TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 8 (2023): 1491-1498
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 8 (2023): 1491-1498
https://doi.org/10.54607/hcmue.js.20.8.3620(2023)

Bài báo nghiên cứu

SỬ DỤNG MÔ HÌNH BERT ĐỂ PHÂN TÍCH THÁI ĐỘ NGƯỜI DÙNG QUA CÁC BÌNH LUẬN

[1]*Phòng Lab NLP & KD, Khoa Công nghệ Thông tin, Trường Đại học Tôn Đức Thắng, Việt Nam*
[2]*Phòng Khoa học Công nghệ, Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh, Việt Nam*
[3]*Khoa Công nghệ Thông tin, Trường Đại học Phan Thiết, Việt Nam*
[*]*Tác giả liên hệ: Trần Thanh Phước – Email: tranthanhphuoc@tdtu.edu.vn*
Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 08-11-2022; ngày duyệt đăng: 21-02-2023

TÓM TẮT

Trong xã hội ngày nay, sự phát triển các trang thương mại điện tử, mạng xã hội ngày càng tăng trưởng mạnh, đi kèm với thương mại điện tử, mạng xã hội chắc chắn không thể thiếu những bình luận thể hiện thái độ của người dung đối với một sản phẩm, vấn đề. Các doanh nghiệp luôn mong muốn có thể nắm bắt được nhu cầu, thái độ của người tiêu dùng với sản phẩm của họ đưa ra thị trường. Đây là động lực để chúng tôi nghiên cứu và ứng dụng phân tích thái độ người dùng qua các bình luận. Chúng tôi sử dụng mô hình BERT để huấn luyện dữ liệu; dữ liệu bình luận được thu thập thực tế trên trang Shopee với nhãn hàng Unilever. Bên cạnh đó, chúng tôi cũng đã so sánh giữa PhoBERT và BERT với 2 mô hình học máy học sâu khác là KNN và LSTM. Ngoài ra, chúng tôi còn tích hợp một số công nghệ tiên tiến như ReactJS cho Frontend và FastAPI cho Backend để triển khai ứng dụng lên một website thực tế nhằm tăng sự trải nghiệm của người dùng. Bước đầu cho thấy kết quả rất khả quan và có thể áp dụng cho nhiều doanh nghiệp kinh doanh khác.

Từ khóa: BERT; bình luận; thái độ người dùng; PhoBERT; phân tích cảm xúc

1. Giới thiệu

Thời gian gần đây, trên mọi thông tin truyền thông như mạng xã hội, các chương trình truyền hình, kể cả những cuộc đối thoại hằng ngày, bất chợt chúng ta đều có thể nghe thấy những từ như "trí tuệ nhân tạo" hay "AI", điều đó chứng tỏ lĩnh vực trí tuệ nhân tạo đang hết sức thịnh hành và phổ biến trong thời đại 4.0 ngày nay. Việc chọn lọc thông tin quý giá từ lượng dữ liệu khổng lồ này ngày càng có ý nghĩa hơn bao giờ hết, nó đóng vai trò là nền tảng thành công cho sự phát triển của tổ chức, doanh nghiệp, cá nhân. Các thông tin tìm được có thể được vận dụng để cải thiện hiệu quả hoạt động của hệ thống thông tin ban đầu, cải thiện thời gian tìm kiếm, khảo sát, hay đưa ra những dự đoán giúp cải thiện những hoạt động, quyết định trong tương lai. Các kĩ thuật khai thác dữ liệu (data mining), xử lí ngôn ngữ tự nhiên (Natural Language Processing – NLP) ngày càng được quan tâm và ứng dụng rộng rãi

Cite this article as: Nguyen Tu Thanh Duy, Tran Thanh Phuoc, Tran Thanh Tram, & Vo Quoc Tuan (2023). Applying BERT model on customer sentiment analysis through comments. *Ho Chi Minh City University of Education Journal of Science, 20*(8), 1491-1498.

trong nhiều lĩnh vực trong cuộc sống như giáo dục, y tế, kinh tế, giao thông…

Unilever là một thương hiệu chuyên cung cấp các sản phẩm chăm sóc sắc đẹp, hóa mĩ phẩm thiết yếu cho cuộc sống. Chính vì vậy mà lượng người tiêu dùng của Unilever là một con số khổng lồ. Sau mỗi thời điểm giới thiệu một sản phẩm mới ra thị trường thì Unilever đều có các cuộc khảo sát, thu thập ý kiến, đánh giá về chất lượng sản phẩm, độ hài lòng của khách hàng. Việc khảo sát được thực hiện thông qua khảo sát trên giấy, khảo sát trên hệ thống chăm sóc khách hàng, các hoạt động khuyến mãi, tri ân khách hàng, và những bình luận đánh giá cũng tác động phần nào đến chất lượng sản phẩm, những việc làm đó nhằm tìm ra những giải pháp tốt hơn trong quá trình kinh doanh, phục vụ cộng đồng, hoạt động của doanh nghiệp.

Vấn đề đặt ra là làm thế nào để thu thập, khai thác được ý kiến đánh giá của người tiêu dùng, việc đánh giá phân loại trở nên có giá trị mang lại kết quả tốt cho doanh nghiệp và người tiêu dùng. Trong bài báo này, chúng tôi sử dụng mô hình BERT để phân tích thái độ người dùng qua các bình luận.

2. Nội dung

2.1. Kiến thức nền tảng

2.1.1 Phân tích cảm xúc bình luận

Bài toán phân tích thái độ bình luận là một bài toán thuộc lĩnh vực xử lí ngôn ngữ tự nhiên, được sử dụng như để đánh giá những gì đang được nói về thương hiệu trên các phương tiện truyền thông, tốt hay xấu ở điểm nào. Những bình luận được chia thành tiêu cực, tích cực, trung lập nói lên thái độ của người dùng.

Phát biểu theo như góc nhìn của Machine Learning thì phân tích cảm xúc là bài toán phân lớp cảm xúc dựa trên văn bản ngôn ngữ tự nhiên. Đầu vào của bài toán là một hay một đoạn văn bản, còn đầu ra là các giá trị xác suất của N lớp cảm xúc mà ta cần xác định.

2.1.2. Mô hình KNN

KNN là một trong những thuật toán học có giám sát đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning (Guo, 2003). Khi huấn luyện, thuật toán này không học một điều gì từ dữ liệu, mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. KNN có thể áp dụng được vào bài toán học có giám sát là phân lớp.

KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong tập dữ liệu gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu đúng như câu "gần mực thì đen, gần đèn thì sáng".

2.1.3. Mô hình LSTM

LSTM là một kiến trúc đặc biệt của RNN có khả năng học được sự phục thuộc trong dài hạn (*long-term dependencies*) được giới thiệu bởi Sepp và Jurgen (Sepp & Jurgen, 1997) khắc phục được rất nhiều những hạn chế của RNN trước đây.

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc

khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác với nhau.

2.1.4. Mô hình BERT

BERT là pre-trained model (mô hình học sẵn) được Devlin và cộng sự (Devlin et al., 2019) tại Google AI Language phát triển. BERT đào tạo ra các vector đại diện cho văn bản thông qua ngữ cảnh 2 chiều trái và phải. Vector được sinh ra từ mô hình BERT được tinh chỉnh với các lớp đầu ra bổ sung đã tạo ra nhiều kiến trúc đáng kể trong nhiệm vụ xử lí ngôn ngữ tự nhiên.

Mô hình BERT được huấn luyện kết hợp theo hai chiến lược gồm Masked Language Model và Next Sentence Prediction. Mô hình BERT là một trong mô hình được chúng tôi sử dụng trong phần thử nghiệm của Phân tích thái độ người dùng qua các bình luận.

2.1.5. Mô hình PhoBERT

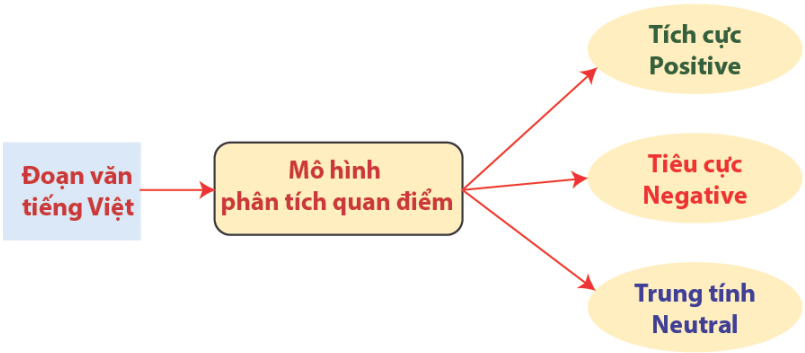
Đây là một pre-trained được huấn luyện monolingual language, tức là chỉ huấn luyện dành riêng cho tiếng Việt được (Nguyen & Nguyen, 2020) tại VINAI phát triển. Tương tự như BERT, PhoBERT cũng có 2 phiên bản là PhoBERT base với 12 transformers block và PhoBERT large với 24 transformers block.

PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT.

2.2. Xây dựng ứng dụng phân tích cảm xúc người dùng qua các bình luận

Phần kết quả và thảo luận có thể được trình bày theo từng phần riêng hoặc kết hợp thành một phần chung và có thể được chia thành các phần nhỏ hơn.

2.2.1. Mô hình hoạt động của ứng dụng phân tích cảm xúc người dùng qua các bình luận



Hình 1. *Mô hình hoạt động khái quát*

Hình 1 thể hiện quá trình hoạt động của ứng dụng phân tích cảm xúc người dùng qua các bình luận, bao gồm các bước sau:

Cho một câu tiếng Việt vào mô hình phân tích quan điểm;

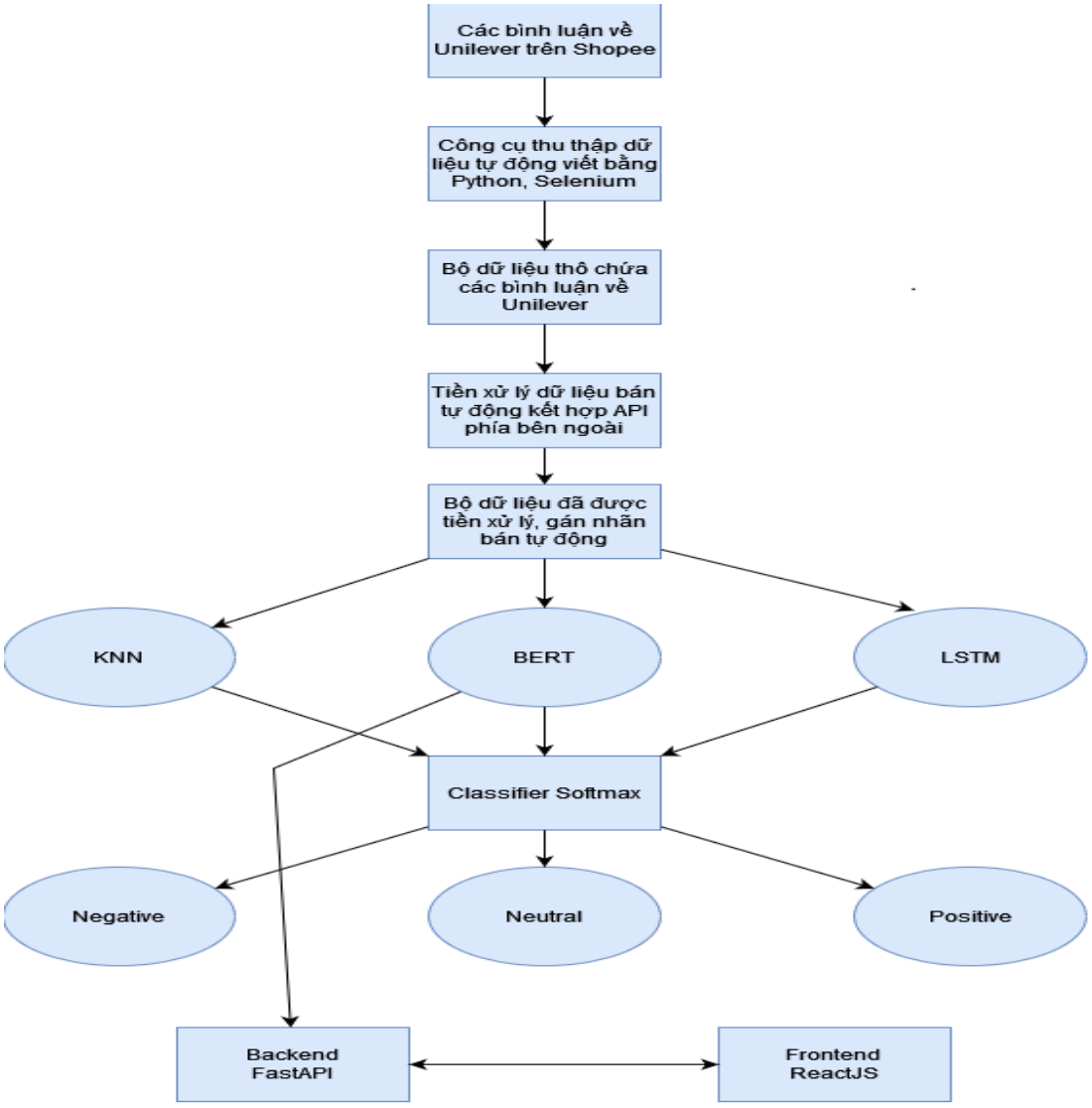
Câu được đưa vào sẽ được chỉnh sửa chính tả, vector hóa trước khi đưa vào mô hình phân tích quan điểm;

Kết quả sau khi đưa vào mô hình là một trong 3 nhãn: tích cực, tiêu cực, trung tính.

2.2.2. Kiến trúc của ứng dụng phân tích cảm xúc người dùng qua các bình luận

Hình 2 thể hiện kiến trúc của ứng dụng phân tích cảm xúc người dùng qua các bình luận. Được xây dựng trên môi trường web với hai thành phần chính Frontend và Backend. Quy trình thực hiện theo thứ tự như sau:

- Người dùng sẽ nhập một câu đánh giá, bình luận thông qua giao diện web (Frontend);
- Phía Frontend sẽ xử lí sơ chính tả trước khi đưa xuống Backend;
- Phía Backend sẽ tách từ sau đó gửi câu đã tách sang API bên ngoài để xử lí chính tả sau đó tách từ một lần nữa;
- Đưa câu nhận được vào trong mô hình phân tích cảm xúc người dùng;
- Lấy ra được kết quả từ mô hình;
- Phía Backend sẽ phản hồi kết quả lên cho Frontend hiển thị cho người dùng.



Hình 2. *Kiến trúc của ứng dụng*

2.3. Thử nghiệm

2.3.1. Dữ liệu thử nghiệm

Bộ dữ liệu được thu thập và tiền xử lí gồm 3317 câu được gán nhãn tự động thông qua thư viện hỗ trợ sau đó được kiểm tra lại và gán nhãn lại thủ công là tiêu cực, tích cực, trung lập.

Gồm:

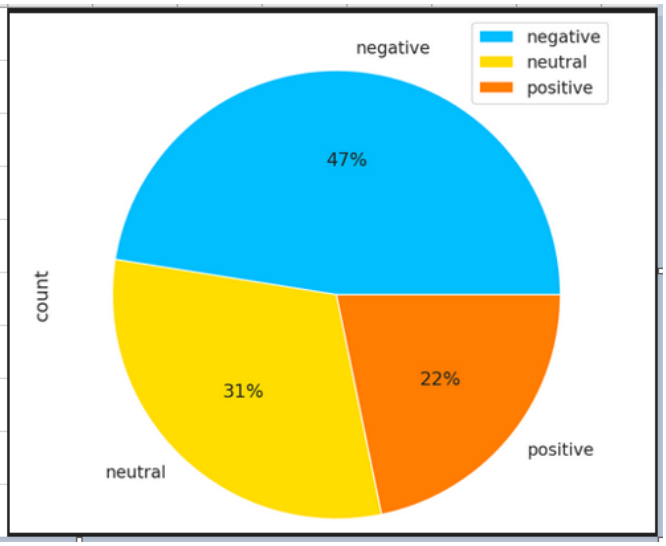
- 1575 bình luận tiêu cực chiếm 47,48%;
- 1020 bình luận trung lập chiếm 30,75%;
- 722 bình luận tích cực chiếm 21,77%.

Sau đó chia theo tỉ lệ 80% để huấn luyện, 10% để phát triển, 10% để kiểm tra

Độ dài tối đa của 1 câu trong tập dữ liệu là 250.

Số lượng từ trong tập dữ liệu là 3050.

Bộ dữ liệu	
Tổng	3317
Tiêu cực	1575
Trung lập	1020
Tích cực	722
Độ dài tối đa	250
Tổng số từ	3046
Train	80%
Test	10%
Validation	10%

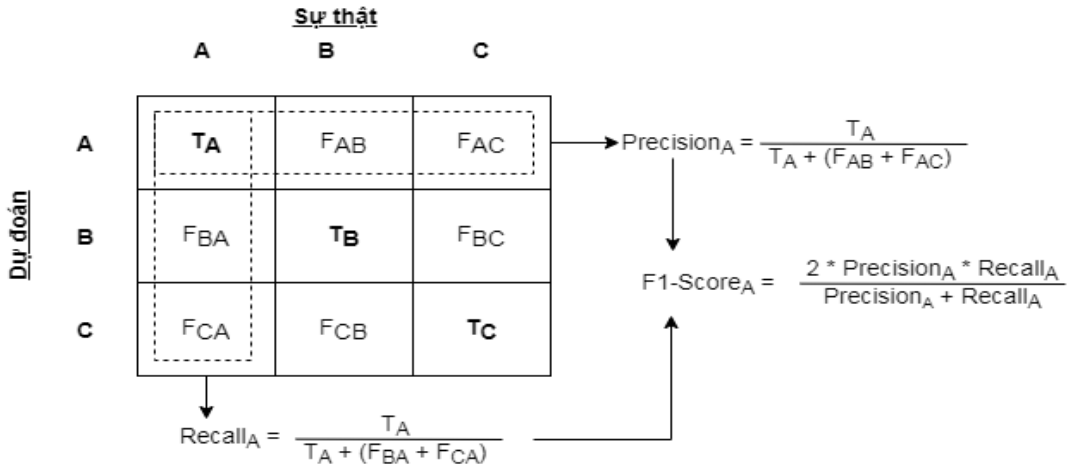


Hình 4. *Thống kê dữ liệu thử nghiệm*

2.3.2. Công cụ đánh giá

Để so sánh và đánh giá hiệu quả của các phương pháp với nhau, đề tài sử dụng các độ đo Macro Average F1-score và độ chính xác (accuracy). Trong đó, giá trị F1-score phụ thuộc vào Precison và Recall.

Macro-average F1 score (F1-Score) được tính độc lập từng lớp rồi tính trung bình lại theo công thức sau:



$$\text{Average F1-score} = \frac{N_A * \text{F1-score}_A + N_B * \text{F1-score}_B + N_C * \text{F1-score}_C}{N_A + N_B + N_C}$$

$$\text{Accuracy} = \frac{T_A + T_B + T_C}{N_A + N_B + N_C}$$

2.3.3. Cài đặt và thử nghiệm

- **Các siêu tham số trong mô hình**

Tham số	KNN	LSTM	BERT	PhoBERT
Epochs		10	10	10
Batch Size		256	16	16
Embedding		128		
Learning Rate		0.001	2e - 5	2e - 5
K	5			
Max Length		250	250	80

- **Kết quả thử nghiệm**

Mô hình	Precision	Recall	F1-Score	Accuracy
KNN	32%	43%	29%	42%
LSTM	35%	34%	34%	37%
BERT	65%	65%	65%	69%
PhoBERT	72%	71%	71%	74%

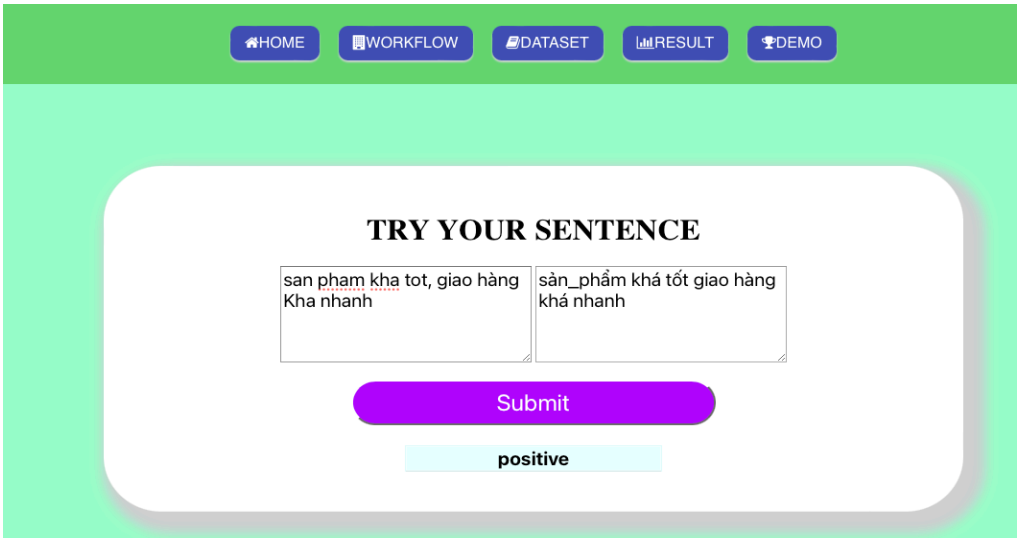
2.3.4. Thảo luận và ứng dụng

Từ kết quả thực nghiệm, chúng tôi nhận thấy rằng mô hình PhoBERT cho kết quả tốt nhất. Đó cũng là minh chứng giải thích cho sự vượt trội của mô hình PhoBERT đối với tiếng Việt. Vì vậy, chúng tôi đã sử dụng mô hình PhoBERT cho ứng dụng phân tích cảm xúc người dùng qua các bình luận của nhãn hàng Unilever trên trang thương mại điện tử Shopee.

Ứng dụng sử dụng Frontend ReactJS và Backend FastAPI.

Trang web sẽ gồm 5 lựa chọn lần lượt là:

- Home: Hiển thị thông tin chung của ứng dụng;
- Workflow: Mô hình hoạt động;
- Dataset: thống kê dữ liệu sử dụng trong ứng dụng;
- Result: kết quả đánh giá;
- Demo: trải nghiệm ứng dụng.



Hình 3. *Ảnh trang demo của ứng dụng*

3. Kết luận

Bài báo này đã đóng góp bộ dữ liệu dành cho bài toán phân tích thái độ người dùng. Đặc biệt ở việc thu thập dữ liệu một cách tự động bằng Selenium. Từ việc thu thập dữ liệu, phân tích dữ liệu, nhìn thấy được các phản hồi mang tính góp ý, mặt chưa tốt được gán nhãn tiêu cực cao hơn các hai nhãn còn lại. Điều này giúp cho doanh nghiệp xem xét cải thiện và phát triển tốt hơn.

Bài báo đề xuất mô hình BERT giải quyết tốt bài toán trên bộ dữ liệu tiếng Việt cho các kết quả cao hơn hẳn so với những mô hình trước khi BERT ra đời như KNN, LSTM.

Bài báo cũng tích hợp xây dựng website trải nghiệm mô hình BERT sử dụng các công nghệ tiên tiến hiện nay như FastAPI, ReactJS cũng như ứng dụng các API được các doanh nghiệp, tổ chức công bố ra bên ngoài phục vụ cho việc làm dữ liệu.

Trong bài báo này, dữ liệu được xử lí bán tự động, tuy nhiên với ngữ pháp tiếng Việt phức tạp, phải mất nhiều thời gian tái xử lí thủ công cho việc sửa lỗi chính tả dù đã qua một bước sửa lỗi chính tả tự động, viết tắt, loại bỏ icon, kí tự, việc gán nhãn tự động bằng thư viện không đem lại độ chính xác cao. Ngoài ra số lượng nhãn cảm xúc còn hạn chế (3 nhãn), chưa được chi tiết cụ thể hơn (nhiều hơn 3).

Trong tương lai, chúng tôi sẽ cải thiện quá trình thu thập dữ liệu liên tục theo thời gian thực, chỉnh sửa lỗi chính tả tốt hơn, gán nhãn tự động chính xác hơn để giảm thiểu việc phải tác động thủ công khi dữ liệu ngày một lớn. Nghiên cứu cải thiện, tinh chỉnh mô hình BERT để cho được kết quả tốt hơn dưới nhiều tập dữ liệu phức tạp và lớn hơn.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, 4171-4186.

Guo, G., Wang, H., David, A. B.l, & Yaxin, B. (2004). KNN Model-Based Approach in Classification. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 986-996.

Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. *Findings of the Association for Computational Linguistics*: *EMNLP 2020*, 1037-1042.

Sepp, H., Jurgen S. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

APPLYING BERT MODEL ON CUSTOMER SENTIMENT ANALYSIS THROUGH COMMENTS

Nguyen Tu Thanh Duy[1], Tran Thanh Phuoc[1*], Tran Thanh Tram[2], Vo Quoc Tuan[3]

[1]Lab NLP & KD, Information Technology Faculty, Ton Duc Thang University, Vietnam

[2]Science & Technology Department, HCMC University of Food Industry, Vietnam

[3]Information Technology Faculty, Phan Thiet University, Vietnam

[*]Corresponding author: Tran Thanh Phuoc – Email: tranthanhphuoc@tdtu.edu.vn

Received: October 11, 2022; Revised: November 08, 2022; Accepted: February 21, 2023

ABSTRACT

In today's society, the development of e-commerce sites and social networks is growing strongly, along with e-commerce and social networks, it is certainly indispensable for comments expressing the people's attitudes for a product or problem. Businesses always tried their best to capture the needs and attitudes of consumers with their products to the market. That is why in this study an application was built to analyze users' attitudes through comments using the BERT model, comments were collected on Shopee with Unilever brand. Besides comprising between PhoBERT and BERT with two other machine learning and deep learning models, KNN and LSTM were also used. In addition, the study also integrated some advanced technologies such as ReactJS for Frontend and FastAPI for Backend to deploy the application to a real website to increase the experience of a user. The initial results are very positive and can be applied to many other businesses.

Keywords: BERT; comments; customer attitudes; PhoBERT; sentiment analysis