PHOBERT FOR SENTIMENT ANALYSIS OF CUSTOMER FEEDBACK IN VIETNAMESE TEXT

Trần Văn Tịnh

Đại học công nghệ thông tin –DHQG TPHCM

Mục tiêu

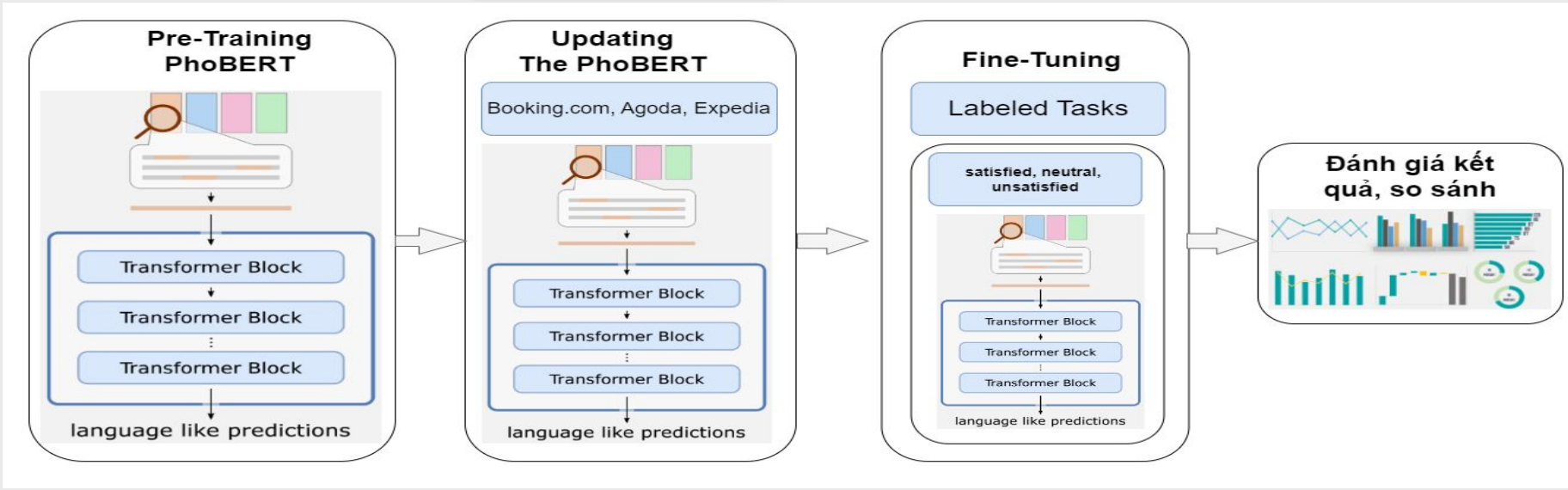
- Mô hình phân loại ý kiến khách hàng tiếng việt dựa vào PhoBERT vào các lớp cho trước.
- Công cụ hỗ trợ các doanh nghiệp, tổ chức tự động phân tích ý kiến khách hàng.
- Nghiên cứu sâu ứng dụng xử lý ngôn ngữ tự nhiên vào đời sống, xã hội.

Lý do đề tài ?

- Tự động phân loại ý kiến, đánh giá khách hàng vào các lớp cho trước, rút ngắn thời gian và nâng cao hiệu suất kinh doanh.
- Ứng dụng sự phát triển của xử lý ngôn ngữ tự nhiên vào các vấn đề đời sống, xã hội.

Overview

Pre-Training PhoBERT → Thu thập dữ liệu và huấn luyện → Tinh chỉnh → Đánh giá và so sánh



Description

1. Pre-training PhoBERT

- Mô hình PhoBERT đã được huấn luyện trên dữ liệu tiếng Việt.

2. Thu thập dữ liệu và huấn luyện mô hình

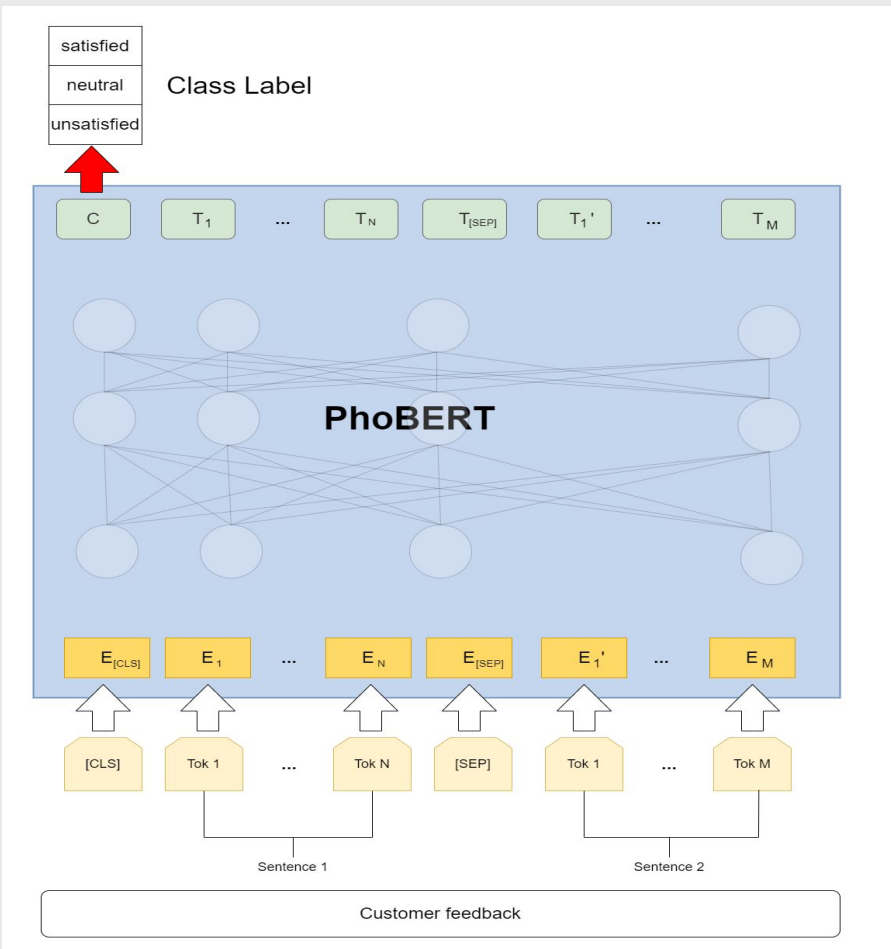
- Thu thập lượng lớn dữ liệu từ các trang như Booking.com, Agoda, Expedia,.. Làm dữ liệu huấn luyện cho mô hình phân loại ý kiến khách hàng.
- Thu thập một tập dữ liệu test để đánh giá mô hình.
- Xử lý dữ liệu thu thập: loại bỏ dấu câu, ký tự đặc biệt, icon,..
- Mã hóa dữ liệu thành các token phù hợp. [CLS] bắt đầu câu, [SEP] kết thúc câu. Token [PAD] để các câu có độ dài đồng nhất phù hợp với yêu cầu đầu vào của mô hình PhoBERT

3. Tinh chỉnh mô hình

- Thêm lớp nhãn dự đoán output và áp dụng các kỹ thuật fine-tuning như Gradient, Regularization,Cross-Validation.

4. Đánh giá và so sánh

- Đánh giá độ chính xác của mô hình trên tập dữ liệu test.
- So sánh kết quả với các mô hình truyền thống SVM, Random Forest và các mô hình học sâu khác CNN, RNN, BERT.



NII

Trần Văn Tịnh– Trường Đại học công nghệ thông tin

TEL : 0358444463 Email : tinhtv1996@gmail.com