Kỹ thuật làm tăng dữ liệu trong phân tích cảm xúc trên ngôn ngữ tiếng Việt

Text data augmentation techniques for sentiment analysis based on Vietnamese language

Hồ Hướng Thiên[1*]

[1]Trường Đại học Mở Thành phố Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ, Email: thien.hh@ou.edu.vn

THÔNG TIN

DOI:10.46223/HCMCOUJS.tech.vi.17.1.2202.2022

Ngày nhận: 04/03/2022

Ngày nhận lại: 15/04/2022

Duyệt đăng: 18/04/2022

Từ khóa:

đánh giá sản phẩm; khai thác văn bản; kỹ thuật tăng dữ liệu; phân tích cảm xúc; xử lý ngôn ngữ tự nhiên

Keywords:

product comments; text mining; text data augmentation; sentiment analysis; natural language processing

TÓM TẮT

Những bình luận phản hồi trong các hệ thống trực tuyến là một nguồn dữ liệu mang nhiều thông tin, cảm xúc của khách hàng về những sản phẩm hoặc dịch vụ. Những thông tin này được khai thác nhằm đem lại những ích lợi trong việc hoạch định chiến lược, quản trị khách hàng. Để đạt được những kết quả tốt đối với mô hình phân tích cảm xúc, đòi hỏi một lượng lớn dữ liệu được gán nhãn. Chi phí cho việc gán nhãn dữ liệu huấn luyện bởi con người là rất lớn. Trong nghiên cứu này chúng tôi đề xuất một mô hình làm tăng dữ liệu văn bản dựa trên các câu bình luận áp dụng cho ngôn ngữ tiếng Việt. Một số kỹ thuật cơ bản được sử dụng nhằm sinh thêm số lượng bình luận như chèn từ, thay thế từ, xóa từ. Kết quả thực nghiệm đã cho thấy hiệu quả của mô hình này.

ABSTRACT

Comments from online system are used as a data source that exist in relevant information about customer sentiment. These include sentiments toward a product or service. This is useful for making a specific decision for customers and management. In order to building a high accuracy prediction model, it requires much more labeled data. In this paper, we have investigated a simple approach for augmenting text data based on Vietnamese language comments. Four basic techniques are used to generate more new sentences such as random insertion, random swap, word replacement, word deletion. The results of experimental shows that the proposed approach is efficient.

1. Giới thiệu

Trong thời đại số hóa như hiện nay, ngày càng có nhiều người dùng đưa ra những ý kiến đóng góp trên các website thương mại, mạng xã hội. Những bình luận này rất quan trọng đối với nhiều doanh nghiệp và dịch vụ, bởi những ý kiến đó cung cấp một số lượng lớn thông tin nhằm hỗ trợ doanh nghiệp, giúp họ nâng cao chất lượng sản phẩm và dịch vụ. Do vậy, các quyết định của các công ty đối với khách hàng dựa nhiều vào những đánh giá này (Pang & Lee, 2008). Tuy nhiên, sử dụng các cách thủ công áp dụng cho việc phân tích những bình luận này sẽ mất rất

nhiều thời gian và việc tổng quát hóa các kết quả cũng rất khó khăn. Phân tích cảm xúc là một chủ đề nghiên cứu dựa trên phương pháp học máy nhằm tìm ra ý kiến của con người thông qua những câu bình luận. Thời gian gần đây, phân tích cảm xúc nhận được sự quan tâm rất lớn và đã được áp dụng rộng rãi vào các lĩnh vực như phân tích thị trường (Chopra & Sharma, 2021), phân tích tỷ lệ đánh giá sản phẩm (Sayyed & Samara, 2020), lĩnh vực chính trị (Costa, Aparicio, & Aparicio, 2021; Matalon, Magdaci, Almozlino, & Yarim, 2021), truyền thông xã hội (Drus & Khalid, 2019).

Phân tích cảm xúc có thể được xem là một bài toán trong khai thác văn bản thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Do phải hiểu được ngữ nghĩa trong bối cảnh nhất định, cho nên việc phân tích trên những đoạn văn bản ngắn khó khăn hơn nhiều so với những đoạn văn bản dài. Dựa trên mục đích của việc phân lớp, cảm xúc của một bình luận có thể được phân ra thành nhiều loại khác nhau như: Tiêu cực, tích cực, trung lập. Như vậy, việc thu thập một số lượng lớn dữ liệu không có nhãn từ các hệ thống mạng xã hội là tương đối đơn giản nhưng việc gán nhãn đầy đủ loại cảm xúc cho các câu bình luận rất tốn chi phí. Kết quả phân lớp dựa rất nhiều vào dữ liệu được gán nhãn, đồng thời yêu cầu số lượng dữ liệu đủ lớn có nhãn cho việc xây dựng mô hình. Phương pháp làm tăng thêm dữ liệu đầu vào cho mô hình là một trong những phương pháp ít tốn kém nhưng hiệu quả để giải quyết vấn đề này. Việc làm tăng thêm dữ liệu này được áp dụng rộng rãi trong các bài toán thị giác máy tính (Wang & Luis, 2017) bằng cách sử dụng những kỹ thuật đơn giản như lật hình, xoay hình, cắt hình, thay đổi tỷ lệ ảnh hoặc biến đổi màu sắc (Duong & Truong, 2019b) nhằm thay đổi hình ảnh ban đầu. Do sự phức tạp về mặt ngữ nghĩa, sự đa dạng về mặt ngữ pháp và ngữ cảnh của ngôn ngữ, cho nên phương pháp làm tăng thêm dữ liệu đối với bài toán sử dụng dữ liệu văn bản vẫn còn là vấn đề nhiều thách thức.

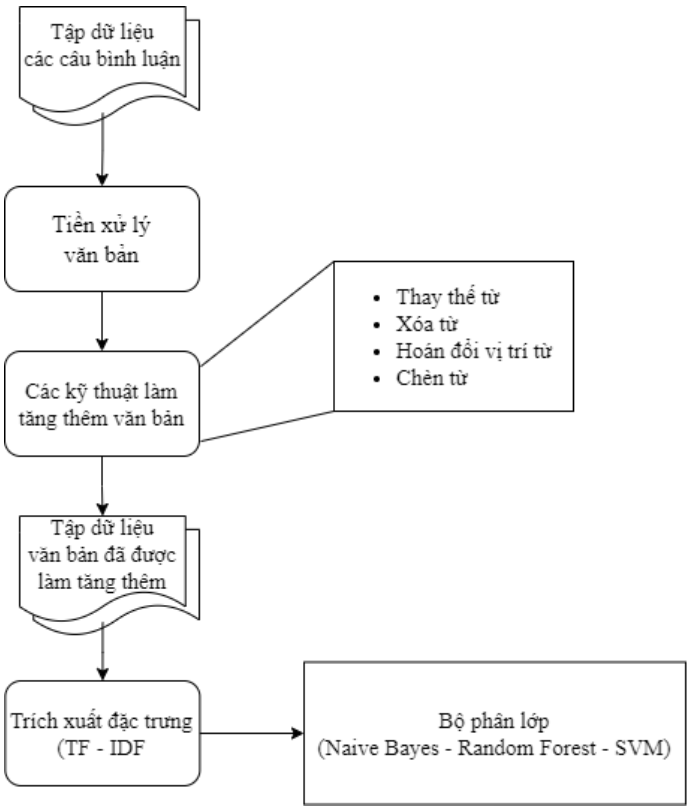
Một số nghiên cứu trong việc sinh thêm dữ liệu cho mô hình huấn luyện dựa trên phương pháp học máy nửa giám sát đã được đề xuất. Trong nghiên cứu của Lu, Zheng, Velivelli, và Zhai (2006) áp dụng phương pháp lan truyền để tạo ra dữ liệu không được gán nhãn thông qua trọng số của đồ thị vô hướng. Lee, Lay, Gan, Tan, và Abdullah (2019) kết hợp có thứ tự hai mô hình học máy giám sát và học máy không giám sát để xử lý một số lượng nhỏ dữ liệu được gán nhãn. Trong công trình nghiên cứu của Shakeel, Asim, và Imdadullah (2020), phương pháp làm tăng thêm dữ liệu và mô hình nhiều tầng nhằm phát hiện những lời diễn giải trong những đoạn văn bản ngắn được các tác giả đề xuất. Cách tiếp cận này dựa trên mối liên hệ giữa tập các văn bản với khái niệm lý thuyết đồ thị nhằm tạo ra những cặp văn bản có diễn giải và không có diễn giải. Wei và Kai (2019) giới thiệu một số kỹ thuật đơn giản cho việc tăng thêm dữ liệu văn bản với tên gọi Easy Data Augmentation (EDA), bao gồm bốn kỹ thuật cơ bản như thay thế từ, chèn từ ngẫu nhiên, thay đổi vị trí từ ngẫu nhiên, và xóa từ ngẫu nhiên. Thêm vào đó, trong công trình này phương pháp thay thế từ đồng nghĩa và phương pháp làm nhiễu ngẫu nhiên dựa trên không gian véc-tơ từ cũng được nghiên cứu áp dụng. Bài báo của Giridhara, Chinmaya, Reddy, Syed, và Andreas (2019) cũng sử dụng từ đồng nghĩa nhằm thay thế từ trong câu, nhưng từ được giới hạn ở ba từ loại là danh từ, tính từ và trạng từ. Từ ngữ dùng để thay thế trong câu được chọn lựa dựa vào việc tính toán các giá trị của mỗi từ cùng nghĩa và từ cùng nghĩa có giá trị cao nhất sẽ được sử dụng trong phương pháp này. Kết quả thực nghiệm chỉ ra rằng, trong một số trường hợp việc thay thế từ đồng nghĩa đối với động từ hoặc giới từ có thể đem lại sự sai sót ngữ pháp của câu đồng thời có thể sai ý nghĩa so với câu gốc ban đầu. Tuy nhiên, việc sai sót này không xảy ra đối với trường hợp thay thế từ đồng nghĩa là các loại từ được kể ở trên.

Phân tích cảm xúc đối với tiếng Việt trong các công trình nghiên cứu nhận được ít sự quan tâm. Trong công trình nghiên cứu (Nguyen & Duong, 2019) của mình, các tác giả đã áp

dụng những kỹ thuật cơ bản nhằm sinh ra thêm nhiều dữ liệu văn bản như thay thế từ đồng nghĩa và hoán đổi vị trí từ ngẫu nhiên trong trường hợp học nửa giám sát. Trong bài báo này, chúng tôi tập trung nghiên cứu về vấn đề trên bằng cách sử dụng một số kỹ thuật thay thế từ đồng nghĩa hoặc gần nghĩa đối với văn bản tiếng Việt. Từ gần nghĩa trong không gian véc-tơ nhúng từ được tính toán dựa trên độ đo khoảng cách cosine (Mikolov, Chen, Corrado, & Dean, 2013). Hai từ có độ đo gần bằng nhau về khoảng cách cosine thì được tính là gần nghĩa với nhau. Bài báo có bố cục được thể hiện ở các mục như sau. Phần 2 trình bày phương pháp, phần 3 mô tả thực nghiệm và những kết quả đạt được, phần cuối cùng là tổng kết một số thảo luận.

2. Mô hình làm tăng dữ liệu văn bản

Toàn bộ mô hình làm tăng dữ liệu văn bản được thể hiện ở Hình 1. Những bình luận trên các hệ thống trực tuyến đối với sản phẩm được sử dụng như là cơ sở để vạch ra nhiều quyết định về mặt quản lý. Những đánh giá này thể hiện ở những cách khác nhau, đôi khi là hình ảnh, biểu tượng, thông thường là những câu văn bản ngắn. Vì vậy, quá trình tiền xử lý văn bản là một trong những bước chính nhằm làm "sạch" cho những bình luận này. Các ký tự trong dữ liệu văn bản không mang ý nghĩa cảm xúc sẽ được loại bỏ khỏi tập dữ liệu huấn luyện. Một số việc cơ bản trong quá trình tiền xử lý như tách từ, loại bỏ URL, hashtag, địa chỉ email, các biểu tượng, số, những ký tự trùng, loại bỏ dấu câu (dấu chấm, dấu phẩy, dấu hai chấm, …); chuyển tất cả văn bản về ký tự thường. Trong công trình nghiên cứu này, chúng tôi tập trung vào việc tiền xử lý bao gồm tách từ, loại bỏ từ stopword và xử lý những từ phủ định.



Hình 1. Các bước thực hiện trong mô hình

Do sự phức tạp về mặt ngữ pháp đối với ngôn ngữ tiếng Việt, cho nên việc phân đoạn là một việc cần thiết và quan trọng. Trong tiếng Việt có hai loại từ là từ đơn và từ ghép. Điều này có nghĩa là một từ khi đứng riêng một mình sẽ mang một ý nghĩa và khi ghép chung với một từ khác lại mang một ý hoàn toàn khác khi đứng riêng. Ví dụ từ "*quê*" và từ "*hương*", khi đứng riêng lẻ hai từ này có ý nghĩa khác hoàn toàn so với khi được ghép chung ("*quê hương*") với nhau. Do vậy,

chúng ta cần một thư viện xử lý đủ tốt, có độ chính xác cao để thực hiện việc phân đoạn này. Trong nghiên cứu này, thư viện pyvi được sử dụng để phân đoạn toàn bộ dữ liệu văn bản. Pyvi là một thư viện xử lý ngôn ngữ tự nhiên dành cho tiếng Việt, được viết bằng ngôn ngữ Python. Bên cạnh đó, ở bước tiền xử lý, những từ stopword cũng được chúng tôi loại ra khỏi tập dữ liệu đầu vào. Từ stopword là từ có mặt nhiều ở các văn bản mặc dù chúng không mang ý nghĩa về nội dung, chúng chỉ có ý nghĩa về mặt ngữ pháp. Một số từ stopword trong ngôn ngữ tiếng Việt như: *thì, là, vì, vậy, mà, cho nên,* … Những từ stopword này được lựa chọn dựa trên việc tính toán giá trị TF-IDF (Term Frequency - Invert Document Frequency) của mỗi từ. Đối với những từ mang ý nghĩa phủ định, dựa vào công trình nghiên cứu (Bui, 2014), chúng tôi xây dựng một danh sách các từ phủ định thường có trong ngôn ngữ tiếng Việt. Ví dụ một số từ mang ý nghĩa phủ định trong ngôn ngữ tiếng Việt: *không, chẳng, chưa, chả, đâu, đâu có, nào, nào có, khỏi, ứ*. Trước tiên chúng xác định các từ phủ định trong các câu, sau đó kết hợp với những từ mang ý nghĩa cảm xúc (Vu & Park, 2014), thêm vào một từ NOT_ để nhận biết là tích cực hoặc tiêu cực.

Tất cả các bình luận sau khi phân đoạn sẽ được mã hóa thành véc-tơ đặc trưng từ. Hai phương pháp trích xuất đặc trưng được xem xét sử dụng như túi từ (Bag of Words) và TF-IDF. Hai phương pháp này đơn giản nhưng hiệu quả đối với việc biểu diễn dữ liệu văn bản (Ahuja, Chung, Kohli, Gupta, & Ahuja, 2019). TF (Term Frequency) là tần suất xuất hiện của từ trong một đoạn văn bản. TF của một từ được tính bằng cách lấy số lần xuất hiện của từ đó chia cho tổng số từ có trong trong đoạn văn.

$$TF(t) = \frac{f(t,d)}{T} \quad (1)$$

Với: t là từ trong đoạn văn, $f(t,d)$ là số lần có mặt của từ, *T* là tổng số từ của đoạn văn.

Mặc dù có nhiều từ có mặt trong hầu hết các văn bản, nhưng những từ này không chứa đựng ý nghĩa trong việc nhận dạng cảm xúc chứa đựng bên trong từ đó. Ví dụ những 1từ như *thì, là, mà, vậy,* ... Qua đó, ta thấy mức độ quan trọng của mỗi từ trong văn bản là khác nhau. Có những từ xuất hiện nhiều nhưng không quan trọng, ngược lại có nhiều từ xuất hiện ít nhưng lại quan trọng. Vì vậy, tính IDF (Invert Document Frequency) nhằm tìm ra mức độ quan trọng của một từ đối với văn bản. Giá trị này được tính bằng cách lấy logarit của tổng số văn bản có trong bộ dữ liệu chia cho số lượng văn bản có từ *t* xuất hiện.

$$IDF\ (t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (2)$$

Với: N là tổng số văn bản trong bộ dữ liệu và mẫu số là số lượng văn bản có chứa từ *t*. Như vậy, giá trị TF-IDF được tính như công thức bên dưới:

$$TF - IDF(t,d,D) = TF(\mathrm{t})\ x\ IDF(t,D) \quad (3)$$

Bốn kỹ thuật cơ bản nhằm tăng dữ liệu văn bản (Wei & Kai, 2019) được chi tiết như sau:

(1) Thay thế từ: Nhiều công trình nghiên cứu đã sử dụng WordNet cho việc thay thế từ đồng nghĩa. Nhưng đối với ngôn ngữ tiếng Việt, không có bộ WordNet đủ tốt cho việc thay thế này. Vì vậy, từ gần nghĩa sẽ được dựa trên khoảng cách cosine trong không gian vec-tơ nhúng từ Word2vec (Mikolov & ctg., 2013). Trong bài báo này, chúng tôi sử dụng mô hình tiền huấn luyện (Vu, 2016) cho việc thực nghiệm kết quả.

(2) Chèn từ: Kỹ thuật này được sử dụng để tìm ra những từ đồng nghĩa có trong câu, sau đó chèn những từ đồng nghĩa này vào cuối câu.

(3) Thay đổi vị trí từ: Kỹ thuật này sẽ được thực hiện hoán đổi *n* lần, với *n* bằng số lượng từ có trong câu trừ đi một.

(4) Xóa từ: Một câu mới sẽ được tạo ra từ câu gốc ban đầu bằng cách xóa đi các từ thuộc loại từ động từ, trạng từ, giới từ.

Bảng 1 thể hiện các câu của một bình luận sau khi áp dụng các kỹ thuật làm tăng dữ liệu nói trên. Sau khi áp dụng bốn kỹ thuật làm tăng văn bản này, cấu trúc ngữ pháp và ý nghĩa câu có thể bị thay đổi. Tuy nhiên cảm xúc trong câu vẫn không thay đổi. Trong bài toán phân loại này, chúng ta chỉ tập trung vào cảm xúc của câu bình luận, bỏ qua việc phân tích cấu trúc ngữ pháp và bối cảnh ngữ nghĩa.

Bảng 1

Câu bình luận được tạo ra sau khi áp dụng bốn kỹ thuật làm tăng văn bản

Kỹ thuật	Câu bình luận
Câu gốc	Nhân_viên phục_vụ nhiệt_tình và lịch_sự
Thay thế từ	Nhân_viên cao_cấp hoan_nghênh và lịch_sự
Chèn từ	Nhân_vien phục_vụ nhiệt_tình và lịch_sự chuyên_viên cao_cấp
Xóa từ	Nhân_viên nhiệt_tình lịch_sự
Hoán đổi vị trí từ	nhiệt_tình Nhân_viên và phục_vụ lịch_sự

Nguồn: Đây là kết quả của công trình nghiên cứu

3. Thực nghiệm và kết quả

Trong phần thực nghiệm, để tìm ra kết qua cho mô hình đề xuất, chúng tôi sử dụng bộ dữ liệu trong công trình nghiên cứu (Nguyen & Duong, 2019). Bộ dữ liệu 1 và bộ dữ liệu 2 là hai bộ dữ liệu ngôn ngữ tiếng Việt về lĩnh vực thức ăn được thu thập tại trang web streetcodevn.com. Bộ dữ liệu 3 được thu thập tại cuộc thi AI về phân tích cảm xúc ở Việt Nam. Những đặc điểm chi tiết của ba bộ dữ liệu này được trình bày ở Bảng 2. Toàn bộ các nhận xét được chia thành hai phân lớp: tích cực và tiêu cực.

Bảng 2

Chi tiết các bộ dữ liệu sử dụng trong việc thực nghiệm

Tên bộ dữ liệu	Phân lớp	Số lượng câu bình luận	Số lượng từ
Bộ dữ liệu 1	Tích cực	15.000	2.962.235
	Tiêu cực	15.000	
Bộ dữ liệu 2	Tích cực	5.000	1.003.237
	Tiêu cực	5.000	
Bộ dữ liệu 3	Tích cực	7.383	347.733
	Tiêu cực	8.690	

Nguồn: Đây là bộ dữ liệu từ công trình nghiên cứu

Có nhiều bộ phân lớp đã đạt được kết quả tốt trong lĩnh vực xử lý ngôn ngữ tự nhiên. Trong công trình nghiên cứu (Duong & Truong, 2019a; Tun, Johnny, & Ling, 2021) các tác giả đã cho chúng ta thấy sự so sánh về tính hiệu quả của các bộ phân lớp được áp dụng đối với phương pháp làm tăng dữ liệu văn bản. Vì vậy, chúng tôi thực nghiệm trên ba bộ phân lớp được sử dụng phổ biến là Naïve Bayes, Random Forest và Support Vector Machine. Việc thực nghiệm được thực hiện trên ngôn ngữ Python với cấu hình máy tính CPU Core I7, bộ nhớ RAM 8Gb.

Sau khi thực hiện việc tiền xử lý, các kỹ thuật làm tăng dữ liệu được áp dụng nhằm tạo ra thêm nhiều câu bình luận. Bảng 3 thể hiện tổng số lượng câu nhận xét và tổng số lượng từ đối với mỗi bộ dữ liệu trong hai thời điểm trước và sau khi áp dụng các bước làm tăng dữ liệu. Số lượng từ trong bộ dữ liệu 1 sau khi áp dụng các kỹ thuật là hơn mười triệu từ. Ba bộ phân lớp phổ biến (Duong & Truong, 2019a) được áp dụng trong nghiên cứu này bao gồm Naïve Bayes (NB), Random Forest (RF) và Support Vector Machine (SVM). Kết quả phân loại trong hai tình huống có áp dụng và không áp dụng các kỹ thuật làm tăng dữ liệu văn bản được thể hiện ở Bảng 4. Kết quả trung bình bộ phân lớp Naïve Bayes đạt được là 84% trong cả hai tình huống trước và sau khi áp dụng các kỹ thuật làm tăng văn bản. Bộ phân lớp Support Vector Machine đạt được ở mức 87%. Độ chính xác cao nhất ở bộ phân lớp Random Forest với kết quả 95% sau khi áp dụng các kỹ thuật làm tăng dữ liệu, tăng gần 10% so với trước khi áp dụng. Với kết quả này, chứng tỏ các kỹ thuật làm tăng dữ liệu văn bản được đề xuất trong nghiên cứu này đã đạt được hiệu quả.

Bảng 3

Tổng số câu bình luận và tổng số từ trong mỗi bộ dữ liệu sau khi áp dụng các kỹ thuật kể trên

Bộ dữ liệu	Bộ dữ liệu 1		Bộ dữ liệu 2		Bộ dữ liệu 3	
	Trước khi áp dụng	Sau khi áp dụng	Trước khi áp dụng	Sau khi áp dụng	Trước khi áp dụng	Sau khi áp dụng
Số lượng câu bình luận	30.000	120.000	10.000	40.000	16.073	64.292
Số lượng từ	2.962.235	10.438.550	1.003.237	3.534.413	347.733	1.287.089

Nguồn: Đây là kết quả của công trình nghiên cứu

Bảng 4

Kết quả trước và sau khi áp dụng các kỹ thuật làm tăng văn bản

Bộ phân lớp	Trước khi áp dụng			Sau khi áp dụng		
	Naïve Bayes	Random Forest	SVM	Naïve Bayes	Random Forest	SVM
Bộ dữ liệu 1	82	84	86	83	97	86
Bộ dữ liệu 2	82	83	85	84	97	89
Bộ dữ liệu 3	88	91	91	85	92	88
Trung bình	**84**	**86**	**87**	**84**	**95**	**87**

Nguồn: Đây là kết quả của công trình nghiên cứu

4. Phần kết luận

Phương pháp tạo thêm dữ liệu văn bản dựa trên bốn kỹ thuật đã được chúng tôi trình bày và áp dụng đối với bài toán phân tích cảm xúc. Kết quả thực nghiệm trên ba bộ dữ liệu cho thấy hiệu

quả của mô hình được chúng tôi đề xuất. Bằng cách sử dụng các kỹ thuật đơn giản cho việc thay thế và chèn từ, cùng với việc sử dụng những bộ phân lớp phổ biến, độ chính xác của mô hình đã cải thiện được gần 10%. Với công trình nghiên cứu này, chúng tôi sẽ tiếp tục xây dựng bộ từ vựng về cảm xúc đối với ngôn ngữ tiếng Việt nhằm làm cho phương pháp này nâng co thêm hiệu quả.

Tài liệu tham khảo

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science, 152*, 341-348. doi:10.1016/j.procs.2019.05.008

Bui, H. T. (2014). Nhóm hư từ mang ý nghĩa phủ định trong tiếng Việt [Function words of negation in Vietnamese]. *Tạp chí Ngôn Ngữ & Đời Sống, 4*(222), 12-20.

Chopra, R., & Sharma, G. (2021). Application of artificial intelligence in stock market forecasting: A critique, review, and research agenda. *Journal of Risk and Financial Management, 14*(11), Article 256. doi:10.3390/jrfm14110526

Costa, C., Aparicio, M., & Aparicio, J. (2021, October). Sentiment analysis of portuguese political parties communication. *The 39th ACM International Conference on Design of Communication,* 63-69. doi:10.1145/3472714.3473624

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, *161*, 707-714. doi:10.1016/j.procs.2019.11.174

Duong, T. H., & Truong, V. H. (2019a). A survey on the multiple classifier for new benchmark dataset of Vietnamese news classification. *11th International Conference on Knowledge and Smart Technology (KST),* 23-28. doi:10.1109/KST.2019.8687509

Duong, T. H., & Truong, V. H. (2019b). Data augmentation based on color features for limited training texture classification. *4th International Conference on Information Technology (InCIT),* 208-211. doi:10.1109/INCIT.2019.8911934

Giridhara, P. K. B., Chinmaya, M., Reddy, K. M. V., Syed, S. B., & Andreas, R. D. (2019, February). A study of various text augmentation techniques for relation classification in free text. *8th International Conference on Pattern Recognition Applications and Methods,* 360-367. doi:10.5220/0007311003600367

Lee, S., Lay, V., Gan, K. H., Tan, T. P., & Abdullah, R. (2019). Semi-supervised learning for sentiment classification using small number of labeled data. *Procedia Computer Science, 161*, 577-584. doi:10.1016/j.procs.2019.11.159

Lu, X., Zheng, B., Velivelli, A., & Zhai, C. (2006). Enhancing text categorization with semantic-enriched representation and training data augmentation. *Journal of the American Medical Informatics Association: JAMIA, 13*(5), 526-535. doi:10.1197/jamia.M2051

Matalon, Y., Magdaci, O., Almozlino, A., & Yarim, D. (2021). Using sentiment analysis to predict opinion inversion in tweets of political communication. *Scientific Reports, 11*(1), Article 7250. doi:10.1038/s41598-021-86510-w

Mikolov, T., Chen, K., Corrado, G., & Dean, Y. (2013). *Efficient estimation of word representations in vector space*. Truy cập ngày 10/10/2021 tại https://arxiv.org/pdf/1301.3781.pdf

Nguyen, K. N. D., & Duong, T. H. (2019). One-document training for Vietnamese sentiment analysis. *Computational Data and Social Networks, 11917*, 189-200. doi:10.1007/978-3-030-34980-6_21

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1/2), 1-135. doi:10.1561/1500000011

Sayyed, J., & Samara, M. (2020). Sentiment analysis on large scale Amazon product reviews. *International Journal of Scientific Research in Computer Science and Engineering, 8*(1), 7-15.

Shakeel, M. H., Asim, K., & Imdadullah, K. (2020). A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. *Information Processing & Management, 57*(3), Article 102204. doi:10.1016/j.ipm.2020.102204

Tun, W., Johnny, K. W. W., & Ling, S. H. (2021). Hybrid random forest and support vector machine modeling for HVAC fault detection and diagnosis. *Sensors, 21*(24), Article 8163. doi:10.3390/s21248163

Vu, S. (2016). *Pre-trained word2vec models for Vietnamese*. Truy cập ngày 10/10/2021 tại https://github.com/sonvx/word2vecVN

Vu, S., & Park, S. B. (2014). Construction of Vietnamese sentiwordnet by using Vietnamese dictionary. *The 40th Conference of the Korea Information Processing Society*, 745-748. doi:10.48550/arXiv.1412.8010

Wang, J., & Perez, L. (2017). *The effectiveness of data augmentation in image classification using deep learning.* Truy cập ngày 10/10/2021 tại https://arxiv.org/pdf/1712.04621.pdf

Wei, J., & Kai, Z. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6381-6387. doi:10.48550/arXiv.1901.11196