Đại học quốc gia TP.HCM
Trường đại học công nghệ thông tin

KHOA KHOA HỌC MÁY TÍNH

CS221.O22 - XỬ LÝ NGÔN NGỮ TỰ NHIÊN

BÁO CÁO ĐỒ ÁN CUỐI KÌ

MULTI-STAGE DOCUMENT RANKING FOR VIETNAMESE NEWS RETRIEVAL

GV hướng dẫn: TS. Nguyễn Trọng Chỉnh, ThS. Nguyễn Đức Vũ, ThS. Đặng Văn Thìn

Nhóm thực hiện

21521992 - Võ Đức Dương

21521465 - Trần Ngọc Thiện

21520617 - Nguyễn Thúc Bảo

Mục lục

1 Lý do lựa chọn đề tài **2**

2 Giới thiệu bài toán **3**
2.1 Mô tả bài toán . . . 3
2.2 Mô tả Input/Output . . . 3
2.3 Mô tả ngữ liệu . . . 3

3 Giới thiệu phương pháp chính **5**
3.1 Quy trình xử lý của Multi-stage Document Ranking . . . 5
3.2 Các phương pháp term-based . . . 6
3.2.1 TF-IDF . . . 6
3.2.2 BM25 . . . 7
3.3 Pre-trained language model . . . 8
3.3.1 PhoBERT . . . 8
3.3.2 Vietnamese-SBERT . . . 8

4 Cài đặt phương pháp **9**
4.1 Tiền xử lý dữ liệu . . . 9
4.2 Đánh giá bài báo bằng Cosine Similarity . . . 9

5 Đánh giá phương pháp **10**
5.1 Tập đánh giá . . . 11
5.2 Kết quả thu được . . . 12

6 Tài liệu tham khảo **13**

1 Lý do lựa chọn đề tài

Trong thời đại cách mạng công nghiệp 4.0, Internet trở thành nguồn tài nguyên thông tin khổng lồ. Việc sử dụng các công cụ tìm kiếm như Google, Bing, ... để tra cứu thông tin trở nên phổ biến và quen thuộc với con người. Thế nhưng, **làm thế nào mà các công cụ tìm kiếm này có thể hiểu được câu truy vấn và trả về các thông tin liên quan?** Bắt nguồn từ câu hỏi này, nhóm quyết định tìm hiểu để xây dựng và tái hiện lại một hệ thống tương tự.

Theo ý tưởng đó, nhóm tiếp tục xem xét: **Đâu là thông tin mà người đọc thường sẽ chú ý nhiều hơn?** Trên thực tế, với mỗi một câu truy vấn chứa thông tin cần tra cứu, công cụ tìm kiếm có thể trả về vô số bài báo, bài viết với vô số đường dẫn chứa thông tin liên quan. Mỗi một câu truy vấn sẽ yêu cầu thông tin về một lĩnh vực, và ứng với mỗi lĩnh vực đó, những bài viết hoặc trang web ưu tiên trả về sẽ khác nhau.

Khi tìm kiếm *Thông tin hôm nay*, 10 kết quả trả về đầu tiên đều là các trang web báo điện tử. Lý giải điều này, nhóm cho rằng đây là các trang báo điện tử phổ biến ở Việt Nam, cập nhật các thông tin mới nhất theo từng ngày, dẫn đến số lượng truy cập đông đảo. Cùng với các dòng văn bản như *mới nhất 24h hôm nay, tin nhanh, tin nóng, Thời sự hôm nay, ...* đã định hướng để trình tìm kiếm trả về kết quả như trên.

Trong phạm vi đồ án môn học, nhóm chỉ thực hiện trong giới hạn là các thông tin văn bản, cụ thể hơn là các bài báo điện tử vì các trang báo điện tử luôn được người đọc đón nhận và truy cập hằng ngày. Tất cả những điều trên là lý do để nhóm thực hiện đề tài.

Đồ án này thể hiện những gì mà chúng em học được từ lớp **CS221.O22**. Chúng em xin gửi lời cảm ơn **TS. Nguyễn Trọng Chỉnh, ThS. Nguyễn Đức Vũ, ThS. Đặng Văn Thìn** đã nhiệt tình giảng dạy, giải đáp thắc mắc trong quá trình học tập và thực hiện đồ án cuối kỳ. Đồ án chắc chắn không tránh khỏi những sai sót, chúng em hi vọng nhận được những góp ý của các thầy về đồ án để cải thiện và hoàn chỉnh hơn.

2 Giới thiệu bài toán

2.1 Mô tả bài toán

Cho một truy vấn **q** và một ngữ liệu **D** là tập hợp nhiều bài báo khác nhau. Bài toán đặt ra là xây dựng một hệ thống nhận vào truy vấn q và trả về các bài báo từ D, trong đó:

- Chứa các ngữ cảnh quan trọng liên quan đến q.
- Các bài báo được xếp hạng từ trên xuống dưới theo **mức độ liên quan** đến truy vấn. Các bài báo có độ liên quan cao hơn thì có xếp hạng (rank) cao hơn.

2.2 Mô tả Input/Output

a) Input:

- 1 tập ngữ liệu D: chứa thông tin các bài báo trên các trang báo điện tử. Thông tin chi tiết về tập ngữ liệu sẽ được mô tả ở mục **2.3**
- 1 câu truy vấn q bằng Tiếng Việt, chứa thông tin về bài báo cần tìm kiếm

b) Output: Một danh sách các bài báo trong ngữ liệu D, được sắp xếp giảm dần về mức độ liên quan của bài báo với truy vấn.

2.3 Mô tả ngữ liệu

Tập ngữ liệu chứa thông tin về các bài báo được thu thập từ 5 trang báo điện tử của Việt Nam, bao gồm: báo Lao Động, báo Dân Trí, báo VnExpresses, báo VTC và báo Đảng Cộng Sản. Với mỗi bài báo, ta thu thập 5 đặc trưng (feature) bao gồm: **title** (tiêu đề), **abstract** (tổng quan), **source** (nguồn bài báo), **link** (đường dẫn bài báo) và **topic** (chủ đề).

Hà Nội công bố điểm thi lớp 10

Điểm thi Toán, Văn, Ngoại ngữ vào lớp 10 của hơn 105.000 thí sinh được công bố lúc 17h.

Hình 1: Một bài báo Giáo dục từ báo VnExpress

Ví dụ, với một bài báo như *Hình 1*, ta có các thông tin sau:

- **title**: Hà Nội công bố điểm thi lớp 10
- **abstract**: Điểm thi Toán, Văn, Ngoại ngữ vào lớp 10 của hơn 105.000 thí sinh được công bố lúc 17h.
- **source**: Báo VnExpress
- **link**: https://vnexpress.net/tra-cuu-diem-thi-lop-10-ha-noi-nam-2024-4762421.html
- **topic**: Giáo dục

Tổng số lượng thu thập là 48357 bài báo. Ở đây, nhóm chỉ tập trung vào mức độ liên quan của tiêu đề và tổng quan đối với truy vấn nên thời gian của bài báo sẽ không được thu thập.

title	abstract	source	link	topic
Chưa giàu đã "khó" vì thuế thu nhập cá nhâ	(ĐCSVN) – Có nhiều ý kiến cho rằng, mức giảm trừ gia cảnh trong cách t	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/chua-giau-da-kho-vi-thue-thu-nhap-ca-nhan	Cùng bàn luận
Băn khoăn trước sự "chuyển mình" chậm c	(ĐCSVN)- Nhiều đại biểu HĐND TP Hà Nội bày tỏ "băn khoăn" khi các d	Báo Đảng Cộng Sản	https://dangcongsan.vn/thoi-su/ban-khoan-truoc-su-chuyen-minh-cham-chap-cua-	Cùng bàn luận
Phát huy hiệu quả việc phối hợp giữa gia đì	(ĐCSVN) - Học sinh ở nhà trường phổ thông đang trong lứa tuổi cần được	Báo Đảng Cộng Sản	https://dangcongsan.vn/giao-duc/phat-huy-hieu-qua-viec-phoi-hop-giua-gia-dinh-nh	Cùng bàn luận
Nỗ lực trả lại đúng nghĩa công viên ở Thủ đ	(ĐCSVN) - Thành phố Hà Nội sẽ tìm mô hình đầu tư để công viên không c	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/no-luc-tra-lai-dung-nghia-cong-vien-o-thu-d	Cùng bàn luận
Hoàn thiện thể chế - Khâu đột phá chiến lư	LTS: Quyết tâm hoàn thiện thể chế trong công tác xây dựng Đảng, phòng	Báo Đảng Cộng Sản	https://dangcongsan.vn/tieu-diem/hoan-thien-the-che-khau-dot-pha-chien-luoc-tro	Cùng bàn luận
Giải pháp nào cho câu chuyện "biết rồi, kh	(ĐCSVN) - Nhiều năm qua, ùn tắc giao thông đã trở thành vấn nạn, trở th	Báo Đảng Cộng Sản	https://dangcongsan.vn/thoi-su/giai-phap-nao-cho-cau-chuyen-biet-roi-kho-lam-no	Cùng bàn luận
Nhân văn, nhân đạo trong xử lý cán bộ vi ph	(ĐCSVN) - Tổng Bí thư Nguyễn Phú Trọng nhiều lần khẳng định các vụ án	Báo Đảng Cộng Sản	https://dangcongsan.vn/xay-dung-dang/nhan-van-nhan-dao-trong-xu-ly-can-bo-vi-p	Cùng bàn luận
Phòng chống lãng phí: "Xử lý" vấn đề phải t	(ĐCSVN) - Lãng phí gây nguy hại cho xã hội không kém gì tham nhũng. Tr	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/phong-chong-lang-phi-xu-ly-van-de-phai-tu-	Cùng bàn luận
Hội nghị COP27: Kỳ vọng vượt qua thực tại	(ĐCSVN) – Hội nghị lần thứ 27 các bên tham gia Công ước khung của Liê	Báo Đảng Cộng Sản	https://dangcongsan.vn/the-gioi/tin-tuc/hoi-nghi-cop27-ky-vong-vuot-qua-thuc-tai-6	Cùng bàn luận
Xóa bỏ sản xuất manh mún, nhỏ lẻ	(ĐCSVN) - Sản xuất manh mún, nhỏ lẻ là vấn đề lớn mỗi khi nhắc đến hạ	Báo Đảng Cộng Sản	https://dangcongsan.vn/kinh-te/xoa-bo-san-xuat-manh-mun-nho-le-626035.html	Cùng bàn luận
Lại "nóng" chuyện cá độ bóng đá mùa Wor	(ĐCSVN) – Liên tiếp trong những ngày vừa qua, Bộ Công an, cơ quan côn	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/lai-nong-chuyen-ca-do-bong-da-mua-world-	Cùng bàn luận
Đảm bảo an toàn thực phẩm trong các bếp	(ĐCSVN) – Những ngày gần đây, thông tin về học sinh tại trường iSchool N	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/dam-bao-an-toan-thuc-pham-trong-cac-bep	Cùng bàn luận
Quyền lực của khán giả	ĐCSVN) - Khán giả thể hiện quyền lực của mình thông qua việc tỏ rõ thái	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/quyen-luc-cua-khan-gia-625270.html	Cùng bàn luận
Gỡ bài toán giảm áp lực đào hạn trái phiếu	(ĐCSVN) - Luỹ kế trong 9 tháng đầu năm 2022, tổng giá trị trái phiếu đã đ	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/go-bai-toan-giam-ap-luc-dao-han-trai-phieu	Cùng bàn luận
TP Hồ Chí Minh đẩy mạnh chuyển đổi số trc	(ĐCSVN) - Những năm qua, quy mô ngành giáo dục và đào tạo của TP Hồ	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/tp-ho-chi-minh-day-manh-chuyen-doi-so-trc	Cùng bàn luận
Vì sự an toàn trên mọi con đường!	(ĐCSVN) - Tai nạn giao thông có thể phòng tránh được. Để không phải hồ	Báo Đảng Cộng Sản	https://dangcongsan.vn/tieu-diem/vi-su-an-toan-tren-moi-con-duong-624918.html	Cùng bàn luận
Cơ hội kích cầu tiêu dùng cuối năm	(ĐCSVN) - Được thực hiện trên phạm vi toàn quốc từ ngày 15/11 - 22/12	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/co-hoi-kich-cau-tieu-dung-cuoi-nam-624877	Cùng bàn luận
Lời tâm huyết của "Tư lệnh" Ngành giáo dụ	(ĐCSVN) - Liên quan đến việc tăng lương, phụ cấp cho giáo viên yên tâm	Báo Đảng Cộng Sản	https://dangcongsan.vn/tieu-diem/loi-tam-huyet-cua-tu-lenh-nganh-giao-duc-truoc-	Cùng bàn luận
"Sàng lọc" các dự án FDI tiềm ẩn rủi ro	(ĐCSVN) - Dù dòng vốn đầu tư nước ngoài (FDI) là một trong các nguồn lự	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/sang-loc-cac-du-an-fdi-tiem-an-rui-ro-6247	Cùng bàn luận
Khẩn trương, trách nhiệm, sáng suốt	(ĐCSVN) - Sau thời gian 21 ngày làm việc, Kỳ họp thứ tư Quốc hội khóa XI	Báo Đảng Cộng Sản	https://dangcongsan.vn/ban-doc/y-kien-ban-doc/khan-truong-trach-nhiem-sang-su	Cùng bàn luận
Tâm lực, trí lực trong từng quyết sách	(ĐCSVN) – Kỳ họp thứ tư, Quốc hội khóa XV đã khép lại vào chiều ngày 15	Báo Đảng Cộng Sản	https://dangcongsan.vn/tieu-diem/tam-luc-tri-luc-trong-tung-quyet-sach-624595.ht	Cùng bàn luận
Cá thể hóa trách nhiệm người đứng đầu trc	(ĐCSVN) – Cá thể hoá trách nhiệm người đứng đầu trong giải ngân vốn đ	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/ca-the-hoa-trach-nhiem-nguoi-dung-dau-trc	Cùng bàn luận
Thu phí thôi, chưa đủ!	(ĐCSVN) - Thông tin về việc TP Hà Nội sẽ lấy ý kiến rộng rãi về việc thu ph	Báo Đảng Cộng Sản	https://dangcongsan.vn/cung-ban-luan/thu-phi-thoi-chua-du-624502.html	Cùng bàn luận
Đấu thầu thuốc: Đừng chỉ chăm chăm vào	(ĐCSVN) - Các đại biểu Quốc hội cho rằng, nên tôn trọng ý kiến của các	Báo Đảng Cộng Sản	https://dangcongsan.vn/y-te/dau-thau-thuoc-dung-chi-cham-cham-vao-loai-nao-re-	Cùng bàn luận
Cần có sự thẩm tra, giám sát kỹ lưỡng về xá	(ĐCSVN) - Nhất trí cao với sự cần thiết phải sửa đổi Luật đất đai, song nh	Báo Đảng Cộng Sản	https://dangcongsan.vn/thoi-su/can-co-su-tham-tra-giam-sat-ky-luong-ve-xac-dinh-	Cùng bàn luận
Không nóng vội thông qua luật, không làm l	(ĐCSVN) - Không nóng vội chạy theo tiến độ, không vì "thành tích" mà là	Báo Đảng Cộng Sản	https://dangcongsan.vn/tieu-diem/khong-nong-voi-thong-qua-luat-khong-lam-luat-c	Cùng bàn luận
Quy định rõ điều kiện thu hồi đất, tránh áp	(ĐCSVN) - Thực tế cho thấy, vấn đề thu hồi đất luôn là vấn đề phức tạp, n	Báo Đảng Cộng Sản	https://dangcongsan.vn/thoi-su/quy-dinh-ro-dieu-kien-thu-hoi-dat-tranh-ap-dung-ch	Cùng bàn luận

Hình 2: Một số bài báo trong tập ngữ liệu

3 Giới thiệu phương pháp chính

Thay vì sử dụng một mô hình đơn lẻ để xếp hạng tất cả các tài liệu, phương pháp **Multi-stage Document Ranking** áp dụng một chuỗi các mô hình khác nhau, mỗi mô hình thực hiện một chức năng cụ thể.

3.1 Quy trình xử lý của Multi-stage Document Ranking

Đầu tiên, ta sử dụng một **phương pháp term-based** như TF-IDF hay BM25 để lọc các bài báo có nội dung không liên quan.

Ở đây, TF-IDF được sử dụng với mục đích chuyển đổi các query cùng với corpus (nội dung bài báo bao gồm title và abstrac) về dạng feature vector. Tiếp đó, các bài báo này sẽ được xếp hạng dựa trên độ đo **cosine similarity**.

Khác với TF-IDF, BM25 không tính toán feature vector. Thay vào đó, thuật toán trực tiếp xếp hạng các bài báo dựa trên **BM25 score**.

Sau khi xếp hạng, ta chọn ra 50 bài báo có xếp hạng cao nhất và những bài báo này sẽ được tiếp tục đưa qua các pre-trained language model để thực hiện **reranking** - sắp xếp lại theo mức độ liên quan với câu truy vấn, cả về mặt ngữ cảnh và ngữ nghĩa.

Đối với PhoBERT, vector ẩn của token [CLS] ở đầu ra của lớp transformer cuối cùng được đưa qua một lớp dense (fully connected layer) để giảm chiều, tạo ra vector **pooler output** và ta sẽ sử dụng vector này sẽ được sử dụng như một feature vector.

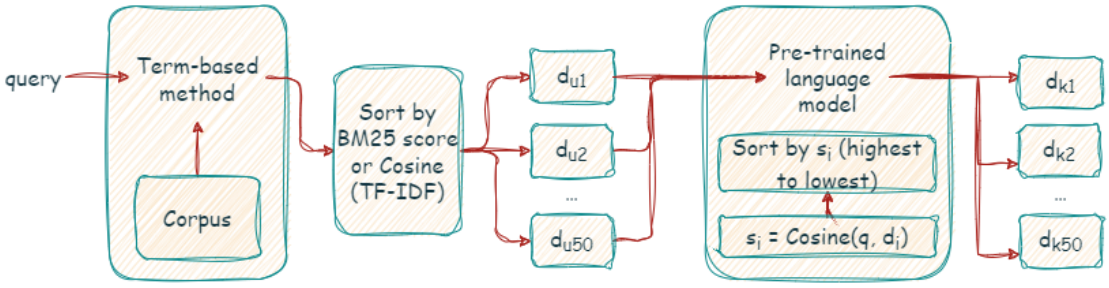
Tương tự như PhoBERT, trong Vietnamese-SBERT, ta sẽ sử dụng vector tương ứng với token [CLS] ở đầu ra của lớp transformer cuối cùng làm feature vector.

Sau khi có biểu diễn feature vector của câu query và 50 bài báo trước đó, quá trình reranking sẽ được thực hiện, sử dụng độ đo **cosine similarity**.

Về vai trò, các phương pháp term-based TF-IDF hay BM25 được sử dụng trước để lọc bớt những bài báo có nội dung không liên quan đến truy vấn, rút ngắn thời gian xử lý cho các pre-trained language model. Tiếp đó, pre-trained language model được sử dụng kết hợp nhằm mục đích khắc phục nhược điểm không nắm bắt được ngữ nghĩa văn bản của các phương pháp term-based. Như

vậy, Multi-stage Document Ranking kết hợp, phát huy ưu điểm và khắc phục nhược điểm của từng phương pháp.

Tổng kết lại, ngữ liệu và truy vấn sau khi được tiền xử lý dữ liệu sẽ đi qua pineline như hình dưới đây.



Hình 3: Pineline xử lý bài toán đặt ra

3.2 Các phương pháp term-based

3.2.1 TF-IDF

TF-IDF (term frequency–inverse document frequency) là độ đo để đánh giá mức độ quan trọng của một từ trong một văn bản. TF-IDF bao gồm 2 thành phần: **TF** - Term Frequency và **IDF** - Inverse Document Frequency.

Cụ thể, với từ khóa t, tài liệu d, ta có:

$$\text{TF}(t, d) = \frac{\text{Số lần từ } t \text{ xuất hiện trong tài liệu } d}{\text{Tổng số từ trong tài liệu } d}$$

Giá trị TF đo lường tần suất xuất hiện của 1 từ trong 1 văn bản. Nếu giá trị TF càng cao thì mức độ liên quan càng cao.

$$\text{IDF}(t) = 1 + \log\left(\frac{\text{Tổng số lượng tài liệu } N}{1 + \text{Số tài liệu chứa từ } t}\right)$$

Trong công thức của IDF, 1 được thêm vào mẫu để tránh lỗi chia cho 0. Còn giá trị 1 ở ngoài là để điều chỉnh cho giá trị IDF của một từ vẫn bằng 1, kể cả khi từ đó xuất hiện trong toàn bộ tập tài liệu, tránh việc giảm quá nhiều mức độ quan trọng của từ đó.

Giá trị IDF phản ánh mức độ phổ biến của từ trong toàn bộ tập tài liệu. Nếu

một từ xuất hiện trong nhiều tài liệu (phổ biến) thì giá trị IDF của nó sẽ thấp và ngược lại.

Kết hợp TF và IDF, ta sẽ xác định được mức độ quan trọng thực sự của từ trong toàn bộ tập tài liệu:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

Thực chất, việc dùng TF-IDF là để biểu diễn từng văn bản và truy vấn dưới dạng một vector, sau đó mới sử dụng một phương thức như L2 hay Cosine similarity để tính toán mức độ tương đồng. Trong đồ án, **cosine similarity** là độ đo được sử dụng.

3.2.2 BM25

BM25 là một hàm xếp hạng giúp xếp hạng một tập các tài liệu dựa vào việc tính xác xuất xuất hiện của các từ khóa trong câu truy vấn. BM25 được cải thiện dựa trên nền tảng của TF-IDF, xem xét thêm một số yếu tố như độ dài tài liệu, các tham số điều chỉnh, ...

Với một truy vấn q chứa các từ khóa $q_1, q_2, ..., q_n$, BM25 score của một tài liệu D được xác định bởi:

$$\text{BM25}(q, D) = \sum_{i=1}^{n} IDF(q_i) \frac{\text{TF}(q_i, D)(k_1 + 1)}{\text{TF}(q_i, D) + k_1(1 - b + b\frac{|D|}{avgdl})}$$

- $f(q_i, D)$ là tần số q_i xuất hiện trong tài liệu D.
- $|D|$ là số lượng các từ trong tài liệu D.
- avgdl là độ dài trung bình số lượng từ của các tài liệu trong ngữ liệu.
- k_1 và b là các tham số điều chỉnh. Thông thường, k_1 nằm trong khoảng $[1.2; 2.5]$ và b được gán bằng 0.75.

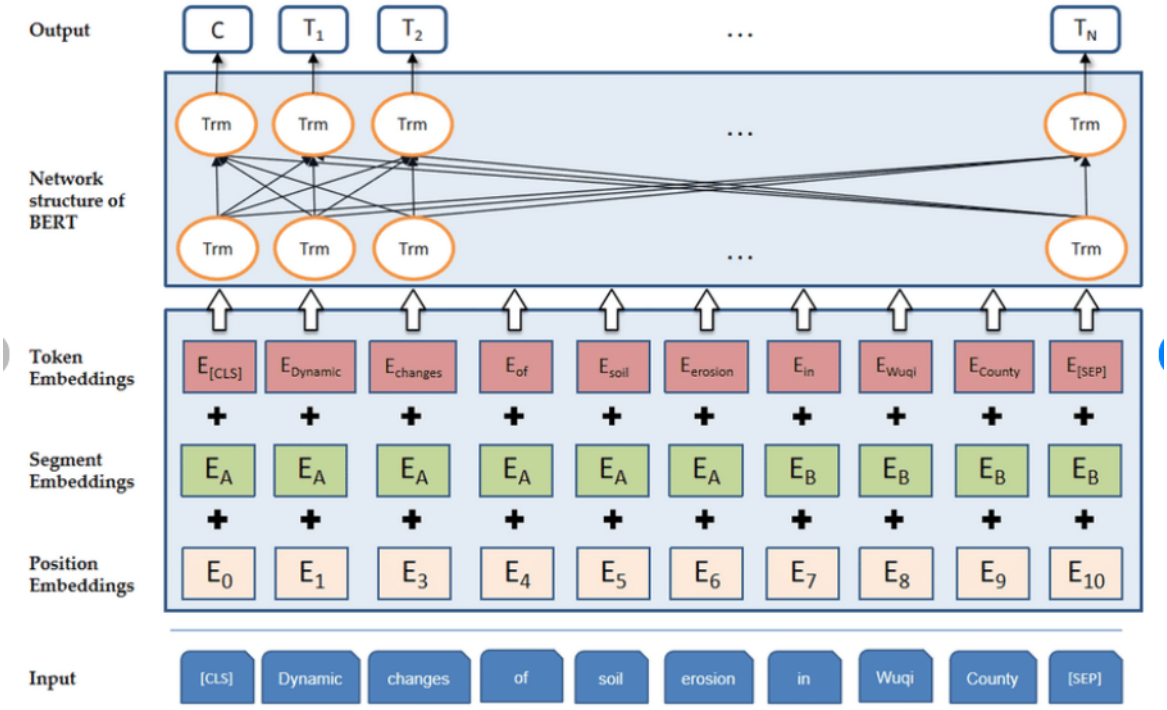
Các bài báo có BM25 score cao có nghĩa là chúng chứa những từ khóa có liên quan đến truy vấn. Ngược lại, các bài báo có BM25 score thấp chứa nội dung không liên quan đến truy vấn và sẽ không được xử lý trong pre-trained language model.

3.3 Pre-trained language model

3.3.1 PhoBERT

PhoBERT là một pre-trained language model được phát triển dựa trên BERT, nhưng được fine tune và tối ưu hóa cho tiếng Việt.

Ta sẽ sử dụng pooled_output của token [CLS] như feature vector cho câu truy vấn và nội dung của các bài báo.

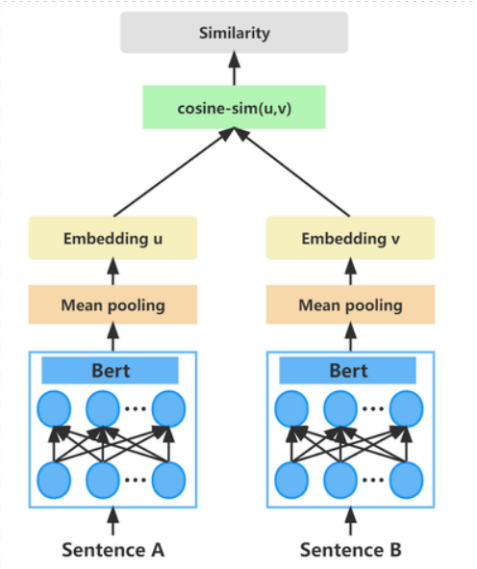


Hình 4: Cấu trúc tổng quát của BERT

3.3.2 Vietnamese-SBERT

Vietnamese-SBERT là một pre-trained language model được sử dụng để biểu diễn văn bản tiếng Việt dưới dạng feature vector, được xây dựng dựa trên mô hình SentenceBERT, cũng là một mô hình biến thể được phát triển dựa trên BERT.

Mục đích chính của mô hình này là để nhận ra được các văn bản đồng nghĩa nhau (nội dung bài báo nào gần nghĩa nhất so với câu truy vấn) thông qua các feature vector (minh họa bởi hình 5).



Hình 5: Quy trình cơ bản của SentenceBERT

4 Cài đặt phương pháp

4.1 Tiền xử lý dữ liệu

Với từng bài báo trong ngữ liệu, dữ liệu dạng text tương ứng với bài báo đó sẽ là kết hợp của title và abstract. Các truy vấn cũng như dữ liệu của các bài báo sẽ đi qua 4 bước tiền xử lý dữ liệu, bao gồm:

- Lowercasing (Di sản → di sản)
- Word segmentation (di sản → di_sản)
- Loại bỏ stopwords
- Loại bỏ các ký tự đặc biệt (tp.hcm → tphcm)

4.2 Đánh giá bài báo bằng Cosine Similarity

Để đánh giá và so sánh độ tương đồng giữa truy vấn và nội dung bài báo, nhóm sử dụng phép đo **cosine similarity** cho tác vụ này. Cụ thể như sau:

- Xét 2 vector u và v là các [CLS] đã được padding về kích thước 768x1 sau khi được xử lý qua multi-stage document ranking, độ tương đồng giữa q và u được tính bằng cách tính cosine của 2 vector này.

$$\textbf{Cosine Similarity}(u, v) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

- Trong đó: $\mathbf{u} \cdot \mathbf{v}$ là tích vô hướng của u và v, $\|\mathbf{u}\|$ và $\|\mathbf{v}\|$ lần lượt là chuẩn 2 của vector u và v. Nếu giá trị CosSimilarity của 2 vector tương ứng với 2 văn bản càng gần 1 thì 2 văn bản này càng đồng nghĩa với nhau và ngược lại, nếu càng gần -1 thì 2 văn bản này càng có ý nghĩa ngược nhau.

5 Đánh giá phương pháp

Thực hiện so sánh độ hiệu quả của hướng tiếp cận Multi-stage document ranking cho truy vấn các bài báo tiếng Việt bằng cách kết hợp một trong các phương pháp term-based (BM25, TF-IDF) cùng với một trong các pre-trained language model (PhoBERT, Vietnamese-SBERT). So sánh tất cả các cặp có thể tạo ra, ghi lại hiệu quả và tốc độ xử lý của từng phương pháp.

Với truy vấn q, ta gọi D_{pred} là danh sách các bài báo đã được sắp xếp từ trên xuống dưới là liên quan với q sau khi sử dụng multi-stage document ranking, D_{truth} là danh sách các bài báo đã được sắp xếp từ trên xuống dưới mà ta thật sự mong muốn. Ta so sánh kết quả giữa mô hình và thực tế dựa trên 10 bài báo đầu tiên bằng $nDCG_{10}$.

Công thức của $nDCG_{10}$ được tính như sau:

$$nDCG_{10} = \frac{DCG_{10}}{IDCG_{10}}$$

trong đó:

- $DCG_{10} = \sum_{i=1}^{10} \frac{rel_i}{\log_2(i+1)}$, rel_i là độ liên quan của bài báo thứ i trong D_{pred}.
- $IDCG_{10} = \sum_{i=1}^{10} \frac{rel_i}{\log_2(i+1)}$, rel_i là độ liên quan của bài báo thứ i trong D_{truth}.
- Giá trị $nDCG_{10}$ nằm trong đoạn $[0; 1]$, càng gần 1 thì mô hình càng tốt.

5.1 Tập đánh giá

Tập đánh giá gồm 300 bài báo đã được xếp hạng và gán nhãn mức độ liên quan (dành cho việc tính $nDCG_{10}$) với truy vấn: "Giá lăn bánh của Vinfast Vf8 năm 2024?"

tittle	abstract	link	rank	score
Giá xe VinFast VF 8 mới nhất tháng 7/2024	Cập nhật giá xe VinFast VF 8 2024 kèm tin khuyến mại, thông	https://oto.com.vn/bang-gia-xe-o-to-vinfast-vf-8-moi-nl	1	4
Bảng giá xe VinFast kèm ưu đãi mới nhất tháng 7/2024	Bảng giá xe VinFast 2024 kèm tin khuyến mại, hình ảnh, videc	https://oto.com.vn/bang-gia-xe-o-to-vinfast-moi-nhat	2	4
Bảng giá xe VinFast tháng 4/2023 kèm ưu đãi	Cập nhật bảng giá xe VinFast mới nhất tháng 4/2023 tại Việt N	https://vinfast.vn/bang-gia-xe-vinfast-thang-4-2023-ker	3	4
Vinfast VF8 2024: Giá xe Vinfast VF8 mới nhất và những thông tin c	Được trang bị nhiều công nghệ hiện đại, Vinfast VF8 hứa hẹn	https://tinxe.vn/gia-xe-vinfast-vf8	4	4
Giá xe VinFast VF 8 tháng 11/2023: TSKT & đánh giá chi tiết	Tham khảo ngay giá lăn bánh mới nhất của mẫu xe Vinfast VF	https://thethao247.vn/364-gia-xe-vinfast-vf-8-d250055.	5	4
Xe Điện VinFast VF8 2024 : Bảng Giá Lăn Bánh, Thông số kĩ thuật &	Cùng VinFast Sài Gòn cập nhập các thông tin mới nhất về xe c	https://oto-vinfastsaigon.com/xe-vinfast-vf8/?gad_sourc	6	4
So sánh Hyundai Ioniq 5 hay VinFast VF8 tầm giá 1,5 tỷ đồng	Hai mẫu xe điện lắp ráp tại thị trường Việt Nam đang được nh	https://www.24h.com.vn/o-to/so-sanh-hyundai-ioniq-5	7	4
Sau 1 năm lăn bánh, VinFast VF8 được rao bán giá khó tin	Những chiếc VinFast VF8 đang được rao bán khá nhiều hiện	https://danviet.vn/sau-1-nam-lan-banh-vinfast-vf8-duo	8	3
Xe điện VinFast VF 8 lăn bánh hơn 1 năm "hét giá" chỉ chưa đến 80	Ở thời điểm hiện tại, những chiếc xe điện đã qua sử dụng của	https://auto5.vn/325-xe-dien-vinfast-vf-8-lan-banh-hon	9	3
SUV điện VinFast VF 8 rớt giá bao nhiêu sau 2 năm lăn bánh?	Một chiếc SUV điện VinFast VF 8 hiện đang được một cơ sở k	https://thethao247.vn/450-suv-dien-vinfast-vf-8-rot-gia	10	3
VinFast VF8 Plus lần đầu xuống giá chỉ hơn 800 triệu đồng sau 1 n	VinFast VF8 Plus là bản cao cấp từng có giá hơn 1 tỷ đồng trê	https://etime.danviet.vn/vinfast-vf8-plus-lan-dau-xuong	11	3
VinFast VF8 mất giá khó tin sau 1 năm lăn bánh	Sau 1 năm lăn bánh, VinFast VF8 hiện đang được rao bán khá	https://auto5.vn/326-vinfast-vf8-mat-gia-kho-tin-sau-1-	12	3
Sau 2 năm lăn bánh, VinFast VF 8 "lượt" lên sàn xe cũ với mức giá r	Theo tham khảo, một chiếc VinFast VF 8 sản xuất năm 2022 c	https://thethao247.vn/419-sau-2-nam-lan-banh-vinfast	13	3

Hình 6: Dữ liệu tập đánh giá

Quy tắc gán nhãn mức độ liên quan được quy định như sau:

- 0: Bài báo không chứa nội dung liên quan đến truy vấn.
- 1: Bài báo chứa thông tin giúp ta có được các nội dung bên lề (một phần các chủ đề) liên quan đến truy vấn.
- 2: Bài báo chứa thông tin hữu ích giúp ta có được các nội dung chứa khía cạnh liên quan đến các chủ đề của truy vấn.
- 3: Bài báo chứa các thông tin mang tính cụ thể hơn so với mức 2 (tức là hơn cả về mặt chủ đề, nó mang đến các thông tin quan trọng về người, sự kiện nào đó...), sao cho giải thích được một phần truy vấn.
- 4: Bài báo chứa các ngữ cảnh quan trọng chứa đựng hoặc giải thích được đầy đủ các thông tin liên quan đến truy vấn.

Mức độ liên quan giữa bài báo ứng với truy vấn *"Giá lăn bánh của vinfast vf8 năm 2024"* được thể hiện trên cột "score"của tập đánh giá. Trong tập đánh giá, có 7 bài báo có relevance score là 4, 8 bài báo là 3, 7 bài báo là 2, 15 bài báo là 1 và 13 bài còn lại là 0.

5.2 Kết quả thu được

Giá trị $nDCG_{10}$ của từng hướng tiếp cận trong bảng dưới (kết quả được làm tròn đến chữ số thập phân thứ 4). Kết quả cao nhất của bảng được **in đậm**. Các kết quả được lấy trung bình trên **10 lần chạy** và **không sử dụng GPU**.

Bảng 1: Kết quả $nDCG_{10}$ của một số phương pháp

Hướng tiếp cận	**Phương pháp**	**nDCG$_{10}$**	**Thời gian thực thi**
Term-based	TF-IDF	0.9090	0.02 s
	BM25	0.8541	0.04 s
Pre-trained language model	Vietnamese-SBERT	0.8399	58.85 s
	PhoBERT	0.9170	60.91 s
Multi-stage document ranking	TF-IDF + Vietnamese-SBERT	0.8399	10.34 s
	TF-IDF + PhoBERT	**0.9513**	11.43 s
	BM25 + Vietnamese-SBERT	0.8002	11.86 s
	BM25 + PhoBERT	0.8268	11.33 s

Nhận xét:

- Hướng tiếp cận **TF-IDF kết hợp PhoBERT** cho kết quả tốt nhất với điểm $nDCG_{10}$ đạt đến **0.9513**
- Có thể thấy khi sử dụng multi-stage document ranking thay vì chỉ sử dụng đơn lẻ pre-trained language model là PhoBERT thì kết quả và tốc độ xử lý đều tốt hơn.
- Phương pháp **TF-IDF** cho điểm $nDCG_{10}$ khá tốt với giá trị **0.91**, đặc biệt hiệu quả khi tốc độ thực thi rất nhanh (chỉ mất **0.02s**).
- Từ đây, ta có nhận xét: khi cần ưu tiên tính real-time (tức là tốc độ trả về output truy vấn phải gần như ngang bằng tốc độ đưa vào input), ta có thể đánh đổi độ chính xác như phương pháp TF-IDF. Ngược lại, nếu cần ưu tiên độ chính xác của kết quả truy vấn, việc kết hợp các pre-trained language model sẽ cho kết quả tối ưu.

6 Tài liệu tham khảo

1. PhoBERT: Pre-trained language models for Vietnamese - https://aclanthology.org/2020.findings-emnlp.92.pdf

2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding -https://arxiv.org/pdf/1810.04805.pdf

3. Improvements to BM25 and Language Models Examined - https://www.cs.otago.ac.nz/homepages/andrew/papers/2014-2.pdf

4. Rank-BM25: A two line search engine - https://pypi.org/project/rank-bm25/

5. IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles -https://arxiv.org/abs/2007.12603

6. Ad-hoc retrieval with BERT -https://arxiv.org/abs/2007.12603

7. Vietnamese-SBERT: https://huggingface.co/keepitreal/vietnamese-sbert

8. TF-IDF là gì?: https://vi.wikipedia.org/wiki/Tf%E2%80%93idf

9. Thực thi TF-IDF: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html