PHÂN LOẠI Ý KIẾN KHÁCH HÀNG VĂN BẢN TIẾNG VIỆT VỚI MÔ HÌNH PhoBERT

Trần Văn Tịnh - 220101039

Tóm tắt

- Lớp: CS2205.APR2023
- Link Github: https://github.com/TranVanTinhUIT/CS2205.APR2023
- Link YouTube video:
- Ảnh + Họ và Tên: Trần Văn Tịnh
- MSHV: 220101039

Giới thiệu

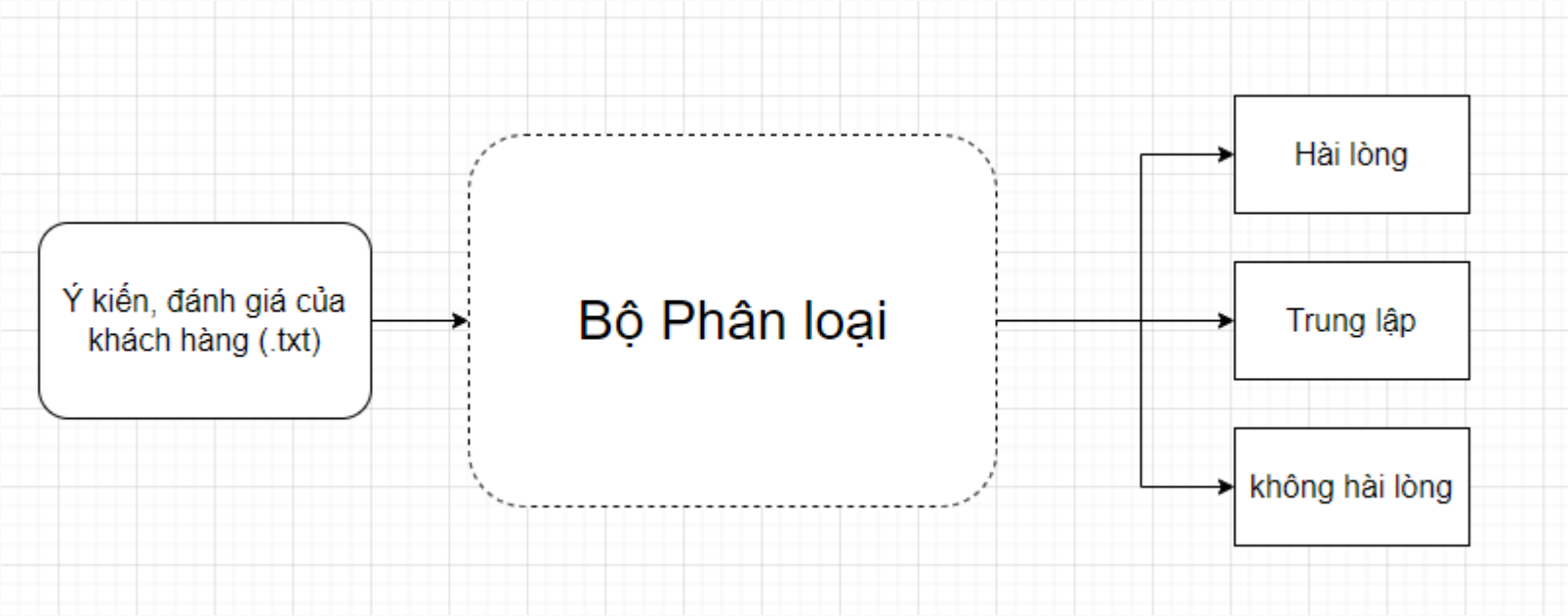
- Ý kiến khách hàng đóng vai trò thiết yếu cho việc cải thiện chất lượng dịch vụ và nâng cao sự hài lòng của khách hàng.
- Các cách thủ công để phân tích các ý kiến, đánh giá mất rất nhiều thời gian và việc tổng quát hóa các kết quả cũng rất khó khăn.

=> Công cụ tự động phân tích các ý kiến, đánh giá của khách hàng.

Giới thiệu

- BERT, một mô hình NLP sử dụng kiến trúc Transformer và huấn luyện trên một lượng dữ liệu lớn để hiểu ngữ nghĩa từ và ngữ cảnh trong câu.
- PhoBERT, mô hình NLP huấn luyện cho dữ liệu Tiếng Việt dựa trên BERT.

Giới thiệu


Ý kiến, đánh giá của khách hàng (.txt)
Bộ Phân loại
Hài lòng
Trung lập
không hài lòng

Mục tiêu

- Xây dựng mô hình phân loại ý kiến khách hàng vào các lớp cho trước với hiệu suất và độ chính xác cao.
- Ứng dụng kết quả xây dựng công cụ hỗ trợ doanh nghiệp, tổ chức tự động phân loại ý kiến của khách hàng.
- Đặt tiền đề cho các nghiên cứu sâu hơn (phân tích và thống kê những đóng góp mà khách hàng đề cập và mong muốn cải thiện).

Nội dung và Phương pháp

Thu thập và tiền xử lý dữ liệu

- Thu thập dữ liệu huấn luyện từ các trang web như Booking.com, Agoda, Expedia,... và một bộ dữ liệu test để đánh giá mô hình.
- Tiền xử lý: loại bỏ dấu câu không cần thiết, ký tự đặc biệt, icon,..
- Mã hóa dữ liệu thành các token phù hợp với mô hình. [CLS] bắt đầu câu, [SEP] kết thúc câu. Ngoài ra token [PAD] được thêm để câu có độ dài phù hợp nhằm cải thiện hiệu suất.

Nội dung và Phương pháp

Huấn luyện và tinh chỉnh

- Thêm lớp output và tiếp tục huấn luyện PhoBERT trên tập dữ liệu đã thu thập cho tác vụ phân loại ý kiến.
- Tinh chỉnh mô hình bằng các kỹ thuật fine-tuning như Gradient, Regularization, Cross-Validation.

Nội dung và Phương pháp

Đánh giá mô hình

- Đánh giá độ chính xác và hiệu suất của mô hình trên dữ liệu test.
- So sánh kết quả với các phương pháp truyền thống như SVM, Naive Bayes, Random Forest và các mô hình học sâu như CNN, RNN, BERT.

Kết quả dự kiến

- Mô hình có độ chính xác và hiệu suất vượt trội hơn các mô hình cũ.
- Khả năng hiểu và xử lý ngôn ngữ tự nhiên một cách hiệu quả và chính xác.
- Ứng dụng vào thực tế nhằm tối ưu hóa chiến lược kinh doanh, chất lượng sản phẩm.

Tài liệu tham khảo

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
- [2] Dat Quoc Nguyen, Anh Tuan Nguyen: PhoBERT: Pre-trained language models for Vietnamese. EMNLP (Findings) 2020: 1037-1042
- [3] Nguyen Phuc Minh, Tran Hoang Vu, Vu Hoang, Ta Duc Huy, Trung Huu Bui, Steven Quoc Hung Truong:ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining. LREC 2022: 328-337
- [4] Philip Kenneweg, Sarah Schröder, Barbara Hammer:Neural Architecture Search for Sentence Classification with BERT. ESANN 2022
- [5] Jinghui Lu, Maeve Henchion, Ivan Bacher, Brian Mac Namee: A Sentence-level Hierarchical BERT Model for Document Classification with Limited Labelled Data. CoRR abs/2106.06738 (2021)
- [6] Philip Kenneweg, Sarah Schröder, Barbara Hammer:Neural Architecture Search for Sentence Classification with BERT. ESANN 2022
- [7] Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, Manabu Okumura:Fusing Label Embedding into BERT: An Efficient Improvement for Text Classification. ACL/IJCNLP (Findings) 2021: 1743-1750
- [8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need. NIPS 2017: 5998-6008