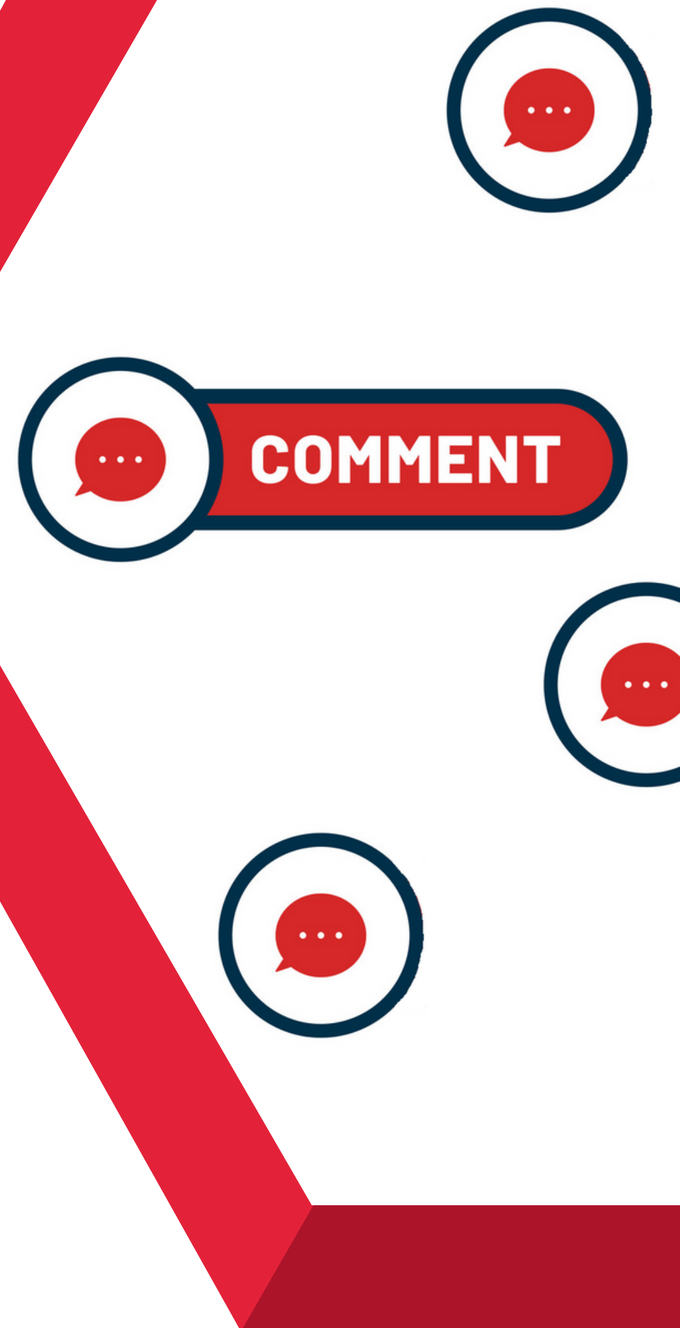
ĐẠI HỌC FPT

KHÓA LUẬN TỐT NGHIỆP

SỬ DỤNG PHƯƠNG PHÁP HỌC ĐỐI LẬP TRONG PHÂN LOẠI KHÍA CẠNH BÌNH LUẬN VỚI DỮ LIỆU TIẾNG VIỆT



SINH VIÊN: LÊ PHƯỚC CƯỜNG, TRỊNH HOÀNG NAM
GIẢNG VIÊN HƯỚNG DẪN : TS. BÙI VĂN HIỆU

MỤC LỤC

ĐẶT VẤN ĐỀ

Tầm quan trọng của thông tin phản hồi của khách hàng trong kinh doanh:

- Là một tài sản quý giá để xây dựng uy tín thương hiệu và tăng lòng tin của người tiêu dùng.
- Tạo cơ hội cho doanh nghiệp giải quyết vấn đề và cải thiện trải nghiệm khách hàng của họ.
- Làm nổi bật những khía cạnh tích cực trong trải nghiệm của khách hàng.
- Tận dụng đánh giá của khách hàng cho SEO và khả năng hiển thị tìm kiếm.

Cô Hằng - Bánh Đa Trộn

Buổi trưa lượn lờ đang tìm kiếm quán ăn thì gặp được ngay quán này. Mình có thể ngồi trong nhà hoặc ngoài vỉa hè này, nhưng mình thích ngồi vỉa hè hơn vì nó thoáng 😂😂 Ở đây có miến trộn, bánh đa trộn, miến nước, bánh đa nước, giá chỉ từ 25k nhá siêu rẻ 💰💰💰 Buổi trưa hơi đông nên gọi đồ phải chờ xíu, nhưng mà bù lại nhân viên rất thân thiện. Một bát miến hoặc bánh đa siêu nhiều, có thịt, đậu, rau muống, giò, xíu gạch, hành khô và không thể thiếu bánh đa (miến)

Số người: 4+ | Chi phí: 50,000đ + | Sẽ quay lại: Chắc chắn

- Đây là nhận xét từ Thành Viên trên Foody, không phải từ Foody Corp. -

Thích Thảo luận Báo lỗi

1. TỔNG QUAN BÀI TOÁN

1.1 Tổng quan bài toán (1)

- Đầu vào: Dữ liệu nhận xét của người dùng trên các nền tảng review online.

- Đầu ra: 1 tập hợp các khía cạnh $Y = \{Y_1, Y_2, ..., Y_k\}$.

  Trong đó:

  - **k** là số lượng các khía cạnh biểu thị nội dung của nhận xét đầu vào.
  - **Y** là tập con của các khía cạnh tiêu chuẩn.

- Trong lĩnh vực đánh giá trên trang đặt đồ ăn online, một số khía cạnh trong câu nhận xét có thể kể đến như: **đồ ăn, không gian, phục vụ, giá cả, khuyến mãi**, **vị trí**,...

1. TỔNG QUAN BÀI TOÁN

1.1 Tổng quan bài toán (2)

Đây là một địa điểm tuyệt vời cho các bạn nào thích ăn vặt nhé. Quán có rất nhiều món món nào cũng ngon, giá cả lại rẻ. Quán có trưng bày nhiều món ăn phía trước tha hồ cho các bạn chọn lựa. Đồ ăn có rất nhanh không phải đợi lâu. Chủ quán cũng rất dễ thương nhiệt tình.

→ **MÔ HÌNH HỌC TƯƠNG PHẢN** →

- ✓ Đồ ăn
- ✓ Giá cả
- ✓ Phục vụ
- ✗ Không gian
- ✗ Vị trí

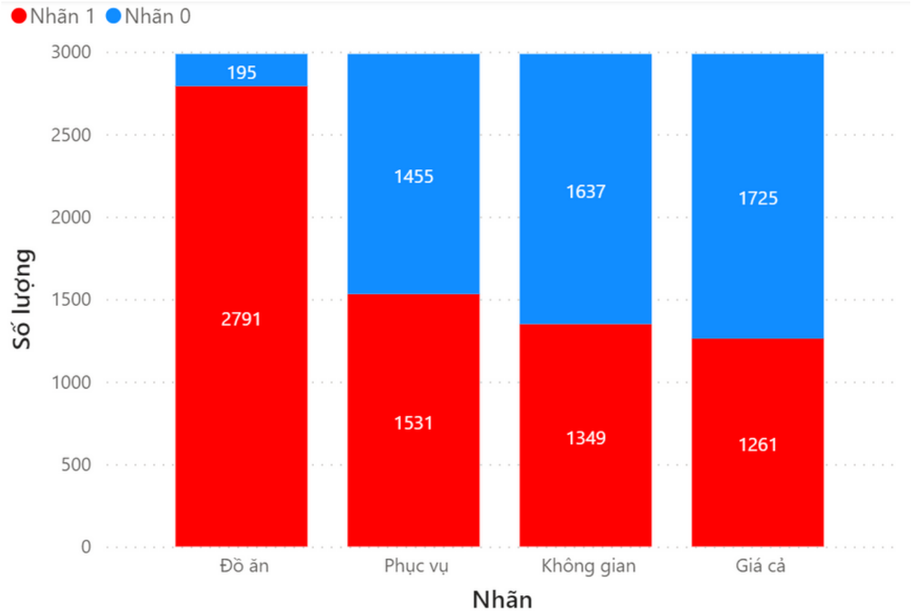
1. TỔNG QUAN BÀI TOÁN

1.2 Tổng quan dữ liệu

- Dữ liệu huấn luyện : ~200.000 bình luận không nhãn.
- Dữ liệu kiểm thử: ~3.000 dữ liệu được đánh nhãn.

Bình luận	Đồ ăn	Giá cả	Không gian	Phục vụ
Cuối tuần đi ăn ở đây mà đông thật, giá cũng hơi cao nhưng chất lượng, món nào cũng tươi á, nhân viên nhiệt tình thân thiện thích nhất là món bạch tuột nướng hihi sẽ ghé thường xuyên	1	1		1
Nem nướng Nha Trang giá rẻ 35k ngon 2 người ăn vừa đủ. Phục vụ có tâm nhiệt tình, quán sạch sẽ	1	1	1	1
Mình có ghé nhà hàng cách đây vài hôm. Nhân viên nhà hàng nhanh nhẹn, không gian sạch thoáng mát. Cảm ơn các bạn đã mang đến cho mình bữa ăn ngon	1		1	1

Một đoạn dữ liệu

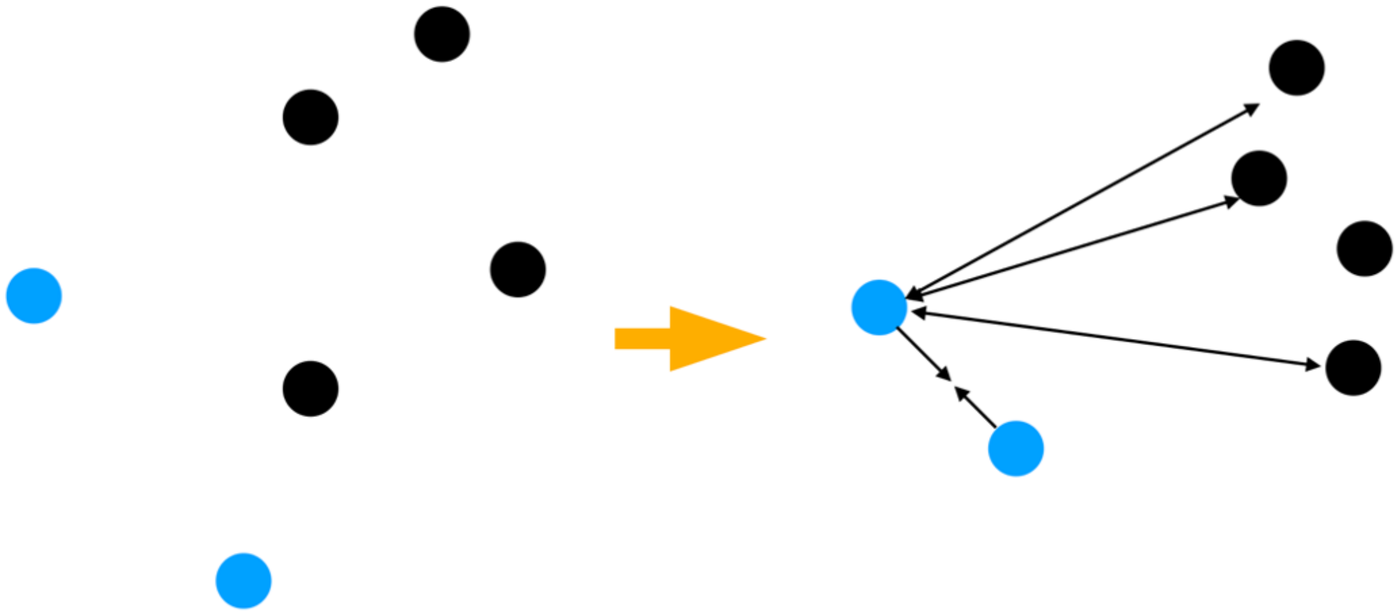


Phân bố dữ liệu

1. TỔNG QUAN BÀI TOÁN

1.3 Tổng quan học tương phản

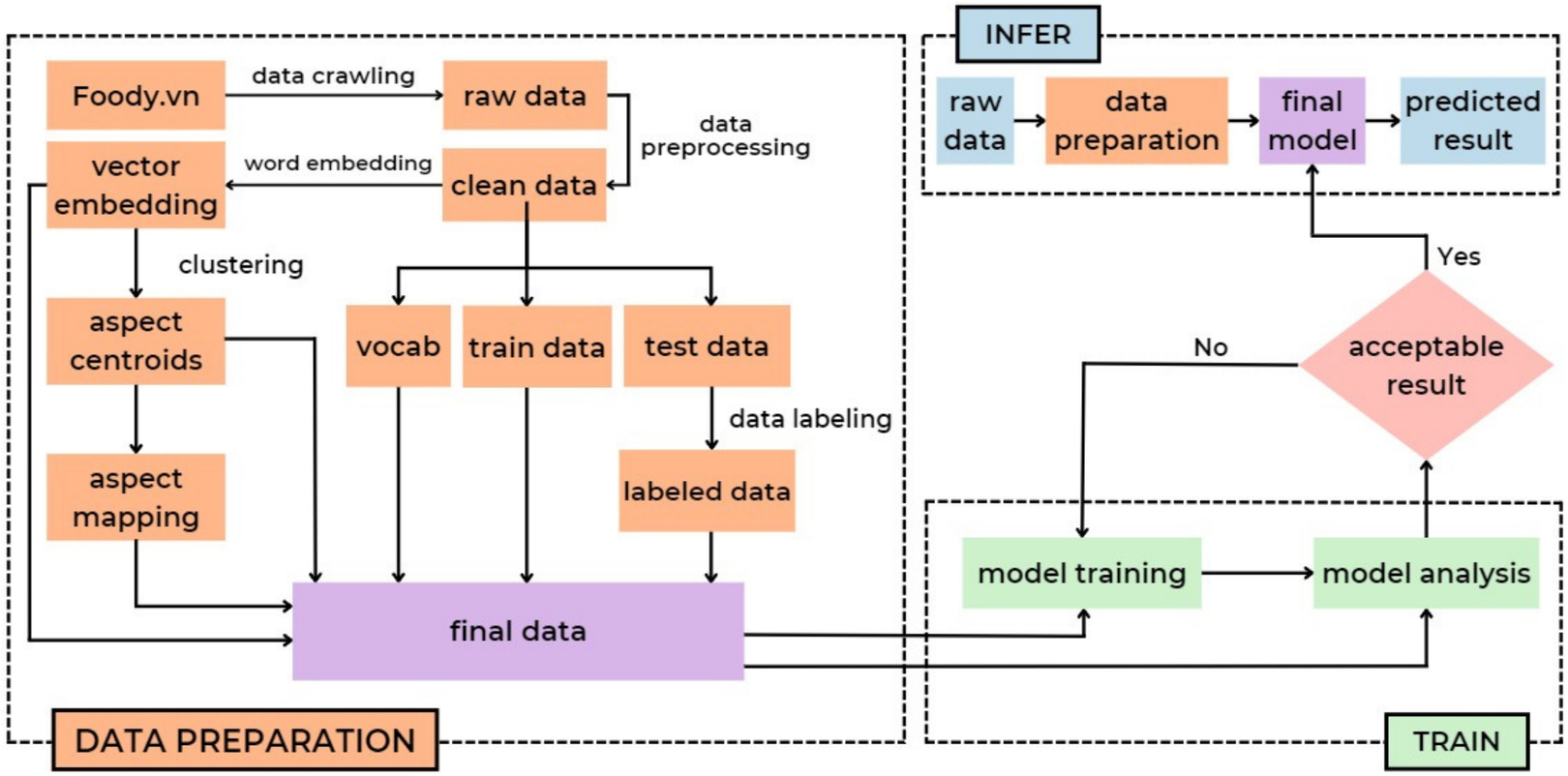
- "Kéo" các dữ liệu tương đồng lại gần nhau.
- "Đẩy" các dữ liệu khác nhau ra xa nhau.



Hình minh họa

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

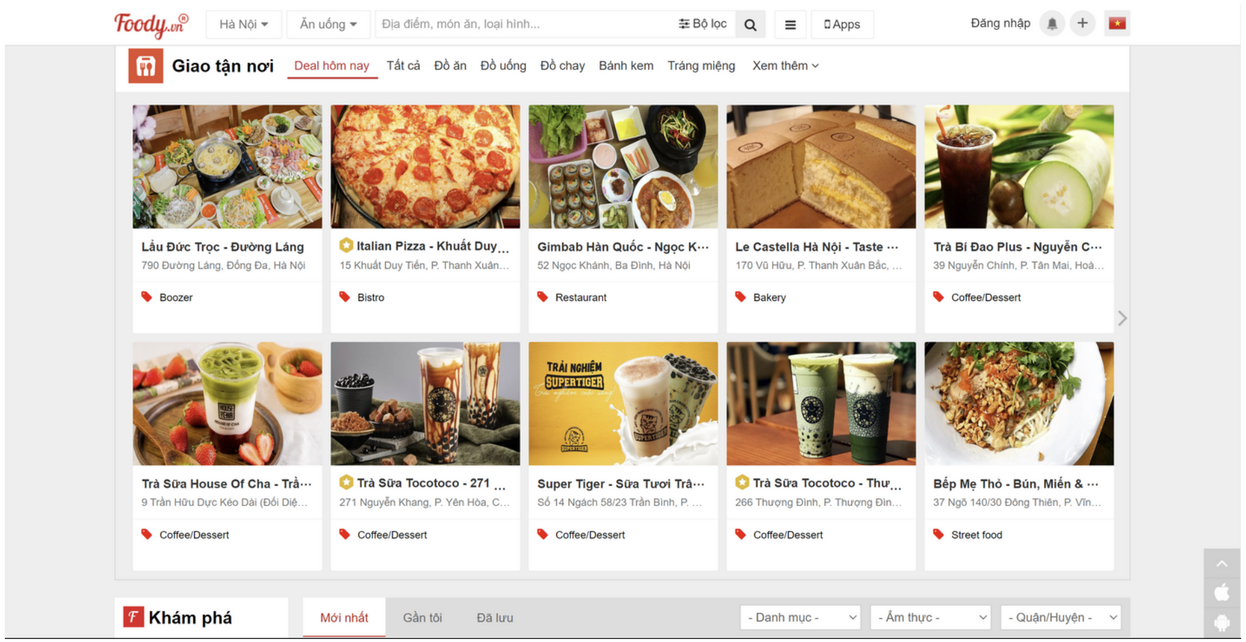
1. Thu thập và xử lý dữ liệu
2. Biểu diễn bộ dữ liệu
3. Biểu diễn bộ khía cạnh
4. Huấn luyện mô hình dựa trên hàm mất mát tương phản
5. Tính toán lại bộ khía cạnh và "ánh xạ" kết quả với bộ khía cạnh tiêu chuẩn

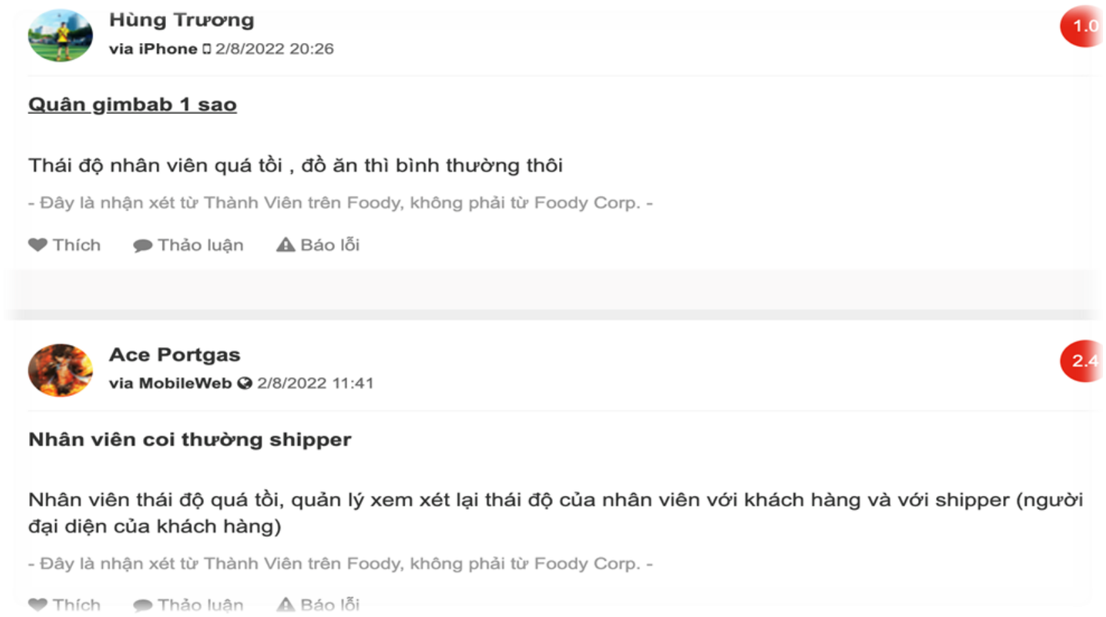


2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.1 Thu thập và xử lý dữ liệu - Thu thập

- Các bình luận về các quán ăn ở trang web ***foody.vn***.
- Là các nhận xét về nhiều chủ đề liên quan đến một quán cụ thể.





2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.1 Thu thập và xử lý dữ liệu - Xử lý (1)

- Đầu vào: Dữ liệu *"raw"* của người dùng bao gồm rất nhiều nhiễu như lỗi chính tả, emoji, hastag, viết tắt, teen code, stop words,...

- Dữ liệu được đưa qua các lớp tiền xử lý nhằm loại bỏ nhiễu trong câu để có thể thực hiện chuyển đổi và đưa vào huấn luyện mô hình nhằm đạt kết quả chính xác hơn.

- Đầu ra: Dữ liệu *"clean"* sẵn sàng cho các bước tiếp theo.

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.1 Thu thập và xử lý dữ liệu - Xử lý (2)

"Ấn tượng cực kỳ với không gian hoài cổ của quán , giống kiểu thời còn bao cấp ấy 😋 Tầng 2 view đẹp , mát mẻ . ❤️
Nhân viên nhanh nhẹn nhiệt tình 😘😍
Đồ uống giá hơi cao , chất lượng khá bình thường 😓
Mình gọi trà quất bạc hà gì đó không nhớ rõ nhưng uống khá nhạt , đc cái mùi bạc hà thơm 👌👌"

Độ dài: 71

- Remove emoji
- Remove hastags
- Remove stopwords
- Format punctuation
- Format price
- Word segmentation

ấn_tượng cực_kỳ không_gian hoài cổ quán kiểu thời bao_cấp tầng 2 view đẹp mát_mẻ nhân_viên nhanh_nhẹn nhiệt_tình đồ uống giá hơi chất_lượng bình_thường gọi trà quất bạc_hà uống nhạt mùi bạc hà_thơm

Độ dài: 31

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.2 Biểu diễn bộ dữ liệu

- Sử dụng các mô hình nhúng từ (embedding) để biểu diễn các từ.
- Một vài mô hình: FastText, Word2Vec, BERT, .v.v.
- Tính vector biểu diễn của cả câu : $s_{x,E} = \sum_{t=1}^{T} \alpha_t E_{x_t}$
- Trong đó α_t là trọng số của các từ : $\alpha_t = \frac{e^{u_t}}{\sum_{t=1}^{T} e^{u_t}}$

với : $u_t = \lambda \tanh\left(q^\intercal \left(W_E E_{x_t} + b_E\right)\right)$

- **W** và **b** là các tham số được huấn luyện.

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.3 Xác định và biểu diễn bộ khía cạnh - Xác định (1)

	BỘ KHÍA CẠNH DỰ ĐOÁN	BỘ KHÍA CẠNH TIÊU CHUẨN
Ý NGHĨA	Xây dựng với mục đích huấn luyện	Là đầu ra mong muốn
CÁCH XÁC ĐỊNH	Tiến hành phân cụm bộ từ điển	Lọc phân cụm bộ khía cạnh dự đoán
SỐ LƯỢNG	Tương đương số cụm sau khi tiến hành phân cụm. Có thể trùng lặp nhiều cụm thể hiện chung 1 khía cạnh.	Bộ khía cạnh duy nhất trong k cụm
VÍ DỤ	Đồ ăn, Giá cả, Đồ ăn, None, Giá cả, None	Đồ ăn, Giá cả

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.3 Xác định và biểu diễn bộ khía cạnh - Xác định (2)

- Phân cụm dữ liệu biểu diễn của bộ từ điển với **k=30**.
- Gán khía cạnh phù hợp cho các cụm.

Hình minh họa các cụm

K=1	Đồ ăn
K=2	Không gian
K=3	None
K=4	None
K=5	Đồ ăn
K=6	Phục vụ
K=7	Giá cả

Hình minh họa gán nhãn các cụm

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.3 Xác định và biểu diễn bộ khía cạnh - Biểu diễn

- Xây dựng biểu diễn khía cạnh cho mỗi bình luận.
- Tính vector biểu diễn của cả câu : $s_{x,A} = \sum_{n=1}^{N} \beta_n A_n$
- Trong đó β_n là trọng số của các khía cạnh dự đoán:

$$\beta_n = \frac{\exp\left(v_{n,A}^{\top} s_{x,E} + b_{n,A}\right)}{\sum_{\eta=1}^{N} \exp\left(v_{\eta,A}^{\top} s_{x,E} + b_{\eta,A}\right)}$$

- $v_{n,A}$ và $b_{n,A}$ là các tham số được huấn luyện.

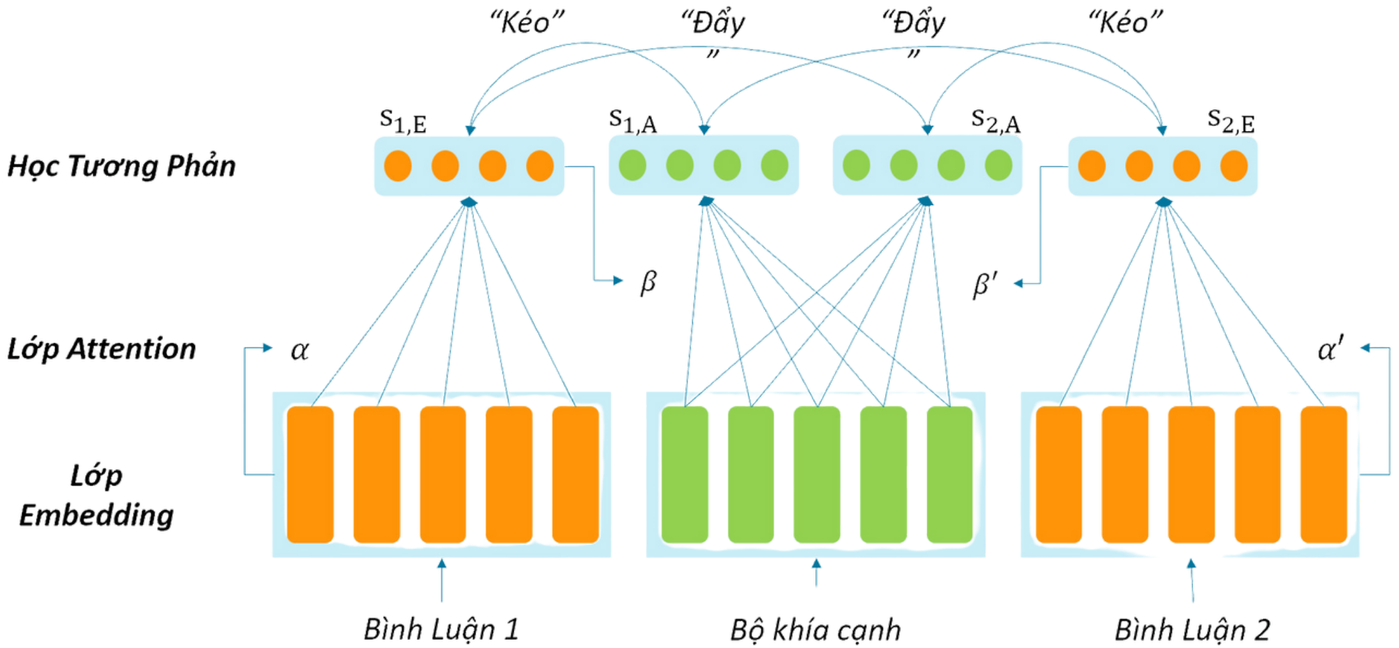
2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.4 Huấn luyện mô hình (1)

Hàm mất mát tương phản:

$$\mathcal{L} = -\log\left(\frac{\exp\left(sim(p, p_2)\right)}{\sum_{k=1}^{n} \exp\left(sim(p, N_k)\right)}\right)$$

- p : điểm dữ liệu đang xét
- p_2 : dữ liệu tương đồng với p
- N_k : các điểm dữ liệu tương phản



2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.4 Huấn luyện mô hình (2)

- Sử dụng độ đo tương tự Cosine để tối ưu độ tương tự giữa bình luận và bộ khía cạnh:

$$\mathrm{sim}(s_{j,E}, s_{i,A}) = \frac{(s_{j,E})^\top s_{i,A}}{\|s_{j,E}\| \|s_{i,A}\|}$$

- Tính giá trị mất mát bằng hàm mất mát tương phản:

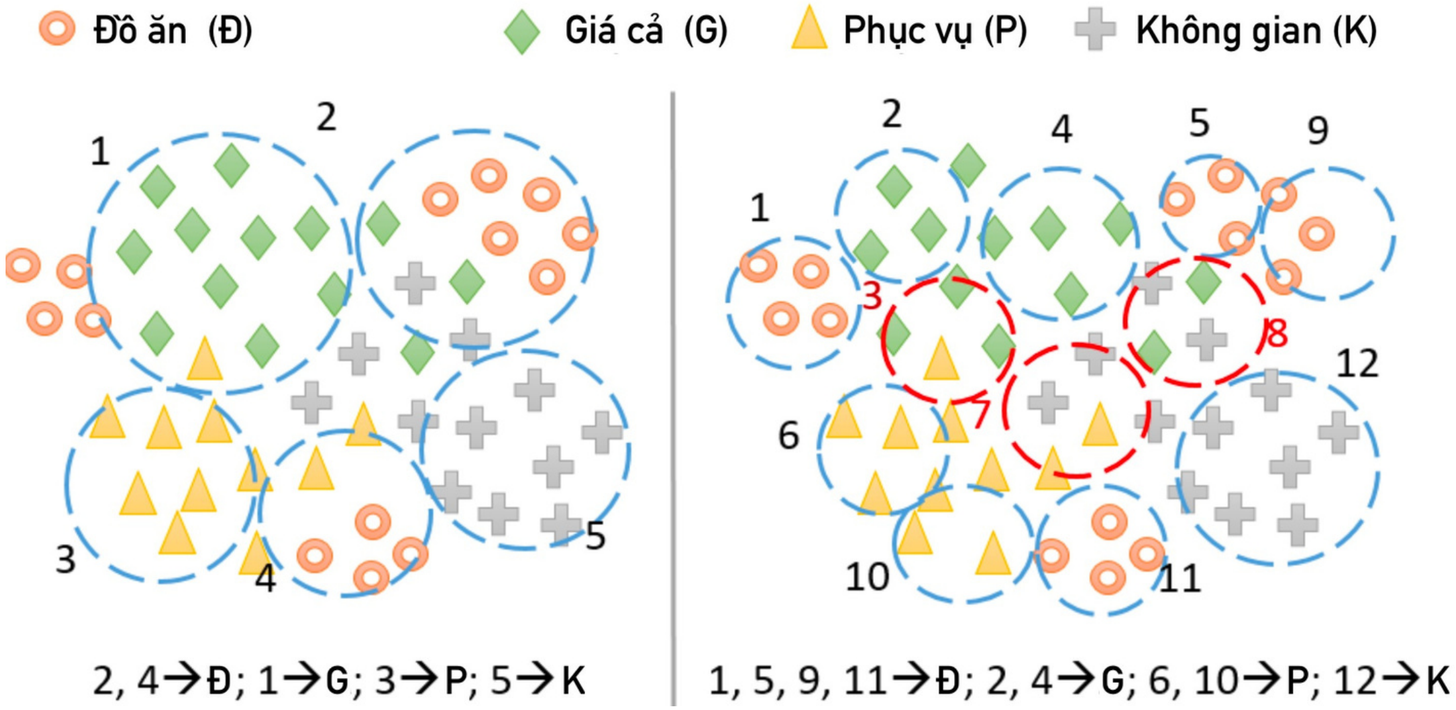
$$l_i = -\log \frac{\exp\left(\mathrm{sim}(s_{i,E}, s_{i,A})/\mu\right)}{\sum_{j=1}^{X} \mathbb{I}_{[j \neq i]} \exp\left(\mathrm{sim}(s_{j,E}, s_{i,A})/\mu\right)}$$

- Cập nhật các tham số của bài toán.

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.5 Ánh xạ kết quả dự đoán

- Mô hình dự đoán nhãn là các bộ khía cạnh dự đoán (Các cụm sau khi tiến hành phân cụm).
- Tiến hành "***ánh xạ***" các cụm về bộ khía cạnh tiêu chuẩn.



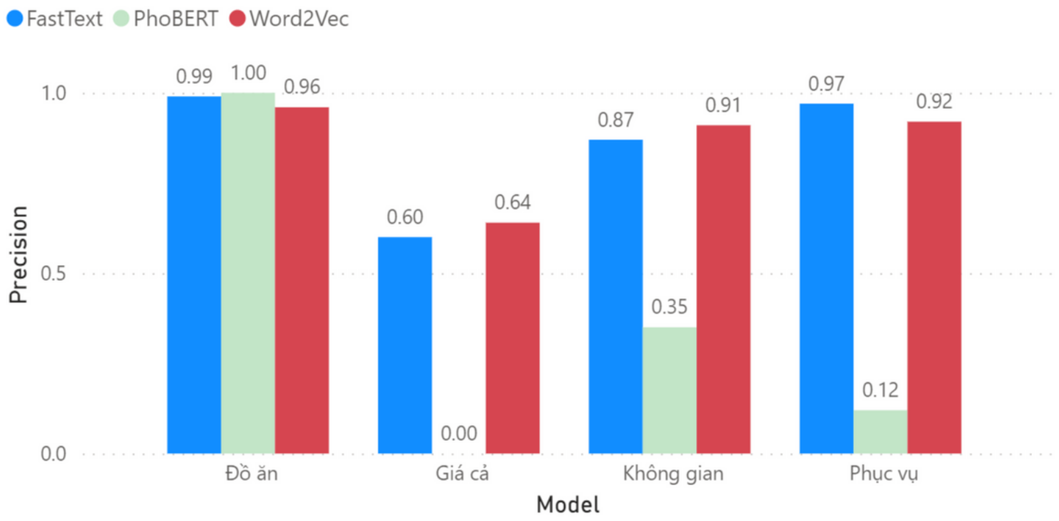
3. KẾT QUẢ

Bộ tham số huấn luyện:

ĐỘ ĐO	Precision, Recall, F1-Score	
BỘ THAM SỐ HUẤN LUYỆN	Epoch	[2, 5, 10]
	Batch size	[32, 64, 128]
	Learning rate	[0.0001, 0.0002, 0.0003, 0.0004, 0.0005]
	Smooth factor	[0.75, 0.8, 0.85, 0.9, 0.95]
	Threshold	[0.5, 1, 1.5, 2, 2.5, 3, 3.5]

3. KẾT QUẢ

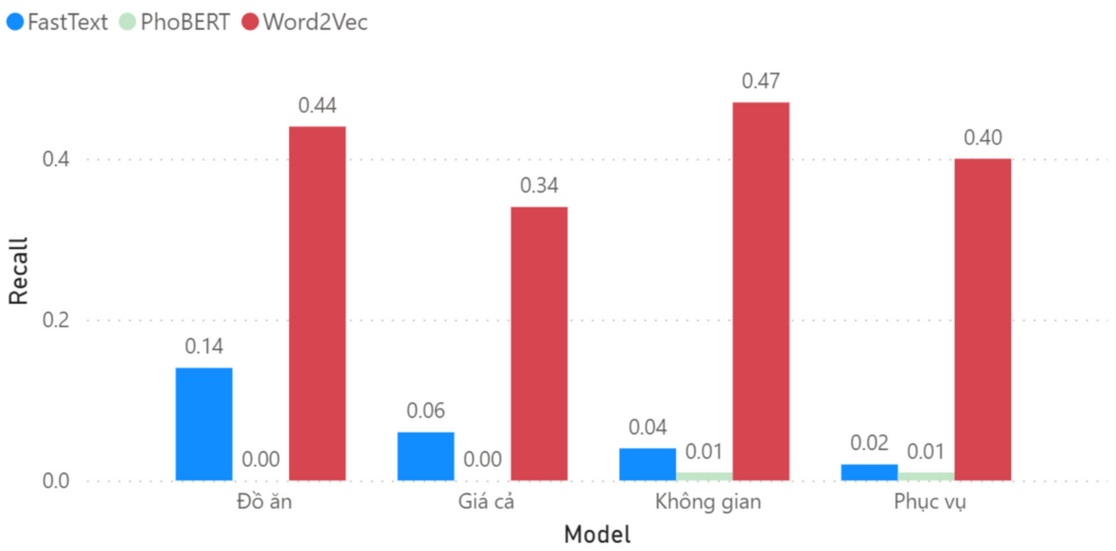
Các mô hình nhúng từ (1)



- Cả **3 mô hình** đều đạt kết quả ***precision*** rất cao với khía cạnh "**Đồ ăn**".
- ***Precision*** của **PhoBERT** đạt kết quả **thấp nhất** trong cả 3 mô hình.
- **Word2Vec** và **FastText** đều đạt ***precision*** **rất cao**, có đa số khía cạnh khoảng 0.9.

3. KẾT QUẢ

Các mô hình nhúng từ (2)



- Trái với *precision*, ***recall*** của **PhoBERT** và **FastText** cho kết quả **rất thấp**.
- Tương tự với *precision*, ***recall*** của **PhoBERT** đạt quả **thấp nhất** trong cả 3 mô hình.
- **Word2Vec** đạt kết ***recall*** **tương đối ổn** với trung bình khoảng 0.41.

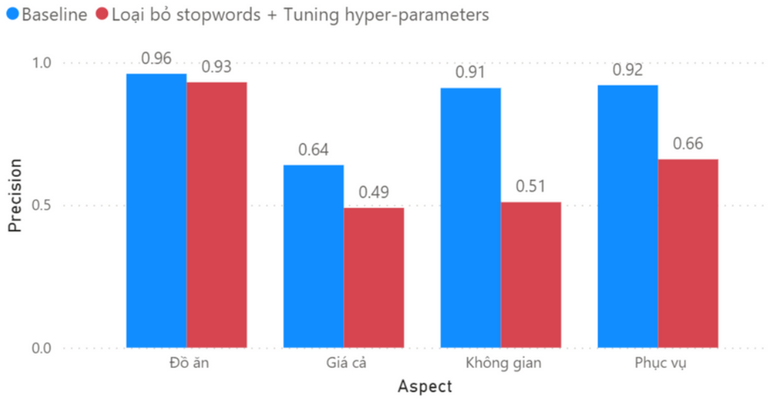
3. KẾT QUẢ

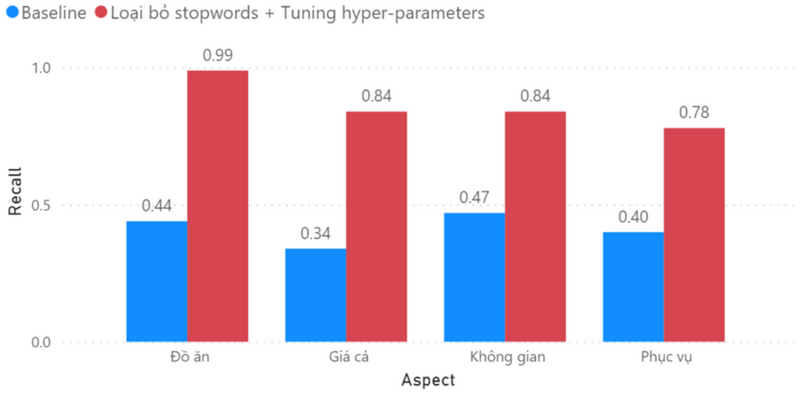
Các mô hình nhúng từ (3)

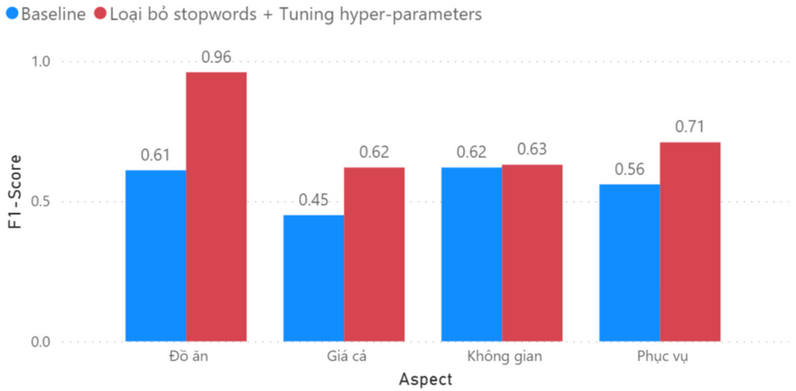
- Với *F1-score*, chúng ta có thể thấy được kết quả tổng thể của 3 mô hình.
- **Word2Vec** cho kết quả **vượt trội hơn** so với 2 mô hình còn lại với trung bình *F1-score* khoảng **0.56**.
- **PhoBERT** và **FastText** đều cho kết quả **rất thấp** trong đó PhoBERT có *F1-score* đạt 0.

3. KẾT QUẢ

Tinh chỉnh các yếu tố và tham số



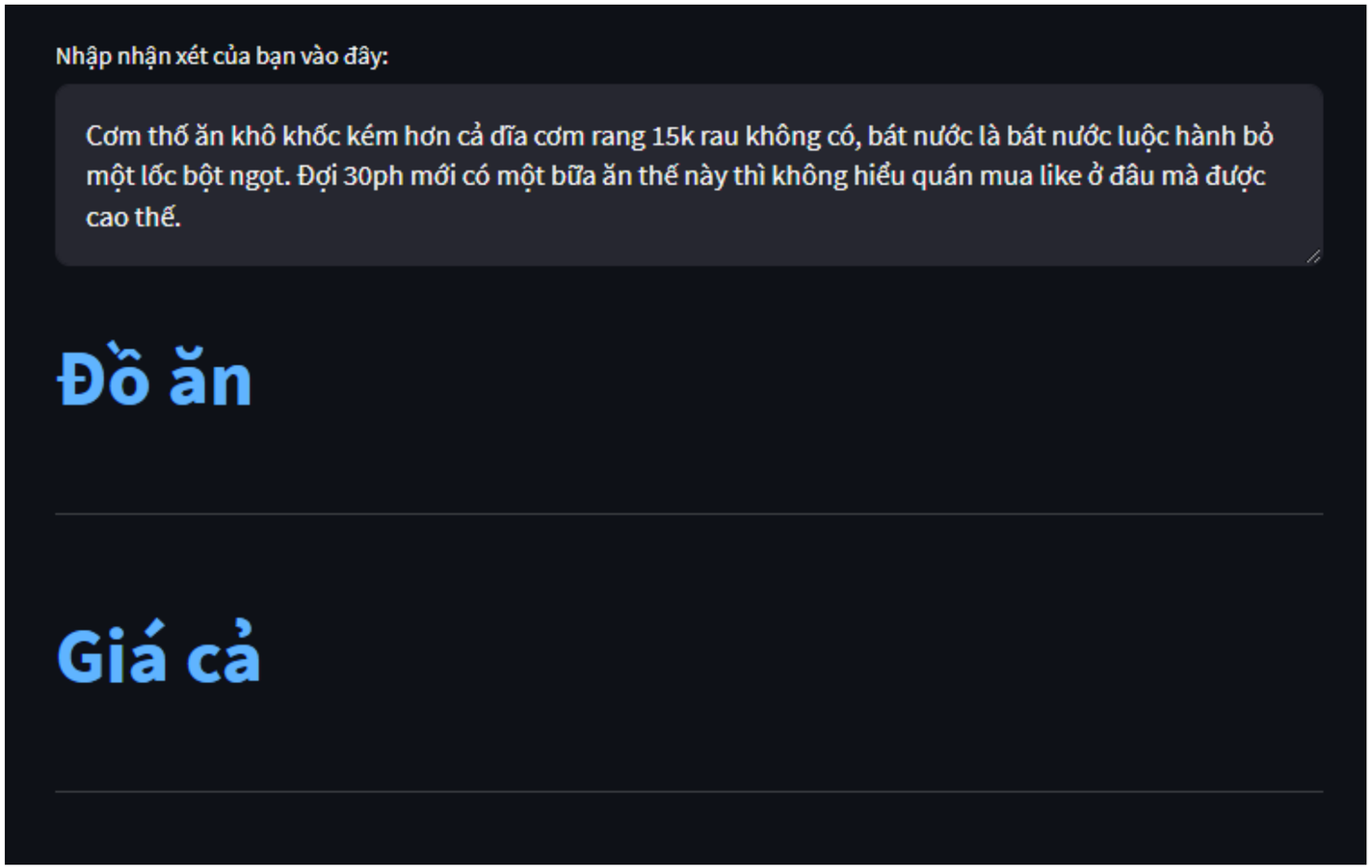




- 2 bước chính: ***Loại bỏ stopwords*** và ***tinh chỉnh bộ siêu tham số*** của mô hình.
- Chọn ***F1-score*** làm độ đo cuối, do đó sau khi tinh chỉnh thì mặc dù ***precision*** của mô hình có giảm nhưng ***recall*** của mô hình tăng lên rất nhiều.
- Kết quả ***F1-score*** trung bình tăng từ khoảng **0.56** lên **0.73**.

3. KẾT QUẢ

Triển khai mô hình thử nghiệm



https://huggingface.co/spaces/strongpear/Vietnamese-aspect-detection

4. KẾT LUẬN

- Khóa luận trình bày về phương pháp *học tương phản* và các kỹ thuật liên quan cần sử dụng trong việc ứng dụng học tương phản vào bài toán với dữ liệu không nhãn.

- Việc nhúng dữ liệu đầu vào có ảnh hưởng lớn đến kết quả dự đoán của mô hình.

- Phương pháp *học tương phản* hoàn toàn có thể được sử dụng rộng rãi hơn cho các bài toán với dữ liệu *không nhãn* khi tập dữ liệu đầu vào đủ lớn.

DEMO

ĐẠI HỌC FPT

KHÓA LUẬN TỐT NGHIỆP

SỬ DỤNG PHƯƠNG PHÁP HỌC ĐỐI LẬP TRONG PHÂN LOẠI KHÍA CẠNH BÌNH LUẬN VỚI DỮ LIỆU TIẾNG VIỆT

SINH VIÊN: LÊ PHƯỚC CƯỜNG, TRỊNH HOÀNG NAM
GIẢNG VIÊN HƯỚNG DẪN : TS. BÙI VĂN HIỆU

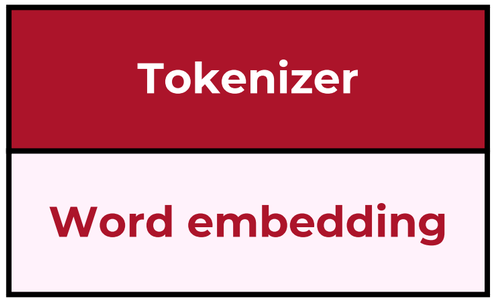
PHỤ LỤC

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.1 Thu thập và xử lý dữ liệu - Xử lý

Tokenizer + Word embedding

ngon sạch_sẽ giá_cả
phải_chăng hơn_nữa
gia_vị bột ngọt_thức uống
quán món yêu_cầu quán
lưu_ý nhân_viên si_rô quán
hầu_như nấu đặc_biệt ép
hề pha nha



[6.39828728e-01
6.68903037e-01
2.96384699e+01
8.51439517e+01
.....
2.57813867e+01
7.18937947e-01]

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.3 Xác định và biểu diễn bộ khía cạnh - Xác định

Minh hoạ các cụm

FastText

PhoBERT

2. QUY TRÌNH XỬ LÝ BÀI TOÁN

2.3 Xác định và biểu diễn bộ khía cạnh - Xác định

Xác định bộ khía cạnh dự đoán

K=1	Đồ ăn
K=2	Không gian
K=3	None
K=4	None
K=5	Đồ ăn
K=6	Phục vụ
K=7	Giá cả

Word2Vec

K=1	None
K=2	Đồ ăn
K=3	None
K=4	None
K=5	Phục vụ
K=6	None
K=7	None

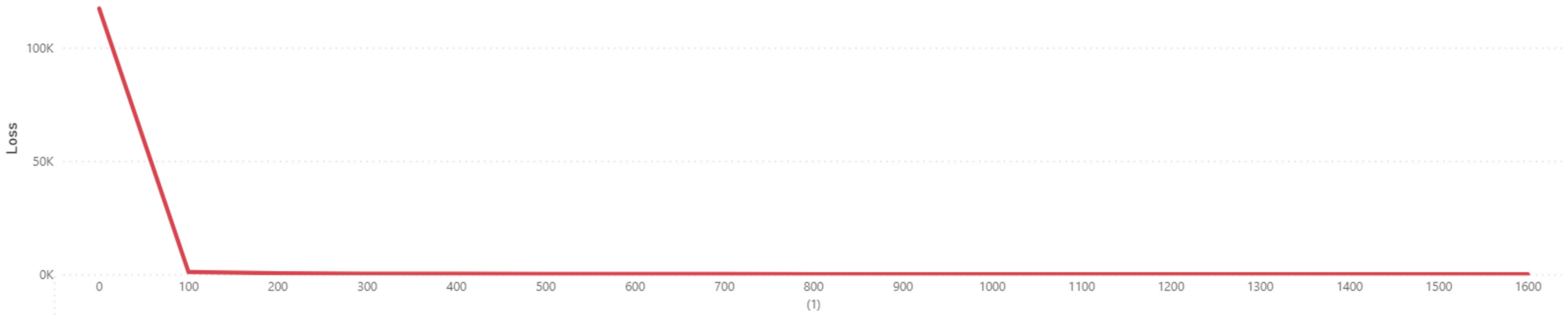
FastText

K=1	Đồ ăn
K=2	None
K=3	None
K=4	None
K=5	Giá cả
K=6	None
K=7	None

PhoBERT

3. KẾT QUẢ

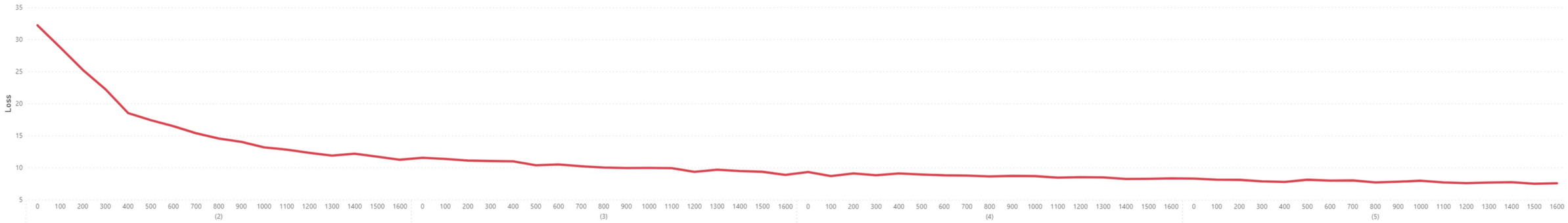
Biến thiên giá trị Loss (1)



*Hình minh họa **Loss** của epoch đầu tiên*

3. KẾT QUẢ

Biến thiên giá trị Loss (2)



*Hình minh họa **Loss** của 4 epoch còn lại*

3. KẾT QUẢ

Loại bỏ stopwords