THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/WVTEu_tz8f8
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/PhuongNGT/CS2205.APR2023
- *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: NGUYỄN GIANG THANH PHƯƠNG • MSSV: 230201050 	• Lớp: CS2205.APR2023 • Tự đánh giá (điểm tổng kết môn): 6/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 0 • Link youtube: https://youtu.be/JHTZZx8N7s4 • Link Github: https://github.com/PhuongNGT/CS2205.APR2023

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN TÍCH BÌNH LUẬN GOOGLE RATING VỀ BỆNH VIỆN ĐẠI HỌC Y DƯỢC TP.HCM TRÊN NHIỀU KHÍA CẠNH BẰNG MÔ HÌNH HSD

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ANALYSIS OF GOOGLE RATING COMMENTS ABOUT UNIVERSITY MEDICAL CENTER HCM ON MANY ASPECTS USING THE HSD MODEL

TÓM TẮT *(Tối đa 400 từ)*

Bệnh viện Đại học Y Dược TPHCM là bệnh viện đại học dẫn đầu về Công nghệ thông tin tại Việt Nam và với tiêu chí hết lòng vì người bệnh nên việc khảo sát, thu thập ý kiến đóng vai trò quan trọng đặc biệt là ý kiến phản hồi thông qua người bệnh để nâng cao việc khám chữa bệnh là một yêu cầu cấp thiết. Bệnh viện đã có nhiều nghiên cứu hỗ trợ việc khảo sát, thu thập ý kiến nhưng cũng giống các nghiên cứu hiện nay trong lĩnh vực này vẫn phải đối mặt với nhiều thiếu sót lớn có thể kể đến như thiếu sót trong quá trình xử lý dữ liệu và thiếu ứng dụng thực tế. Thông qua nghiên cứu dụng mô hình HSD trong phân tích bình luận google rating về bệnh viện đại học y dược tp.hcm trên nhiều khía cạnh sẽ tập trung vào phát triển một hệ thống thông minh có khả năng giải quyết những thiếu sót này. Trước tiên, mô hình HSD là sự kết hợp giữa mô hình PhoBERT được huấn luyện trước và mô hình Text-CNN, đã được đề xuất để xử lý các bài toán phân loại tiếng Việt. Thứ hai, Bộ dữ liệu bao gồm các bình luận được lọc ra từ Google Rating với hơn 10000 câu được gán nhãn theo ba cảm xúc (tích cực, tiêu cực và trung tính). Thứ ba, sẽ áp dụng một kỹ thuật tiền xử lý hiệu quả để làm sạch bộ dữ liệu bình luận trên. Ngoài ra, nhiều thí nghiệm khác nhau đã được tiến hành làm cơ sở để so sánh và điều tra hiệu suất của mô hình đề xuất với các phương pháp tiên tiến nhất. Kết quả thực nghiệm cho thấy mô hình PhoBERT - CNN được đề xuất vượt trội hơn so với phương pháp khác và cuối cùng chứng minh cho tính thực tế của hệ thống được đề xuất.

GIỚI THIỆU (Tối đa 1 trang A4)

Trong các năm gần đây, việc khảo sát để lấy ý kiến phản hồi đóng vai trò quan trọng. Các ý kiến này được phân tích một cách thủ công bởi các nhân viên thuộc đơn vị trung tâm quản lý truyền thông và việc phân tích ý kiến phản hồi theo cách thủ công sẽ làm mất nhiều thời gian và không tổng hợp được một cách chính xác các vấn đề được đề cập đến. Bài toán phân tích bình luận phản hồi theo khía cạnh được các nhà nghiên cứu đặt ra với mục đích giảm thời gian chọn lọc, xử lý và tính chính xác. Vì vậy, nghiên cứu này góp phần hoàn thiện mô hình HSD trên ngôn ngữ tiếng Việt. Nghiên cứu này đề xuất một hệ thống mới áp dụng xử lý ngôn ngữ tự nhiên tiên tiến, kỹ thuật phân loại những bình luận. Nghiên cứu này có thể xử lý các vấn đề nhỏ từ xử lý những bình luận nhận xét đơn lẻ đến liên tục với số lượng lớn.

Về mặt hình thức, nhiệm vụ của mô hình HSD được mô tả như hình sau.

Hình 1: Tổng quan về phương pháp HSD

-Đầu vào: những bình luận tiếng Việt trên google rating.

- Đầu ra: Một trong ba nhãn khác nhau được các bộ phân loại dự đoán.

- (Tiêu Cực) chứa ngôn ngữ thường mang tính chất xúc phạm có thể bao gồm lời nói xúc phạm hay thể hiện ý định nhằm kích động hoặc gây bất lợi (3).
- (Trung Tính) có thể chứa những từ ngữ xúc phạm nhưng không kích động hoặc gây bất lợi
- (Tích cực) là cuộc trò chuyện và thể hiện cảm xúc tốt và không chứa đựng ngôn ngữ xúc phạm hoặc lời nói căm thù là một điều bình thường

MỤC TIÊU

- Triển khai, đánh giá hiệu quả mô hình kết hợp PhoBERT và CNN trong việc phân loại các bình luận thu thập từ Google Rating.
- So sánh mô hình kết hợp PhoBERT và CNN với các mô hình khác về nhiệm vụ phân loại văn bản tiếng Việt.

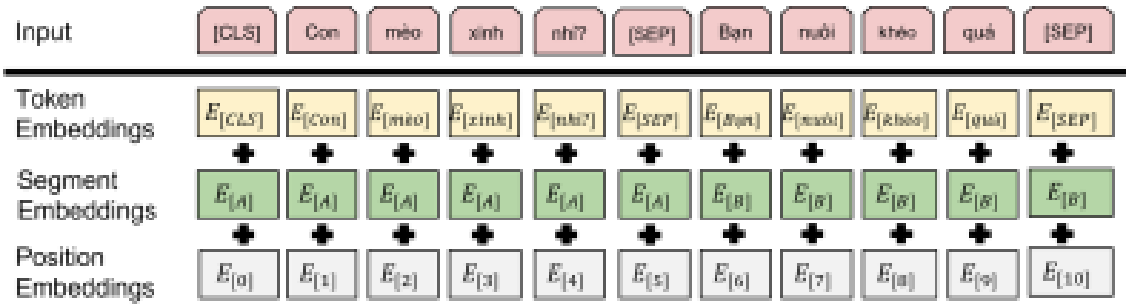
NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Trình bày kiến trúc của mô hình PhoBERT [1], xây dựng bộ dữ liệu và cách thức mô hình PhoBERT được điều chỉnh để hoạt động như một lớp nhúng từ để trích xuất thông tin từ bộ dữ liệu.

từ dữ liệu,

Phương Pháp:

- Tìm hiểu mô hình kiến trúc của mô hình PhoBERT phát hiện và nhận dạng văn bản ngôn ngữ được đào tạo trước như đơn ngữ và đa ngữ.
- Tìm hiểu cách xây dựng bộ dữ liệu bình luận theo tiếng Việt gồm 10000 bình luận theo google rating.
- Tìm hiểu kiến trúc đa lớp của mô hình PhoBERT như Hình 2. Tìm hiểu cách xử lý, mã hóa dữ liệu khi đưa vào của mô hình PhoBERT đào tạo trước khi được thay thế xử lý bằng kiến trúc mô hình CNN [3].



Hình 2: Quá trình dữ liệu đầu vào của mô hình PhoBERT

Nội dung 2: Trình bày kiến trúc của các lớp của mô hình CNN dùng trong nghiên cứu này.

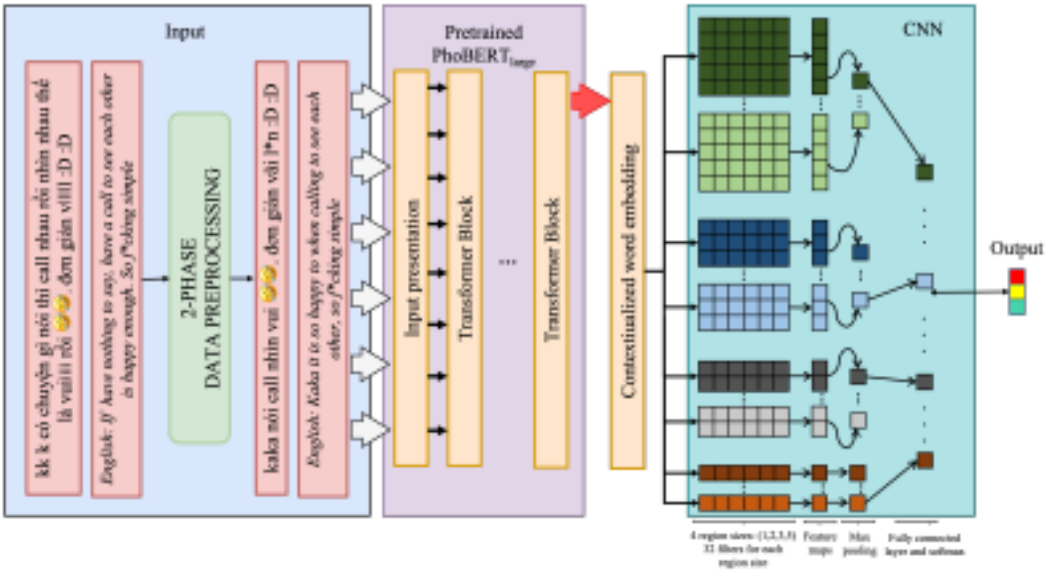
Phương Pháp:

- Phân tích kỹ thuật tích chập và gộp của CNN trong việc trích xuất các khái niệm và từ khóa chính của văn bản.
- Phân tích tại sao lại sử dụng CNN là mô hình để giải quyết các nhiệm vụ phân loại văn bản ngắn [2] mà không phải là các mạng nơ-ron sâu điển hình khác như LSTM, Bi-LSTM và GRU [4–6].
- Phân tích những mặt hạn chế cuả mạng CNN đối với các bình luận dạng chuỗi

[2, 3]

- Tìm hiểu các thành phần kiến trúc INPUT, CONV1D, POOLING, DROPOUT và FC dùng để đánh giá trong bài toán.

Nội dung 3: Trình bày bài toán phân loại tiếng Việt áp dụng mô hình HSD là sự kết hợp giữa mô hình PhoBERT được huấn luyện trước và mô hình Text-CNN như hình 3.



Hình 3: Tổng quan mô hình HSD

Phương Pháp:

- Tìm hiểu cách kết hợp giữa mô hình PhoBERT có nhiệm vụ trích xuất các đặc điểm từ các câu cho đầu vào của mô hình Text-CNN và việc nhúng từ theo ngữ cảnh của các nhận xét từ PhoBERT sẽ được đưa vào mô hình Text-CNN để lấy bản đồ tính năng.
- Cài đặt, huấn luyện mô hình PhoBERT được huấn luyện trước và mô hình Text-CNN trên bộ dữ liệu tự xây dựng.

KẾT QUẢ MONG ĐỢI

Việc ứng dụng PhoBERT-CNN trong phân loại bình luận nhiều khía cạnh mang lại nhiều kết quả mong đợi bao gồm:

- Đề xuất một giải pháp tiên tiến mới cho mô hình HSD với quy trình xử lý dữ liệu gồm hai giai đoạn để làm sạch tập dữ liệu.
- Tìm ra được một cách tiếp cận hiệu quả và đơn giản để phân tích bình luận

bằng tiếng Việt dựa trên mô hình kết hợp PhoBERT-CNN.

- Xây dựng thành công hệ thống phân tích bình luận bằng tiếng Việt dựa trên Google Rating theo thời gian thực.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Nguyen, D.Q., Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037–1042. Association for Computational Linguistics, Online (2020)

[2]. He, C., Chen, S., Huang, S., Zhang, J., Song, X.: Using convolutional neural network with bert for intent determination. In: 2019 International Conference on Asian Language Processing (IALP), pp. 65–70 (2019). IEEE

[3]. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (2014).

[4]. Pham-Hong, B.-T., Chokshi, S.: Pgsg at semeval-2020 task 12: Bert-lstm with tweets' pretrained model and noisy student training method. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 2111–2116 (2020).

[5]. Li, X., Bing, L., Zhang, W., Lam, W.: Exploiting BERT for end-to-end aspect-based sentiment analysis. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 34–41. Association for Computational Linguistics, Hong Kong, China (2019).

[6]. Yi, R., Hu, W.: Pre-trained bert-gru model for relation extraction. In: Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, pp. 453–457 (2019)