BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS2205 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS2205.CH183

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

UIT-HCM

PHÂN TÍCH CẢM XÚC DỰA TRÊN KHÍA CẠNH CỦA LAPTOP

Phan Lực Lượng - 230101011

Tóm tắt

- Link Github của nhóm:

https://github.com/LuongPhan/CS2205.CH183

- Link YouTube video:

https://youtu.be/1SQCUXqPx5A

- Ảnh + Họ và Tên của các thành viên:

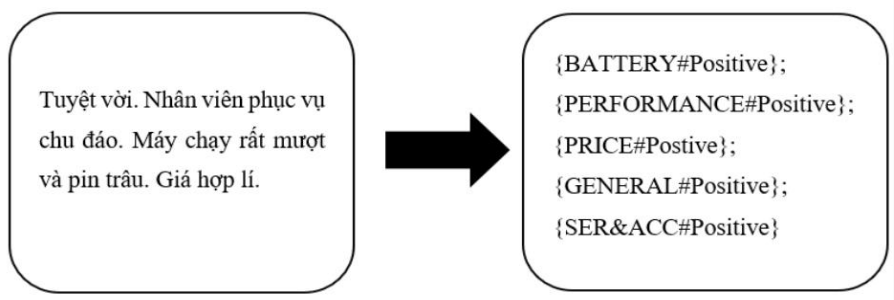
Phan Lực Lượng - 230101011

Giới thiệu

- Trên Shopee, Lazada, Tiktok có hàng triệu bình luận phản hồi của khách hàng về sản phẩm, những dữ liệu này cung cấp thông tin quan trọng cho nhà sản xuất, nhưng chúng chưa được gán nhãn và phân tích tự động hiệu quả.
- Phần lớn các bộ dữ liệu về phân tích cảm xúc đều được xây dựng trên tiếng Anh, số lượng bộ dữ liệu cho tiếng Việt còn ít.

Giới thiệu

- Aspect Based Sentiment Analysis (ABSA) là phương pháp phân tích cảm xúc chuyên sâu, giúp nhận diện cảm xúc của người dùng không chỉ ở mức tổng thể mà còn theo từng khía cạnh cụ thể của sản phẩm hoặc dịch vụ.



Input Output

Mục tiêu

- Xây dựng bộ dữ liệu phản hồi khách hàng bằng tiếng Việt cho Laptop: Thu thập, xử lý và gán nhãn dữ liệu từ các nền tảng thương mại điện tử, đảm bảo tính đa dạng và chất lượng cao.

- Ứng dụng mô hình học máy tiên tiến: Áp dụng phoBERT để phân tích cảm xúc khía cạnh, nâng cao độ chính xác và so sánh với các phương pháp truyền thống.

Nội dung và Phương pháp

1. Xây dựng bộ dữ liệu:

- Thu thập dữ liệu từ các nền tảng thương mại điện tử và mạng xã hội: Thế giới đi động, Shopee, Tiki, Tiktok.

- Tiền xử lý và gán nhãn dữ liệu theo từng khía cạnh sản phẩm.

- Kiểm định chất lượng dữ liệu để đảm bảo tính chính xác và độ tin cậy: sử dụng độ đo Cohen`s Kappa.

Nội dung và Phương pháp

2. Phát triển mô hình phân tích cảm xúc:

- Huấn luyện mô hình phoBERT và so sánh với các phương pháp truyền thống: Naïve Bayes, Random Forest, Logistic Regression.

- Đánh giá hiệu suất dựa trên Precision, Recall, F1-score.

Nội dung và Phương pháp

3. Kiểm thử và ứng dụng thực tế:

- Kiểm thử mô hình trên tập dữ liệu thực tế để đảm bảo tính tổng quát.

- Công khai bộ dữ liệu để hỗ trợ nghiên cứu và phát triển NLP tiếng Việt.

Kết quả dự kiến

- Bộ dữ liệu phản hồi khách hàng chuẩn hóa, đảm bảo tính đa dạng và độ chính xác cao, hỗ trợ nghiên cứu và ứng dụng NLP trong tiếng Việt.
- Mô hình học máy tiên tiến, tối ưu hóa cho phân tích cảm xúc khía cạnh, đạt độ chính xác cao và có thể triển khai thực tế.
- Công khai bộ dữ liệu và mô hình nghiên cứu, tạo điều kiện cho cộng đồng khoa học và doanh nghiệp ứng dụng vào các giải pháp phân tích dữ liệu hiệu quả hơn.

Tài liệu tham khảo

[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.

[4] Mohammad, R. M., McCluskey, T. L., & Thabtah, F. "An assessment of features related to phishing websites using an automated technique." In *Proceedings of the 2014 International Conference on Security and Management*, 2014.