BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

CS2205 - PHƯƠNG PHÁP LUẬN NGHIÊN CỨU KHOA HỌC

Lớp học

CS2205.xxx

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

09/2024 - 12/2024

----- Trang này cố tình để trống -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  (ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)
- Link slides (dạng .pdf đặt trên Github của nhóm):
  (ví dụ: https://github.com/mynameuit/CS2205.xxxTenDeTai.pdf)
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

- Họ và Tên: Hoàng Sơn Kim
- MSSV: 12345

- Lớp: 2205.xxx
- Tự đánh giá (điểm tổng kết môn): 7.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 15
- Link Github:
  https://github.com/mynameuit/CS2205.xxx/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

ĐẾM SỐ NGƯỜI ĐEO KHẨU TRANG DÙNG DEEP LEARNING

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

TÓM TẮT *(Tối đa 400 từ)*

GIỚI THIỆU *(Tối đa 1 trang A4)*

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

NỘI DUNG VÀ PHƯƠNG PHÁP

KẾT QUẢ MONG ĐỢI

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, Pheng-Ann Heng:
Single-Stage Instance Shadow Detection With Bidirectional Relation Learning.
CVPR 2021: 1-11

----- Trang này cố tình để trống - Các nhóm copy & paste bài làm của mình vào trang tiếp theo -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/1SQCUXqPx5A
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/LuongPhan/CS2205.CH183/Slide.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: **Phan Lực Lượng**
- MSSV: **230101011**

- Lớp: **CS2205.CH183**
- Tự đánh giá (điểm tổng kết môn): 8.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 5
- Link Github: https://github.com/LuongPhan/CS2205.CH183

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN TÍCH CẢM XÚC DỰA TRÊN KHÍA CẠNH CỦA LAPTOP

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ASPECT-BASED SENTIMENT ANALYSIS OF LAPTOP

TÓM TẮT *(Tối đa 400 từ)*

Trong thời đại bùng nổ dữ liệu trực tuyến, việc phân tích và khai thác thông tin từ phản hồi của khách hàng trên các nền tảng thương mại điện tử đóng vai trò quan trọng trong việc hiểu rõ nhu cầu và xu hướng tiêu dùng. Dự án này hướng đến việc xây dựng một bộ dữ liệu chất lượng cao phục vụ cho bài toán phân tích cảm xúc khía cạnh (Aspect-Based Sentiment Analysis - ABSA), nhằm hỗ trợ nghiên cứu xử lý ngôn ngữ tự nhiên (NLP) và ứng dụng trong lĩnh vực trí tuệ nhân tạo. Dự án tập trung vào hai nhiệm vụ cốt lõi: (1) Xây dựng bộ dữ liệu từ phản hồi thực tế của khách hàng, trải qua quy trình thu thập, làm sạch và gán nhãn một cách có hệ thống để đảm bảo độ chính xác và tính đa dạng của dữ liệu; (2) Ứng dụng các kỹ thuật học máy tiên tiến, bao gồm mô hình phoBERT để nâng cao hiệu suất phân tích cảm xúc khía cạnh, tối ưu hóa độ chính xác trong nhận diện các quan điểm và đánh giá của người dùng trong tiếng Việt. Với cách tiếp cận kết hợp giữa việc xây dựng bộ dữ liệu chất lượng cao và khai thác sức mạnh của học máy, nghiên cứu này không chỉ cung cấp một tập dữ liệu tiêu chuẩn cho cộng đồng khoa học mà còn tạo tiền đề cho các ứng dụng AI trong phân tích hành vi khách hàng. Kết quả từ dự án sẽ đóng góp đáng kể vào sự phát triển của hệ sinh thái AI tại Việt Nam, đặc biệt trong lĩnh vực thương mại điện tử và dịch vụ khách hàng, đồng thời thúc đẩy các nghiên cứu chuyên sâu trong xử lý ngôn ngữ tự nhiên.

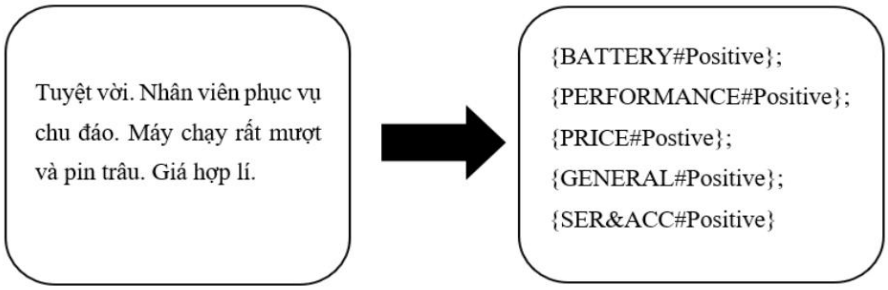
GIỚI THIỆU *(Tối đa 1 trang A4)*

Phân tích cảm xúc (Sentiment Analysis - SA) là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), đóng vai trò thiết yếu trong nhiều ứng dụng thực tế như đánh giá sản phẩm, dịch vụ khách hàng, phân tích thị trường và giám sát thương hiệu. Các phương pháp SA truyền thống thường chỉ tập trung vào việc phân loại cảm xúc chung của một văn bản mà không thể xác định được cảm xúc cụ thể đối với từng khía cạnh của sản phẩm hay dịch vụ. Điều này dẫn đến những hạn chế trong việc hiểu sâu hơn về phản hồi của khách hàng.

Phân tích cảm xúc khía cạnh (Aspect-Based Sentiment Analysis - ABSA) [1] mở rộng khả năng của SA bằng cách phân tích cảm xúc đối với từng thành phần riêng biệt của một đối tượng. Ví dụ, một khách hàng có thể đánh giá tích cực về hiệu suất của một chiếc laptop nhưng lại không hài lòng về thời lượng pin của nó. Việc tách biệt cảm xúc theo từng khía cạnh giúp doanh nghiệp có cái nhìn chi tiết hơn về những ưu điểm và nhược điểm của sản phẩm. Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, việc phân tích phản hồi khách hàng trở thành một công cụ quan trọng giúp doanh nghiệp điều chỉnh chiến lược kinh doanh, cải thiện sản phẩm và dịch vụ để đáp ứng tốt hơn nhu cầu thị trường. Tuy nhiên, nghiên cứu về ABSA trong tiếng Việt vẫn còn nhiều thách thức do thiếu hụt các bộ dữ liệu chuẩn hóa và hệ thống phân tích phù hợp với đặc trưng ngôn ngữ. Dự án này nhằm khắc phục những hạn chế đó bằng cách xây dựng một bộ dữ liệu phản hồi khách hàng chuẩn hóa bằng tiếng Việt phục vụ cho nghiên cứu ABSA

Input: Một câu bình luận về laptop

Output: Khía cạnh được nhắc đến trong câu bình luận và cảm xúc của khía cạnh đó



MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Xây dựng bộ dữ liệu phản hồi khách hàng chuẩn hóa bằng tiếng Việt: Thu thập, xử lý và gán nhãn bộ dữ liệu phản hồi khách hàng từ các nền tảng thương mại điện tử, giúp chuẩn hóa dữ liệu phục vụ cho nghiên cứu ABSA. Bộ dữ liệu này sẽ đảm bảo tính đa dạng và chất lượng cao để đáp ứng yêu cầu của các mô hình học máy.
- Ứng dụng các mô hình học máy tiên tiến để phân tích cảm xúc khía cạnh: Triển khai và tối ưu hóa các thuật toán học sâu như phoBERT để nâng cao hiệu suất trong nhận diện cảm xúc trên từng khía cạnh cụ thể. Đồng thời, so sánh với các mô hình truyền thống nhằm đánh giá hiệu quả của phương pháp đề xuất.
- Đóng góp vào cộng đồng nghiên cứu NLP và ứng dụng thực tế: Công khai bộ dữ liệu và mô hình nghiên cứu để hỗ trợ cộng đồng nghiên cứu trong lĩnh vực NLP. Bên cạnh đó, cung cấp các giải pháp phân tích phản hồi khách hàng giúp doanh nghiệp cải thiện sản phẩm và tối ưu hóa chiến lược kinh doanh dựa trên dữ liệu thực tế.

NỘI DUNG VÀ PHƯƠNG PHÁP

1 Xây dựng bộ dữ liệu phản hồi khách hàng

Nội dung: Xây dựng một tập dữ liệu chuẩn hóa phản hồi khách hàng để phục vụ bài toán phân tích cảm xúc khía cạnh.

Phương pháp thực hiện:

- Thu thập dữ liệu: Lấy dữ liệu từ các nền tảng thương mại điện tử như Shopee, Lazada, Tiki,... và các mạng xã hội.
- Tiền xử lý dữ liệu: Loại bỏ dữ liệu nhiễu và gán nhãn cho dữ liệu.
- Gán nhãn dữ liệu: Xác định các khía cạnh chính trong phản hồi khách hàng như hiệu suất, màn hình, bàn phím, giá cả, pin,...

- Kiểm định chất lượng dữ liệu: Kiểm tra mức độ đồng thuận giữa các chuyên gia gán nhãn để đảm bảo độ tin cậy của dữ liệu bằng độ đo Cohen`s Kappa [3].

2 Ứng dụng học máy trong phân tích cảm xúc khía cạnh

Nội dung: Xây dựng mô hình phân tích cảm xúc dựa trên học máy và học sâu để xác định cảm xúc của từng khía cạnh trong phản hồi khách hàng.

Phương pháp thực hiện:

- Huấn luyện mô hình học máy: Sử dụng phoBERT [2] để trích xuất đặc trưng từ phản hồi khách hàng.
- So sánh với các phương pháp khác: Đánh giá hiệu suất mô hình dựa trên các chỉ số như Precision, Recall, F1-score.
- Cải thiện mô hình: Tinh chỉnh siêu tham số, điều chỉnh tập dữ liệu đầu vào nhằm nâng cao độ chính xác của hệ thống.

3.3 Đánh giá và ứng dụng thực tế

Nội dung: Kiểm tra hiệu suất mô hình và áp dụng vào các ứng dụng thực tế.

Phương pháp thực hiện:

- Kiểm thử mô hình trên tập dữ liệu thực tế để đánh giá độ chính xác và khả năng tổng quát hóa của hệ thống.
- Ứng dụng trong thương mại điện tử: Phát triển công cụ hỗ trợ doanh nghiệp theo dõi phản hồi khách hàng và cải thiện chất lượng sản phẩm.
- Phát hành bộ dữ liệu: Công khai bộ dữ liệu nhằm hỗ trợ cộng đồng nghiên cứu trong lĩnh vực xử lý ngôn ngữ tự nhiên.

KẾT QUẢ MONG ĐỢI

Dự án này hướng đến những kết quả mang lại giá trị thực tiễn và khoa học cao:

- Xây dựng một bộ dữ liệu chuẩn hóa có chất lượng cao**,** phản ánh chân thực phản hồi của khách hàng, giúp cải thiện nghiên cứu và ứng dụng NLP trong tiếng Việt.
- Phát triển mô hình học máy tiên tiến**,** tối ưu hóa cho bài toán phân tích cảm xúc khía cạnh, đảm bảo tính chính xác cao và có thể áp dụng rộng rãi trong thực tế.

- Công khai bộ dữ liệu và mô hình nghiên cứu**,** đóng góp cho cộng đồng khoa học, giúp các nhà nghiên cứu và doanh nghiệp tận dụng tài nguyên này để phát triển các giải pháp phân tích dữ liệu hiệu quả hơn.
- Ứng dụng thực tiễn mạnh mẽ**,** giúp doanh nghiệp thương mại điện tử và các ngành dịch vụ nâng cao chất lượng sản phẩm, dịch vụ, từ đó tối ưu hóa chiến lược kinh doanh dựa trên dữ liệu thực tế.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Nguyen, H. T. M.; Nguyen, H. V.; Ngo, Q. T.; Vu, L. X.; Tran, V. M.; Ngo, B. X.; Le, C. A. VLSP SHARED TASK: SENTIMENT ANALYSIS. *J. Comput. Sci. Cybern.* **2019**, *34*, 295-310.

[2] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

[3] McHugh, Mary. (2012). Interrater reliability: The kappa statistic. Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB. 22. 276-82. 10.11613/BM.2012.031.