Lựa chọn công nghệ phân loại văn bản

Dựa trên nghiên cứu đánh giá hiệu năng các model phân loại văn bản, sử dụng tập dữ liệu bình luận spam, đánh giá trên 2 nhiệm vụ: nhiệm vụ phân loại nhị phân để phát hiện xem đánh giá có phải là thư rác hay không và nhiệm vụ phân loại nhiều lớp để xác định loại thư rác.

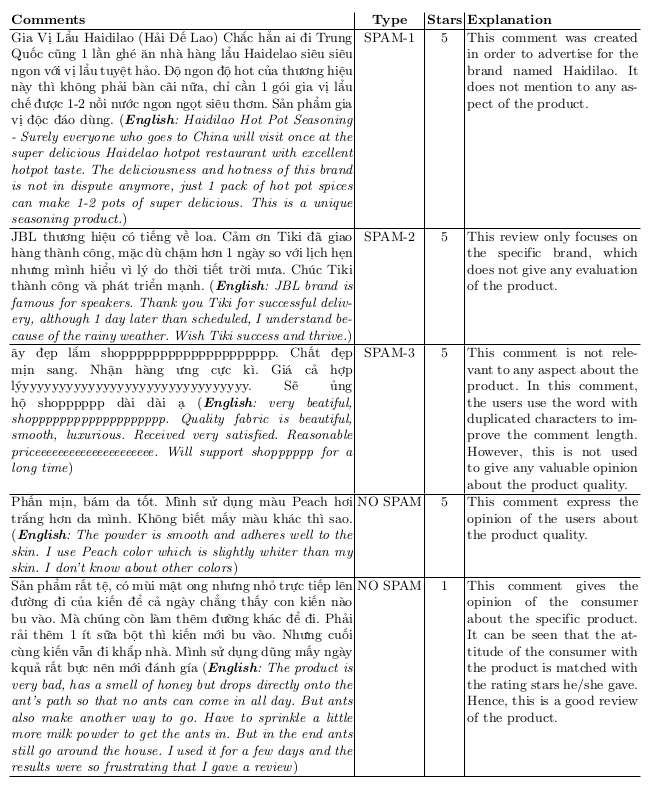
Tài liệu này đánh giá 5 mô hình bao gồm: Text-CNN, LSTM, GRU, PhoBERT, BERT4News. PhoBERT thu được kết quả cao nhất trong cả hai nhiệm vụ, lần lượt là 86,89% và 72,17% theo điểm F1 trung bình.

# 1.Bộ dữ liệu được dùng để đánh giá

### 1.1.Quy trình tạo bộ dữ liệu

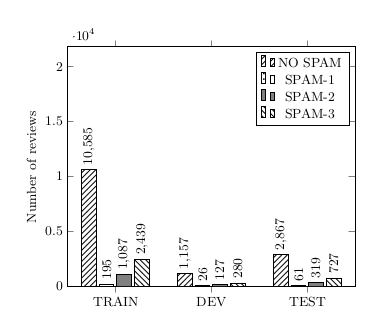
Thu thập dữ liệu từ các nền tảng mua sắm trực tuyến hàng đầu tại Việt Nam. Sau đó, chọn một số sản phẩm bán chạy nhất cho từng danh mục sản phẩm và thu thập tối đa 15 bài đánh giá cho mỗi sản phẩm. Sau khi thu thập, nhận được tập dữ liệu gồm 19.868 đánh giá sản phẩm, trong đó có số lượng đánh giá theo sao, nhận xét về sản phẩm và đường dẫn đến sản phẩm đó. Sau đó, xây dựng hướng dẫn chú thích và chú thích cho kho văn bản.

Quá trình chú thích bao gồm hai giai đoạn. Đầu tiên là giai đoạn đào tạo cho người chú thích. Người chú thích đọc hướng dẫn mô tả ý nghĩa của các nhãn, đánh giá mẫu và một số ví dụ về các trường hợp cụ thể. Người chú thích sẽ đọc hướng dẫn và chú thích 300 mẫu ngẫu nhiên trong tập dữ liệu. Sau đó, tính toán và đánh giá mức độ thỏa thuận trung bình giữa các nhà chú thích. Nếu thỏa thuận giữa các nhà chú thích được thỏa mãn, chúng ta sẽ chuyển sang giai đoạn thứ hai, đó là giai đoạn chú thích. Ngược lại, chúng tôi đào tạo lại người chú thích và cập nhật nguyên tắc chú thích. Ở giai đoạn thứ hai, người chú thích sẽ được cung cấp một tập dữ liệu hoàn chỉnh và chú thích trên tập dữ liệu này. Có cả ba người chú thích trong các giai đoạn chú thích và nhãn cuối cùng của tập dữ liệu sẽ được quyết định bằng cách bỏ phiếu cho nhãn được chỉ định nhiều nhất.



*Bảng 1. Một số ví dụ về đánh giá từ người dùng và giải thích cách chọn nhãn cho mỗi đánh giá. Đối với mỗi đánh giá, người chú thích sẽ chọn một nhãn phù hợp.*

### 1.2.Tổng quan về tập dữ liệu



Sau khi chú thích tập dữ liệu, có gần 19.000 đánh giá từ người dùng, trong đó mỗi đánh giá được phân loại là spam hoặc không phải spam. Nếu các bài đánh giá là thư rác, chúng bao gồm các loại thư rác. Sau đó, chia tập dữ liệu thành các tập huấn luyện, phát triển và kiểm tra với tỷ lệ 7-1-2.

# 2.Phương pháp

## 2.1.Định nghĩa nhiệm vụ

Vấn đề phát hiện đánh giá thư rác được xếp vào nhiệm vụ phân loại văn bản. Vấn đề này bao gồm hai nhiệm vụ: Nhiệm vụ 1 là nhiệm vụ phân loại nhị phân để phân loại xem đánh giá có phải là thư rác hay không phải là thư rác và Nhiệm vụ 2 là nhiệm vụ phân loại nhiều lớp để xác định loại thư rác, một trong ba loại như đã đề cập.

## 2.2.Các mô hình được sử dụng

### 2.2.1.Text-CNN

Mạng nơ ron tích chập (CNN) là mô hình được kết hợp giữa nhiều lớp khác nhau. CNN thường được ứng dụng trong thị giác máy tính để trích xuất các đặc điểm của ảnh phục vụ phân loại ảnh và đã đạt được hiệu suất cao so với các phương pháp truyền thống. Ngoài ra, CNN còn được ứng dụng trong các bài toán xử lý ngôn ngữ tự nhiên, điển hình là phân loại văn bản với mô hình Text-CNN. Mô hình này dựa trên kiến trúc tích chập để trích xuất các đặc điểm có giá trị từ các văn bản tự nhiên.

### 2.2.2.LSTM

Bộ nhớ ngắn hạn dài (LSTM) là mô hình cải tiến từ Mạng thần kinh tái phát (RNN). LSTM giúp mô hình ghi nhớ lâu các thông tin trước đó, đây là hạn chế mà mô hình RNN gặp phải. LSTM bao gồm ba cổng: cổng vào, cổng ra và cổng quên. Cổng đầu vào chọn thông tin để thêm vào ngữ cảnh, cổng đầu ra quyết định xem đầu vào có cần thiết cho hiện tại hay không và cổng quên được sử dụng để xóa thông tin khỏi ngữ cảnh khi không còn cần thiết. Mô hình này giúp mô hình phân loại văn bản tốt hơn vì có thể nắm bắt được thông tin ngữ cảnh trong toàn bộ văn bản.

### 2.2.3.GRU

Đơn vị định kỳ có cổng (GRU) là một biến thể của mô hình LSTM. Mô hình này có độ phức tạp thấp hơn LSTM. Trong khi LSTM có ba cổng thì GRU chỉ có hai cổng: cổng cập nhật và cổng đặt lại. Cổng cập nhật xác định xem có bất kỳ thông tin nào trong quá khứ được giữ lại và sử dụng trong tương lai hay không và cổng đặt lại quyết định rằng thông tin trong quá khứ nên được giữ lại và mọi thông tin bị lãng quên. Ưu điểm của GRU là sử dụng ít tham số hơn trong quá trình đào tạo và do đó sử dụng ít bộ nhớ hơn và thời gian đào tạo nhanh hơn LSTM.

### 2.2.4.Transformers

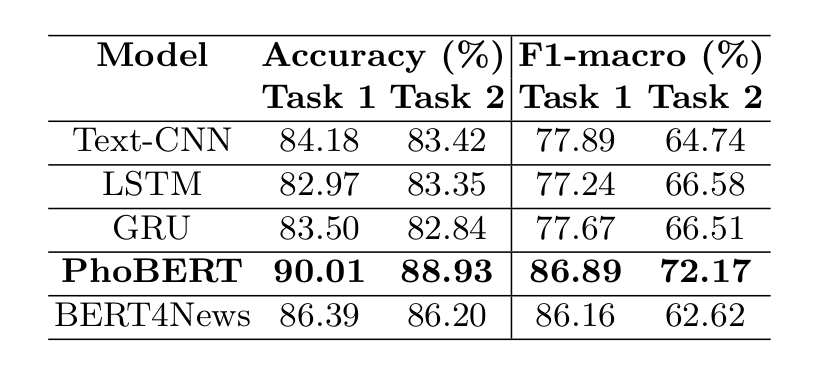
Một kiến trúc đã được đề xuất trong những năm gần đây và hiện đang được sử dụng rộng rãi. Sự xuất hiện của BERT giúp nhiều tác vụ hạ nguồn trong NLP đạt được kết quả hiệu suất cao khi đào tạo trên một tập dữ liệu nhỏ. BERT và các phương sai của nó trở thành các phương pháp tiếp cận cơ bản trong nhiều nhiệm vụ NLP, được gọi là BERTology.

Trong tiếng Việt, có hai loại cách tiếp cận BERTology: mô hình đa ngôn ngữ và mô hình đơn ngữ. Kết quả là mô hình đơn ngữ thu được kết quả tốt hơn mô hình đa ngôn ngữ cho nhiệm vụ phân loại văn bản và nhiệm vụ tuần tự. Vì vậy, đánh giá này áp dụng hai mô hình BERT đơn ngữ là PhoBERT và BERT4News cho bài toán phát hiện các đánh giá spam.

# 3. Kết quả thực nghiệm

Thực hiện hai nhiệm vụ cho vấn đề phát hiện thư rác. Đầu tiên, triển khai bộ phân loại nhị phân để phân loại các bài đánh giá là spam hoặc không phải spam. Thứ hai, xây dựng mô hình phân loại để xác định loại thư rác.

Áp dụng các mô hình Text CNN, LSTM, GRU và Transformer với PhoBERT và BERT4News cho các nhiệm vụ. Cuối cùng, sử dụng điểm F1 để đánh giá hiệu năng của model.



*Bảng 2. Kết quả thực nghiệm của các mô hình phân loại trên tập dữ liệu*

Về phân loại đánh giá là spam hay không, mô hình Text CNN cho kết quả trên bộ thử nghiệm với Độ chính xác và điểm F1 là 84,18% và 77,89%, tốt hơn hai mô hình LSTM và GRU. Ngoài ra, ở nhiệm vụ này, PhoBERT đạt kết quả tốt hơn BERT4News với độ chính xác và điểm F1 lần lượt là 90,01% và 86,89%. Đối với tác vụ phát hiện loại thư rác, mô hình PhoBERT đạt 88,93% cho Độ chính xác và 72,17% cho điểm F1, cao hơn BERT4News.

Dựa trên kết quả huấn luyện và đánh giá các mô hình, có thể thấy rằng việc phân loại các đánh giá bằng mô hình Transformer cho hiệu suất tốt hơn so với các mô hình mạng nơ-ron sâu. Mô hình PhoBERT đạt được hiệu suất tốt nhất trong hai nhiệm vụ. Còn đối với việc phát hiện các loại thư rác, kết quả Độ chính xác và điểm F1 khác biệt nhiều hơn so với việc phân loại đánh giá có phải là thư rác hay không do sự mất cân bằng giữa các nhãn dữ liệu. Ngoài ra, việc phân loại đánh giá có phải là thư rác hay không cũng ít phức tạp hơn so với việc phát hiện các loại thư rác nên kết quả của nhiệm vụ này cao hơn nhiệm vụ phát hiện các loại thư rác.

# 4.Kết luận

Tài liệu này áp dụng các mô hình phân loại mạnh mẽ cho tập dữ liệu ViSpamReviews - bộ dữ liệu phát hiện đánh giá spam trên các trang mua sắm trực tuyến của Việt Nam với hơn 19.000 đánh giá được chú thích bởi con người, và mô hình PhoBERT thu được kết quả cao nhất với 86,89% điểm F1 cho nhiệm vụ phân loại thư rác và 72,17% điểm F1 cho nhiệm vụ phát hiện loại thư rác.

# 5.Tài liệu tham khảo:

<https://arxiv.org/abs/2301.10186>

<https://arxiv.org/abs/2207.14636>