COMMENT CLASSIFICATION IN SHOPEE

Phạm Huỳnh Tấn Đạt

Trường ĐH CNTT TPHCM

What ?

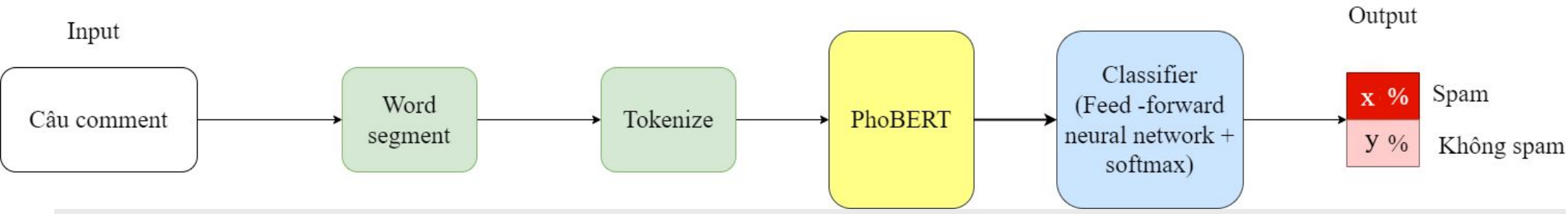
Chúng tôi nghiên cứu mô hình PhoBERT nhăm:

- Giải quyết bài toán phân loại câu comment spam trên shopee.
- Xây dựng một bộ dữ liệu gồm 10000 câu comment về các sản phẩm điện tử trên shopee.
- Xây dựng một api và một ứng dụng sử dụng mô hình sau khi huấn luyện để phân loại các câu comment.

Why ?

- Với các cửa hàng trên sàn thương mại điện tử, các comment có tác dụng giúp họ nâng cao chất lượng dịch vụ. Tuy nhiên, có nhiều câu comment spam làm tốn thời gian lọc cũng như không đem lại giá trị gì cho cửa hàng.
- Các phương pháp máy học truyền thống như Support Vector Machine,... hay học sâu như RNN, LSTM,.. chưa biển diễn tốt nội dung, ngữ nghĩa của các câu dài, phức tạp.

Overview



Description

1. Câu comment

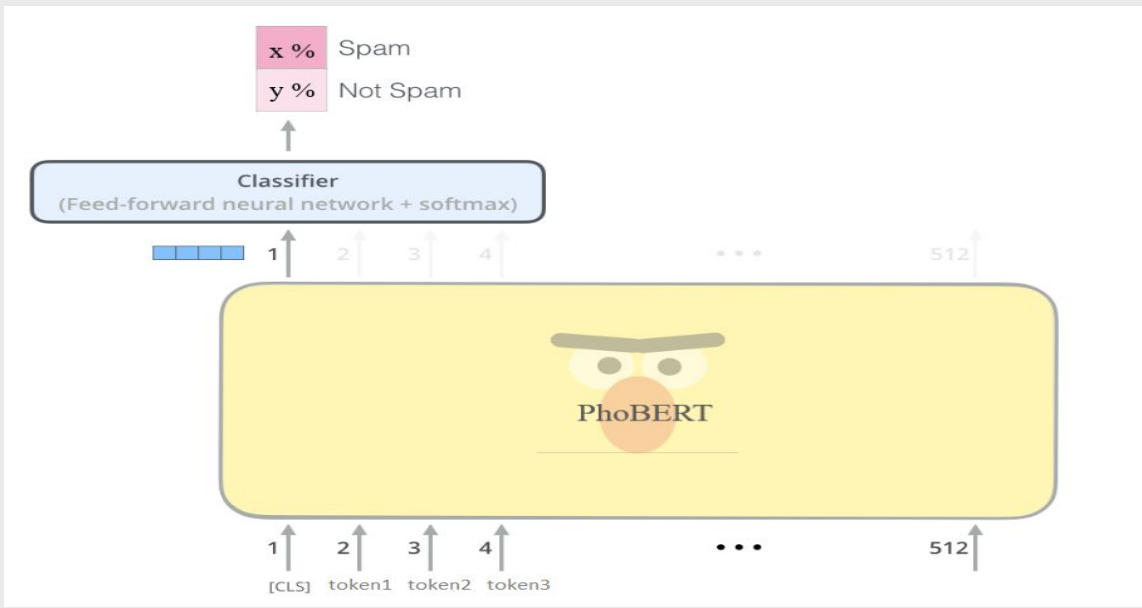
- Câu comment về một sản phẩm điện tử trên shopee.

2. Word segment

- Câu comment sẽ được word segment, nối các cặp từ có nghĩa thành một token duy nhất.
- Sử dụng mô hình word segment của Underthesea, một bộ công cụ hỗ trợ cho việc nghiên cứu và phát triển xử lý ngôn ngữ tự nhiên tiếng Việt.

3. Tokenize

- Sau đó câu comment sẽ được tokenize, mã hóa các văn bản thành các index dạng số mang thông tin của văn bản để cho máy tính có thể huấn luyện được.
- Sử dụng 2 tokenizer của VinAI tương ứng với từng phiên bản của PhoBERT base và PhoBERT large.



Ảnh 1: Câu comment sau khi được word segment và tokenize được đưa vào mô hình để phân loại.

2. PhoBERT

- Sử dụng 2 mô hình pretrained PhoBERT base và PhoBERT large của VinAI từ Hugging Face.
- Fine-tuning 2 phiên bản mô hình PhoBERT base và PhoBERT large với bộ dữ liệu 10000 câu comment từ shopee.

3. Classifier

- Mô hình feed forward neural network kết hợp softmax có tác dụng đưa kết quả của model về dạng xác suất cho từng lớp.

NII

Phạm Huỳnh Tấn Đạt– Trường Đại học Công nghệ thông tin
TEL : 0907839696 Email : datpht.18@grad.uit.edu.vn