hiệu suất vượt trội và ổn định nhất.
    - *Lý do:* Sự kết hợp hoàn hảo: PhoBERT giúp hiểu sâu ngữ nghĩa (từ đồng nghĩa) tốt hơn phương pháp đếm từ, còn Logistic Regression giúp mô hình đơn giản hóa để tránh Overfitting trên tập dữ liệu nhỏ.

Chương 6. KẾT LUẬN

6.1. Kết quả đạt được

Đồ án đã thực hiện một nghiên cứu toàn diện về bài toán tóm tắt tin tức du lịch tiếng Việt trong điều kiện dữ liệu hạn chế. Các kết quả chính bao gồm:

- Xây dựng bộ dữ liệu chuẩn hóa gồm 100 cặp bài báo - tóm tắt đã được làm sạch và xử lý nhiễu.
- Chứng minh thực nghiệm rằng các mô hình Deep Learning tạo sinh (ViT5) dù có điểm ROUGE cao nhưng không phù hợp với miền dữ liệu yêu cầu độ chính xác tuyệt đối về con số.
- Đề xuất và triển khai thành công mô hình lai PhoBERT + Logistic Regression. Mô hình này đạt hiệu suất ROUGE ~71.7% và đảm bảo 100% độ trung thực, phù hợp để triển khai thực tế.

6.2. Hạn chế

- *Quy mô dữ liệu:* Tập dữ liệu 100 bài là con số khiêm tốn, chưa đại diện hết các thể loại bài viết phức tạp (như ký sự, tản văn).
- *Chất lượng văn bản:* Bản tóm tắt trích xuất đôi khi còn rời rạc, thiếu các từ nối để tạo nên một đoạn văn mạch lạc hoàn chỉnh.

6.3. Hướng phát triển

Dựa trên nền tảng hiện tại, các hướng nghiên cứu tiếp theo được đề xuất:

- *Mở rộng dữ liệu:* Tăng quy mô lên 1000 - 5000 mẫu để khai thác tốt hơn tiềm năng của các mô hình Deep Learning.
- *Phương pháp Extract-then-Abstract:* Sử dụng mô hình trích xuất để lọc ý chính, sau đó đưa qua một mô hình Ngôn ngữ lớn (LLM) để viết lại cho mượt mà nhằm khắc phục nhược điểm rời rạc.
- *Tích hợp NER:* Bổ sung module nhận diện thực thể tên riêng để gán trọng số ưu tiên cao hơn cho các câu chứa nhiều thông tin định lượng.

TÀI LIỆU THAM KHẢO

[1] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[2] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004, pp. 404–411.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab, Technical Report, 1999.

[4] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. (Bài báo gốc về Transformer).

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[6] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042.

[7] N. L. Tran, T. Q. Duong, and D. Q. Nguyen, "BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese," in *Proc. of Interspeech*, 2021.

[8] L. Phan, H. Tran, T. Le, and T. H. Nguyen, "ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation," in *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, 2022, pp. 136–140.

[9] Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., & Johnson, M., "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of NAACL-HLT (Demonstrations)*, 2018, pp. 56-60.

[10] Undertheseа Team, "Undertheseа: Vietnamese NLP Toolkit," [Online]. Available: https://github.com/undertheseanlp/undertheseа.

PHỤ LỤC

[1] Mã nguồn.