Aspect Category Sentiment Analysis using Real-time Big Data Processing

1st Nguyễn Hoàng Quý
University of Information Technology, Ho Chi Minh City
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
20521815@gm.uit.edu.vn

2nd Trương Phước Bảo Khanh
University of Information Technology, Ho Chi Minh City
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
20520579@gm.uit.edu.vn

***Tóm tắt nội dung*—Trong thời đại công nghệ số hiện nay, lượng dữ liệu sinh ra từ các nền tảng trực tuyến là vô cùng to lớn. Muốn tận dụng, khai thác hết nguồn tài nguyên dữ liệu này, cần phải xây dựng các mô hình dự đoán, cũng như các hệ thống lưu trữ dữ liệu. Trong bài báo này, chúng tôi đề xuất các mô hình phân tích phân loại đa khía cạnh ACSA (Aspect-Category based Sentiment Analysis - ACSA) trên dữ liệu đánh giá dịch vụ khách sạn, đồng thời xây dựng hệ thống giao diện dashboard theo thời gian thực Spark, giúp cho các nhà phân tích dựa trên tình hình thực tế, kết hợp cung luc cac dư liêu vê thời điểm, giao dịch, truyền thông xã hội, dự báo thời tiết để xác định chinh xac sản phẩm phù hợp đê cung ưng cho khach hang.**

***Index Terms*—Aspect-Based Sentiment Analysis, Sentiment Analysis, Pre-trained Language Models, Big Data.**

I. Giới thiệu

Trong thời đại bùng nổ thông tin trên các nền tảng trực tuyến, nguồn thông tin người dùng tạo ra đã trở thành tài nguyên vô giá đối với doanh nghiệp và tổ chức nhằm hiểu và cải thiện sản phẩm hoặc dịch vụ của mình. Trong số các hình thức thông tin người dùng tạo ra, những đánh giá trực tuyến đóng vai trò quan trọng trong hình thành ý kiến của người tiêu dùng và ảnh hưởng đến quyết định mua hàng.

Trong lĩnh vực phân tích cảm xúc, các nhà nghiên cứu đã khám phá các kỹ thuật sáng tạo để trích xuất thông tin ý nghĩa từ những kho dữ liệu lớn này. Một phương pháp tiếp cận như vậy là phân tích cảm xúc dựa trên khía cạnh và danh mục (Aspect-Category based Sentiment Analysis - ACSA), tập trung vào việc phân tích cảm xúc được diễn đạt về các khía cạnh hoặc danh mục cụ thể của một sản phẩm hoặc dịch vụ.

Trong bài viết này, chúng tôi sẽ đào sâu vào ứng dụng của ACSA trên một bộ dữ liệu đánh giá khách sạn, nhằm khám phá các mẫu cảm xúc liên quan đến các khía cạnh khác nhau của trải nghiệm lưu trú. Bằng cách phân loại các khía cạnh như chất lượng phòng, dịch vụ, vị trí, tiện nghi và nhiều hơn nữa, chúng ta có thể hiểu rõ hơn về cách khách hàng nhận thức và đánh giá các khía cạnh khác nhau của chuyến lưu trú của họ.

Việc sử dụng kỹ thuật ACSA trong bài toán đánh giá khách sạn cho phép doanh nghiệp trong ngành du lịch nhận thức được điểm mạnh và điểm yếu của mình, cải thiện chúng để nâng cao sự hài lòng của khách hàng và trải nghiệm tổng thể.

Trong khuôn khổ bài báo này, chúng tôi tiến hành khám phá phương pháp ACSA, bao gồm quá trình trích xuất khía cạnh và phân loại cảm xúc. Chúng tôi cũng sẽ thảo luận về lợi ích và thách thức liên quan đến việc áp dụng kỹ thuật này vào bộ dữ liệu đánh giá khách sạn, nhấn mạnh các ứng dụng tiềm năng và tác động của ACSA trong bối cảnh rộng hơn của phân tích phản hồi khách hàng và quyết định kinh doanh.

II. Các nghiên cứu liên quan

Aspect Category Sentiment Analysis (ACSA) trên dữ liệu tiếng Việt là một lĩnh vực nghiên cứu phát triển trong thời gian gần đây, nhằm phân tích và xác định quan điểm hoặc cảm xúc của người dùng đối với từng khía cạnh cụ thể trong một văn bản, thường là các đánh giá hoặc bình luận.(Graves and Graves, 2012) tiến hành định nghĩa lại các tác vụ trong bài toán ABSA, trong đó có tác vụ ACSA. Một số công trình nghiên cứu quan trọng về ACSA trên các bộ dữ liệu tiếng Việt có thể kể đến như (Ho et al., 2023) đề xuất một mô hình kết hợp giữa mạng neural học từ khóa và mạng neural tích chập (CNN) để phân tích ACSA trên các đánh giá sản phẩm tiếng Việt. Mô hình này có khả năng xác định khía cạnh và quan điểm từ ngữ cảm xúc tương ứng trong văn bản. (Van Thin et al., 2021) công bố hai bộ dữ liệu về domain nhà hàng, khách sạn cho hai tác vụ Aspect Category Detection và Aspect Polarity Classification. Bên cạnh đó, sự xuất hiện cũng như phát triển của các mô hình pre-trained Bert-based trên dữ liệu tiếng Việt (Nguyen and Nguyen, 2020) cũng là một kỹ thuật NLP mới bên cạnh các mô hình truyền thống như RNN, LSTM, etc.

III. Mô hình

Ở đây, chúng tôi đề xuất các mô hình ở dưới áp dụng cho bài toán ACSA.

A. Bi-directional long short term memory (Bi-LSTM)

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) là một kiến trúc đặc biệt của mạng RNN có khả năng học được sự phụ thuộc dài hạn. LSTM đã khắc phục các vấn đề của mạng RNN trước đây là Vanishing Gradient và Exploding Gradient. Tuy nhiên LSTM có cấu trúc phức tạp hơn mặc dù vẫn giữ được ý tưởng chính của RNN là sao chép các kiến trúc theo dạng chuỗi. LSTM được trang bị một ô trạng thái (Cell State) chạy xuyên suốt toàn bộ chuỗi với chỉ một vài tương tác tuyến tính nhỏ giúp cho thông tin có thể truyền trên mạng một cách ổn định. LSTM có khả năng

xóa và thêm thông tin vào ô trạng thái và điều chỉnh các luồng thông tin này. Ô trạng thái sẽ được cập nhật trong quá trình huấn luyện mô hình, tỉ lệ thông tin đi qua được quyết định bởi ba cổng chính là: input gate, forget gate, output gate.

Dù cho LSTM có thể giải quyết bài toán phụ thuộc dài hạn, nó vẫn mất một vài thông tin ngữ nghĩa trong quá trình truyền đi (forward). Do đó, việc sử dụng mô hình Bi-LSTM là cần thiết vì mô hình này có thể biểu diễn 2 chiều thông tin ngữ cảnh của từ.

B. BERT-Multilingual

BERT (Devlin et al., 2018) là mô hình transformer được sử dụng để giải quyết sự hạn chế của mạng LSTM về vấn đề phụ thuộc xa và tốc độ huấn luyện. Mô hình BERT được huấn luyện trước trên một tập dữ liệu văn bản lớn sử dụng 2 cơ chế là Masked Language Prediction và Next Sentence Prediction. Bằng cách này, BERT có thể học được các biểu diễn từ trong dữ liệu mà sau đó có thể được sử dụng để trích xuất các đặc trưng hữu ích cho các tác vụ liên quan tới ngữ cảnh từ. BERT-multilingual là một mô hình đa ngôn ngữ được huấn luyện trước trên 104 ngôn ngữ trên Wikipedia chỉ sử dụng cơ chế Masked language Prediction.

C. PhoBERT

PhoBERT (Nguyen and Nguyen, 2020) là một mô hình ngôn ngữ dựa trên kiến trúc Transformer, được đào tạo đặc biệt cho tiếng Việt. Được xây dựng trên cơ sở của RoBERTa (A Robustly Optimized BERT Pretraining Approach), PhoBERT có khả năng hiểu và biểu diễn các ngữ cảnh ngôn ngữ tiếng Việt phức tạp.

Mô hình PhoBERT được huấn luyện trước (pretrain) trên một lượng lớn dữ liệu tiếng Việt từ các trang wiki và các bài báo trên internet. Quá trình pretrain này giúp PhoBERT nắm bắt được nhiều thông tin ngôn ngữ, từ đó tạo ra các biểu diễn từ và câu có khả năng áp dụng cho nhiều tác vụ ngôn ngữ khác nhau.

PhoBERTv2 là phiên bản cải tiến của PhoBERT, nâng cao khả năng biểu diễn ngôn ngữ tiếng Việt. PhoBERTv2 được huấn luyện thêm trên 120GB dữ liệu từ bộ OSCAR-2301, điều này giúp nắm bắt nhiều ngữ cảnh và thông tin quan trọng hơn, từ đó giúp tạo ra các vector biểu diễn từ chính xác.

D. XLM-RoBERTa

XLM-RoBERTa (Conneau et al., 2019) là một mô hình đa ngôn ngữ, được huấn luyện trên nhiều ngôn ngữ khác nhau. Mô hình này là một phiên bản mở rộng từ RoBERTa (một biến thể của BERT) và có khả năng biểu diễn ngôn ngữ đa ngôn ngữ một cách hiệu quả. XLM-RoBERTa đã được áp dụng rộng rãi trong các nhiệm vụ xử lý ngôn ngữ tự nhiên, bao gồm cả nhận diện và trích xuất cảm xúc trong đoạn văn bản Tiếng Việt. Mô hình này cung cấp khả năng biểu diễn linh hoạt nhiều ngôn ngữ với hiệu suất tốt trong việc biểu diễn đặc trưng các từ.

IV. Kết quả thực nghiệm

A. Bộ dữ liệu

Trong bài toán này, chúng tôi sử dụng bộ dữ liệu VLSP 2018 (Nguyen et al., 2018) thuộc chủ đề khách sạn. Bộ dữ liệu chứa các bình luận về đánh giá của người dùng và được thu nhập trên trang web https://www.booking.com/. Bộ dữ liệu chứa 5600 bình luận được gán nhãn dựa trên các khía cạnh và 3 phân cực cảm xúc tương ứng là Positive, Neutral và Negative. Các nhãn được thể hiện ở bảng I.

Bảng I: Các khía cạnh và thuộc tính tương ứng. Dấu được biểu thị có là ✓ và không có là ×.

	General	Prices	Design & Features	Cleanliness	Comfort	Quality	Style & Options	Miscellaneous
Hotel	✓	✓	✓	✓	✓	✓	×	✓
Rooms	✓	✓	✓	✓	✓	✓	×	✓
Room_Amenities	✓	✓	✓	✓	✓	✓	×	✓
Facilities	✓	✓	✓	✓	✓	✓	×	✓
Service	✓	×	×	×	×	×	×	×
Location	✓	×	×	×	×	×	×	×
Food & Drinks	×	✓	×	×	×	✓	✓	✓

B. Tiền xử lí dữ liệu

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, tiền xử lí dữ liệu đóng vai trò vô cùng quan trọng. Đặc biệt, trong trường hợp ngôn ngữ Tiếng Việt, tiền xử lí dữ liệu trở nên càng quan trọng hơn do tồn tại nhiều biến thể của cùng một từ, đôi khi người dùng viết sai chính tả hoặc viết tắt.

Việc tiền xử lí tốt có thể giúp cải thiện đáng kể hiệu suất của mô hình. Quá trình tiền xử lí được chúng tôi mô tả ở hình 1.

Hình 1: Các bước tiền xử lí dữ liệu

Đầu tiên, chúng tôi tiến hành kiểm tra xem từ đó có phải là từ Tiếng Việt hay không. Nếu đó là từ Tiếng Việt, chúng tôi sẽ chuẩn hóa các thanh điệu về dạng chuẩn theo các quy tắc về nguyên âm và phụ âm của Tiếng Việt (ví dụ: "maý" => "máy"). Tiếp theo, chúng tôi chuyển các câu bình luận sang kiểu viết thường, rút gọn các từ kéo dài ("xịnnnn" => "xịn"), sau đó chuẩn hóa các từ viết tắt thông dụng. Cuối cùng, chúng tôi thực hiện tách từ để tạo đầu vào cho các mô hình được

Bảng II: Input và Output tương ứng của bài toán.

Input	Output
Rộng rãi KS mới nhưng rất vắng. Các dịch vụ chất lượng chưa cao và thiếu.	{HOTEL#DESIGN&FEATURES,positive}, {HOTEL#GENERAL, negative}

huấn luyện trước trên Tiếng Việt. Bảng II mô tả đầu vào và đầu ra của quá trình tiền xử lí.

C. Độ đo đánh giá

Bộ dữ liệu được chúng tôi sử dụng bị mất cân bằng nhãn đối với các khía cạnh và phân cực cảm xúc, do đó chúng tôi sử dụng các phương pháp đánh giá là Precision-macro, Recall-macro và F1-macro. Phương pháp này sẽ tính toán trên từng lớp trước khi tính trung bình tổng thể.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

D. Thiết kế thực nghiệm

Đối với mô hình Bi-LSTM, chúng tôi sử dụng PhoW2V được huấn luyện trước trên từng từ làm vector embedding với số chiều là 100. Chúng tôi sử dụng 512 cell cho lớp LSTM theo sau đó là 512 cell cho lớp Convolution.

Đối với các mô hình đa ngôn ngữ, chúng tôi không thực hiện tách từ và lấy trực tiếp vector biểu diễn đặc trưng đưa vào lớp dự đoán.

E. Kết quả thực nghiệm

Kết quả thực nghiệm được mô tả ở bảng III. Kết quả cho thấy các mô hình pre-trained đạt được kết quả tốt hơn khá nhiều so với mô hình BiLSTM với lớp embedding PhoW2V (F1-score chỉ đạt được 34.78%). Trong khi đó, mô hình PhoBERTv2 sử dụng tách từ đạt kết quả tốt nhất trên 3 thang đo với Precision, Recall và F1-score lần lượt là 56.81%, 53.96%, 53.48%.

Bảng IV là kết quả đánh giá các nhãn cảm xúc của mô hình PhoBERTv2. Nhãn Positive đạt kết quả tốt nhất trên cả ba thang đo với Precision, Recall và F1-score lần lượt là 81.42%, 80.79%, 81.10%. Bởi vì bộ dữ liệu có sự mất cân bằng giữa các nhãn do đó nhãn Positive chiếm tỷ lệ nhiều hơn hai nhãn còn lại dẫn đến kết quả không đồng đều giữa các nhãn.

Bảng III: Kết quả trên 3 thang đo của các mô hình.

	Precision	Recall	F1-score
BiLSTM-PhoW2V	36.85	34.77	34.78
Bert-multilingual	47.09	47.69	46.08
PhoBERT	50.93	47.99	48.21
XLM-roberta	53.82	50.73	50.18
PhoBERTv2	**56.81**	**53.96**	**53.48**

Bảng IV: Kết quả chi tiết các nhãn cảm xúc của mô hình PhoBERTv2.

	Precision	Recall	F1-score
Negative	66.14	58.14	61.88
Neutral	51.43	13.53	21.43
Positive	81.42	80.79	81.10

V. Big data

A. Xử lí dữ liệu lớn trong thời gian thực

1) Apache Kafka: (Martín et al., 2022) là kho lưu trữ dữ liệu phân tán được tối ưu hóa để nhập và xử lý dữ liệu trong thời gian thực. Kafka kết hợp hai mô hình nhắn tin, queuing và publish-subscribe, để cung cấp những lợi ích chính của từng mô hình cho các consumers. Queuing cho phép xử lý dữ liệu được phân phối trên nhiều phiên bản của consumers, làm cho nó có khả năng mở rộng cao. Trong bài báo này, Kafka được sử dụng để nhập, nhận dữ liệu liên tục từ trang web và gửi đến thành phần xử lý dữ liệu trực tuyến để có kết quả theo thời gian thực.

2) Apache Spark: (Shaikh et al., 2019) là một trong những mã nguồn mở được sử dụng rộng rãi nhất khi xử lý và làm việc với dữ liệu lớn. Spark không chỉ có thể xử lý khối lượng dữ liệu khổng lồ mà còn xử lý và cho ra kết quả theo thời gian thực. Trong thực tế, lượng dữ liệu lớn liên tục đổ về từ các trang mạng xã hội như Facebook, Twitter, Youtube hoặc các trang bán hàng cần được xử lý ngay lập tức. Các mô hình Social Listening có thể được xây dựng để tích hợp vào một thành phần có tên là Spark Structured Streaming bên trong Spark để xử lý dữ liệu trực tuyến từ các mạng xã hội và đưa ra kết quả theo thời gian thực.

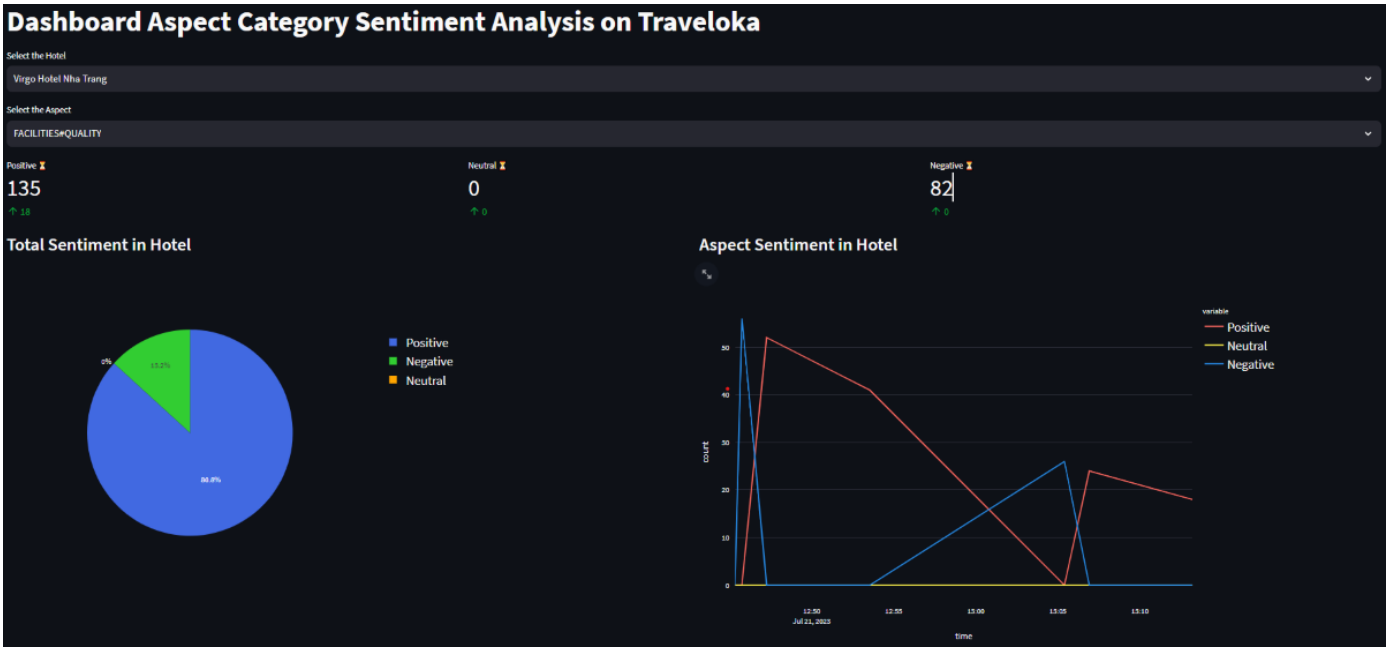
3) Streamlit: Công cụ này được phát triển để tạo ra các giao diện trực tuyến tương tự như Jupyter Notebook. Streamlit là một web framework dựa trên python để trực quan và phân tích dữ liệu một cách linh hoạt và hiệu quả. Trong bài báo này, Streamlit được sử dụng để tạo dashboard hiển thị kết quả phân tích của các khía cạnh dự đoán.

B. Hệ thống

Hệ thống của chúng tôi được xây dựng dựa trên mô hình đã đạt kết quả tốt nhất ở phần đánh giá và tích hợp nó vào Spark streaming. Các bình luận của người dùng được thu thập trực tiếp từ Traveloka. Sau đó dữ liệu sẽ được đưa vào Spark Streaming để tiến hành tiền xử lý bằng các phương pháp được mô tả ở mục IV-B.Cuối cùng, dữ liệu sau khi được xử lý sẽ được truyền vào mô hình để tiến hành dự đoán. Kết quả thống kê sẽ được trực quan hóa thông qua giao diện website.

Để tiến hành thử nghiệm, chúng tôi đã thu thập dữ liệu liên tục từ người dùng thông qua các từ khóa tên khách sạn. Các bình luận được thu thập sẽ được lưu trữ bao gồm Timestamp (thời điểm bình luận được thu thập, hotel_name (tên khách sạn) và comment (nội dung bình luận)

Giao diện website được xây dựng bằng framework Streamlit. Các kết quả dự đoán được truyền vào dashboard bao gồm các



Hình 2: Dashboard trực quan hóa các nhãn cảm xúc của từng khách sạn.

thống kê về bình luận trên từng aspect của từng khách sạn khác nhau. Người dùng có thể tùy chỉnh để theo dõi khách sạn mà mình quan tâm. Bên cạnh đó còn có biểu đồ tròn thể hiện tổng quan của khách sạn. Biểu đồ đường cho thấy số lượng bình luận theo nhãn cảm xúc qua từng thời gian cụ thể.

VI. Kết luận

Trong bài báo này, chúng tôi đã xây dựng hệ thống phân loại cảm xúc dựa trên khía cạnh theo thời gian thực sử dụng các kĩ thuật của Big Data là Spark và Kafka. Đồng thời, chúng tôi đã tiến hành thử nghiệm trên 5 mô hình và thấy rằng mô hình PhoBERTv2 đạt kết quả tốt nhất với F1-score là 53.48%. Chúng tôi cũng đã xây dựng một dashboard để thống kê và hiển thị nhãn cảm xúc của từng khách sạn theo thời gian thực. Chúng tôi thấy rằng việc xây dựng hệ thống phân loại cảm xúc theo thời gian thực là bước quan trọng trong việc nắm bắt ý kiến của người dùng và áp dụng trong nhiều lĩnh vực như dịch vụ khách hàng, quảng cáo và phân tích thị trường.

Hướng phát triển

Tài liệu

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

Thanh Trung Ho, Hien Minh Bui, and Phung Kim Thai. A hybrid model for aspect-based sentiment analysis on customer feedback: research on the mobile commerce sector in vietnam. *International Journal of Advances in Intelligent Informatics*, 9(2):273–285, 2023.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Cristian Martín, Peter Langendoerfer, Pouya Soltani Zarrin, Manuel Díaz, and Bartolomé Rubio. Kafka-ml: Connecting the data stream with ml/ai frameworks. *Future Generation Computer Systems*, 126:15–33, 2022.

Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.

Huyen TM Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310, 2018.

Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan, and Irum Nahvi. Apache spark: A big data processing engine. In *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, pages 1–6. IEEE, 2019.

Dang Van Thin, Ngan Luu-Thuy Nguyen, Tri Minh Truong, Lac Si Le, and Duy Tin Vo. Two new large corpora for vietnamese aspect-based sentiment analysis at sentence level. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–22, 2021.