KHẢO SÁT CÁC MÔ HÌNH PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

NGUYỄN CHÍ HIẾU
Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh
nguyenchihieu@iuh.edu.vn
DOIs: https://doi.org/10.46242/jstiuh.v57i03.4395

Tóm tắt: Phân loại văn bản là một trong những nhiệm vụ cơ bản của Xử lý ngôn ngữ tự nhiên, được ứng dụng rộng rãi trong phân tích tình cảm, phát hiện spam, gắn nhãn chủ đề, phát hiện ý định... Với sự bùng nổ của các nguồn thông tin trên Web, mạng xã hội… làm cho nó ngày càng trở nên quan trọng và thu hút nhiều nhà nghiên cứu. Nhiều phương pháp lựa chọn đặc trưng và thuật toán phân loại đã được đề xuất sử dụng. Tuy nhiên, sự gia tăng nhanh chóng của dữ liệu lớn đang tạo ra thách thức đối với việc phân loại văn bản nói chung và tiếng Việt nói riêng, chẳng hạn như vấn đề mở rộng ứng dụng, khả năng phân loại các vấn đề xã hội... Mục đích của báo cáo này là khảo sát các nghiên cứu về phân loại văn bản, trong đó có tiếng Việt, nhằm cung cấp cho bạn đọc một cái nhìn tổng quan về các công nghệ phân loại văn bản hiện có và đề xuất cách giải quyết vấn đề thách thức trong phân loại văn bản.

Từ khóa: Phân loại văn bản, tiếng Việt, học có giám sát, học bán giám sát

1. GIỚI THIỆU

Phân loại văn bản (*Text Classification*) là một kỹ thuật máy học (*Machine Learning*) tự động gán các nhãn (*tags*) hoặc danh mục (*categories*) cho văn bản. Sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên *(NLP: Natural Language Processing)* và máy học, bộ phân loại văn bản có thể phân tích và sắp xếp văn bản theo danh mục, chủ đề và ý định của khách hàng… nhanh hơn và chính xác hơn con người. Với dữ liệu đổ về từ nhiều nguồn khác nhau, bao gồm email, chát, web, phương tiện truyền thông xã hội, đánh giá trực tuyến, phiếu hỗ trợ, phản hồi, khảo sát… Nếu làm thủ công, con người khó theo kịp được yêu cầu. Chỉ riêng trên Facebook Messenger, 20 tỷ tin nhắn được trao đổi giữa doanh nghiệp và người dùng hàng tháng [1]. Để giải quyết vấn đề này, các kỹ thuật của trí tuệ nhân tạo đã được áp dụng, cụ thể là các kỹ thuật máy học để phân loại văn bản là một kỹ thuật quan trọng để tổ chức và quản lý thông tin. Các nghiên cứu sử dụng nhiều loại kỹ thuật phân loại khác nhau, bao gồm mạng nơ-ron, cây quyết định, k-láng giềng gần nhất, hỗ trợ vectơ máy, Naïve Bayes, phương pháp dựa trên luật sinh…, đã được phát triển [2–3]. Nhiều ứng dụng phân loại văn bản hiệu quả và thiết thực trong các lĩnh vực như truy xuất thông tin, lọc văn bản, phân loại tin bài, phân loại thư điện tử, phân loại các trang web, phân loại các bài báo học thuật sử dụng các lĩnh vực kỹ thuật và tên miền phụ, lọc thư rác và khiêu dâm, tin sinh học, tự động hóa dịch vụ khách hàng, phân loại chủ đề và phân tích tình cảm… [4–11].

Tuy nhiên, sự gia tăng nhanh chóng của dữ liệu lớn đang tạo ra thách thức đối với việc phân loại văn bản nói chung và tiếng Việt nói riêng, chẳng hạn như vấn đề mở rộng ứng dụng, khả năng phân loại các vấn đề xã hội... Mục đích của báo cáo này là khảo sát các nghiên cứu về phân loại văn bản, trong đó có tiếng Việt, nhằm cung cấp cho bạn đọc một cái nhìn tổng quan về các công nghệ phân loại văn bản hiện có và đề xuất cách giải quyết vấn đề thách thức trong phân loại văn bản.

Bài báo được tổ chức như sau: Phần 2 bắt đầu với phần giới thiệu ngắn gọn về phân loại văn bản để cung cấp khái niệm cơ bản và kiến thức nền tảng. Trong Phần 3, chúng tôi xem xét một số phương pháp phân loại văn bản phổ biến. Phần 4 khảo sát về phân loại văn bản tiếng Việt và những giơi hạn của các phương pháp đang áp dụng và đề xuất một số hướng tiếp cập với phân loại văn bản dữ liệu lớn. Phần 5 kết thúc nghiên cứu này.

2. PHÂN LOẠI VĂN BẢN

Phân loại văn bản là quá trình phân loại một luồng tài liệu đến thành các loại tài liệu theo yêu cầu, bằng cách sử dụng các bộ phân loại học được từ các mẫu huấn luyện. Cách tiếp cận Máy học để phân loại văn bản đã trở nên phổ biến và cuối cùng đã trở thành phương pháp phổ biến [13]. Sử dụng học máy, là tìm hiểu các bộ phân loại từ các ví dụ tự động sau đó thực hiện phân loại tài liệu. Đầu vào cho bộ phân loại là một tập hợp các bản ghi huấn luyện, mỗi bản ghi trong số đó được gắn nhãn lớp (loại). Một tập hợp các giá

trị thuộc tính xác định mỗi bản ghi. Mục đích là tạo ra một mô hình hoặc mô tả cho mỗi lớp về các thuộc tính. Sau đó, mô hình được sử dụng để phân loại các bản ghi trong tương lai mà các lớp của chúng chưa được biết đến. Cụ thể hơn, bộ phân loại văn bản gán giá trị *Boolean* cho mỗi cặp $(d_i, c_j) \in (D \times C)$, trong đó D là miền tài liệu và C là tập hợp các danh mục được xác định trước [13]. Nhiệm vụ là làm gần đúng hàm ϕ: $D \times C \rightarrow (1, 0)$ bằng hàm $\phi\hat{}$: $D \times C \rightarrow (1, 0)$, sao cho ϕ và $\phi\hat{}$ trùng nhau càng nhiều càng tốt. Hàm $\phi\hat{}$ được gọi là bộ phân loại. Mục tiêu của bộ phân loại là xác định và ước lượng chính xác sự trùng hợp này.

Nói chung, bài toán phân loại văn bản có thể là bài toán phân loại "nhị phân". Nếu có chính xác hai lớp hoặc bài toán "nhiều lớp" nếu có nhiều hơn hai lớp và mỗi tài liệu thuộc đúng một lớp, hoặc "phân loại nhiều nhãn" nếu một tài liệu có thể có nhiều hơn một danh mục liên quan trong một sơ đồ phân loại [4]. Hình thức cơ bản của phân loại văn bản là phân loại nhị phân, trong đó một tài liệu văn bản được cho một trong hai nhãn, thường được gọi là tích cực và tiêu cực. Các tác vụ nhiều nhãn và nhiều lớp thường được xử lý bằng cách giảm chúng thành k nhiệm vụ phân loại nhị phân, một tác vụ cho mỗi loại [4, 13]. Ví dụ, trong [15], một bài toán phân loại nhiều nhãn đã được chuyển đổi thành một tập hợp nhiều bài toán phân loại nhị phân và sau đó áp dụng mô hình mạng nơ-ron tích chập phức hợp (*CNN: Convolutional Neural Networks*) cho việc phân loại văn bản. Các bộ phân loại văn bản hiện tại không thể mô tả rõ ràng ranh giới quyết định giữa các đối tượng tích cực và tiêu cực, do sự không chắc chắn gây ra bởi việc lựa chọn đối tượng văn bản và quá trình học. Để khắc phục vấn đề này, một mô hình quyết định ba chiều đã được đề xuất gần đây. Mục tiêu của mô hình mới là giải quyết ranh giới không chắc chắn để cải thiện hiệu suất phân loại văn bản nhị phân dựa trên các kỹ thuật thiết lập thô và giải pháp trọng tâm (*centroid*) [16].

Một số thuật toán đã được đề xuất bao gồm mạng nơ-ron, cây quyết định, *K-Nearest Neighbor*, bộ phân loại *Naive Bayes*, bộ phân loại dựa trên tập hợp thô và hỗ trợ véc tơ máy (*Support Vector Machines*) [17–19]. Các thuật toán này có thể được mở rộng một cách tự nhiên cho phân loại nhiều lớp. Một cách khác để giải bài toán nhiều lớp là chuyển bài toán phân lớp nhiều lớp thành một tập các bài toán phân lớp nhị phân [20]. Đối với nhiều nhãn, nó phải được chuyển đổi thành nhãn đơn trước khi được xử lý trong phân loại nhị phân. Ít nhất bốn cách tiếp cận chuyển đổi từ tập dữ liệu nhiều nhãn thành tập dữ liệu một nhãn đã được trình bày trong [21]. Đó là gán nhãn tất cả (*ALA: All Label Assignment*), không chỉ định nhãn (*NLA: No Label Assignment*), chỉ định nhãn lớn nhất (*LLA: Largest Label Assignment*) và chỉ định nhãn nhỏ nhất (*SLA: Smallest Label Assignment*). Trong số những cách tiếp cận này, ALA thường là tốt nhất; tuy nhiên, thực tế là các tài liệu trùng lặp với các nhãn khác nhau gây nhiễu và giảm hiệu quả phân loại. ALA là chuyển đổi vấn đề 5 (PT5: Problem Transformation 5) trong [22, 23]. Một chuyển đổi mới chỉ định nhãn dựa trên Entropy (*ELA: Entropy-based Label Assignment*) đã sửa đổi ALA đã được đề xuất trong [21]. Trong [24], cung cấp thêm chi tiết về phân loại đa nhãn.

Máy học cho các nhiệm vụ phân loại văn bản có thể được phân loại thành nhiệm vụ học tập có giám sát, bán giám sát và không giám sát. Trong học có giám sát, máy học được trình bày với các mẫu dữ liệu huấn luyện bao gồm các cặp đầu vào và đầu ra, trong đó nó được yêu cầu dự đoán giá trị đầu ra của các mẫu mới dựa trên giá trị đầu vào của chúng. Học có giám sát yêu cầu một tập mẫu huấn luyện. Tuy nhiên, đôi khi các mẫu huấn luyện có thể bị thiếu hoặc không đủ nhãn cần thiết mặc dù có sẵn. Bài toán như vậy được gọi là phân loại văn bản bán giám sát.

Phương pháp tiếp cận bán giám sát đã được đề xuất trong [25] để tìm hiểu các bộ phân loại chỉ từ các mẫu một phần được gán nhãn (các tài liệu huấn luyện được phân loại trước thành một tập hợp các lớp khả thi với chỉ một lớp chính xác). Các kỹ thuật phân loại văn bản có giám sát và bán giám sát ít nhiều dựa vào các mẫu được phân loại trước để tìm hiểu các bộ phân loại. Học không giám sát đề cập đến vấn đề cố gắng tìm kiếm cấu trúc ẩn trong dữ liệu không được gắn nhãn. Các tác giả của [26] đề xuất xây dựng mô hình phân loại cho một lớp mục tiêu không có các mẫu huấn luyện liên quan, bằng cách phân tích các lớp bổ trợ tương quan.

3. THUẬT TOÁN PHÂN LOẠI VĂN BẢN

Phân lớp văn bản là quá trình phân loại văn bản thành nhóm các từ. Bằng cách sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên, phân loại văn bản có thể tự động phân tích văn bản và sau đó gán các thẻ hoặc danh mục được xác định trước dựa trên ngữ cảnh của nó. Hệ thống phân loại văn bản chủ yếu dựa trên ba cách tiếp cận: luật sinh (Rule-based) [27], máy học (*Machine Learning*) [3] và hệ thống lai (*Hybrid System*) [28].

Trong cách tiếp cận dựa trên luật sinh, các văn bản được tách thành một nhóm có tổ chức bằng cách sử dụng một tập hợp các quy tắc (luật sinh) của ngôn ngữ bằng phương pháp thủ công. Các luật sinh đó dùng để xác định danh sách các từ được đặc trưng bởi các nhóm cần phân loại trong văn bản. Ví dụ, những từ như Donald Trump và Boris Johnson sẽ được phân vào nhóm chính trị. Những người như Nadal và Ronaldo sẽ được xếp vào nhóm thể thao…

Hệ thống phân loại dựa trên máy học thực hiện phân loại dựa trên quan sát trước đây từ các tập dữ liệu huấn luyện có gán nhãn (label). Hệ thống sẽ học từ dữ liệu có nhãn, sau đó nó sử dụng tri thức đã học được để dự đoán nhãn cho các dữ liệu mới.

Hệ thống lai là hệ thống kết hợp bộ phân loại máy học với luật sinh, để cải thiện kết quả của hệ thống. Có thể dễ dàng tinh chỉnh các hệ thống kết hợp này bằng cách thêm các luật cụ thể cho các thẻ (tags) bị xung đột mà hệ thống máy học chưa phân loại chính xác được.

Có nhiều thuật toán phân loại được áp dụng vào phân loại văn bản [29], trong mục này chúng tôi giới thiệu một số thuật toán phân loại văn bản phổ biến nhất bao gồm thuật toán Naive Bayes [30], Máy hỗ trợ véctơ (*SVM: Support Vector Machines*) [31], K-Láng giềng gần nhất (*KNN: K-Nearest Neighbour*)[32] và Học sâu (*Deep Learning*) [33].

3.1 Thuật toán Naive Bayes

Phương pháp Naive Bayes là một tập hợp các thuật toán học có giám sát dựa trên việc áp dụng định lý Bayes với giả định "ngây thơ" (*Naive*) về sự độc lập có điều kiện giữa mọi cặp đặc trưng cho giá trị của biến lớp. Định lý Bayes phát biểu mối quan hệ giữa biến lớp cho trước *y* và vectơ đặc trưng phụ thuộc x_1 đến x_n, theo công thức (1).

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1,\dots,x_n)} \quad (1)$$

Sử dụng giả thiết độc lập có điều kiện *Naive* rằng:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (2)$$

với tất cả *i*, mối quan hệ này được đơn giản hóa thành

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\dots,x_n)} \quad (3)$$

Vì P (x_1,…, x_n) là hằng số cho đầu vào, chúng ta có thể sử dụng quy tắc phân loại sau:

$$\begin{gathered} P(y \mid x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y) \\ \Downarrow \\ \hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y) \end{gathered} \quad (4)$$

Và chúng ta có thể sử dụng ước lượng xác suất lớn nhất *(MLE: Maximum Likelihood Estimation)* hoặc tối đa Posteriori *(MAP: Maximum A Posteriori)* để ước lượng P (y) và $P(x_i \mid y)$; sau đó là tần xuất tương đối của lớp *y* trong tập huấn luyện. Các bộ phân loại Naive Bayes khác nhau chủ yếu khác nhau bởi các giả định đưa ra liên quan đến phân phối $P(x_i \mid y)$. Bất chấp những giả định được đơn giản hóa, các bộ phân loại Naive Bayes đã hoạt động khá tốt trong nhiều tình huống thực tế, nổi tiếng là phân loại tài liệu và lọc thư rác. Naive Bayes yêu cầu một lượng nhỏ dữ liệu huấn luyện để ước tính các thông số cần thiết. Naive Bayes học và phân loại rất nhanh so với các phương pháp phức tạp khác. Việc tách và phân bố các đặc trưng có điều kiện của lớp có nghĩa là mỗi phân bố có thể được ước tính độc lập như phân bố một chiều. Điều này giúp giảm bớt các vấn đề của dữ liệu đa chiều. Ở một khía cạnh khác, cho dù Naive Bayes được biết đến như một công cụ phân loại tốt, nó vẫn được coi là một công cụ ước lượng tồi, vì kết quả xác suất dự đoán không được coi trọng (do giảm số chiều của dữ liệu). Việc tính $P(x_i \mid y)$ phụ thuộc vào loại dữ liệu. Có ba loại phân bố được sử dụng phổ biến là: *Gaussian Naive Bayes, Multinomial Naive Bayes* và *Bernoulli Naive*.

Mô hình Gaussian Naive Bayes được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục. Với mỗi chiều dữ liệu *i* và lớp *y* , x_i tuân theo một phân phối chuẩn có kỳ vọng σ_y phương sai μ_y, lấy xác suất tối đa theo phân bố Gaussian như công thức (5):

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (5)$$

Multinomial Naive Bayes khai triển thuật toán Naive Bayes cho dữ liệu theo phân bố đa thức và là một trong hai biến thể của Naive Bayes cổ điển được sử dụng trong phân loại văn bản (trong đó dữ liệu thường được biểu diễn dưới dạng đếm số lượng vectơ đặc trưng với các phần tử nguyên có giá trị là tần suất xuất hiện của từ đó trong tài liệu, mặc dù vectơ *tf-idf* (*tf-idf*: *Term Frequency-Inverse Document Frequency)* cũng hoạt động tốt trong thực tế). Phân phối được tham số hóa bởi vectơ $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$ cho mỗi lớp *y*, trong đó *n* là số đặc trưng *(features)* và θ_{yi} là xác suất $P(x_i \mid y)$ của đặc trưng *i* xuất hiện trong một mẫu thuộc lớp *y*. Các tham số *θy* được ước tính bằng một phiên bản làm mịn của xác xuất tối đa, tức là đếm tần số tương đối theo công thức (6):

$$\hat{\theta}_{yi} = \frac{N_{yi}+\alpha}{N_y+\alpha n} \tag{6}$$

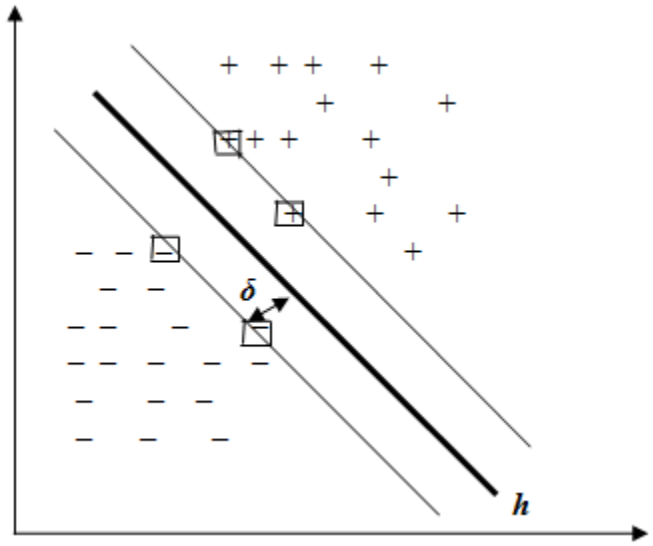
trong đó, $N_{yi} = \sum_{x \in T} x_i$ là số lần đặc trưng *i* xuất hiện trong một mẫu của lớp *y* trong tập huấn luyện T, và $N_y = \sum_{i=1}^{n} N_{yi}$ là tổng số của tất cả các đặc trưng cho lớp *y*. Thông số làm mịn α≥0 giải thích cho các đặc trưng không có trong các mẫu huấn luyện và ngăn ngừa xác suất bằng không trong các tính toán tiếp theo. Khi cho α = 1 được gọi là làm mịn *Laplace*, trong khi α <1 được gọi là làm mịn *Lidstone*.
Bernoulli Naive Bayes khai triển các thuật toán huấn luyện và phân loại Naive Bayes cho dữ liệu được phân bố theo các phân bố Bernoulli đa biến; tức là, có thể có nhiều đặc trưng nhưng mỗi đặc trưng được giả định là biến có giá trị nhị phân (tài liệu được biểu diễn bằng một vector đặc trưng với các phần tử nhị phân nhận giá trị 1 nếu từ tương ứng có trong tài liệu và 0 nếu từ không có trong tài liệu). Do đó, lớp này yêu cầu các mẫu phải được biểu diễn dưới dạng vectơ đặc trưng có giá trị nhị phân. Luật quyết định cho phân bố *Bernoulli Naive Bayes* dựa trên công thức (7):

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i) \tag{7}$$

Trong phân loại văn bản, vectơ xuất hiện từ (chứ không phải vectơ đếm từ) có thể được sử dụng để huấn luyện và sử dụng bộ phân loại này. *Bernoulli Naive Bayes* có thể hoạt động tốt hơn trên một số bộ dữ liệu, đặc biệt là những dữ liệu có kích thước nhỏ.

3.2 Máy hỗ trợ véctơ (SVM: Support Vector Machines)

Phương pháp này đã được giới thiệu bởi Joachims [31] để phân loại văn bản và V. Vapnik hình thức hóa trong tài liệu [34]. SVM dựa trên nguyên tắc giảm thiểu rủi ro cấu trúc từ lý thuyết thống kê. SVM có thể được sử dụng để tìm một siêu phẳng h có lề lớn nhất là ranh giới quyết định trong không gian tuyến tính. Nhiệm vụ phân loại thường liên quan đến dữ liệu huấn luyện và kiểm tra bao gồm một số trường hợp dữ liệu. Mỗi cá thể trong tập huấn luyện chứa một nhãn lớp và một số đặc trưng (thuộc tính). Đối với phân loại tuyến tính, SVM học các luật quyết định tuyến tính, $h(\vec{x}) = \text{sign } \vec{w}.\ \vec{x} + b$ được mô tả bởi một vectơ trọng số $\vec{w}$ và một ngưỡng b. Cho một tập huấn luyên gồm *n* mẫu huấn luyện, $T_n = ((\vec{x}1, y1), \dots (\vec{x}_n, y_n), x_i \in R^N$, and $y_i \in -1, +1$ là phân loại của T_n (dấu '+' cho mẫu tích cực và dấu '-' cho mẫu tiêu cực của lớp đã cho). SVM tìm siêu phẳng với khoảng cách *Euclid* tối đa đến các mẫu huấn luyện gần nhất để phân tách tuyến tính T_n. Khoảng cách này được gọi là biên *δ*. Đối với tập huấn luyện không phân tách, số lượng lỗi huấn luyện được đo bằng cách sử dụng biến "*slack*" ξi. Hình 1 mô tả bộ phân loại vectơ hỗ trợ nhị phân. Trong Hình 1, "+" và "-" đại diện cho các mẫu huấn luyện tích cực và tiêu cực, mặt phẳng *h* được biểu diễn bằng một đường kẻ dày hơn (giữa), nó phân tách các mẫu huấn luyện dương và âm với lề lớn nhất *(maximum)*. Các mẫu gần với siêu phẳng *h* nhất là các vectơ hỗ trợ được biểu diễn bằng các hình vuông nhỏ.



Hình 1. Phân loại vectơ hỗ trợ nhị phân

SVM yêu cầu giải pháp tối ưu hóa sau [34]:

$$\forall_{i=1}^{n} : \xi_i > 0$$
$$minimize : V(\overrightarrow{w},b,\overrightarrow{\xi}) = \frac{1}{2}\overrightarrow{w}.\overrightarrow{w} + C\sum_{i=1}^{n}\xi_i$$
$$subjectto : \forall_{i=1}^{n} : y_i[\overrightarrow{w}.\overrightarrow{x} + b] \geqslant 1 - \xi_i \quad (8)$$

Trong công thức (8), C> 0 là tham số của mẫu huấn luyện lỗi. Biên của siêu phẳng thu được là δ = 1/|| $\overrightarrow{w}$||. Bề mặt quyết định chỉ được xác định bởi các điểm dữ liệu có khoảng cách chính xác 1/|| $\overrightarrow{w}$|| từ mặt phẳng quyết định. Những điểm đó được gọi là vectơ hỗ trợ và là những phần tử hiệu quả duy nhất trong tập huấn luyện. Tuy nhiên, SVM không thích hợp để phân loại các tập dữ liệu lớn hoặc kho ngữ liệu văn bản vì độ phức tạp huấn luyện của SVM phụ thuộc nhiều vào kích thước đầu vào. Một SVM đa nhân mới được phát triển trong [35] để xử lý dữ liệu nhiều chiều. Kết quả kiểm tra của họ đã chứng minh rằng mô hình phân loại SVM đa nhân mới có độ chính xác tốt hơn so với SVM cổ điển, trong khi việc huấn luyện nhanh hơn đáng kể so với một số mô hình phân loại SVM khác.

3.3 K-Láng giềng gần nhất (KNN: K-Nearest Neighbor)

Phân loại K-Láng giềng gần nhất là một phương pháp thống kê nổi tiếng đã được nghiên cứu chuyên sâu về đối chiếu mẫu trong hơn bốn thập kỷ [32]. KNN đã được áp dụng để phân loại văn bản ngay từ giai đoạn đầu của nghiên cứu về phân loại văn bản [36]. Phương pháp KNN trở lên được sử dụng phổ biến do tính đơn giản và độ chính xác của dự đoán. Với một tài liệu đầu vào tùy ý, hệ thống xếp hạng các láng giềng gần nhất của nó trong số các tài liệu huấn luyện và sử dụng các danh mục của K- láng giềng xếp hạng cao nhất để dự đoán các danh mục của tài liệu đầu vào. Điểm tương tự của mỗi tài liệu láng giềng với tài liệu mới đang được phân loại được sử dụng làm trọng số của từng danh mục của nó; tổng trọng số của danh mục trên k lân cận gần nhất được sử dụng để xếp hạng danh mục. Độ phức tạp của mô hình được kiểm soát bởi sự lựa chọn hệ số k. Về mặt hình thức, bậc tự do của một mô hình phân loại KNN được định nghĩa là d.f. = n / k với *n* là số tài liệu trong tập huấn luyện. Khi k = 1, KNN có độ phức tạp lớn nhất và có xu hướng quá phù hợp với tập huấn luyện; khi k tăng thì độ phức tạp của mô hình giảm tương ứng [4].

3.4 Học sâu (Deep Learning)

Học sâu [37] là một tập hợp các thuật toán và kỹ thuật học máy dựa trên việc học dữ liệu huấn luyện, hệ thống tự động học và khám phá các đặc trưng cần thiết để phân loại thông qua việc xử lý nhiều lớp dữ liệu đầu vào. Học sâu đã trở thành một kỹ thuật học máy chủ đạo với khả năng thực hiện các nhiệm vụ mô hình hóa phi tuyến khác nhau bao gồm phân tích tình cảm, phân loại tin tức, trả lời câu hỏi, suy luận ngôn ngữ

tự nhiên, phân loại và trích xuất đặc trưng từ các bộ dữ liệu phức tạp. Trong những năm gần đây, nhiều kỹ thuật học sâu đã được khảo sát trong [32]. Hai kiến trúc học sâu chính để phân loại văn bản là mạng nơ-ron phức hợp (CNN: *Convolutional Neural Networks*) [38, 39] và mạng nơ-ron hồi quy (RNN: *Recurrent Neural Networks*) [40], đã được khám phá để phân loại văn bản. Học sâu đã được chứng minh là có hiệu quả để thực hiện việc học từ đầu đến cuối các biểu diễn đặc trưng phân cấp. Nó đã chứng tỏ hiệu suất vượt trội đối với phân loại văn bản phẳng [41]. Năm 2020, Shervin Minaee và các đồng nghiệp [33] đã khảo sát và đánh giá toàn diện hơn 150 mô hình dựa trên học sâu để phân loại văn bản được phát triển trong những năm gần đây và thảo luận về những đóng góp kỹ thuật, điểm tương đồng và điểm mạnh của từng mô hình. Các thuật toán học sâu yêu cầu nhiều dữ liệu huấn luyện hơn các thuật toán học máy truyền thống (ít nhất hàng triệu ví dụ được gắn thẻ). Tuy nhiên, chúng không có ngưỡng học từ dữ liệu huấn luyện, giống như các thuật toán máy học truyền thống, chẳng hạn như bộ phân loại học tập SVM và NB sẽ tiếp tục cải thiện tốt hơn khi bạn cung cấp cho chúng nhiều dữ liệu để huấn luyện hơn.

4. PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

Theo các nhà ngôn ngữ học ước lượng trên thế giới hiện nay có ít nhất 7.099 ngôn ngữ, trong đó tiếng Việt đứng thứ 23 về số lượng người sử dụng. Với sự phát triển mạnh mẽ của công nghệ thông tin, số lượng văn bản bằng các thứ tiếng được xuất bản không ngừng gia tăng, trong đó có tiếng Việt. Việc phân loại văn bản phục vụ cho các bài toán ứng dụng NLP ngày càng được hoàn thiện, nhất là các ứng dụng tiếng Anh. Do có sự khác biệt về loại hình ngôn ngữ, làm thế nào để có thể áp dụng được các nghiên cứu về phân loại văn bản của tiếng Anh cho tiếng Việt, hoặc tìm giải pháp tốt nhất cho bài toán phân loại văn bản tiếng Việt luôn là câu hỏi cần tìm lời giải đáp. Trong mục này, chúng tôi nêu những đặc trưng cơ bản của tiếng Việt, khảo sát sơ lược một số nghiên cứu về phân loại tiếng Việt, nêu ra một số hạn chế của một số thuật toán phân loại hiện hành và đề suất một số giải pháp nghiên cứu cho bài toán phân loại tiếng Việt.

4.1 Đặc trưng của ngôn ngữ tiếng Việt

Tiếng Việt thuộc loại hình ngôn ngữ đơn lập với các đặc điểm nổi bật là: đơn vị cơ sở của ngữ pháp là tiếng, từ không biến đổi hình thái, ý nghĩa ngữ pháp được biểu thị bằng trật tự từ và hư từ. Tiếng Việt có một số đặc trưng chính sau [42]:

- Ngôn ngữ duy nhất thuộc hệ chữ Hán chuyển sang La tinh
- Nhiều âm nhất: Tiếng Việt ước khoảng 15.000 âm (29 ký tự và 6 dấu giọng, có phụ âm đơn hay đôi ở đầu và cuối, có nguyên âm đơn, đôi, ba).
- Phức tạp trong xưng hô, tùy theo quan hệ, vị thế của những người có quan hệ, cách dùng từ mang tính tương đối tùy theo vị trí giữa người và ngoại cảnh hay vật…
- Dễ đảo ngữ trong câu và khi đảo ngữ thì nghĩa của câu có thể bị thay đổi,
- Nói láy, là giao hoán âm đầu vần và thanh điệu hoặc trật tự của 2 âm tiết để tạo nghĩa khác hẳn, như: đại phong - lọ tương, đầu tiên - tiền đâu...
- Từ đôi, như: có "một, hai, ba…" lại có "nhất, nhị, tam..." (song song nhau khoảng 70%). Văn bình dân có khoảng 70% từ Nôm, văn bác học có khoảng 70% từ Hán-Việt.
- Văn phạm đôi, như: nhà trắng - bạch ốc, viện bảo tàng - bảo tàng viện; không chia động từ, tính từ...
- Loại chữ đơn âm tiết, mỗi chữ 1 âm, nhưng có một số ký tự thì lại đa âm như: l, m, n, x, y...

4.2 Khảo sát các nghiên cứu về phân loại văn bản tiếng Việt

Trong bài báo [43], các tác giả đã giải quyết vấn đề tự động phân loại vấn đề, đưa các nguồn văn bản vào các danh mục chuẩn bị trước. So sánh giữa mô hình ngôn ngữ N-Gram thống kê và phương pháp tiếp cận túi từ (sử dụng các giải thuật Naïve Bayes, K-Nearest Neighbour (KNN) và Support Vector Machine (SVM)). Một số tác giả đã áp dụng ý tưởng lọc thư rác vào các nguồn văn bản tiếng Việt [44]. Các tài liệu ngắn như văn bản hội thoại cũng đã được khai thác bằng cách giải quyết bài toán xác định từ mục tiêu [45]. Các tác giả đã dùng ý định đề xuất của người dùng qua các văn bản hội thoại tại làm một đơn vị phân đoạn chức năng. Một số nghiên cứu bài toán phân loại văn bản chú trọng so sánh hiệu suất của trọng số thuật ngữ hơn là phân tích bài toán phân loại thực tế [46]. Về dữ liệu tiếng Việt, biểu diễn toàn văn đã được nhiều

tài liệu nghiên cứu khác khai thác [47- 49]. Tìm hiểu các bài báo, chúng tôi thấy rằng đây là nỗ lực sử dụng ý tưởng về các từ khóa đại diện trong phân loại văn bản tiếng Việt. Trong bài báo [50], nhóm tác giả đã đề xuất phương pháp phân loại văn bản tiếng Việt dựa vào thuật toán TextRank và hệ số tương tự Jaccard. TextRank xếp hạng các từ và câu theo giá trị đóng góp của chúng và trích xuất các từ khóa tiêu biểu nhất. Các tác giả thu thập văn bản từ các trang web tin tức Việt Nam, tiến hành bước tiền xử lý dữ liệu, trích xuất từ khóa bằng thuật toán TextRank, sau đó đo điểm tương tự theo khoảng cách Jaccard và dự đoán các danh mục.

Về kết quả đạt được cũng khó để đánh giá so sánh, do các nghiên cứu thực hiện trên các tập dữ liệu và giải thuật khác nhau. Nhưng nhìn chung cũng là những nghiên cứu đáng khích lệ, đặc biêt trong tài liệu [51], tác giả đã cho thấy việc làm quen với phân loại văn bản là không khó với các công cụ hỗ trợ có sẵn cho người có hứng thú nghiên cứu đến lĩnh vực này.

4.3 Một số hạn chế của thuật toán phân loại văn bản hiên hành

Để nghiên cứu sâu hơn và ứng dụng vào phân loại văn bản tiếng Việt, ngay cả các thuật toán áp dụng cho tiếng Anh như: K-Láng giềng gần nhất (KNN), Naïve Bayes, Máy hỗ trợ véctơ (SVM) và Học sâu (Deep Learning) còn có một số hạn chế như phân tích trong tài liệu [52]:

- KNN là một phương pháp phân loại dễ thực hiện và thích ứng với bất kỳ loại không gian đặc trưng nào. Mô hình này cũng tự nhiên xử lý các trường hợp nhiều lớp. Tuy nhiên, KNN bị giới hạn bởi các ràng buộc về lưu trữ dữ liệu đối với các bài toán tìm kiếm lớn để tìm các láng giềng gần nhất. Ngoài ra, hiệu suất của KNN phụ thuộc vào việc tìm một hàm khoảng cách có ý nghĩa, do đó, nó làm cho kỹ thuật này trở thành một thuật toán rất phụ thuộc vào dữ liệu.
- Thuật toán Naïve Bayes (NB) cũng có một số hạn chế. NB đưa ra giả định mạnh mẽ về hình dạng của phân bố dữ liệu. NB cũng bị giới hạn bởi sự khan hiếm dữ liệu mà bất kỳ giá trị nào có thể có trong không gian đặc trưng, giá trị xác suất phải được ước tính thường xuyên bởi con người.
- Máy hỗ trợ véctơ là một trong những thuật toán học máy hiệu quả nhất kể từ khi được giới thiệu vào những năm 1990. Tuy nhiên, các thuật toán Máy hỗ trợ véctơ để phân loại văn bản bị hạn chế bởi sự thiếu minh bạch trong kết quả do số lượng thứ nguyên cao. Do đó, nó không thể hiển thị điểm số của công ty dưới dạng một hàm tham số dựa trên các tỷ số tài chính hoặc bất kỳ dạng hàm nào khác khi áp dụng vào phân tích chứng khoán.
- Khả năng diễn giải của mô hình học sâu, đặc biệt là DNN *(Deep neural networks)*, luôn là một yếu tố hạn chế đối với các trường hợp sử dụng yêu cầu giải thích về các tính năng liên quan đến mô hình hóa trong các ứng dụng về chăm sóc sức khỏe. Vấn đề này là do các nhà khoa học thích sử dụng các kỹ thuật truyền thống như mô hình tuyến tính, Mô hình Bayes, SVM, cây quyết định... cho các công trình của họ. Trọng số trong mạng nơ-ron là thước đo mức độ mạnh mẽ của mỗi kết nối giữa mỗi nơ-ron để tìm ra không gian đặc trưng quan trọng. Ngoài ra, các thuật toán học sâu rất phức tạp và khó hiểu. Học sâu là một trong những kỹ thuật mạnh mẽ nhất trong trí tuệ nhân tạo (AI), nhiều nhà nghiên cứu và nhà khoa học tập trung vào kiến trúc học sâu để cải thiện sức mạnh tính toán của công cụ này. Tuy nhiên, các kiến trúc học sâu cũng có một số nhược điểm và hạn chế khi áp dụng cho bài toán phân loại văn bản. Một trong những vấn đề chính của mô hình này là học sâu không tạo điều kiện thuận lợi cho việc hiểu biết toàn diện về mặt lý thuyết. Một nhược điểm nổi tiếng của các phương pháp học sâu là bản chất "hộp đen" của chúng. Có nghĩa là, phương pháp mà các phương thức học sâu đưa ra với đầu ra được biến đổi không dễ hiểu. Một hạn chế khác của học sâu là nó thường yêu cầu nhiều dữ liệu hơn các thuật toán học máy truyền thống, có nghĩa là kỹ thuật này không thể áp dụng cho các tác vụ phân loại trên các tập dữ liệu nhỏ. Ngoài ra, lượng dữ liệu khổng lồ cần thiết cho các thuật toán phân loại bằng học sâu càng làm trầm trọng thêm độ phức tạp tính toán trong bước huấn luyện.

4.4 Đề xuất một số giải pháp nghiên cứu cho bài toán phân loại văn bản tiếng Việt

Mặc dù nghiên cứu và ứng dụng về phân loại văn bản tiếng Anh và các ngôn ngữ Ấn – Âu khá hoàn chỉnh và có nhiều ứng dụng vào thực tế như phân loại thư rác, phân tích xã hội, phân loại tin tức, tài chính, chứng khoán… [53]. Ngoài ra còn có nhiều ứng dụng phân loại văn bản hiệu quả và thiết thực trong các lĩnh vực như truy xuất thông tin, lọc văn bản, phân loại tin bài, phân loại thư điện tử và các bản ghi nhớ, phân loại các trang web, phân loại các bài báo học thuật sử dụng các lĩnh vực kỹ thuật và tên miền phụ, lọc thư khiêu dâm, tin sinh học, tự động hóa dịch vụ khách hàng, phân loại chủ đề và phân tích tình cảm... Một số nghiên cứu đã tập trung vào việc xử lý thông tin dạng văn bản có sẵn trong bộ dữ liệu chăm sóc sức khỏe để cải

thiện việc chăm sóc y tế, đồng thời giảm chi phí; hoặc sử dụng công nghệ khai thác văn bản để phát triển hệ thống hỗ trợ quyết định chẩn đoán dựa trên máy tính nhằm giúp các bác sĩ đưa ra quyết định y tế tốt hơn hoặc áp dụng công nghệ khai thác dữ liệu y tế để phát hiện các tác dụng phụ của thuốc [54]…

Tuy nhiên, do khác biệt về đặc tính ngôn ngữ, phương thức ngữ pháp và phương thức cấu tạo từ; sự bùng nổ của nguồn thông tin trên Internet với dữ liệu lớn, nên việc nghiên cứu về phân loại văn bản tiếng Việt vẫn còn là lĩnh vực nghiên cứu hấp dẫn.

Trên cơ sở khảo sát một số nghiên cứu về các phương pháp phân loại văn bản nói chung và phân loại văn bản tiếng Việt nói riêng, chúng tôi đưa ra một số gợi ý như sau:

- Đối với ứng dụng phải xử lý trên nguồn dữ liệu lớn, chúng ta nên sử dụng kỹ thuật học sâu (DL), học không giám sát [26, 55, 56] hoặc học bán giám sát (*Semi-Supervised Learning*) [25, 57, 62] để xây dựng mô hình. Học bán giám sát là một loại bài toán học có giám sát sử dụng dữ liệu không được gắn nhãn để huấn luyện một mô hình. Thông thường, các nhà nghiên cứu và nhà khoa học thích sử dụng kỹ thuật bán giám sát khi một phần nhỏ của tập dữ liệu chứa các điểm dữ liệu được gắn nhãn và một lượng lớn tập dữ liệu không gán nhãn. Hầu hết các thuật toán học bán giám sát cho bài toán phân loại sử dụng kỹ thuật phân cụm như sau: Ban đầu, kỹ thuật phân nhóm được áp dụng trên tập dữ liệu D^T (một tập con của dữ liệu đã gán nhãn thêm vào một số dữ liệu được chọn ngẫu nhiên từ tập dữ liệu không gán nhãn để gán nhãn) với K số lớp đã gán nhãn [58]. Nếu một phân vùng P_i có các mẫu được gắn nhãn, thì tất cả các điểm dữ liệu trên cụm đó thuộc về nhãn đó. Mục tiêu nghiên cứu của kỹ thuật phân cụm là xác định xem chúng ta có nhiều hơn một lớp được gắn nhãn trên một cụm hay không và điều gì sẽ xảy ra nếu chúng ta không có điểm dữ liệu được gắn nhãn trong một cụm [59].
- Đối với ứng dụng trên tập dữ liệu vừa và nhỏ, hoặc muốn cải thiện những giới hạn của các thuật toán học có giám sát đã nêu trong mục 4.3, chúng ta có thể sử dụng dữ liệu trong các tài liệu [52] với tiếng Anh, tài liệu [60, 61] với tiếng Việt và các công cụ trong các tài liệu [51, 62, 63].
- Thực hiện các bước tiền xử lý tiếng Việt trước khi sử dụng các thuật toán trong thư viện máy học [62] như: loại bỏ '*stopword*', phân đoạn từ tiếng Việt, thay thế từ đồng nghĩa…

5. KẾT LUẬN

Mục đích của cuộc khảo sát này là để mô tả và phân tích hiện trạng của việc phân loại văn bản nói chung và phân loại văn bản tiếng Việt nói riêng, đồng thời mong muốn truyền tải cho người đọc một cảm giác hứng thú về sự phong phú và ứng dụng rộng rãi của trí tuệ nhân tạo trong lĩnh vực này. Trong những năm gần đây, nhiều nhóm nghiên cứu đã đầu tư nhiều công sức vào phân tích và phân loại văn bản tự động và đã đạt được nhiều thành tựu to lớn. Tuy nhiên, những vấn đề thách thức vẫn còn tồn tại trong những lĩnh vực này. Đặc biệt, các vấn đề nghiên cứu về cách tạo đột phá về phân loại văn bản để giải các bài toán phân loại văn bản quy mô lớn, khả năng mở rộng của mô hình ứng dụng hiện có và làm thế nào để xây dựng mô hình lựa chọn đặc trưng hiệu quả hơn đã thu hút rất nhiều sự quan tâm nghiên cứu. Chúng tôi hy vọng đã cung cấp một số thông tin hữu ích cho bạn đọc, những người được khuyến khích chấp nhận nhiều thách thức còn tồn tại trong lĩnh vực phân loại văn bản.

TÀI LIỆU THAM KHẢO

[1] D. Georgiev, "20+ Incredible Facebook Messenger Statistics in 2022", Created 2021. [Online]. Available at: https://review42.com/resources/facebook-messenger-statistics/ [Accessed 04 July 2021].
[2] H. Shimodaira, "Text Classification using Naive Bayes", Created January-March 2020. [Online]. Available at: https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn07-notes-nup.pdf [Accessed 05 January 2021].
[3] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol.34, no.1, 1–47, (2002). DOI: https://doi.org/10.1145/505282.505283.
[4] Y. Yang and T. Joachims, "Text categorization", *Scholarpedia*, vol.3, no.5, 42-42, 2008.
[5] S. Ye et al., "Clustering web pages about persons and organizations", *Web Intelligence and Agent Systems: An International Journal*, vol 3, no.4, 203–216, 2005.
[6] A. Díaz and P. Gervás, "Personalisation in news deliv ery systems: Item summarization and multi-tier item selection using relevance feedback", *Web Intelligence and Agent Systems: An International Journal*, vol.3, no.3, 135–154, (2005).
[7] Y. Li et al., "A two-stage text mining model for information filtering", *Proceedings of the 17th ACM conference on Information and knowledge management, ACM*, 2008, 1023–1032.
[8] Y. Gao et al., "Pattern-based topics for document modelling in information filtering", *IEEE Transactions on Knowledge and Data Engineering*, vol.2, no.6, 1629–1642, 2014.

[9] X. Zhou et al., "Coupling topic modelling in opinion mining for social media analysis", *Proceedings of the International Conference on Web Intelligence*, ACM, 2017, 533–540.
[10] X. Zhou et al., "Sentiment analysis on tweets for social events", *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2013, 557–562.
[11] X. Tao et al., "Sentiment analysis for depression detection on social networks", *International Conference on Advanced Data Mining and Applications*, Springer, 2016, 807–810.
[12] B. Liu et al., "Building text classifiers using positive and unlabeled examples", *ICDM03*, 2003, 179–186.
[13] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol.34, no.1, 1–47, 2002.
[14] G.P.C. Fung et al., "Text Classification without Negative Examples Revisit", *IEEE transactions on Knowledge and Data Engineering*, vol.18, no.1, 6–20, 2006.
[15] J. Liu et al., "Deep learning for extreme multi-label text classification", *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2017, 115–124.
[16] Y. Li et al., "Enhancing binary classification by modeling uncertain boundary in three-way decisions", *IEEE Transactions on Knowledge and Data Engineering, vol.*29, no.7, 1438–1451, 2017.
[17] X. Zhou et al., "Rough sets based reasoning and pattern mining for a two-stage information filtering system", *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, 2010, 1429–1432.
[18] L. Zhang et al., "Rough set based approach to text classification", *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 03, IEEE Computer Society, 2013, 245–252.
[19] L. Zhang et al., "Centroid training to achieve effective text classification", *International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2014, 406–412.
[20] E.L. Allwein et al., "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers", *Journal of Machine Learning Research*, vol.1, 113–141, 2000.
[21] W. Chen et al., "Document Transformation for Multi-label Feature Selection in Text Categorization", *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, 2007, 451–456. ISBN 0-7695-3018-4.
[22] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview", *Int J Data Warehousing and Mining*, 2007, 1–13.
[23] G. Tsoumakas, I. Katakis and I. Vlahavas, "Mining Multi-label Data", *Transformation*, 2010, 1–19.
24 [24] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms", *IEEE transactions on knowledge and data engineering*, vol.26, no.8, 1819–1837, 2013.
[25] N. Nguyen and R. Caruana, "Classification with partial labels", *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, ACM, New York, NY, USA, 2008, 551–559.
[26] T. Yang et al., "Unsupervised transfer classification: application to text categorization", *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2010, 1159–1168.
[27] A. M. Aubaid, A. Mishra, *A Rule-Based Approach to Embedding Techniques for Text Document Classification*, Applied Sciences, 2020, 10, 4009. DOI: https://doi.org/10.3390/app10114009
[28] S. M. Kamruzzaman and Farhana Haider, "A HYBRID LEARNING ALGORITHM FOR TEXT CLASSIFICATION", *3rd International Conference on Electrical & Computer Engineering*, ICECE, 2004, 28-30.
[29] X. Zhou et al., "A survey on text classification and its applications", *Web Intelligence*, vol. 18, no. 3, 205-216, 2020.
[30] Scikit-learn: Machine Learning in Python, Pedregosa et al., "Naive Bayes", JMLR 12, 2825-2830, 2011. Available at: https://scikit-learn.org/dev/modules/naive_bayes.html#naive-bayes, [Accessed 02 January 2021].
[31] T. Joachims, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", ECML, 1998, 137–142.
[32] B.V. Dasarathy, "Nearest neighbor (NN) Norms: NN pattern classification techniques", *IEEE Computer Society Tutorial*, 1991.
[33] Shervin Minaee et al., "Deep Learning Based Text Classification: A Comprehensive Review", 2020. Available at: https://arxiv.org/pdf/2004.03705.pdf.
[34] V. Vapnik, *The nature of statistical learning theory*, NY: Springer New York, 2013.
[35] R. Romero, E. Iglesias and L. Borrajo, A linear-RBF multikernel SVM to classify big text corpora, BioMed Research International, 2015. DOI: https://doi.org/10.1155/2015/87829.
[36] Y. Yang and C.G. Chute, "An Example-Based Mapping Method for Text Categorization and Retrieval", ACM TOIS, vol.12, no.3, 1994, 252-277.
[37] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", Nature, 2015, 436–444.

[38] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition", *Neural computation*, vol.1, no.4, 541–551, 1989.
[39] Y. LeCun et al., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no.11, 2278–2324, 1998.
[40] Y. Bengio et al., "Aneural probabilistic language model", *Journal of machine learning research 3*, 1137–1155, 2003.
[41] Y. Meng et al., "Weakly supervised hierarchical text classification*", Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, 6826–6833.
[42] N.T.Can, *Ngữ pháp tiếng Việt*. Việt Nam: Nhà xuất bản Đại học Quốc gia Hà nội, 1999.
[43] V. C. D. Hoang et al., "A comparative study on vietnamese text classification methods", *International Conference on Research, Innovation and Vision for the Future*, 267–273, IEEE, 2007. DOI:10.1109/RIVF.2007.369167.
[44] T.H. Pham, P. Le-Hong, "Content-based approach for Vietnamese spam SMS filtering,", *International Conference on Asian Language Processing(IALP)*, 41–44, 2016. DOI:10.1109/IALP.2016.7875930.
[45] T.L. Ngo et al. "On the identification of suggestion intents from vietnamese conversational texts," *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 417–424, 2017. DOI:10.1145/3155133.3155201.
[46] V. T. Nguyen et al. "A Term Weighting Scheme Approach for Vietnamese Text Classification," *International Conference on Future Data and Security Engineering*, 46–53, Springer, 2015. DOI:10.1007/978-3-319-26135-5 4.
[47] N. H. D. Tri et al. "Xay dung mo hinh phan tan cho phan lop khoi luong lon van ban theo chu de (in English: building distributed model for classification massive text data by topic)," *PRO-CEEDING of Publishing House for Science and Technology*, 2017, DOI:10.15625/vap.2016.000104.
[48] B. K. Linh et al. "Phan loai van ban tieng Viet dua tren mo hinh chu de (in English: vietnamese text classification based on topic modeling)," *PROCEEDING of Publishing House for Science and Technology*, 2017. DOI:10.15625/vap.2016.00065.
[49] T. Ngoc Phuc et al. "Phan loai noi dung tai lieu Web tieng viet (in English: classification of vietnamese texts on the web)", *Vietnam Journal of Science and Technology*, vol.51, no.6, 669–680, 2020. DOI:10.15625/2525-2518/51/6/11629.
[50] H.T. Huynh et al. "Vietnamese Text Classification with TextRank and Jaccard Similarity Coefficient", Advances in Science, *Technology and Engineering Systems Journal*, vol.5, no. 6, 363-369, 2020.
[51] N.V. Hieu, "Phân loại văn bản tiếng Việt sử dụng machine learning", Trực tuyến, Địa chỉ: https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/
[52] Kowsari, et al., "Text Classification Algorithms: A Survey", *Information*, vol. 10, no. 4, 150, April 23, 2019.
[53] Qian Li, et al., "A Survey on Text Classification: From Shallow to Deep Learning", *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, vol.31, no.11, 2020.
[54] Xujuan Zhou, et al., "A Survey on Text Classification and Its Application", Created 2019, Available: https://www.researchgate.net/publication/346646048
[55] X. Zhou, X. Tao, J. Yong and Z. Yang, "Sentiment analysis on tweets for social events", *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2013, 557–562.
[56] X. Tao, X. Zhou, J. Zhang and J. Yong, "Sentiment analysis for depression detection on social networks", *International Conference on Advanced Data Mining and Applications*, Springer, 2016, 807–810.
[57] N. Nguyen and R. Caruana, "Classification with partial labels", *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, ACM, New York, NY, USA, 2008, 551–559.
[58] Gowda, H.S. et al. "Semi-supervised text categorization using recursive K-means clustering". *In International Conference on Recent Trends in Image Processing and Pattern Recognition;* Springer: Berlin/Heidelberg, Germany, 2016, 217–227.
[59] Kowsari, K. et al. "Construction of fuzzyfind dictionary using golay coding transformation for searching applications". Department of Computer Science, School of Engineering and Applied Sciences at The George Washington University, Washington DC, 2015. Available: https://arxiv.org/ftp/arxiv/papers/1503/1503.06483.pdf.
[60] V.Q. Binh, "New Corpus". Created 2021, Available: https://github.com/binhvq/news-corpus#%C4%91%E1%BB%8Bnh-d%E1%BA%A1ng-mongodb-dump.
[61] V. Duy, "A Large-scale Vietnamese News Text Classification Corpus". Created 2019. Available: https://github.com/duyvuleo/VNTC/.
[62] Scikit-learn, "Machine Learning in Python". Available: https://scikit-learn.org/stable/index.html.

[63] M.F. Zafra, "Text Classification in Python.", Created June 2019. Available: https://www.mfz.es/machine-learning/an-end-to-end-machine-learning-project-part-i-text-classification-in-python/.

A SURVEY ON VIETNAMESE TEXT CLASSIFICATION MODELS

NGUYEN CHI HIEU
Faculty of Information Technology, Industrial University of Ho Chi Minh City
nguyenchihieu@iuh.edi.vn

Abtract: Text classification is one of the basic tasks of Natural Language Processing, widely applied in sentiment analysis, spam detection, topic labeling, intent detection, etc. The explosion of information sources on the Web, social networks... makes it more and more important and attracts many researchers. Many feature selection methods and classification algorithms have been proposed to use. However, the rapid increase of big data is creating challenges for text classification in general and Vietnamese language in particular, such as the problem of application expansion, the ability to classify social problems, etc. The purpose of this report is to examine the research on text classification including Vietnamese, in order to provide readers with an overview of existing text classification technologies and topics. propose ways to solve challenging problems in text classification.

Keyword: Text classification, Vietnamese, supervised learning, semi-supervised learning

Ngày nhận bài:04/10/2021

Ngày chấp nhận đăng: 05/12/2021