Sentiment Analys

**NHÓM INT3405E\_41-NJR**

Điểm số hiện tại 0.88432

1st Vũ Minh Nhật   
University of Engineering and Technology  
20021407@vnu.edu.vn 2nd Trần Văn Duy  
University of Engineering and Technology   
20021319@vnu.edu.vn 3rd Nguyễn Quang Minh  
University of Engineering and Technology  
20020146@vnu.edu.vn

# **GIỚI THIỆU**

Ngày nay, với sự phát triển mạnh mẽ của Internet và Công nghệ thông tin, các trang website thương mại điện tử không còn xa lạ gì với chúng ta nữa, chúng ra đời như là một trợ thủ đắc lực, giúp cho việc mua hàng online thuận tiện hơn. Và trong những năm gần đây, chúng ta chứng kiến được sự phát triển mạnh mẽ của thị trường giao đồ ăn trực tuyến khi mà các ứng dụng giao đồ ăn ngày càng hoàn thiện hơn, thanh toán tiện dụng hơn như. Mặt khác các mạng xã hội chuyên review về đồ ăn rất được nhiều người dùng truy cập như Foody, ShoppeFood, GrabFood, Gojek, hay gần đây nhất là Baemin,…có rất nhiều dữ liệu về các bình luận, đánh giá về đồ ăn của người tiêu dùng. Các thương hiệu đồ uống như trà sữa TocoToco, DingTea, hay cà phê như Highlands, Starbucks,…rất được người dùng chú ý. Ý kiến khách hàng là những phản hồi mà khách hàng cảm nhận được sau khi dùng sản phẩm, sử dụng dịch vụ của doanh nghiệp. Vì là ý kiến cá nhân nên những đánh giá của khách hàng có thể tích cực hoặc tiêu cực. Dựa theo những nhận xét của khách hàng, doanh nghiệp sẽ biết được ưu và nhược điểm của sản phẩm, từ đó đưa ra chiến lược để nâng cao giá trị sản phẩm. Không những thế, những ý kiến đó của khách hàng có thể được dùng để quảng bá hay truyền thông. Bởi vậy các doanh nghiệp luôn luôn cải thiện chất lượng dịch vụ để có thể dẫn đầu trên thị trường.

Cùng với sự ra đời của rất nhiều nhãn hàng nổi tiếng, sự cạnh tranh giữa các doanh nghiệp ngày càng tăng và khốc liệt. Theo kiến thức đã học từ môn nguyên lí Marketing, để chinh phục khách hàng thì không thể không tìm hiểu về nhu cầu của họ. Và một trong những cách để biết khách hàng có phù hợp hay không đó là thu hút khách hàng trải nghiệm sản phẩm. Sau đó đánh giá sự thỏa mãn của khách hàng với sản phẩm hay dịch vụ. Tuy nhiên, vấn đề này làm sao doanh nghiệp có thể biết được khách hàng đang hài lòng hay không hài lòng về vấn đề này hay thương hiệu đang được người dùng sử dụng nhiều. Để giải quyết được bài toán này, nghiên cứu đề xuất giải pháp khai thác các bình luận của khách hàng về sản phẩm của các cửa hàng để lại trên trang web Foody. Tuy nhiên dữ liệu chỉ ở mức độ sơ cấp, và lượng dữ liệu rất lớn, các doanh nghiệp không thể dựa vào dữ liệu thô này để đưa ra quyết định được, họ cần biết được các tri thức được phân tích từ tập dữ liệu này. Do đó, nhóm chúng em đã áp dụng phương pháp học sâu để phân loại dữ liệu, dựa vào bình luận để phân tích được bình luận tích cực, bình luận tiêu cực và dùng các phương pháp phân tích và dự đoán. Kết quả nghiên cứu sẽ giúp các cửa hàng, nhà quản lý doanh nghiệp nắm bắt thông tin một cách dễ dàng và nhanh chóng, từ đó việc phát triển kinh doanh được cải thiện và nâng cao, chẳng hạn việc nâng cao sự hài lòng của khách hàng và giữ chân khách hàng tốt hơn.

Trong phạm vi của tài liệu này, nhóm chúng em sẽ phân tích về cơ sở lý thuyết, các phương pháp, kết quả thực nghiệm của mô hình học sâu.

**Chương 2:** Cách tiếp cận, lựa chọn mô hình.

**Chương 3:** Phương pháp

**Chương 4:** Thực nghiệm và kết quả thực nghiệm.

**Chương 5:** Phân chia công việc

**Chương 6:** Tài liệu tham khảo.

# **CÁCH TIẾP CẬN, LỰA CHỌN MODEL**

Như chúng ta được biết việc phân tích cảm xúc của khách hàng thông qua những đánh giá đã được viết, đó là công việc của Xử lí ngôn ngữ tự nhiên (Natural Language Processing – NLP). NLP là một nhánh của trí tuệ nhân tạo tập trung vào việc giải quyết các bài toán liên quan đến ngôn ngữ tự nhiên. Đây không phải là một thuật ngữ mới, tuy nhiên nó mới chỉ được biết tới rộng rãi hơn trong những năm gần đây do sự phát triển mạnh mẽ của Trí tuệ nhân tạo.

Một số bài toán xử lý ngôn ngữ tự nhiên kinh điển mà ta thường thấy đó là: Truy xuất thông tin, Dịch máy, Tóm tắt văn bản, Hệ hỏi đáp, Phân tích cảm xúc,…Những vấn đề, bài toán mà xử lý ngôn ngữ tự nhiên giải quyết có vài trò, ứng dụng to lớn trong sự phát triển của xã hội.

Có rất nhiều mô hình học sâu giúp cho chúng ta xử lý được ngôn ngữ tự nhiên đó là RNN, LSTM, SVM, … sau đó, sự xuất hiện của bài báo Attention Is All You Need đã đưa ra một mô hình đó là Transformer dựa trên cơ chế Attention và Encoder, Decoder đã làm mưa làm gió trong xử lý ngôn ngữ tự nhiên. Nhưng cách đây gần 4 năm, sự ra đời của BERT (Birectional Encoder Representations from Transformers), một nghiên cứu mới mang đầy tính đột phá, một bước nhảy vọt thực sự của Google trong lĩnh vực xử lý ngôn ngữ tự nhiên. Sự ra đời của pre-trained BERT đã kéo theo sự cải tiến đáng kể cho rất nhiều bài toán như Question Answering, Sentiment Analysis,…Nhiều năm sau đó, BERT vẫn chưa từng nguội đi sức nóng của mình và thực tết là nó ngày càng nóng hơn bao giờ hết. Với các phiên bản cải tiến, biến thể như RoBERTa, ALBERT, DistilBERT, … BERT đã càn quét các tác vụ xử lý ngôn ngữ tự nhiên, trở nên áp đảo trong các nền tảng thi đấu như Kaggle, AIVIVN cũng như shared task của nhiều hội nghị.

Các mô hình trên đều được tạo ra để xử lý ngôn ngữ tiếng Anh, các ngoại ngữ khác nhưng với tiếng Việt ( một loại ngôn ngữ có dấu và khá khó so với tiếng Anh ) thì chưa có mô hình nào được nghiên cứu để xử lý nó. Vậy để xử lý các văn bản tiếng Việt thì chúng ta sẽ dùng mô hình nào? Từ ý tưởng này, vào năm 2020 các nhà nghiên cứu của VinAIResearch đã cho ra đời PhoBERT (Pre-trained PhoBERT models are the state-of-the-art language models for Vietnamese)[1]. Từ đó chúng ta đủ để tự tin trong việc sử dụng một pre-trained BERT để giải quyết các bài toán ngôn ngữ tự nhiên cho Tiếng Việt.

PhoBERT khá dễ dùng, nó được build để sử dụng luôn trong các thư viện siêu dễ dùng như [FAIRSeq](https://github.com/pytorch/fairseq) của Facebook hay [Transformers](https://github.com/huggingface/transformers) của Hugging Face nên giờ đây BERT lại càng phổ biến ngay cả với ngôn ngữ tiếng Việt hay tiếng Anh.

Từ những ưu điểm cũng như sự thuận tiện khi áp dụng, nên nhóm chúng em sẽ chọn mô hình PhoBERT của VinAIResearch để tiến hành phân tích cảm xúc của khách hàng trong đánh giá các sản phẩm dịch vụ trên trang website Foody.vn.

# **PHƯƠNG PHÁP**

**1. Tiếp cận**

Dựa vào bài báo tham khảo của VinAIResearch thì PhoBERT là một pre-trained được huấn luyện monolingual language, tức là mô hình này chỉ dành riêng cho tiếng Việt của chúng ta.

**Kiến trúc mô hình:**

PhoBERT gồm 2 version PhoBERT base và PhoBERT large, được dùng giống như là BERT base và BERT large. Phương pháp đào tạo PhoBERT dựa trên trên RoBERTa (Liu et al., 2019) giúp tối ưu cho BERT pre-training hiệu suất mạnh mẽ hơn.

**Đôi chút về RoBERTa:**

Được giới thiệu tại Facebook, BERT được tối ưu hóa mạnh mẽ RoBERTa, là quá trình đào tạo lại BERT với phương pháp đào tạo được cải tiến, 1000% dữ liệu và sức mạnh tính toán.

Để cải tiến quá trình huấn luyện, RoBERTa xóa nhiệm vụ dự đoán Sentence tiếp theo từ BERT’s pre-training và thêm dynamic masking nhờ đó token thay đổi trong suốt các giai đoạn đào tạo. Kích thước batch-training lớn hơn cũng được cho là hữu ích hơn trong quy trình đào tạo.

Và quan trọng hơn, RoBERTa đã dùng 160Gb dữ liệu dạng text cho pre-training, bao gồm 16GB Book Corpus và English Wikipedia được dùng trong BERT. Dữ liệu bổ sung bao gồm bộ dữ liệu CommonCrawl News (63 triệu bài viết, 76GB), kho văn bản Web(38GB) và Stories từ CommonCrawl (31GB). Điều này cùng với con số khổng lồ 1024 GPU V100 Tesla đã chạy trong 1 ngày, dẫn đến việc đào tạo trước của RoBERTa.

**PhoBERT:**

Dựa trên RoBERTa, VinAIResearch cũng đã sử dụng 20GB dữ liệu dạng text cho pre-training. Bộ dữ liệu này bao gồm 2 corpora: thứ nhất là Vietnamese Wikipedia corpus (~1GB), và corpus thứ hai (~19GB) được tạo ra bằng cách loại bỏ các bài viết tương tự và trùng lặp khỏi kho tin tức tiếng Việt (50GB). Để xử lí được Corpus thứ hai, chúng ta sẽ triển khai RDRSegmenter từ VNCoreNLP (Vu et al., 2018) để biểu diễn phân đoạn từ và câu trên tập dữ liệu pre-training.

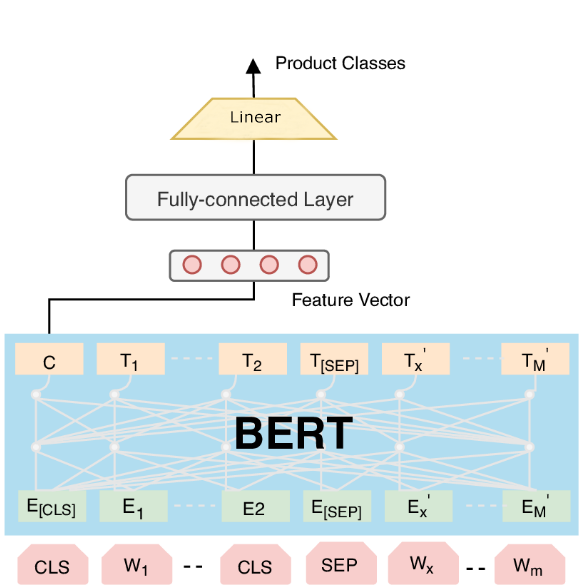
Mặc dù dựa trên RoBERTa, nhưng PhoBERT cũng có điểm khác. PhoBERT đã áp dụng fastBPE(Sennrich et al., 2016) để phân đoạn câu với các subword khi sử dụng một bộ từ vựng chứa 64K loại subword. Ước tính trung bình thì chúng ta sẽ có 24.4 subword tokens trong mỗi câu.

**Tối ưu hóa:**

Chúng ta triển khai cài đặt mô hình RoBERTa trong fairseq(Ott et al., 2019). Sau đó đặt độ dài tối đa là 256 subword tokens, do đó sẽ tạo ra 145M x 24.4 / 256 = 13.8M sentence blocks. Dựa theo bài báo Liu et al. (2019), ta đi đến tối ưu hóa mô hình bằng việc dùng Adam (Kingma an Ba, 2014).

**Fine-tuning:**

Dựa theo Devlin et al.(2019), đối với phân loại văn bản, ta sẽ thêm 1 lớp dự đoán tuyến tính trên cùng của kiến trúc PhoBERT (cho vào phần layer Transformer cuối cùng của PhoBERT). Sau đó, chúng ta sẽ sử dụng transformers (Wolf et al., 2019) để fine-tune PhoBERT cho mỗi task và từng tập dữ liệu một cách độc lập. Tiếp theo, ta dùng AdamW(Loshchilov and Hulter,2019) để cố định learning rate = 1.e-5 và batch size = 32 (Liu et al., 2019).



PhoBERT Architecture

**2. Áp dụng cho đề tài:**

Trong bài tập lớn cuối kỳ, nhóm đã thực hiện dự đoán cảm xúc tích cực hoặc tiêu cực dựa trên 1 câu đánh giá bình luận của khách hàng, với tập dữ liệu gồm có 9071 data, trong đó bao gồm:

* **RevId:** Review Id
* **UserId:** Id người dùng, khách hàng
* **Comment:** đánh giá, bình luận của khách hàng
* **Image\_urls:** chứa các đường link hình ảnh khách đánh giá, gửi lên
* **Rating:** số điểm mà khách đánh giá

Dựa vào lượng dữ liệu sẵn có, ban đầu chúng ta sẽ đi vào việc xử lý dữ liệu trước khi đưa vào mô hình để tránh làm dữ liệu có nhiều dư thừa gây nhiễu trong quá trình Training. Vì dữ liệu tiếng Việt khác so với tiếng Anh là chúng có thêm dấu. Do đó việc gõ chữ sai vị trí dấu như là “Đaị học” hay là “qúa đáng” là điều dễ xảy ra. Để xử lý được vấn đề này, nhóm đã sử dụng thư viện underthesea[2] để chuẩn hóa lại vị trí của dấu câu cho đúng với chuẩn tiếng Việt, sau đó nhóm chúng em đã chuyển về dạng chữ viết thường bằng hàm lower() để cho các kí tự được thống nhất về một dạng.

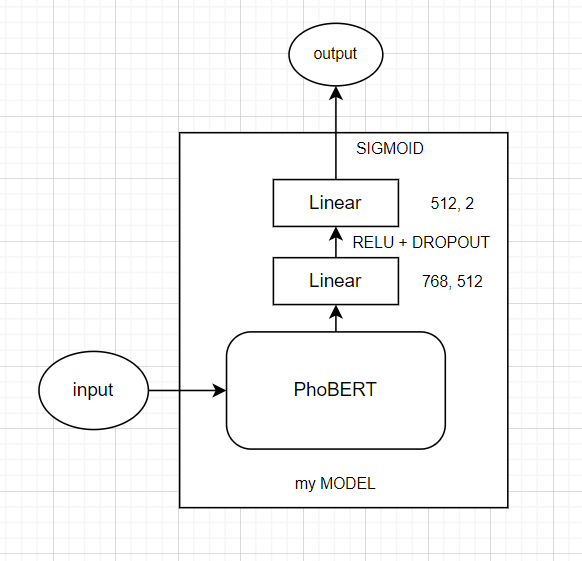
Khi thực hiện việc xử lý dữ liệu dạng văn bản, nhóm còn nhận thấy trong đó còn có các biểu tượng cảm xúc (emoji), ngoài ra còn các từ được viết dưới dạng teencode như là “kaka”, “iu”, “tuỵt zời”,…

Trong xử lý ngôn ngữ tự nhiên, chúng ta chắc hẳn không thể không nhớ đến việc xử lý các stopwords – là những từ thừa, không góp phần vào ngữ cảnh của câu từ, nếu giữ những từ này thì số lượng tính toán sẽ nhiều hơn, ngoài ra còn gây nhiễu cho việc học của máy trên model. Từ vấn đề này, nhóm đã thu thập được một danh sách những từ stopwords trong tiếng Việt[3] để làm gọn đi câu từ trong mỗi bình luận, đánh giá của khách hàng.

Do đầu vào của mô hình PhoBERT là các dữ liệu đã được nén, thế nên, nhóm chúng em đã sử dụng hàm encode() của bpe trong thư viện fastBPE để nén dữ liệu. Rồi dữ liệu thu được sẽ được chuyển về dạng tensor và sử dụng DataLoader của torch để tạo dataloader.

Do fine-tuning có thể hiểu là lấy 1 pre-trained model, lược bớt một số layer, đóng băng nó lại, sau đó thêm – sửa - xóa các layer khác ở tầng trên để xây dựng model mới. Dựa vào việc fine-tuning, nhóm đã áp dụng PhoBERT của bert-transformers nhưng lược bỏ một số layer đầu tiên. Giải thích cho việc đóng băng một vài layer đầu là bởi vì các layer này đã có khả năng trích xuất thông tin mức trừu tượng thấp, khả năng này được học từ quá trình training trước đó. Ta đóng băng lại để tận dụng được khả năng này và giúp việc training diễn ra nhanh hơn (model chỉ phải update trọng số ở các layer cao).

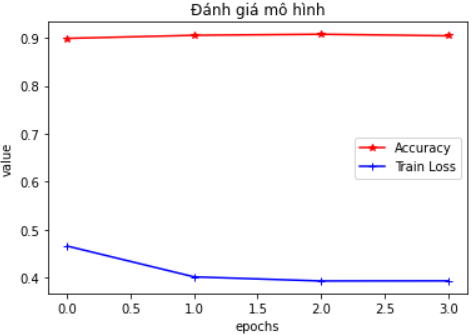
Sau đó lớp trên cùng của mô hình, nhóm chúng em thêm 2 lớp phân loại tuyến tính bằng Linear Regression để trích xuất đặc trưng, sau đó chúng em dùng hàm Sigmoid để nhận được đầu ra 0 và 1 tương ứng với tiêu cực và tích cực.



# **KẾT QUẢ THỰC NGHIỆM**

Đi đến thực nghiệp, nhóm chúng em đã thực hiện fine-tuning trong 4 training epochs, để đánh giá hiệu suất sau mỗi epoch trên tập validation, sau đó chúng em sẽ lưu lại các tham số tốt nhất cho mô hình, và dùng mô hình đó để đưa ra kết quả dự đoán tập test.csv trên Kaggle.

Sau khi dữ liệu được đưa vào mô hình, nhóm đã thu được kết quả khá là khả quan với độ chính xác trong dữ liệu training và training loss:



Training hoàn thành, nhóm chúng em đã nhận được accuracy trên tập validation khá cao và đều hơn 90% trong 4 epochs.

Và kết quả nhóm thu được trên Kaggle được 88.404% ở public test.

# **PHÂN CHIA CÔNG VIỆC**

* Vũ Minh Nhật:
  + Tham khảo Roberta, phoBERT, BERT cho bài toán xử lý Ngôn ngữ tự nhiên.
  + Tìm hiểu về underthesea.
  + Tiền xử lý dữ liệu: Chuẩn hóa lại dấu câu, xóa emoji.
  + Xây dựng mô hình fine-tuning phoBERT.
  + Viết báo cáo.
* Trần Văn Duy:
  + Tham khảo mô hình Transformers với Attentions cho bài toán xử lý ngôn ngữ tự nhiên.
  + Tìm hiểu về stopwords.
  + Tiền xử lý dữ liệu: Xóa stopwords, xóa từ lặp.
* Nguyễn Quang Minh:
  + Tham khảo mô hình LSTM, SVM cho bài toán xử lý ngôn ngữ tự nhiên.

# **tài liệu tham khảo**

[1] PhoBERT, VinAIResearch: https://aclanthology.org/2020.findings-emnlp.92.pdf

[2] Thư viện underthesea: https://github.com/undertheseanlp/underthesea

[3] Stopwords: https://github.com/stopwords/vietnamese-stopwords/blob/master/vietnamese-stopwords.txt

Roberta: https://www.kdnuggets.com/2019/09/bert-roberta-distilbert-xlnet-one-use.html