**CLASSIFICATION OF VISITORS’ VIEWPOINTS TOWARDS PHONG NHA - KE BANG TOURIST ATTRACTION IN QUANG BINH PROVINCE**

PHÂN LOẠI QUAN ĐIỂM CỦA DU KHÁCH ĐỐI VỚI ĐIỂM DU LỊCH PHONG NHA KẺ BÀNG TỈNH QUẢNG BÌNH

**Nguyễn Văn Hướng1, TS. Đậu Mạnh Hoàn2**

1Cục Dự trữ Nhà nước Quảng Bình

2Trường Đại Học Quảng Bình

***ABSTRACT:****Sentiment classification is a special kind of text classification in which a document is classified to predict automatically sentiment polarity (positive or negative). With the aim of studying the emotional feedback from visitors after visiting tourism destination of Phong Nha - Ke Bang, the research was carried out by the data collecting processof the visitors’ ranking and voting on websites for this attraction and viewpoints classification using machine learning techniques to get the information through that data. In this article, the author presents the results of the classification process of visitors' views of the Phong Nha - Ke Bang tourist attraction. As a result, some useful conclusions for the classification process and information collection about visitors' feelings to this attractive destination are suggested.*

***Keywords:*** *Phong Nha - Ke Bang , sentiment classification, Data mining.*

***TÓM TẮT:*** *Phân loại quan điểm là một dạng đặc biệt của phân loại văn bản, trong đó một tài liệu được phân loại để dự đoán tình cảm tự động phân cực (tích cực hay tiêu cực). Với mục đích nắm bắt lại những cảm xúc của du khách sau khi trải nghiệm du lịch đối với điểm du lịch Phong Nha - Kẻ Bàng trong thời gian qua, chúng tôi thực hiện nghiên cứu dựa trên quá trình biên tập dữ liệu bình chọn của du khách trên các trang web, sau đó sử dụng các kỹ thuật học máy để phân loại quan điểm nhằm nắm bắt các thông tin thông qua dữ liệu đó. Trong bài viết, tác giả trình bày kết quả thực hiện biên tập dữ liệu và phân loại quan điểm của du khách đối với điểm du lịch Phong Nha - Kẻ Bàng, từ đó rút ra những kết luận hữu ích cho quá trình phân loại và các thông tin về cảm xúc của du khách đối với điểm du lịch này.*

***Từ khóa:*** *Phong Nha - Kẻ Bàng, phân loại quan điểm, khai phá dữ liệu.*

1. KHAI PHÁ QUAN ĐIỂM VÀ BÀI TOÁN KHAI PHÁ QUAN ĐIỂM DU KHÁCH

**1.1. Khai phá quan điểm**

Khai phá quan điểm là lĩnh vực nghiên cứu chuyên sâu trong các lĩnh vực nghiên cứu khai phá dữ liệu văn bản, xử lý ngôn ngữ tự nhiên với mục đích thu thập được những thông tin mong muốn về quan điểm của người dùng. Khai phá quan điểm là nghiên cứu tính toán các ý kiến của con người, thái độ, cảm xúc và quan điểm của người đó đối với một thực thể. Khai phá quan điểm trợ giúp cho các nhà quản lý giải quyết các công việc trong thực tế vì nó là quan điểm, nhận thức, niềm tin của con người về vấn đề thực tế và nó phụ thuộc vào cách nhìn của mỗi người. Do đó, mục tiêu của khai phá quan điểm là tìm xu hướng quan điểm, xác định những quan điểm mà họ thể hiện theo khuynh hướng nào, từ đó phân loại các quan điểm phân cực của họ đối với thực thể đó. Vì vậy, phân tích quan điểm có thể được xem là một trường hợp đặc biệt trong phân loại văn bản và thường được gọi là phân loại quan điểm.

Ngày nay, người sử dụng thường xuyên thể hiện và chia sẻ tình cảm, quan điểm của mình đối với các đối tượng thông qua trải nghiệm của mình với một thực thể hay một đối tượng nào đó. Những tình cảm đó, quan điểm đó được biểu hiện qua văn bản thông qua các trang Web, các trang mạng xã hội, các Websites, blogs hay forums,... tạo thành một xu thế về cảm xúc thông qua trải nghiệm của con người với một thực thể hay một đối tượng nào đó. Quá trình phân loại các đối tượng có thể chia thành các nhóm khác nhau, mỗi một nhóm thể hiện mối quan hệ giữa các chủ thể và đối tượng phân loại. Như đã phân tích bài toán phân loại quan điểm là một trường hợp đặc biệt của bài toán phân loại văn bản mà giá trị phân loại nhận được theo tính phân cực (tích cực hoặc tiêu cực hoặc trung tính), nó được sử dụng trong các lĩnh vực khác nhau như kinh doanh, chính trị và tâm lý học, dự báo, kinh tế, du lịch,…. Mục đích chính của phân loại quan điểm người sử dụng là xác định sự phân cực của văn bản trong xử lý ngôn ngữ tự nhiên. Trong nghiên cứu này, chúng tôi vận dụng vào bài toán phân loại quan điểm của du khách đối với điểm du lịch Phong Nha - Kẻ Bàng.

**1.2. Bài toán khai phá quan điểm du khách**

Quảng Bình là một trong những địa điểm có tiềm năng du lịch rất đa dạng và phong phú, du lịch Quảng Bình được thiên nhiên ưu ái ban tặng nhiều tài nguyên quý, độc đáo như bãi biển Nhật Lệ, bãi biển Đá nhảy, thế giới biết đến với hệ thống hang động Phong Nha - Kẻ Bàng được xem là kỳ quan thiên nhiên thế giới, những tiềm năng này đã được tỉnh Quảng Bình khai thác và phát huy hiệu quả trong những năm gần đây. Cũng chính những lợi thế này nên Quảng Bình đã xác định phát triển du lịch là một ngành kinh tế mũi nhọn, qua đó nhằm khai thác triệt để tiềm năng sẵn có của thiên nhiên. Doanh thu từ nguồn du lịch ngày một tăng nhờ lượng du khách đến Quảng Bình ngày một nhiều. Với mục đích nắm bắt những cảm xúc của du khách sau khi đến du lịch tại Phong Nha – Kẻ Bàng trong thời gian qua, chúng tôi đã tiến hành biên tập dữ liệu thông qua các trang web sau đó sử dụng các kỹ thuật học máy để phân loại quan điểm nhằm nắm bắt được các thông tin qua dữ liệu đó. Chúng tôi tiến hành biên tập dữ liệu từ các trang Web [8-16], là các trang báo điện tử phổ biến ở trong nước, qua đó lấy các ý kiến đánh giá của du khách về điểm du lịch Phong Nha – Kẻ Bàng.

Bài toán phân loại quan điểm của du khách ở đây sẽ được xác định thuộc một trong các trạng thái tích cực, tiêu cực và trung tính dựa trên việc phân tích và đoán nhận ngữ nghĩa câu văn, đoạn văn. Giải pháp đoán nhận ngữ nghĩa văn bản cho bài toán này là sử dụng kỹ thuật học máy, hệ thống được huấn luyện có giám sát với bộ dữ liệu mẫu, văn bản được phân loại thành 3 mức tích cực, tiêu cực và trung tính. Trong phạm vi nghiên cứu của bài toán đặt ra, việc đoán nhận cảm xúc được giới hạn phạm vi với các ngữ cảnh thuộc về du lịch, danh lam thắng cảnh.

2. DỮ LIỆU PHÂN LOẠI

Để phân loại quan điểm cho bài toán, đầu tiên chúng tôi tiến hành biên tập dữ liệu đó là các ý kiến đánh giá về điểm du lịch Phong Nha – Kẻ Bàng dựa trên các trang báo điện tử lớn ở trong nước, các trang mạng xã hội [8-16]. Dữ liệu thu thập được từ Web sẽ được tiến hành tiền xử lý trên máy tính để phục vụ cho các xử lý của các bước sau có hiệu quả hơn.

Dữ liệu thực nghiệm có 828 file văn bản được tập hợp để xây dựng máy học và kiểm thử. Sau khi tách từ và loại bỏ từ dừng số từ có được là 62745 từ. Như vậy, tập ngữ liệu cần mô hình hóa là một ma trận chứa TF\*IDF của các từ có kích thước là 828×62745 phần tử. Chúng tôi sử dụng 70% dữ liệu có được để làm dữ liệu huấn luyện và dùng 30% dữ liệu còn lại làm dữ liệu kiểm thử. Chi tiết được thể hiện tại Bảng 1.

**Bảng 1.** Dữ liệu sử dụng để thực nghiệm

|  |  |  |
| --- | --- | --- |
| Loại | Số file | |
| Câu văn | Tích cực | 390 |
| Tiêu cực | 208 |
| Đoạn văn, file văn bản | Tích cực | 150 |
| Tiêu cực | 80 |
| Tổng cộng | | 828 |

.

Ngoài ra chúng tôi cũng thu thập dữ liệu từ các trang web [8-16] để làm dữ liệu luyện học trong phạm vi câu văn mô tả về du lịch, danh lam thắng cảnh có 6700 được phân loại theo Bảng 2.

**Bảng 2.** Dữ liệu hỗ trợ quá trình luyện học

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loại | Danh từ | Động từ | Tính từ | Trạng từ | Từ khác |
| Số lượng (Đơn vị: từ) | 2500 | 900 | 1200 | 600 | 1500 |
| Tổng cộng | 6700 | | | | |

Các loại dữ liệu này cũng sẽ được bổ sung trong quá trình thực nghiệm thuật toán để làm cho hệ thống học ngày càng thông minh hơn.

3. CÁC BƯỚC XỬ LÝ

Dữ liệu phân loại quan điểm được thực hiện theo quá trình ở sơ đồ sau:

Dữ liệu

phân loại

Học máy

Tích cực

Tiêu cực

Dữ liệu huấn luyện

Tiền xử lý

dữ liệu

Trung tính

Dữ liệu đã phân loại

**Hình 1.** Sơ đồ tổng quát bài toán phân loại quan điểm.

Trong đó, dữ liệu phân loại trước khi đưa vào học máy sẽ được thực hiện xử lý qua các công đoạn sau:

**3.1 Tiền xử lý dữ liệu**:

Dữ liệu sau khi được chuẩn bị được sử dụng để tiến hành đưa vào xử lý bài toán. Dữ liệu được tiến hành tách đoạn, tách câu, chuẩn hóa chính tả và xác nhận các câu văn có liên quan đến lĩnh vực cũng như mục đích đoán nhận ngữ cảnh cần thực hiện. Quá trình tiền xử lý dữ liệu sẽ tiến hành xử lý các trường hợp chuẩn hóa chính tả luật trên tiếng và sai khác mẫu tự; tiến hành chuẩn hóa dấu chấm câu bao gồm xử lý dấu chấm cuối câu, dấu chấm giữa câu, các chuỗi có chứa nhiều dấu chấm liên tục và xử lý các trường hợp viết tắt.

+ Tách từ: Tách từ là vấn đề quan trọng nhất của chương trình, nó quyết định chương trình có thể thực hiện đúng và chính xác việc phân loại hay không là nhờ kết quả của việc tách từ đúng hay sai. Tất cả các tài liệu qua bước này đều được xử lý thành các từ là đầu vào cho bước xử lý tiếp theo. Trong thực nghiệm của mình chúng tôi chọn phương pháp MMSEG này để giải quyết bài toán tách từ tiếng Việt [3].

+ Loại bỏ từ Stopwords: Chúng là các từ chức năng hay các phụ từ, hư từ, các từ như từ nối, từ chỉ số lượng...những từ không mang tính phân biệt trong khi phân loại. Ngoài ra, còn có rất nhiều từ khác cũng không có giá trị phân loại ví dụ như từ xuất hiện hầu khắp các văn bản hay dùng không phổ biến trong văn bản, những từ gọi là stopword (từ dừng) này sẽ được lược bỏ để tăng hiệu năng cũng như giảm bớt số lượng đặc trưng vốn đã rất lớn trong các mô hình phân loại văn bản. Để loại bỏ từ Stopwords chúng tôi dựa vào danh sách các từ stopword tham khảo từ website [17].

+ Chuyển đổi câu: Văn bản sau khi xử lý được biến đổi từ một câu văn phức tạp thành câu văn cơ bản gồm các thành phần chính như chủ ngữ, vị ngữ, tính từ, trạng từ. Các thành phần này sẽ cấu thành một câu văn mới để nhận biết cảm xúc của văn bản.

+ Trích xuất tập đặc trưng và biểu diễn văn bản:

Ở đây chúng tôi sử dụng công thức tính toán giá trị là dựa vào phương pháp TFxIDF (Terms Frequency và Inverse Document Frequency). Cách định trọng số này cho biết một từ là quan trọng cho một tài liệu nếu nó xuất hiện nhiều lần trong tài liệu đó [1, 2].

Trọng số wij được tính bằng tần số xuất hiện của từ khóa ti trong văn bản dj và độ hiếm của từ khóa ti trong tập văn bản.

Công thức tính wij:

(1)

Trong đó:

weigh (i, j): trọng số của từ thứ i trong văn bản thứ j,

fij (term frequency): số lần xuất hiện của từ thứ i trong văn bản thứ j, fij càng cao thì từ đó càng miêu tả tốt nội dung văn bản,

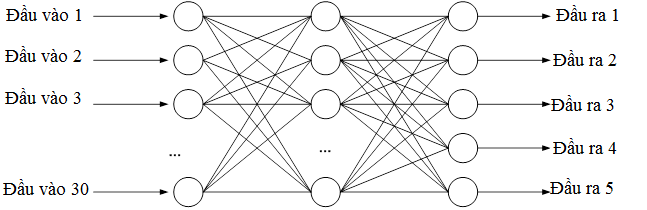
dfi (document frequency): số văn bản có chứa từ thứ i.

Ưu điểm của phương pháp này là mô hình biểu diễn văn bản được sử dụng khá phổ biến trong các hệ xử lý văn bản, nhất là bài toán phân loại, ngoài ra mối quan hệ giữa các văn bản được tính toán dựa trên các vector biểu diễn nên dễ dàng thực hiện hơn.

3.2. Kỹ thuật học máy

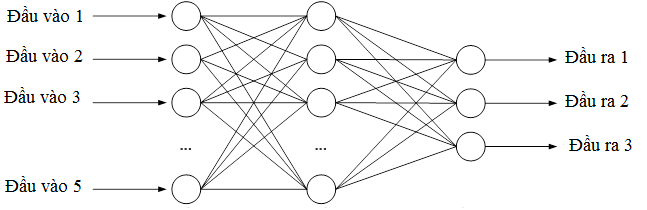
Chúng tôi sử dụng kỹ thuật học máy Mạng noron nhân tạo (ANN: Aritificial Neural Network) [4-7] để cho hệ thống được luyện học cho phân tích, đoán nhận cảm xúc, ở đây mạng neuron nhận tạo được ứng dụng trong phân loại, cấu trúc mạng neuron có cấu trúc 03 lớp (01 lớp ẩn), gồm 02 mạng noron có các chức năng: ANN1- Nhận dạng danh từ, tính từ, trạng từ tham gia cấu trúc câu cơ bản, ANN2- Đoán nhận quan điểm, cấu trúc như sau:

Mạng noron ANN1: Gồm 3 lớp, lớp đầu vào (input) gồm 30 nút (bảo đảm tối đa số từ có trong câu), lớp ẩn (hidden) gồm 12 nút, lớp đầu ra (output) gồm 5 nút.



**Hình 2.** Cấu trúc mạng noron ANN1

Mạng noron ANN2: Gồm 3 lớp, lớp đầu vào (input) gồm 5 nút (vector danh từ, động từ, tính từ, trạng từ và từ khác), lớp ẩn (hidden) gồm 7 nút, lớp đầu ra (output) gồm 3 nút (tích cực, không xác định và tiêu cực).



**Hình 3.** Cấu trúc mạng noron ANN2

Dữ liệu luyện học được thu thập từ các câu văn, văn bản tự nhiên về du lịch, danh lam thắng cảnh gồm các thành phần: danh từ, động từ, tính từ, trạng từ. Trong đó, các từ mô tả mức độ về tích cực và tiêu cực được gán nhãn phục vụ quá trình luyện học.

Kết quả đoán nhận được phân thành 03 lớp tích cực, tiêu cực và trung tính, kết quả này được sử dụng thống kê và làm cơ sở đánh giá cảm xúc của đoạn văn.

3.3. Đánh giá quan điểm đoạn văn

Đoạn văn phân loại được tạo thành từ các câu văn, quá trình phân loại được xử lý theo hai trường hợp, thứ nhất nếu các câu văn đều thể hiện cùng cảm xúc là tích cực hoặc tiêu cực thì kết quả phân loại quan điểm của đoạn văn được xác định theo cùng quan điểm của các câu văn; trường hợp thứ hai nếu các câu văn thể hiện quan điểm trái chiều nhau, đây là trường hợp cần có sự đánh giá theo từng mức độ, có thể dựa vào số lượng quan điểm của các câu văn đánh giá. Xem xét các câu văn có quan điểm ngược nhau nhưng cùng mô tả một vấn đề cụ thể, phân loại và xác định câu văn thể hiện quan điểm mạnh để dẫn dắt xác định. Như vậy, trong đoán nhận quan điểm của câu văn cơ bản, kết quả phân loại cũng cần ghi nhận mức độ quan điểm mạnh, yếu để quyết định cho tình huống có những quan điểm bị xung đột.

4. KẾT QUẢ THỰC NGHIỆM

Trong thực nghiệm có 828 văn bản được sử dụng, chia làm hai loại. Đánh giá độ chính xác được thực hiện theo các tiêu chí: Độ chính xác (precision), độ bao phủ (Recall) và F1, cách tính được thực hiện theo công thức sau:

(2)

(3)

(4)

**Bảng 3.** Kết quả thực nghiệm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Loại** | **Số file**  **huấn luyện** | **Số file**  **kiểm chứng** | **Độ chính xác (Precision)** | **Độ phủ (Recall)** | **F1** |
| Câu văn | 406 | 174 | 77% | 71% | 74% |
| Đoạn văn,  file văn bản | 174 | 74 | 63% | 68% | 65% |

.

5. ĐÁNH GIÁ KẾT QUẢ VÀ BÀI HỌC KINH NGHIỆM

Trong nghiên cứu này, chúng tôi sử dụng kỹ thuật học máy ANN để xử lý và đoán nhận quan điểm của du khách đối với điểm du lịch Phong Nha - Kẻ Bàng bước đầu đã cho kết quả tốt, khẳng định tính đúng đắn của quá trình nghiên cứu.

Từ kết quả thực nghiệm cho thấy độ chính xác của phân loại Câu văn là 74%, hầu hết các trường hợp phân loại theo câu được đánh giá chính xác. Đối với trường hợp đoạn văn gồm nhiều câu hoặc file văn bản gồm nhiều đoạn văn bản và văn bản có nội dung đa dạng, có nhiều từ đa nghĩa thì kết quả đánh giá đạt được chưa cao hoặc quá trình đánh giá không xác định được thuộc về loại nào, kết quả chung của loại này là 65%. Như vậy, kết quả phân loại trung bình trên dữ liệu thực nghiệm đạt độ chính xác 70%, kết quả này chưa cao so với các phương pháp và dữ liệu khác nhưng cũng chấp nhận được.

Qua quá trình nghiên cứu chúng tôi rút ra được 5 nhận xét như sau:

- Thứ nhất: Trường hợp kết quả phân loại chưa cao là do dữ liệu phân loại rơi vào trường hợp nội dung văn bản nhập nhằng, có từ ẩn ý, từ lóng không rõ ràng nên hệ thống không thể phân loại được.

- Thứ hai: Cần phải thực hiện đánh trọng số cho các câu văn để kết quả có ngữ cảnh có độ chính xác cao hơn.

- Thứ ba: Xây dựng bộ dữ liệu đủ lớn, dữ liệu học nhiều sẽ xử lý hết được các tình huống.

- Thứ tư: Một vấn đề quan trọng đó là sự khó khăn của vấn đề phân loại quan điểm trên tiếng Việt, một ngôn ngữ giàu cảm xúc và nhiều từ đa nghĩa. Đây cũng chính là thách thức đối với bài toán phân loại quan điểm trên tiếng Việt hiện nay.

- Thứ năm: Ngữ pháp và ngữ nghĩa tiếng Việt rất phong phú và đa dạng, từ đa nghĩa rất nhiều. Việc thể hiện quan điểm của mỗi cá nhân trên mạng ngày nay cũng đa dạng, trong quá trình bình phẩm mỗi người khác nhau sẽ có những cách viết khác nhau về ngôn từ lẫn cách biểu đạt và cả cách thức sử dụng ngôn ngữ của người đó. Ngoài ra, cách thể hiện tình cảm, quan điểm còn phụ thuộc vào trình độ, tính cách, tuổi tác, cảm xúc, bối cảnh và thời gian lúc bình phẩm của những người đó. Trong thực tế một từ trong một lời bình phẩm được xem là quan điểm tích cực trong tình huống này song vẫn có thể là tiêu cực trong tình huống khác, nhất là trường hợp nói châm biếm, mỉa mai hay cách dùng từ ẩn dụ nên dễ gây ra đoán nhận sai quan điểm. Đôi khi một quan điểm về một vấn đề hoặc chỉ một phần của vấn đề cũng có thể gây hiểu lầm cho hệ thống khai phá quan điểm. Đối với cấp độ Đoạn văn bản hoặc File văn bản thể hiện quan điểm thì có có tình huống tồn tại nhiều quan điểm mâu thuẫn trong cùng một đoạn văn hay tài liệu, tức là cùng bình luận về một vấn đề nhưng bài viết bao gồm cả ý kiến tích cực lẫn ý kiến tiêu cực hoặc chứa nhiều quan điểm mâu thuẫn lẫn nhau trong văn bản. Chính những vấn đề nói trên đã gây khó khăn để phân tích cú pháp và khai phá quan điểm.

6. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề cập đến bài toán phân loại quan điểm của du khách về một địa danh du lịch trong tỉnh Quảng Bình. Để giải quyết được bài toán này chúng tôi đã nghiên cứu các vấn đề khác nhau trong khai phá dữ liệu nói chung, khai phá dữ liệu văn bản nói riêng, và đặc biệt kỹ thuật phân loại quan điểm người dùng, một trường hợp đặc biệt của phân loại văn bản, trong đó một tài liệu được phân loại để dự đoán quan điểm tự động phân cực (tích cực hay tiêu cực). Qua nghiên cứu, khảo sát các phương pháp đánh giá với ngôn ngữ tự nhiên tiếng Việt, trong nghiên cứu này sử dụng kỹ thuật máy học để xử lý và đoán nhận cảm xúc của du khách đối với điểm du lịch Phong Nha - Kẻ Bàng bước đầu đã cho kết quả tốt, khẳng định tính đúng đắn về hướng nghiên cứu.

Qua thực nghiệm trên 828 văn bản cho kết quả cao nhất là 74% đối với câu văn bản sử dụng khi phương pháp học máy. Kết quả này cũng là một cơ sở được chúng tôi minh chứng là thành công trên ngôn ngữ tiếng Việt. Tuy nhiên, kết quả còn có khả năng cải thiện tốt hơn nữa khi giải quyết được các nguyên nhân đã chỉ ra trong phần đánh giá nên cần tiếp tục thử nghiệm và điều chỉnh để nâng cao tỷ lệ chính xác. Quá trình nghiên cứu chúng tôi đã rút ra được 5 bài học lớn cho quá trình phân loại quan điểm trên tiếng Việt. Ngoài ra, chúng tôi cũng nhận thấy quá trình thực hiện còn liên quan đến số đặc trưng được lựa chọn sao cho đảm bảo kỹ thuật mà không mất đi ý nghĩa của văn bản, số đặc trưng không ít quá, nếu ít quá sẽ không phản ảnh hết nội dung mà văn bản thể hiện, ngược lại số đặc trưng nhiều quá sẽ tạo ra không gian đặc trưng lớn, khó xử lý và kiểm soát.

Bài toán phân loại quan điểm vẫn là bài toán phức tạp và cũng rất đa dạng theo nghĩa đặc trưng của ngôn ngữ tự nhiên, nâng cao hiệu quả phân loại văn bản là mục đích mà nhiều nhà nghiên cứu hướng đến. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu để xây dựng các ứng dụng cụ thể tốt hơn cho quá trình phân loại này nhằm khai thác vào mọi lĩnh vực của đời sống xã hội. Với kết quả phân loại quan điểm của du khách đối với điểm du lịch Phong Nha - Kẻ Bàng tỉnh Quảng Bình nói trên cho thấy hệ thống phân loại bước đầu đã hoạt động hiệu quả góp phần trợ giúp cho các nhà quản lý, các đơn vị tổ chức trong tỉnh đặc biệt là các đơn vị hoạt động trong lĩnh vực du lịch sẽ tìm ra các giải pháp nâng cao hiệu quả cho hoạt động du lịch ở Phong Nha - Kẻ Bàng nói riêng và tại tỉnh Quảng Bình nói chung.

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt:**

[1] Đậu Mạnh Hoàn, Các hướng tiếp cận chính trong kỹ thuật phân loại tình cảm. Đại học Quảng Bình. 2017.

[2] Nguyễn Thị Lan Anh, Nghiên cứu thuật toán học máy SVM và ứng dụng trong bài toán khai phá ý kiến phản hồi của khách hàng trên Website. Luận văn Thạc sĩ Hệ thống thông tin. Học viên Công nghệ Bưu chính Viễn Thông. 2013.

**Tiếng Anh:**

[3] Chih-Hao Tsai ,” MMSeg: A Word Identification System for Mandarin Chinese Text Based on two Variants of the Maximum Matching Algorithm” 2000.

[4] Wiener, E. D., Pedersen, J. O., And Weigend, A. S. 1995. A neural network approach to topic spotting. In Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1995), 317–332.

[5] Lam, S. L. And Lee, D. L. 1999. Feature reduction for neural network based text categorization. In Proceedings of DASFAA-99, 6th IEEE International Conference on Database Advanced Systems for Advanced Application (Hsinchu, Taiwan, 1999), 195–202.

[6] Wiener, E. D., Pedersen, J. O., And Weigend, A. S. 1995. A neural network approach to topic spotting. In Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1995), 317–332

[7] Russell, S.J., & Norvig, P. 2016. Artificial Intelligence: A Modern Approach. Malaysia; Pearson Education Limited.

**Website:**

[8] <https://phongnhakebang.vn>

[9] <https://www.quangbinhtravel.vn/dong-phong-nha.html>

[10] <https://www.quangbinhtravel.vn/dong-phong-nha.html>

[11] <https://www.facebook.com/quangbinhtravel.vn/>

[12] <https://www.facebook.com/dlphongnha/>

[13] https://vi.wikipedia.org

[14] https://mytour.vn

[15] http://dulichkhatvongviet.com

[16] <https://www.tripadvisor.com.vn>

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

**Ghi chú:** Công trình này được phát triển từ luận văn thạc sỹ ngành Khoa học máy tính K35 của Trường Đại học Bách khoa, Đại học Đà Nẵng đặt tại Trường Đại học Quảng Bình.

*Liên hệ:*

1. **CH. Nguyễn Văn Hưởng,**

Cục Dự trữ Nhà nước khu vực Bình Trị Thiên,

Địa chỉ: 21 Lý Thường Kiệt, Đồng Hới, Quảng Bình

Email: [huongdtqg@gmail.com](mailto:huongdtqg@gmail.com)

1. **TS. Đậu Mạnh Hoàn,**

Trung tâm Học liệu, Đại học Quảng Bình

Địa chỉ: 312 Lý Thường Kiệt, Đồng Hới Quảng Bình

Email: daumanhhoan@yahoo.com