TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

A red and blue logo

Description automatically generated

**ĐÀO NGỌC BÁCH - 52000737**

**PHAN HUỲNH TOÀN ĐỨC - 52000750**

**XÂY DỰNG HỆ THỐNG SỬ DỤNG MÔ HÌNH HỌC SÂU TÓM TẮT BẢN TIN THỂ THAO**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**KHOA HỌC MÁY TÍNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

A red and blue logo

Description automatically generated

**ĐÀO NGỌC BÁCH - 52000737**

**PHAN HUỲNH TOÀN ĐỨC - 52000750**

**XÂY DỰNG HỆ THỐNG SỬ DỤNG MÔ HÌNH HỌC SÂU TÓM TẮT BẢN TIN THỂ THAO**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**KHOA HỌC MÁY TÍNH**

Người hướng dẫn

**PGS. TS. Lê Anh Cường**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn trường Đại học Tôn Đức Thắng đã giúp em có những khoảng thời gian vui vẻ khi học tại trường và đã đưa môn này vào chương trình học. Đặc biệt cảm ơn thầy Lê Anh cường, người đã hướng đã tận tình chúng em. Trong thời gian học làm dự án, em học thêm được rất nhiều kiến thức và thầy rất vui tính. Em chúc thầy thật nhiều sức khỏe và hạnh phúc.

*TP. Hồ Chí Minh, ngày ... tháng … năm 20..*

*Tác giả*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của PGS. TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày … tháng … năm 20..*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

**MỤC LỤC**

[DANH MỤC HÌNH VẼ viii](#_Toc161327385)

[DANH MỤC BẢNG BIỂU x](#_Toc161327386)

[DANH MỤC CÁC CHỮ VIẾT TẮT xi](#_Toc161327387)

[CHƯƠNG 1. GIỚI THIỆU 1](#_Toc161327388)

[1.1 Tổng quan về bài toán 1](#_Toc161327389)

[1.2 Mục tiêu và phạm vi, phương pháp 2](#_Toc161327390)

[1.2.1 Mục tiêu nghiên cứu 2](#_Toc161327391)

[1.2.2 Phạm vi nghiên cứu 3](#_Toc161327392)

[1.2.3 Phương pháp nghiên cứu 3](#_Toc161327393)

[CHƯƠNG 2. CÁC KIẾN THỨC CƠ BẢN 4](#_Toc161327394)

[2.1 Mô hình T5 4](#_Toc161327395)

[2.1.1 Input and Output Format 4](#_Toc161327396)

[2.1.2 The Colossal Clean Crawled Corpus 5](#_Toc161327397)

[2.1.3 Downstream Tasks 5](#_Toc161327398)

[2.1.4 Text-to-Text Framework 6](#_Toc161327399)

[2.1.5 Model Architecture 6](#_Toc161327400)

[2.1.6 Kết quả 8](#_Toc161327401)

[2.2 ViT5 8](#_Toc161327402)

[2.3 BART 9](#_Toc161327403)

[2.3.1 Abstract 9](#_Toc161327404)

[2.3.2 BERT vs GPT và BART 9](#_Toc161327405)

[2.3.3 Model 10](#_Toc161327406)

[2.4 BARTpho 12](#_Toc161327407)

[2.5 mT5 12](#_Toc161327408)

[CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 13](#_Toc161327409)

[3.1 Overview về các nghiên cứu liên quan 14](#_Toc161327410)

[3.2 Mô hình (hoặc hệ thống) đề xuất 14](#_Toc161327411)

[3.2.1 Kiến trúc, thiết kế engine hệ thống 14](#_Toc161327412)

[3.2.2 Nêu chức năng các thành phần 14](#_Toc161327413)

[3.3 Mô hình xử lý dữ liệu 14](#_Toc161327414)

[3.3.1 Miền dữ liệu 14](#_Toc161327415)

[3.3.2 Xử lý dữ liệu 14](#_Toc161327416)

[3.3.3 Lựa chọn biểu diễn, trích chọn đặc trưng 14](#_Toc161327417)

[CHƯƠNG 4. THỰC NGHIỆM 13](#_Toc161327418)

[4.1 Xây dựng dữ liệu 14](#_Toc161327419)

[4.1.1 Thu thập dữ liệu 14](#_Toc161327420)

[4.1.2 Thống kê về dữ liệu 14](#_Toc161327421)

[4.1.3 Tạo dữ liệu tóm tắt 14](#_Toc161327422)

[4.2 Mô hình thực nghiệm 14](#_Toc161327423)

[4.2.1 Mô hình ViT5 14](#_Toc161327424)

[4.2.2 Mô hình mT5-small 14](#_Toc161327425)

[4.2.3 Mô hình BARTpho 14](#_Toc161327426)

[4.2.4 Các nguồn source tham khảo 14](#_Toc161327427)

[4.3 Các tiêu chí, độ đo đánh giá mô hình 14](#_Toc161327428)

[4.4 Kết quả thực nghiệm 14](#_Toc161327429)

[4.4.1 Bảng kết quả so sánh các mô hình 14](#_Toc161327430)

[4.4.2 Thống kê về huấn luyện mô hình 14](#_Toc161327431)

[4.4.3 Nhận xét, đánh giá về kết quả 14](#_Toc161327432)

[4.5 Demo hệ thống 14](#_Toc161327433)

[4.5.1 Giao diện 14](#_Toc161327434)

[4.5.2 Các user cases thực nghiệm và kết quả 14](#_Toc161327435)

[CHƯƠNG 5. KẾT LUẬN 13](#_Toc161327436)

[5.1 Kết quả đạt được 13](#_Toc161327437)

[5.1.1 Kết quả đạt được 13](#_Toc161327438)

[5.1.2 Ý nghĩa khoa học và/hoặc thực tiễn của kết quả đạt được 13](#_Toc161327439)

[5.2 Tổng quát những hạn chế và hướng phát triển trong tương lai 14](#_Toc161327440)

[5.2.1 Hạn chế 14](#_Toc161327441)

[5.2.2 Hướng phát triển 14](#_Toc161327442)

# DANH MỤC HÌNH VẼ

[Hình 1: Sơ đồ objective sử dụng trong T5 baseline model 4](#_Toc161327300)

[Hình 2: Sở đồ T5 text-to-text framework 6](#_Toc161327301)

[Hình 3: Matrices representing different attention mask patterns. 6](#_Toc161327302)

[Hình 4: ViT5 encoder-decoder architecture 8](#_Toc161327303)

[Hình 5: Sơ đồ so sánh BART với BERT và GPT 9](#_Toc161327304)

[Hình 6: Những hàm tranformation để làm nhiễu đầu vào 11](#_Toc161327305)

[Hình 7: Minh họa kiến trúc hệ thống 14](#_Toc161327306)

[Hình 8: Minh họa cho lỗi tên của tác giả bài báo đôi khi cũng bị thu thập 17](#_Toc161327307)

[Hình 9: Minh họa cho bài báo dạng danh sách 18](#_Toc161327308)

[Hình 10: Minh họa cho bài báo chỉ có hình ảnh 18](#_Toc161327309)

[Hình 11: Minh họa cho lỗi sinh thừa 19](#_Toc161327310)

[Hình 12: Minh họa cho lỗi chuyển sang liệt kê 20](#_Toc161327311)

[Hình 13: Rouge1 trung bình 24](#_Toc161327312)

[Hình 14: Rouge2 trung bình 25](#_Toc161327313)

[Hình 15: RougeL trung bình 25](#_Toc161327314)

[Hình 16: Thống kê của BARTpho 26](#_Toc161327315)

[Hình 17: Thống kê của mT5-small 26](#_Toc161327316)

[Hình 18: Thống kê của ViT5 27](#_Toc161327317)

[Hình 19: Trang chủ 28](#_Toc161327318)

[Hình 20: Nhập link và trả nội dung bài báo 29](#_Toc161327319)

[Hình 21: Trang trả kết quả tóm tắt 29](#_Toc161327320)

[Hình 22: Nhập tay văn bản 30](#_Toc161327321)

[Hình 23: Trang trả kết quả 30](#_Toc161327322)

[Hình 24: Kết quả tóm tắt trả về user case 1 31](#_Toc161327323)

[Hình 25: Kết quả tóm tắt trả về user case 2 32](#_Toc161327324)

[Hình 26: Kết quả tóm tắt trả về user case 3 33](#_Toc161327325)

# DANH MỤC BẢNG BIỂU

[Bảng 1: Hiệu suất của các biến thể kiến trúc 8](#_Toc161327190)

[Bảng 2: Đánh giá của 3 mô hình lần lượt ViT5, mT5 và BARTpho 28](#_Toc161327191)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| BART | Bidirectional and Auto-Regressive Transformers |
| HTML | HyperText Markup Language |
| JS | Javascript |
| MLM | Mask language modeling |
| mT5 | Multilingual T5 |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation |
| T5 | Text-to-Text Transfer Transformer |

# GIỚI THIỆU

## Tổng quan về bài toán

Trong thời đại số hóa hiện nay, việc tiếp cận thông tin nhanh chóng và chính xác đang trở thành một yếu tố quan trọng, đặc biệt trong lĩnh vực thể thao - một ngành đầy sôi động và thường xuyên cập nhật thông tin mới. Đối với người hâm mộ và những người làm việc trong lĩnh vực này, việc nắm bắt và phân tích thông tin thể thao đúng mức, đúng thời điểm có ý nghĩa không chỉ trong việc cập nhật kiến thức mà còn trong việc đưa ra quyết định và phân tích chiến thuật.

Mô hình học sâu, một lĩnh vực nghiên cứu đang phát triển mạnh mẽ trong trí tuệ nhân tạo, mang lại cơ hội không giới hạn để ứng dụng trong việc tự động hóa quá trình tóm tắt bài báo thể thao. Sử dụng mô hình học sâu để tóm tắt tin tức thể thao tiếng Việt không chỉ giúp tiết kiệm thời gian mà còn đảm bảo tính chính xác, độ tin cậy và tính cập nhật nhanh chóng của thông tin. Điều này không chỉ thuận tiện cho người đọc mà còn giúp các nhà báo và những người làm việc trong lĩnh vực thể thao có cái nhìn tổng quan, nhanh chóng từ các nguồn tin đa dạng.

Lý do chọn đề tài này cũng xuất phát từ sự nhận biết về nhu cầu thực tế. Trong thể thao, mỗi chi tiết, thông tin mới, hay phát ngôn của cầu thủ, HLV đều có giá trị. Tuy nhiên, việc theo dõi và xử lý lượng thông tin khổng lồ này đòi hỏi một sự can thiệp nhanh chóng và chính xác, điều mà con người không thể đảm bảo trong mọi tình huống. Với sự hỗ trợ của mô hình học sâu, chúng ta có thể xây dựng những hệ thống tự động, hiệu quả và linh hoạt để phục vụ nhu cầu này.

Tóm lại, việc tóm tắt bài báo tin tức thể thao tiếng Việt sử dụng mô hình học sâu giúp đáp ứng nhu cầu thực tế về việc cập nhật và phân tích thông tin thể thao một cách nhanh chóng và chính xác.

Có hai hướng tiếp cận trong bài toán tóm tắt văn bản:

* Tóm tắt trích xuất:

Trong phương pháp này, các câu hoặc đoạn văn được chọn từ văn bản nguồn và được sắp xếp lại để tạo thành một tóm tắt. Các đoạn văn được chọn có thể được chọn trực tiếp từ văn bản nguồn mà không cần thay đổi hoặc chỉnh sửa. Mục tiêu của tóm tắt trích xuất là duy trì tính chính xác của thông tin từ văn bản nguồn trong tóm tắt.

* + Tóm tắt trừu tượng:

Trái ngược với tóm tắt trích xuất, phương pháp này tạo ra một tóm tắt mới, không chỉ là sự tái sắp xếp của các phần của văn bản nguồn, mà là việc tạo ra các câu mới để diễn đạt ý nghĩa của văn bản nguồn. tóm tắt trừu tượng thường sử dụng các phương pháp ngôn ngữ tự nhiên (NLP) và các mô học sâu như T5 (Raffel et al. 2019), ViT5 (Phan et al. 2022), BART (Lewis et al. 2019), BARTpho (Tran, Le, and Nguyen 2022), mT5 (Xue et al. 2021) phức tạp để tạo ra các câu tóm tắt có thể không xuất hiện trong văn bản nguồn. Mục tiêu của tóm tắt trừu tượng là tạo ra một tóm tắt ngắn gọn nhưng vẫn bảo toàn ý nghĩa và thông tin chính xác của văn bản nguồn.

## Mục tiêu và phạm vi, phương pháp

### Mục tiêu nghiên cứu

Phát triển mô hình tự động hóa: Tạo ra một mô hình học sâu chuyên biệt có khả năng tự động tóm tắt, xử lý và phân tích thông tin từ các bài báo thể thao tiếng Việt một cách chính xác và nhanh chóng.

Nâng cao chất lượng tóm tắt: Đảm bảo rằng các bản tóm tắt được sản xuất từ mô hình đạt được độ chính xác cao, giữ lại những thông tin quan trọng và loại bỏ thông tin không cần thiết, từ đó đảm bảo tính nguyên bản và ý nghĩa của nội dung gốc.

Tích hợp ngữ cảnh và ngữ nghĩa: Xây dựng một hệ thống hiểu và phân tích ngữ cảnh trong các bài báo, giúp mô hình có khả năng hiểu và phân loại thông tin dựa trên ngữ nghĩa và tình huống cụ thể.

Ứng dụng thực tế: Áp dụng mô hình vào các ứng dụng và dịch vụ thực tế như website tin tức, ứng dụng di động, và các nền tảng truyền thông xã hội để cung cấp thông tin nhanh và chính xác đến cộng đồng người yêu thể thao.

### Phạm vi nghiên cứu

Data: Các bài báo thể thao chuyển nhượng tiếng Việt được crawl từ trang web Vietnamnet, VNExpress,…

Các mô hình học sâu: T5, ViT5, BART, BARTpho, mT5.

Các bài báo tin tức thể thao tiếng Việt.

### Phương pháp nghiên cứu

#### Lý thuyết

Trình bày các khái niệm lý thuyết cơ bản liên quan đến tóm tắt văn bản và mô hình học sâu.

Đặc biệt tập trung vào các lý thuyết về attention mechanisms, biến thể mô hình Transformer.

#### Thực nghiệm

Bộ dữ liệu được crawl từ trang web thể thao chất lượng VnExpress, Vietnamnet,...

Sử dụng các mô hình học sâu để tóm tắt: ViT5, BARTpho, mT5.

Mô tả các phương pháp đánh giá hiệu suất bằng độ đo ROUGE.

Triển khai lên môi trường website sử dụng Streamlit.

#### Nội dung công việc

Chương 2: Trình bày lý về các mô hình sẽ được sử dụng trong bài toán tóm tắt văn bản sử dụng mô hình học sâu.

Chương 3: Phương pháp đề xuất.

Chương 4: Thực nghiệm các mô hình đã trình bày ở chương 2, so sánh kết quả bằng độ đo ROUGE.

Chương 5: Kết luận những việc đã làm và kết quả đạt được. Hạn chế và hướng phát triển tương lai.

# CÁC KIẾN THỨC CƠ BẢN

## Mô hình T5

### Input and Output Format

A group of text boxes

Description automatically generated

Hình : Sơ đồ objective sử dụng trong T5 baseline model

Từ Original text, lấy ngẫu nhiên các từ. Trong ví dụ này là “for inviting” và “last”. Sau đó mask chúng với certain placeholders (“last” mask thành <Y> và “for inviting” mask thành <X>). Inputs trên hình chính là inputs của encoder model và ở decoder model mong muốn tạo ra những từ đã được mask, trong ví dụ này là “for inviting” và “last”.

Trong khi mô hình BERT chỉ dự đoán một single token tại một thời điểm, trong ví dụ này không chỉ dự đoán single word mà còn couple of words.

### The Colossal Clean Crawled Corpus

Là Dataset được scraped từ web.

Common Crawl là tập tin web công khai, được cào từ toàn bộ web mỗi tháng. Dung lượng khoảng 20TB. Lưu text bằng cách xóa markup từ HTML file. Những quy tắc thực tế (heuristics) để clean Common Crawl’s web extracted text, chỉ extract high quality điều này giúp mô hình có tính khái quát hóa:

* Xóa những trang có ít hơn 3 câu, giữ lại những dòng ít hơn 1 câu.
* Xóa những dòng chứa mà JS.
* Xóa trang nào chứa những từ tục.
* Những trang cào từ Wikipedia có các dấu trích dẫn [1], …, xóa những phần đó.
* Xóa những trang có chứa dấu “{“.

Và chỉ cào những trang tiếng Anh, sử dụng **langdetect** để filter với tỉ lệ ít nhất 0.99. Cuối cùng thu thập được C4 dataset với 750GB.

### Downstream Tasks

Tác giả kiểm tra hiệu suất của mô hình trên nhiều task khác nhau, bao gồm machine translation, question answering, abstractive summarization, và text classification. Họ đo hiệu suất trên GLUE và SuperGLUE text classification meta-benchmarks; CNN/Daily Mail abstractive summarization, SQuAD question answering, và WMT machine translation.

Tác giả đề xuất một unified hay single model cho nhiều loại tasks. Tác giả đã đưa ra khái niệm task-specific (text) prefix. Chúng được đưa vào tập train và test như một loại hint để model chọn ra subset of the weights, điều này sẽ giúp ích cho nhiệm vụ cụ thể.

### Text-to-Text Framework

A diagram of text and a number

Description automatically generated with medium confidence

Hình : Sở đồ T5 text-to-text framework

T5 model lấy input là text và output cũng là text (sequence to sequence model).

Nếu muốn sử dụng mô hình để translation từ tiếng anh sang tiếng đức thì nó thêm prefix “translate English to German: That is good” và tương tự với các task khác (unified model that can do multitask learning pretraining on the same object and then fine-tune to many different tasks).

### Model Architecture

A diagram of a graph

Description automatically generated with medium confidence

Hình : Matrices representing different attention mask patterns.

Tác giả xem xét 3 mô hình biến thể:

Hình 3 cho biết những attention mask pattern.

Hình bên trái, Encoder sử dụng attention mask “Fully-visible”. Khi generate từng token của output nó xem xét tất cả các token trong input để đưa ra dự đoán. Điều này cho phép mô hình lấy thông tin từ một phạm vi rộng hơn của dữ liệu đầu vào.

Hình ở giữa, Transformer’s decoder sử dụng “causal” masking pattern. Khi tạo ra entry thứ i của output sequence, việc che dấu (masking) causal sẽ ngăn mô hình tham gia entry thứ j của input sequence đối với j > i. Điều này được sử dụng trong quá trình train để mô hình không thể “nhìn thấy tương lai” khi nó tạo ra output. Do đó, bộ Transformer decoder (không có bộ encoder) có thể được sử dụng như một language model (LM), tức là một mô hình được đào tạo chỉ để dự đoán bước tiếp theo.

Hình bên phải, các language models thường được sử dụng để generate sequence. Tuy nhiên, chúng cũng có thể được sử dụng trong text-to-text framework simply chỉ bằng cách nối the inputs and targets. Ví dụ: hãy xem xét trường hợp dịch từ tiếng Anh sang tiếng Đức: Nếu chúng ta có một training datapoint với input sentence “That is good”. Và nhắm target “Das ist gut.”, chúng tôi chỉ cần train model trên next-step prediction qua input sequence được nối “translate English to German: That is good. Target: Das ist gut.” Nếu chúng tôi muốn nhận được dự đoán của mô hình cho ví dụ này, mô hình sẽ được cung cấp tiền tố (prefix) “translate English to German: That is good. Target:” và sẽ được yêu cầu generate phần còn lại của sequence một cách tự động. Fully-visible masking sẽ được áp dụng cho prefix “translate English to German: That is good. Target:” và causal masking sẽ được sử dụng trong quá trình train để dự đoán target “Das ist gut”.

### Kết quả

A table with numbers and letters

Description automatically generated

Bảng : Hiệu suất của các biến thể kiến trúc

Enc-dec shared có hiệu suất gần giống baseline và tốt hơn prefix LM.

Denoising objective > LM objective.

## ViT5

A diagram of a variety of colored squares

Description automatically generated with medium confidence

Hình : ViT5 encoder-decoder architecture

ViT5 là một mô hình encoder-decoder dựa trên Transformer được pretrain cho tiếng Việt. ViT5 được đào tạo trên kho tàng văn bản tiếng Việt đa dạng, chất lượng cao. Benchmark ViT5 trên 2 downstream text generation tasks, Abstractive Text Summarization and Named Entity Recognition (NER).

Pre-trained trên CC100 copus. Vietnamese Corpus với 138GB raw text. Xử lý và lọc ra 69GB đoạn văn ngắn cho mô hình có độ dài 256 và 71GB đoạn văn dài cho mô hình có độ dài 1024.

ViT5 tuân theo kiến trúc encoder-decoder và T5 framework cho Base and Large Model. Huấn luyện các mô hình ViT5 với hai độ dài đầu vào và đầu ra khác nhau: độ dài 256 và độ dài 1024.

## BART

### Abstract

BART là một denoising autodecoder cho việc pretraining sequence-to-sequence models. BART được huấn luyện bằng cách (1) làm hỏng văn bản với một hàm noising tùy ý, và (2) học để tái tạo lại văn bản gốc.

### BERT vs GPT và BART

|  |  |
| --- | --- |
| A diagram of a bidirectional encoder  Description automatically generated  (a) BERT | A diagram of a decoder  Description automatically generated  (b) GPT |
| A diagram of a machine  Description automatically generated  (c) BART | |

Hình : Sơ đồ so sánh BART với BERT và GPT

BERT (Devlin et al. 2019): chỉ là phần Encoder trong kiến trúc Transformer sequence to sequences. Trong đó việc pretraining objective mà nó được train là MLM (mask language modeling) và NSP (next sequence prediction). Trong bài báo chỉ nhắc đến MLM.

Hình bên trái (BERT), chuỗi ban đầu là ABCDE, sau đó mask random tokens từ input bằng dấu gạch ngang, và output là dự đoán những từ bị mask dựa trên các từ khác xuất hiện trong ngữ cảnh đó. Trong ví dụ trên từ B, D có thể quan sát được các từ A, C, E. BERT dự đoán không theo tuần tự, nếu B được dự đoán trước thì D sẽ có thể quan sát được cả B.

Hình bên phải (GPT) là auto-regressive, ví dụ tại thời điểm mô hình generate C thì C chỉ có thể quan sát được B, A và <s> (những token trước C), tương tự với D, E. GPT là unidirectional, nghĩa là không có token output nào có thể là điểm cuối ở bên phải.

Hình ở giữa (BART, kết hợp giữa BERT và GPT): phần Encoder: ta có input là ABCDE và đưa qua noise function, noise function có nhiệm vụ lấy random words ta thu được output là A\_B\_E (masked random tokens). Encoder học the representation of mask tokens and non-mask tokens, sau đó chuyển đến decoder. Phần Decoder: output của decoder có nhiệm vụ tạo ra tất các token tiếp theo bao gồm mask tokens.

### Model

BART là một mô hình denoising autoencoder mà ánh xạ một corrupted document đến original document mà nó được tạo ra từ đó. Nó được thực hiện dưới dạng một sequence-to-sequence model với một bidirectional encoder trên văn bản bị hỏng và một left-to-right autoregressive decoder.

#### Architecture

BART sử dụng kiến trúc Transformer. Phần Decoder (GPT), thay thế hàm kích hoạt ReLU thành GeLU. Đối với base model, sử dụng 6 layers trong the encoder and decoder, còn đối với large model, sử dụng 12 layers. Kiến trúc có liên quan chặt chẽ với kiến trúc được sử dụng trong BERT, với những điểm khác biệt sau: (1) mỗi layer của decoder còn thực hiện thêm cross-attention đối với final hidden layer of the encoder (như trong transformer sequence-to-sequence model trước đây); và (2) BERT sử dụng additional feed-forward network trước khi dự đoán từ, BART thì không. BART chứa nhiều tham số hơn khoảng 10% so với mô hình BERT có kích thước tương đương.

#### Pre-traning BART

BART là denoising autoencoder, nó sẽ có những hàm transformation, dùng để transform input bằng cách thêm noise sau đó đưa vào encoder và tại decoder, được huấn luyện lại trên non-noise để tái tạo lại noise.

A diagram of a diagram

Description automatically generated with medium confidence

Hình : Những hàm tranformation để làm nhiễu đầu vào

Những hàm transformation:

**Token Masking** Giống BERT (Devlin et al., 2019), random tokens được lấy mẫu và thay thế bằng phần tử [MASK].

**Token Deletion** Random tokens sẽ bị xóa khỏi input. Ngược lại với việc token masking, model phải quyết định vị trí nào bị missing inputs.

**Text Infilling** Thay vì chỉ một mask thay cho một word như BERT (Devlin et al., 2019), nó có thể mask nhiều words bằng single mask token.

**Sentence Permutation** A document được chia thành các câu dựa trên các dấu chấm và các sentences này được xáo trộn theo thứ tự ngẫu nhiên.

**Document Rotation** A token được chọn thống nhất một cách ngẫu nhiên và document được xoay để bắt đầu bằng token đó. Nhiệm vụ này huấn luyện mô hình để xác định điểm bắt đầu của tài liệu.

## BARTpho

BARTpho có hai phiên bản,  và , đây là những mô hình tuần tự đơn ngữ quy mô lớn đầu tiên được đào tạo trước cho tiếng Việt. BARTpho sử dụng kiến trúc “large” và pre-training scheme của sequence-to-sequence denoising autoencoder BART, do đó nó đặc biệt phù hợp cho các generative NLP tasks.

Cả  và  đều sử dụng kiến trúc “large” với 12 encoder and decoder layers cùng pre-training scheme của BART. Cụ thể, BART đào tạo trước có hai giai đoạn: (i) làm hỏng văn bản đầu vào với chức năng nhiễu tùy ý và (ii) học cách xây dựng lại văn bản gốc, tức là tối ưu hóa cross-entropy giữa decoder’s output và original text. Giống BART, tác giả sử dụng hai loại noise trong noise function, bao gồm text infilling and sentence permutation. Thêm một layer-normalization layer lên trên cả encoder and decoder.

## mT5

Mô hình ngôn ngữ mT5 (multilingual T5) là một biến thể của mô hình T5 (Text-To-Text Transfer Transformer) được Google Research phát triển. T5 là một kiến trúc mạng nơ-ron biến thể Transformer mạnh mẽ được huấn luyện trên một loạt các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP). Khả năng chính của T5 là chuyển đổi các nhiệm vụ NLP thành dạng "text-to-text", có nghĩa là cả input và output đều được biểu diễn bằng văn bản.

Mô hình mT5 được phát triển với mục tiêu hỗ trợ nhiều ngôn ngữ (hơn 100 ngôn ngữ). Được pre-training trên bộ dữ liệu mC4. mC4 bao gồm văn bản tự nhiên bằng 101 ngôn ngữ được rút ra từ kho lưu trữ web Common Crawl công khai và được clean data giống như bộ dữ liệu C4 của mô hình T5. Sử dụng cơ chế “line length filter” yêu cầu các trang phải chứa ít nhất ba dòng văn bản có 200 ký tự trở lên, phát hiện những ngôn ngữ của mỗi trang bằng cách sử dụng **cld3** và xóa những ngôn ngữ có độ chính xác dưới 70%. Điều này có nghĩa là mT5 có khả năng xử lý và hiểu nội dung của nhiều ngôn ngữ khác nhau mà không cần phải có các mô hình riêng lẻ cho từng ngôn ngữ. Mô hình này được huấn luyện trên một lượng lớn dữ liệu từ nhiều nguồn và nhiều ngôn ngữ khác nhau, giúp nó học được biểu diễn ngôn ngữ phổ quát và có thể áp dụng cho các nhiệm vụ NLP trên nhiều ngôn ngữ.

Mô hình mT5 đã chứng minh được hiệu suất tốt trên nhiều nhiệm vụ NLP khác nhau: translation, summarization, classification…

# PHƯƠNG PHÁP ĐỀ XUẤT

## Overview về các nghiên cứu liên quan

Trong lĩnh vực tóm tắt văn bản sử dụng mô hình học sâu, có nhiều nghiên cứu và công trình đáng chú ý đã được thực hiện. Dưới đây là một cái nhìn tổng quan về những nghiên cứu quan trọng và các công trình đã được công bố trước đó:

- Nghiên Cứu về Attention Mechanisms trong Tóm Tắt Văn Bản.

+ Đã có nhiều nghiên cứu về sử dụng cơ chế attention trong mô hình học sâu để tăng cường khả năng tập trung vào các phần quan trọng của văn bản trong quá trình tóm tắt.

+ Công trình của (Vaswani et al. 2017) về Transformer, với cơ chế attention tiên tiến, đã mở ra nhiều khả năng mới trong xử lý ngôn ngữ tự nhiên.

- Công Trình Áp Dụng Mô Hình GPT, BART, T5, … trong Tóm Tắt Văn Bản.

+ Những mô hình này đã được áp dụng thành công trong nhiều ứng dụng tóm tắt văn bản do khả năng sinh đầu ra mạch lạc và tự động hóa.

- Nghiên Cứu về Đánh Giá và Đo Lường Hiệu Suất Tóm Tắt Văn Bản.

+ Các nghiên cứu về tiêu chí đánh giá hiệu suất của mô hình tóm tắt văn bản là một phần quan trọng để đảm bảo chất lượng kết quả.

+ Công trình của (Lin 2004) về ROUGE (Recall-Oriented Understudy for Gisting Evaluation) là một trong những cơ sở đo lường phổ biến được sử dụng để đánh giá chất lượng tóm tắt.

## Mô hình (hoặc hệ thống) đề xuất

### Kiến trúc, thiết kế engine hệ thống

A diagram of a flowchart

Description automatically generated

Hình : Minh họa kiến trúc hệ thống

Hệ thống được xây dựng trên nền tảng Docker để đảm bảo tính di động và dễ dàng triển khai trên nhiều môi trường.

Hệ thống sẽ chia thành hai phần chính: phần frontend và phần backend.

Phần frontend sẽ là trang web streamlit đơn giản để người dùng có thể nhập link hoặc nhập tay nội dung của bài báo thể thao chuyển nhượng tiếng Việt. Trong trường hợp nhập link thì nếu link đó hợp lệ thì sẽ in nội dung bài báo đó ra màn hình.

Phần backend được xây dựng trên nền tảng Google Cloud Platform sẽ xử lý yêu cầu từ phần frontend và sử dụng mô hình đã được fine-tuning để tạo ra các bản tóm tắt từ nội dung của bài báo đó, sau đó trả về kết quả cho người dùng.

### Nêu chức năng các thành phần

Frontend gồm giao diện Streamlit đơn giản cho phép người dùng nhập link của bài báo hoặc nhập tay nội dung của bài thể thao chuyển nhượng, tích hợp với backend để gửi yêu cầu tóm tắt và hiển thị kết quả tóm tắt.

Backend là Docker container trên GCP nhận yêu cầu tóm tắt từ frontend, sử dụng mô hình đã fine-tuning để tạo ra tóm tắt cho nội dung bài báo nhập từ link hoặc nhập tay được cung cấp sau đó trả về kết quả tóm tắt cho frontend để hiển thị cho người dùng.

## Mô hình xử lý dữ liệu

### Miền dữ liệu

Miền dữ liệu của bài toán là bài báo thể thao thuộc chủ đề chuyển nhượng bóng đá của các trang báo online tại thời điểm hiện tại.

Các bản tóm tắt sẽ cố gắng giữ lại các tiêu chí: cầu thủ, CLB liên quan, giá chuyển nhượng, chi tiết hợp đồng, lý do chuyển nhượng. Nếu nội dung bài báo không đề cập thì có thể bỏ qua.

### Xử lý dữ liệu

Loại bỏ các ký tự đặc biệt: đầu vào là các văn bản tiếng Việt sẽ được loại bỏ các ký tự đặc biệt không mang ý nghĩa.

Chuyển đổi tất cả các chữ cái viết hoa thành viết thường để có sự đồng nhất trong việc xử lý văn bản.

### Lựa chọn biểu diễn, trích chọn đặc trưng

Đặc trưng của các bài báo sau khi tóm tắt phải cố gắng giữ không quá dài và giữ được các thông tin sau đây:

- Cầu thủ và câu lạc bộ liên quan.

- Già chuyển nhượng.

- Chi tiết hợp đồng.

- Lý do chuyển nhượng.

Các thông tin nào mà bài báo không cung cấp thì có thể bỏ qua.

# THỰC NGHIỆM

## Xây dựng dữ liệu

### Thu thập dữ liệu

Bộ dữ liệu dùng để giải quyết bài xây dựng hệ thống tóm tắt văn bản thể thao chuyển nhượng của nhóm có khoảng hơn 12500 bài báo tiếng Việt thuộc thẻ chuyển nhượng được thu thập từ nhiều trang báo online ở Việt Nam thời điểm hiện tại.

Nguồn dữ liệu được thu thập từ các nguồn: các bài báo thuộc thẻ chuyển nhượng của báo “vnexpress” với số lượng 500 bài báo, thẻ chuyển nhượng của báo **“**vietnamnet**”** với số lượng khoảng 8830 bài báo, báo “dantri” với số lượng khoảng 488 bài báo, báo “” với số lượng khoảng 3184 bài báo, báo “tuoitre” với số lượng 8 bài báo, báo “thanh niên” với số lượng 15 bài báo, báo “tiền phong” với số lượng 35 bài báo.

Phương pháp thu thập: Nhóm đã sử dụng kết hợp giữa việc thu thập các bài báo bằng BeautifulSoup và thu thập thủ công. Các bài báo sau khi thu thập thông qua BeautifulSoup sẽ được nhóm xem qua lại một lần nữa để loại bỏ các từ không liên quan đến nội dung bài báo như: tên tác giả bài báo, từ “Nguồn: “ trong những bức ảnh minh họa của bài báo…

A screenshot of a computer

Description automatically generated

Hình : Minh họa cho lỗi tên của tác giả bài báo đôi khi cũng bị thu thập

Tiêu chí chọn lọc các bài báo: Do tài nguyên GPU có hạn nên nhóm đã tập trung vào việc tạo ra các tóm tắt ngắn nhưng vẫn cố gắng giữ được các tiêu chí đề ra nên các bài báo quá dài, dạng danh sách hay dạng tổng hợp hoặc quá ngắn như chỉ có 1 ảnh trong nội dung sẽ bị hạn chế và loại bỏ. Sau khi qua các tiêu chí chọn lọc thì còn được số lượng bài báo như trên.

**A person holding a jersey

Description automatically generated**

Hình : Minh họa cho bài báo dạng danh sách

A screenshot of a football game

Description automatically generated

Hình : Minh họa cho bài báo chỉ có hình ảnh

### Thống kê về dữ liệu

Số lượng: bộ dữ liệu gồm 12531 bài báo thể thao tiếng Việt thuộc thẻ chuyển nhượng được chia theo tỉ lệ 9:1 gồm khoảng 11600 bài báo trong tập train và 907 bài báo trong tập validation.

Đặc điểm của dữ liệu: bao gồm các bài báo và các bản tóm tắt có độ dài vừa phải và cố gắng giữ được cái tiêu chí đề ra.

### Tạo dữ liệu tóm tắt

Các bài báo sau khi được thu thập về sẽ được tóm tắt kết hợp giữa API GPT 3.5 và nhập tóm tắt trực tiếp trên web với lệnh như sau: ”hãy tóm tắt bản tin tức thể thao sau (bằng đoạn văn) chú ý đến những thông tin như cầu thủ/CLB liên quan, giá chuyển nhượng, chi tiết hợp đồng, lý do chuyển nhượng”. Các bản tóm tắt bằng API sẽ được chuyển vào một tập tin csv cùng với bản gốc của nó.

Các bài báo sau khi được tóm tắt qua API sẽ được nhóm xem xét lại 1 lần nữa xem liệu bản tóm tắt đó có giữ lại được các tiêu chí hay không, liệu bản tóm tắt có quá dài hay quá ngắn hay không, có mắc các lỗi như thêm các từ “tóm tắt: ” hay chuyển sang dạng liệt kê hay không. Sau khi có tập dữ liệu hoàn chỉnh thì tiến hành chia ngẫu nhiên thành hai tập train và validation và tải lên trang HuggingFace.

A screenshot of a computer

Description automatically generated

Hình : Minh họa cho lỗi sinh thừa

A close up of a text

Description automatically generated

Hình : Minh họa cho lỗi chuyển sang liệt kê

## Mô hình thực nghiệm

### Mô hình ViT5

Môi trường huấn luyện: Sử dụng GPU T4 trên Google Colab với 15GB RAM.

Thêm prefix “summarize” vào mỗi bài báo gốc.

Đầu vào tối đa là 1024 token.

Đầu ra tối đa là 128 token.

Số epoch là 16 epoch.

Kích thước batch là 4.

Learning rate là 2e-5.

Hệ số weight decay là 0.01.

Lưu model sau mỗi epoch và giữ lại tối đa 3 phiên bản .

Sử dụng predict\_with\_generate để dự đoán kết quả tóm tắt.

Có sử dụng fp16.

### Mô hình mT5-small

Môi trường huấn luyện: Sử dụng GPU T4 trên Google Colab với 15GB RAM.

Thêm prefix “summarize” vào mỗi bài báo gốc.

Đầu vào tối đa là 1024 token.

Đầu ra tối đa là 128 token.

Số epoch là 16 epoch.

Kích thước batch là 4.

Learning rate là 2e-5.

Hệ số weight decay là 0.01.

Lưu model sau mỗi epoch và giữ lại tối đa 3 phiên bản .

Sử dụng predict\_with\_generate để dự đoán kết quả tóm tắt.

Không sử dụng fp16.

### Mô hình BARTpho

Môi trường huấn luyện: Sử dụng GPU T4 trên Google Colab với 15GB RAM.

Thêm prefix “summarize” vào mỗi bài báo gốc.

Đầu vào tối đa là 1024 token.

Đầu ra tối đa là 128 token.

Số epoch là 16 epoch.

Kích thước batch là 4.

Learning rate là 2e-5.

Hệ số weight decay là 0.01.

Lưu model sau mỗi epoch và giữ lại tối đa 3 phiên bản .

Sử dụng predict\_with\_generate để dự đoán kết quả tóm tắt.

Có sử dụng fp16.

### Các nguồn source tham khảo

Tham khảo cách để có thể thực hiện fine-tuning trên mô hình T5 [[1]](#footnote-2).

Tham khảo về mô hình mT5 trên Hugging Face [[2]](#footnote-3).

Tham khảo về mô hình ViT5 [[3]](#footnote-4).

Tham khảo về mô hình BARTpho [[4]](#footnote-5) .

## Các tiêu chí, độ đo đánh giá mô hình

Trong bài toán này, chúng tôi sử dụng độ đo ROUGE[[5]](#footnote-6) để đánh giá các mô hình tóm tắt văn bản.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE là một bộ đo phổ biến được sử dụng để đánh giá chất lượng của các tóm tắt văn bản. ROUGE đo lường độ tương đồng giữa bản tóm tắt sinh ra bởi mô hình và bản tóm tắt tham khảo (thường là của con người) bằng cách sử dụng các phép đo như ROUGE-N (đo lường độ chính xác của N từ liên tiếp), và ROUGE-L (đo lường độ dài của chuỗi con dài nhất chung).

**Rouge-N:** Đo số lượng trùng lặp của n-gram phù hợp giữa văn bản do mô hình tạo ra và tham chiếu do con người tạo ra. n-gram nằm trong khoảng 1 đến vô cùng nhưng chủ yếu chỉ dùng 1-gram và 2-gram.

VD: A là bản tóm tắt do con người tóm tắt, B là bản tóm tắt do mô hình tạo ra.

H: I really like listening to music.

M: I like listening to music.

**Rouge-1:** Ta xét từng từ (1-gram).

H 1-gram: *I*, really, *like*, *listening*, *to*, *music*

M 1-gram: *I*, *like*, *listening*, *to*, *music*

Rouge-1 precision được tính bằng tỷ lệ giữa 1-gram xuất hiện trong cả H và M trên số 1-gram trong H.

Rouge-1 precision = 5/6

Rouge-1 recall được tính bằng tỷ lệ giữa 1-gram xuất hiện trong cả H và M trên số 1-gram trong M.

Rouge-1 recall = 5/5 = 1

Rouge-1 F1-score = 2 (precision x recall) / (precision + recall)

**Rouge-2:** Ta xét 2-gram

H 2-gram: I really, really like, *like listening*, *listening to*, *to music*

M 2-gram: I like, *like listening*, *listening to*, *to music*

Tương tự Rouge-1 ta có kết quả Rouge-2 như sau:

Rouge-2 precision = 3/5

Rouge-2 recall = ¾

**Rouge-L:** Dựa trên chuỗi con chung dài nhất (LCS) giữa M và H, tức chuỗi dài nhất (không cần liên tục nhưng phải theo thứ tự) giữa M và H. Trình tự dài hơn sẽ biểu thị nhiều điểm giống nhau hơn giữa hai trình tự. Trong ví dụ này, LCS là 5-gram “I like listening to music”.

Rouge-L precision là tỷ lệ giữa độ dài của LCS trên số 1-gram trong H.

Rouge-L precision = 5/6

Rouge-L recall là tỷ lệ giữa độ dài của LCS trên số 1-gram trong M.

Rouge-L recall = 5/5

## Kết quả thực nghiệm

### Bảng kết quả so sánh các mô hình

Kết quả độ đo Rouge được đánh giá trên tập test gồm 50 bài báo được đăng lên thẻ chuyển nhượng của trang Tiền Phong và Thanh niên trong tháng 2 và tháng 3 năm 2024 và được tóm tắt bằng Chat GPT 3.5, được nhóm xem và chỉnh sửa lại.

A bar graph with different colors

Description automatically generated

Hình : Rouge1 trung bình

A graph of different colors

Description automatically generated with medium confidence

Hình : Rouge2 trung bình

A graph of different colors

Description automatically generated with medium confidence

Hình : RougeL trung bình

### Thống kê về huấn luyện mô hình

A screenshot of a computer screen

Description automatically generated

Hình : Thống kê của BARTpho

A screenshot of a computer screen

Description automatically generated

Hình : Thống kê của mT5-small

A screenshot of a graph

Description automatically generated

Hình : Thống kê của ViT5

### Nhận xét, đánh giá về kết quả

Mô hình ViT5 cho hiệu suất tốt nhất với tập dữ liệu này.

Với mô hình ViT5 thì Rouge1, Rouge2 và RougeL đều tăng từ epoch 1 đến epoch 4, sau đó giảm nhẹ hoặc duy trì ổn định. Có vẻ mô hình đang học được cách tạo ra các từ và câu phù hợp.

Với mô hình mt5-small thì đang học và cải thiện dần qua các epoch. Các chỉ số Rouge tăng đều đặn, trong khi đó không có dấu hiệu của overfitting từ việc so sánh Training Loss và Validation Loss nhưng hiệu quả còn thấp so với 2 mô hình kia.

Với mô hình BARTpho thì mô hình có dấu hiệu của sự overfitting qua việc tăng dần của Validation Loss nhưng các chỉ số Rouge không thay đổi đáng kể qua các epoch. Mô hình có vẻ không đạt được hiệu suất mong muốn trong việc tạo ra các văn bản phù hợp với dữ liệu thực tế.

BARTpho lại có cải thiện khá nhiều về hiệu suất khi tăng kích thước dữ liệu huấn luyện. Với tập dữ liệu chỉ 1100 bài báo thì BARTpho cho hiệu suất rất kém nhưng khi tăng kích thước dữ liệu huấn luyện thì mô hình cho hiệu suất tốt hơn hẳn.

A table with numbers and letters

Description automatically generated

Bảng : Đánh giá của 3 mô hình lần lượt ViT5, mT5 và BARTpho

với tập dữ liệu chỉ 1100 bài báo

## Demo hệ thống

### Giao diện

A screenshot of a computer

Description automatically generated

Hình : Trang chủ

A screenshot of a computer

Description automatically generated

Hình : Nhập link và trả nội dung bài báo

A screenshot of a computer screen

Description automatically generated

Hình : Trang trả kết quả tóm tắt

A screenshot of a computer

Description automatically generated

Hình : Nhập tay văn bản

A screenshot of a computer

Description automatically generated

Hình : Trang trả kết quả

### Các user cases thực nghiệm và kết quả

**User Case 1**: Thử nghiệm với bài báo chuyển nhượng ngày 13/3/2024 trên trang báo: “Chelsea bán loạt ngôi sao cho đội bóng nhà giàu” - Vietnamnet [[6]](#footnote-7).

A screenshot of a computer

Description automatically generated

Hình : Kết quả tóm tắt trả về user case 1

Các tiêu chí quan trọng đã được bao gồm trong tóm tắt:

- Các cầu thủ và CLB liên quan: Romelu Lukaku, Kepa, Kante, Edouard Mendy, Kalidou Koulibaly và các đối tác ở giải Saudi Pro League.

- Lý do chuyển nhượng: Chelsea cần bán cầu thủ để tuân thủ Luật Công bằng tài chính sau khi công bố khoản lỗ.

- Hoạt động thị trường: Boehly đã thực hiện cuộc đàm phán tại Saudi Arabia để bán các cầu thủ cho các CLB ở đó.

Tóm tắt này giữ lại các tiêu chí chính từ thông tin gốc và truyền đạt đầy đủ về hoạt động thị trường và lý do chuyển nhượng của Chelsea.

**User Case 2**: Thử nghiệm với bài báo chuyển nhượng ngày 9/3/2024 trên trang báo “Giải Brazil mua sắm vượt Ngoại hạng Anh” - vnexpress [[7]](#footnote-8).

A black background with white text

Description automatically generated

Hình : Kết quả tóm tắt trả về user case 2

Các tiêu chí quan trọng đã được bao gồm trong tóm tắt:

- Các cầu thủ và CLB liên quan: Luiz Henrique (Botafogo), De la Cruz (Flamengo), Matias Vina (Flamengo), Leo Ortiz (Flamengo), Anibal Moreno (Palmeiras), Rodrigo Garro (Corinthians), Rafael Santos Borre (Internacional).

- Giá chuyển nhượng: Các giá trị chuyển nhượng cao nhất đã được liệt kê.

- Chi tiết hợp đồng: Các vụ mua sắm đáng chú ý đã được mô tả, bao gồm giá trị và các CLB liên quan.

Một số chi tiết cụ thể như lý do chuyển nhượng hoặc các thông tin chi tiết về các cầu thủ đã được bỏ qua để tóm tắt được ngắn gọn và súc tích hơn.

**User Case 3**: Thử nghiệm với bài báo chuyển nhượng ngày 21/12/2023 trên trang báo “Chuyển nhượng MU, hãy học cách Man City chộp lấy Doku” – Vietnamnet.

A black background with white text

Description automatically generated

Hình : Kết quả tóm tắt trả về user case 3

Hệ thống đã trả về đoạn tóm tắt tuy nhiên vẫn không thể sinh ra hết do nội dung bài báo quá dài.

# KẾT LUẬN

## Kết quả đạt được

### Kết quả đạt được

Xây dựng được một bộ dữ liệu tin tức thể thao chuyển nhượng bằng tiếng Việt và thành công thử nghiệm các mô hình học máy vào tập dữ liệu đó.

Các mô hình cho ra kết quả tương đối ổn trong việc tạo ra các bản tóm tắt ngắn gọn và cố gắng giữ lại được tiêu chí.

Xây dựng được web trên các mô hình đã có.

### Ý nghĩa khoa học và/hoặc thực tiễn của kết quả đạt được

Xây dựng được một bộ dữ liệu nhỏ về đề tài tóm tắt các bài báo thể thao chuyển nhượng tiếng Việt.

Nhóm đã xây dựng được hệ thống tóm tắt văn bản thể thao chuyển nhượng tiếng Việt. Hệ thống sẽ có thể phục vụ những người cần những bản tóm tắt về thông tin thể thao ngắn gọn nhưng vẫn có những nội dung cần thiết.

## Tổng quát những hạn chế và hướng phát triển trong tương lai

### Hạn chế

Tuy mô hình đã cho kết quả tương đối ổn với các bản tóm tắt ngắn, nhưng nếu bài báo đầu vào quá dài hoặc có nhu cầu cho các đoạn tóm tắt dài hơn thì mô hình sẽ không đáp ứng được.

Các đoạn tóm tắt quá ngắn có thể dẫn đến bỏ sót những thông tin quan trọng.

Do hạn chế về mặt tài nguyên nên khó có thể thử nghiệm đa dạng với nhiều độ dài tóm tắt khác nhau.

Phần nhập liên kết để tóm tắt của web hiện tại chỉ nhận của trang Vietnamnet, VnExpress và bongda.com.vn, những trang báo còn lại phải chuyển qua phần “nhập tay” để nhập nội dung.

Không có hạn chế cho người dùng nhập link ngoài các trang báo thể thao chuyển nhượng, phần nhập tay cũng không có hạn chế về việc có hoặc không nhập gì và nội dung thế nào.

Mô hình sinh ra tóm tắt còn tương đối chậm.

### Hướng phát triển

Cố gắng bổ sung thêm vào tập dữ liệu cả về kích thước và độ đa dạng, thu thập thêm từ những trang báo cũ và cả những trang báo mới.

Cố gắng nghiên cứu về các phương pháp mới để tối ưu hóa siêu tham số của mô hình.

Cố gắng nâng cấp phần “nhập link” của trang web để có thể tóm tắt nhiều bài báo ở các trang khác hơn.

**TÀI LIỆU THAM KHẢO**

1.Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.”

2.Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. “BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension.”

3.Lin, Chin-Yew. 2004. “ROUGE: A Package for Automatic Evaluation of Summaries.” Pp. 74–81 in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics.

4.Phan, Long, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. “ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation.”

5.Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.”

6.Tran, Nguyen Luong, Duong Minh Le, and Dat Quoc Nguyen. 2022. “BARTpho: Pre-Trained Sequence-to-Sequence Models for Vietnamese.”

7.Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.”

8.Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. “mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer.”

9.Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

1. <https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/summarization.ipynb?hl=vi> [↑](#footnote-ref-2)
2. <https://huggingface.co/docs/transformers/model_doc/mt5> [↑](#footnote-ref-3)
3. <https://github.com/vietai/ViT5> [↑](#footnote-ref-4)
4. <https://huggingface.co/docs/transformers/model_doc/bartpho> [↑](#footnote-ref-5)
5. https://viblo.asia/p/2-phut-de-hieu-cach-tinh-rouge-metric-5OXLAPEBJGr [↑](#footnote-ref-6)
6. https://vietnamnet.vn/chelsea-ban-loat-ngoi-sao-cho-cac-doi-bong-nha-giau-2259202.html [↑](#footnote-ref-7)
7. https://vnexpress.net/giai-brazil-mua-sam-vuot-ngoai-hang-anh-4720211.html [↑](#footnote-ref-8)