

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**ĐỒ ÁN MÔN HỌC
KHAI KHOÁNG DỮ LIỆU**

Đề tài

**PHÂN LỚP DỮ LIỆU BÌNH LUẬN SẢN PHẨM
TRÊN WEBSITE THEGIOIDIDONG.COM**

Sinh viên thực hiện :

**Nguyễn Duy Phương
Nguyễn Quốc Hưng**

Mã số : B1812294

Mã số : B1812270

Cần Thơ, 6/2021

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**ĐỒ ÁN MÔN HỌC
KHAI KHOÁNG DỮ LIỆU**

Đề tài

**PHÂN LỚP DỮ LIỆU BÌNH LUẬN SẢN PHẨM
TRÊN WEBSITE THEGIOIDIDONG.COM**

**Giáo viên hướng dẫn:
TS. Lưu Tiến Đạo**

**Sinh viên thực hiện:
Nguyễn Duy Phương
Mã số: B1812294
Nguyễn Quốc Hưng
Mã số: B1812270**

Cần Thơ, 6/2021

[illegible]

Cần Thơ, ngày tháng 6 năm 2021
(GVHD ký và ghi rõ họ tên)

LỜI CẢM ƠN

Để có được bài niên luận này, em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến Thầy Lưu Tiến Đạo – người đã trực tiếp tận tình hướng dẫn, giúp đỡ em. Trong suốt quá trình thực hiện niên luận, nhờ những sự chỉ bảo và hướng dẫn quý giá đó mà bài niên luận này được hoàn thành một cách tốt nhất.

Em cũng xin gửi lời cảm ơn chân thành đến các Thầy Cô Giảng viên Đại học Cần Thơ, đặc biệt là các Thầy Cô ở Khoa CNTT & TT, những người đã truyền đạt những kiến thức quý báu trong thời gian qua.

Em cũng xin chân thành cảm ơn bạn bè cùng với gia đình đã luôn động viên, khích lệ và tạo điều kiện giúp đỡ trong suốt quá trình thực hiện để em có thể hoàn thành bài niên luận một cách tốt nhất.

Tuy có nhiều cố gắng trong quá trình thực hiện niên luận, nhưng không thể tránh khỏi những sai sót. Em rất mong nhận được sự đóng góp ý kiến quý báu của quý Thầy Cô và các bạn để bài niên luận hoàn thiện hơn.

Cần Thơ, ngày tháng 6 năm 2021

Người viết

Nguyễn Duy Phương

Nguyễn Quốc Hưng

MỤC LỤC

PHẦN GIỚI THIỆU	5
1. Đặt vấn đề.	5
2. Lịch sử giải quyết vấn đề	5
3. Mục tiêu đề tài	6
4. Đối tượng và phạm vi nghiên cứu	6
5. Phương pháp nghiên cứu	6
6. Kết quả đạt được	7
7. Bố cục niên luận	7
PHẦN NỘI DUNG.....	8
CHƯƠNG 1	8
GIỚI THIỆU VỀ CƠ SỞ LÝ THUYẾT	8
1.1.Mô tả bài toán.	8
1.2.Vấn đề và giải pháp liên quan đến bài toán.	8
1.2.1.Máy học là gì.	8
1.2.2.Định lý Bayes.....	9
1.2.3.Các khái niệm liên quan.....	10
1.2.4.Thu thập dữ liệu.	11
CHƯƠNG 2	13
THIẾT KẾ VÀ CÀI ĐẶT	13
2.1. Thiết kế hệ thống.	13
2.2. Thu thập và tiền xử lý dữ liệu.....	13
2.2.1. Thu thập dữ liệu.....	13
2.2.2. Làm sạch dữ liệu.....	14
2.2.3. Tách từ	14
2.2.4. Chuẩn hoá từ.....	15
2.2.5. Xoá từ dừng	15
2.2.6. Số hoá dữ liệu	16
2.3. Cài đặt mô hình.....	16
2.3.1. Mô hình Naive Bayes.	16
2.3.2. Huấn luyện.....	17
2.3.3. Đánh giá.	18
2.4. Cài đặt giao diện người dùng.....	18

CHƯƠNG 3	19
ĐÁNH GIÁ MÔ HÌNH.....	19
3.1. Môi trường cài đặt.	19
3.2. Kết quả kiểm tra.....	19
PHẦN KẾT LUẬN	21
1. Kết quả đạt được	21
2. Ưu điểm.	21
3. Nhược điểm.	21
4. Hướng phát triển.	21
TÀI LIỆU THAM KHẢO.....	22

DANH MỤC HÌNH

Hình 1 Mô tả máy học	9
Hình 2 Mô tả hệ thống thu thập dữ liệu	12
Hình 3 Mô tả hoạt động của hệ thống	13
Hình 4 làm sạch văn bản.....	14
Hình 5 Mô tả hành động tách từ	15
Hình 6 Mô tả hành động chuẩn hóa từ.....	15
Hình 7 Mô tả hành động xóa từ dừng	16
Hình 8 Giao diện người dùng	18

DANH MỤC BẢNG BIỂU

Bảng 1 Kết quả thực nghiệm dự trên mô hình.....	20
Bảng 2 Độ chính xác của kết quả thực nghiệm	20

ABSTRACT

Commenting on social networks (review comments, quality feedback comments, comments containing product information, ...) is a problem that websites from news, e-commerce, classifieds, blogs ... have to face every day. Therefore, if we can build a system to evaluate these comments, the problem will be solved a lot easier. Meeting the needs of customers exists on a general basis. Therefore, the user comment analysis system was built to solve the above problem, which we would like to apply on a website that can be Thegioididong.com. The system will have the main function of analyzing as well as classifying the comment content of the customers based on the star rating criteria from 1 to 5. From there, showing a customer's attitude (label 0 is positive, label 1 is negative) towards a product specific product. The system uses machine learning algorithms: Naïve Bayes along with data collection methods to proceed to build a classification model for users.

TÓM TẮT

Bình luận trên mạng xã hội (bình luận đánh giá, bình luận phản hồi chất lượng, bình luận có chứa thông tin về sản phẩm,...) là vấn đề mà các website từ tin tức, thương mại điện tử, rao vặt, blog,... phải đối mặt hàng ngày. Do đó, nếu chúng ta có thể xây dựng một hệ thống đánh giá những bình luận này thì vấn đề sẽ được giải quyết nhẹ nhàng hơn rất nhiều. Việc đáp ứng nhu cầu của khách hàng của hiện hữu trên phương diện tổng quát. Do đó hệ thống phân tích bình luận của người dùng được xây dựng để giải quyết vấn đề trên mà ở đây chúng tôi xin được ứng dụng trên một website của thể là Thegioididong.com. Hệ thống sẽ có chức năng chính là phân tích cũng như phân loại nội dung bình luận của các khách hàng dựa trên tiêu chí đánh giá sao từ 1 đến 5. Từ đó đưa ra thể hiện thái độ (nhãn 0 là tích cực, nhãn 1 là tiêu cực) của một khách hàng đối với một sản phẩm cụ thể. Hệ thống sử dụng giải thuật máy học: Naïve Bayes cùng với các phương pháp thu thập dữ liệu để tiến hành xây dựng một mô hình phân loại cho người dùng.

PHẦN GIỚI THIỆU

1. Đặt vấn đề.

Trong thời kì 4.0 như hiện nay, sự gia tăng mạnh mẽ của thông tin dạng văn bản trực tuyến, ví dụ như email, tin tức trực tuyến, các trang web, cũng như một số lượng lớn các nguồn tài nguyên cho các chủ đề như khoa học, xã hội,...làm cho nhu cầu phân loại văn bản ngày càng tăng cao. Phân loại văn bản là vấn đề đã được nhiều nhà nghiên cứu tích cực tìm hiểu từ những năm 1980. Từ quan điểm của khai phá văn bản, phân loại văn bản có thể được coi là công nghệ tiền xử lý để lọc ra các tài liệu liên quan từ một kho ngữ liệu quy mô lớn. Nhiệm vụ của phân loại văn bản là phân bổ tài liệu thành các chủ đề được xác định trước, chẳng hạn như kinh tế, chính trị và thể thao.

Tại sao phải phân loại văn bản tự động? Việc phân loại văn bản sẽ giúp chúng ta tìm kiếm thông tin dễ dàng và nhanh chóng hơn rất nhiều so với việc phải duyệt mọi thứ trong ổ đĩa lưu trữ để tìm kiếm thông tin. Mặt khác, lượng thông tin ngày một tăng lên đáng kể, việc phân loại văn bản tự động sẽ giúp con người tiết kiệm được rất nhiều thời gian và công sức. Chúng ta chỉ cần áp dụng một mô hình toán học, một bộ phân loại, bằng cách nào đó ước tính các điểm tương đồng giữa các văn bản khác nhau dựa trên các ngữ nghĩa và đoán thuộc tính mục tiêu. Do vậy, các phương pháp phân loại văn bản tự động đã ra đời để phục vụ cho nhu cầu chính đáng đó.

Đề tài “Phân lớp dữ liệu bình luận sản phẩm trên website Thegioididong.com” nhằm tìm hiểu và thử nghiệm các phương pháp phân loại văn bản áp dụng trên tiếng Việt. Phân loại văn bản (text classification) là một trong những công cụ khai phá dữ liệu dạng văn bản một cách hữu hiệu, làm nhiệm vụ đưa những văn bản có cùng nội dung chủ đề giống nhau về cùng một lớp có sẵn. Phân loại văn bản giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin cần thiết đồng thời có thể lưu trữ các thông tin theo đúng chủ đề (topic) hay lớp (class) dựa trên các thuật toán phân loại.

2. Lịch sử giải quyết vấn đề

Phân loại văn bản là một vấn đề đã xuất hiện từ rất lâu. Ngay từ những năm 60 của thế kỷ trước, các nghiên cứu đã được thực hiện ngay khi các máy tính đầu tiên ra đời.

Các nghiên cứu về phân loại văn bản tập trung vào việc áp dụng các phương pháp học giám sát, sử dụng các kho dữ liệu lớn là tập các văn bản được phân loại theo các chủ đề khác nhau như:

- Phương pháp Naive Bayes (McCalum, 1998; Ko, 2000)
- Phân loại văn bản dựa trên mô hình xác suất Bayes và áp dụng cho tiếng Việt (Nguyễn Tuấn, Anh, 2003).
- Một số các nghiên cứu phân loại tiếng Việt tập trung vào ứng dụng các phương pháp máy học hoặc áp dụng các phương pháp đã được đề xuất hiệu quả cho tiếng Anh như Phân loại văn bản do nhóm tác giả Phạm Nguyên Khang, Đỗ Thanh Nghị, Francois Poulet đề xuất

3. Mục tiêu đề tài

Đề tài này thực hiện thu thập dữ liệu bình luận từ internet cụ thể là trang Thegioididong sau đó nghiên cứu, tìm hiểu về thuật toán Baiyes thơ ngây (Naïve Bayes), tập trung chủ yếu vào nghiên cứu lý thuyết và ứng dụng thuật toán đồng thời đánh giá tính hiệu quả của thuật toán.

Sau đó thực hiện cài đặt một mô hình huấn luyện về phân loại bình luận trong máy học với các đánh giá bình luận cùng nội dung và sử dụng chúng làm bộ phận phân lớp cho ứng dụng phân loại bình luận trên các thiết bị thông minh như điện thoại, máy tính .

4. Đối tượng và phạm vi nghiên cứu

Về đối tượng nghiên cứu: phương pháp cào dữ liệu từ internet, thuật toán Baiyes thơ ngây là đối tượng chính của đề tài này.

Về phạm vi nghiên cứu: Các bình luận trên trang Thegioididong.

5. Phương pháp nghiên cứu

Phương pháp lý thuyết:

- Phương pháp phân tích điều tra số liệu: thu thập và nghiên cứu các tài liệu có liên quan đến đề tài.
- Phương pháp nghiên cứu tài liệu: các kỹ thuật tiền xử lý văn bản, trí tuệ nhân tạo và đặc biệt là kỹ thuật máy học.

Phương pháp thực nghiệm:

- Nghiên cứu và khai thác dữ liệu từ internet.
- Nghiên cứu và khai thác các mô hình phân loại văn bản.
- Xây dựng chương trình ứng dụng phân loại bình luận và các điểm đặc trưng của bình luận.
- Kiểm tra, thử nghiệm, nhận xét và đánh giá kết quả.

6. Kết quả đạt được

Xây dựng được một website phân loại bình luận trực tuyến hoàn chỉnh. Với một số chức năng dành cho người dùng như có thể tra cứu, phân loại bình luận, ...

Hiệu quả:

- Ít tốn thời gian của người dùng trong việc tham gia vào tham gia tiếp cận trực tiếp mô hình.
- Đơn giản hoá công việc, dễ tiếp cận mọi người dùng do xây dựng trên nền internet.

7. Bố cục niên luận

Phần giới thiệu

Giới thiệu tổng quát về đề tài.

Phần nội dung

Chương 1 : Mô tả bài toán và giới thiệu mạng nơ ron nhân tạo.

Chương 2 : Giới thiệu về cơ sở lí thuyết

Chương 3 : Xây dựng mô hình và triển khai ứng dụng..

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống.

PHẦN NỘI DUNG

CHƯƠNG 1 GIỚI THIỆU VỀ CƠ SỞ LÝ THUYẾT

1.1. Mô tả bài toán.

Trong đề tài này, sử dụng mô hình máy học để xác định mức độ đánh giá một bình luận mới đến, sử dụng mô hình Multinomial Naive Bayes để tìm hiểu các đặc điểm của văn bản như một phương tiện để thực hiện phân loại văn bản. Về mặt kinh nghiệm, phương pháp Naive Bayes một trong những thuật toán rất tiêu biểu cho hướng phân loại dựa trên lý thuyết xác suất.

Nhưng trước tiên ta cần tìm hiểu cơ bản về mô máy học và phương thức thu thập dữ liệu từ internet, sau đó sẽ tiến hành phân tích và xây dựng mô hình Naive Bayes cho phân loại bình luận.

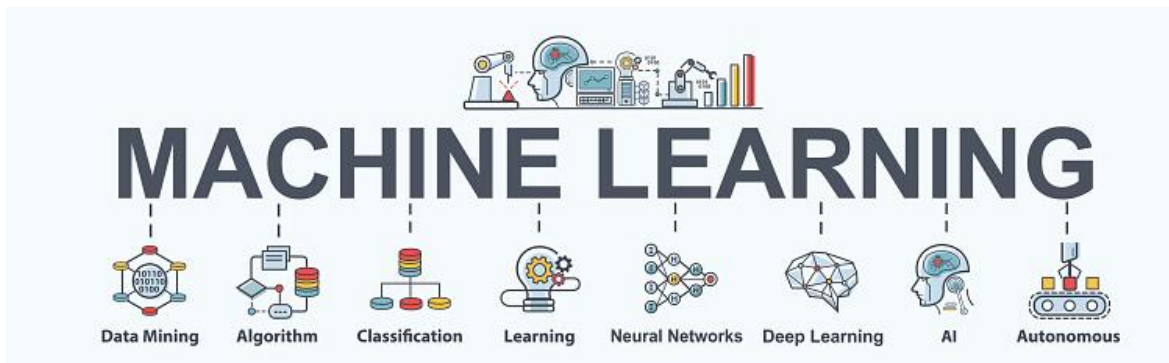
1.2. Vấn đề và giải pháp liên quan đến bài toán.

1.2.1. Máy học là gì.

Máy học hay học máy (Machine learning) là một ứng dụng của *trí tuệ nhân tạo (AI)* cung cấp cho các hệ thống khả năng tự động học hỏi và cải thiện từ kinh nghiệm mà không cần lập trình rõ ràng. Học máy tập trung vào việc phát triển các chương trình máy tính có thể truy cập dữ liệu và sử dụng nó để tự học. Quá trình học bắt đầu bằng các quan sát hoặc dữ liệu.

Ví dụ, để tìm kiếm các mẫu trong dữ liệu và đưa ra quyết định tốt hơn trong tương lai dựa trên các ví dụ mà chúng tôi cung cấp. Mục đích chính là cho phép các máy tính tự động học mà không cần sự can thiệp hay trợ giúp của con người và điều chỉnh các hành động tương ứng.

Học máy có liên quan lớn đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán.



Hình 1 Mô tả máy học

Nhiều bài toán suy luận được xếp vào loại bài toán khó, vì thế một phần của máy học là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được. Các thuật toán học máy được phân loại theo kết quả mong muốn của thuật toán. Các loại thuật toán thường dùng bao gồm:

- Học có giám sát: trong đó, thuật toán tạo ra một hàm ánh xạ dữ liệu vào tới kết quả mong muốn. Một phát biểu chuẩn về một việc học có giám sát là bài toán phân loại: chương trình cần học (cách xấp xỉ biểu hiện của) một hàm ánh xạ một vector tới một vài lớp bằng cách xem xét một số mẫu dữ liệu - kết quả của hàm đó.
- Học không giám sát: mô hình hóa một tập dữ liệu, không có sẵn các ví dụ đã được gắn nhãn.
- Học tăng cường: trong đó, thuật toán học một chính sách hành động tùy theo các quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường, và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán của quá trình học.

1.2.2. Định lý Bayes.

Định lý Bayes (Bayes' Theorem) là một định lý toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Định lý này đặt theo tên nhà toán học Thomas Bayes, người Anh sống ở thế kỷ 18. Đây là một trong những công cụ vô cùng hữu ích, người bạn thân của các nhà khoa học người làm trong ngành khoa học dữ liệu.

Ngây thơ là các thuộc tính (biến) có độ quan trọng như nhau, các thuộc tính (biến) độc lập có điều kiện khi được cho lớp/nhãn

- Ưu điểm của mô hình sử dụng định lý Bayes thơ ngây:

- Cho kết quả tốt trong thực tế mặc dù chịu những giả thiết về tính độc lập có điều kiện (khi được cho nhãn/lớp) của các thuộc tính.
- Phân lớp không yêu cầu phải ước lượng một cách chính xác xác suất.
- Dễ cài đặt, học nhanh, kết quả dễ hiểu.
- Sử dụng trong phân loại text, spam, etc.
- Dữ liệu liên tục có thể không tuân theo phân phối chuẩn.

Tuy nhiên khi dữ liệu có nhiều thuộc tính dư thừa thì Bayes không còn hiệu quả.

Các mô hình thường dùng trong Bayes:

- Multinomial Naive Bayes: Mô hình này chủ yếu được sử dụng trong phân loại văn bản. Đặc trưng đầu vào ở đây chính là tần suất xuất hiện của từ trong văn bản đó.
- Bernoulli Naive Bayes: Mô hình này được sử dụng khi các đặc trưng đầu vào chỉ nhận giá trị nhị phân 0 hoặc 1 (phân bố Bernoulli).
- Gaussian Naive Bayes: Khi các đặc trưng nhận giá trị liên tục, ta giả sử các đặc trưng đó có phân phối Gaussian.

1.2.3. Các khái niệm liên quan.

Observation: kí hiệu là x , input trong các bài toán. Observation thường có dạng một vector $x = (x_1, x_2, \dots, x_n)$, gọi là **feature vector**. Mỗi x_i gọi là một **feature**. Ví dụ bạn muốn đoán xem hôm nay có mưa không dựa vào observation gồm các feature (nhiệt độ, độ ẩm, tốc độ gió).

Label: kí hiệu là y , output của bài toán. Mỗi observation sẽ có một label tương ứng. Ở ví dụ về mưa ở trên label chỉ là "mưa" hoặc "không mưa". Label có thể mang nhiều dạng nhưng đều có thể chuyển đổi thành một số thực hoặc một vector. Trong chương này chủ yếu làm việc với label là số thực.

Model: trong chương này các bạn hiểu là nó là một hàm số $f(x)$, nhận vào một observation x và trả về một label $y=f(x)$.

Parameter: mọi thứ của model được sử dụng để tính toán ra output. Ví dụ model là một hàm đa thức bậc hai: $f(x)=ax_1+bx_2+c$ thì parameter của nó là bộ ba (a,b,c) .

Training set: Đây thường là một tập dữ liệu có kích thước lớn, được dùng để học trong quá trình huấn luyện máy học. Nôm na dễ hiểu là, đây chính là tập dữ liệu máy dùng để học và rút trích được những đặc điểm quan trọng để ghi nhớ lại. Tập training set sẽ gồm 2 phần:

- Input: sẽ là những dữ liệu đầu vào. Ví dụ với bài toán nhận dạng hình ảnh chẳng hạn: input sẽ là những bức hình.

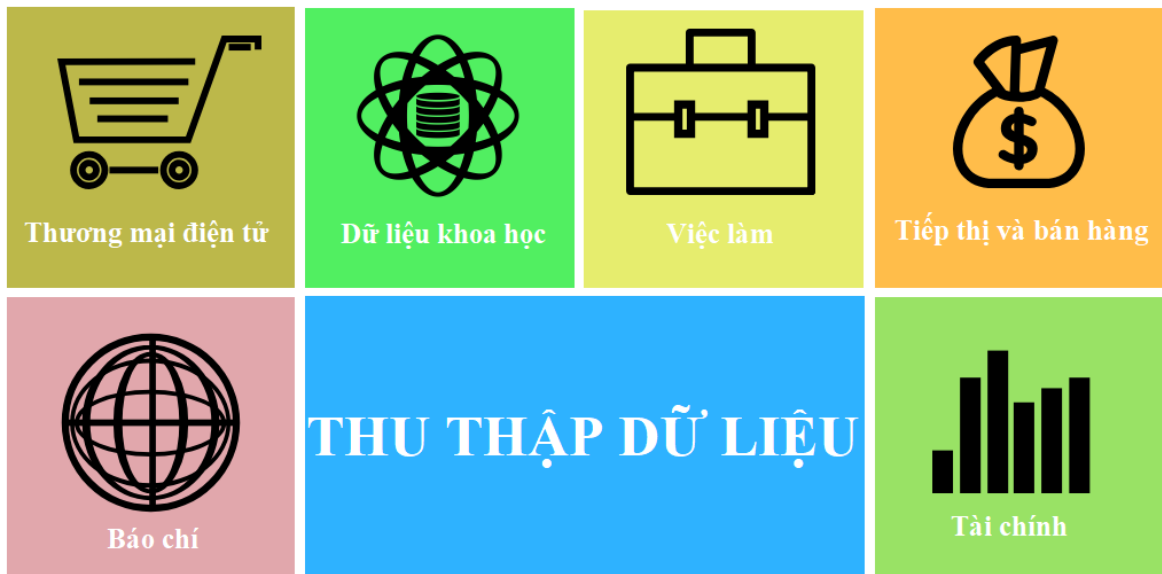
- Output: sẽ là những kết quả tương ứng với tập input. Ví dụ: Nếu input là (có ria, thích bắt chuột, kêu meo, có lông), thì output sẽ là "con mèo".

Tóm lại: Train set là tập các cặp input và output dùng để huấn luyện trong quá trình máy học.

Testing set: là tập dữ liệu dùng để kiểm tra sau khi máy đã học xong. Một mô hình máy học sau khi được huấn luyện, sẽ cần phải được kiểm chứng xem nó có đạt hiệu quả không. Và để kiểm nghiệm được độ chính xác của mô hình này, người ta dùng tập Testing set. Khác với Training set, Testing set chỉ gồm các giá trị input mà không có các giá trị output. Máy tính sẽ nhận những giá trị input này, và xử lý các giá trị, sau đó đưa ra output tương ứng cho giá trị input. Ví dụ, bạn đưa cho máy tính 1 lá bài hình con mèo : đây chính là giá trị input. Máy tính sẽ xử lý các chi tiết trên lá bài này và in ra màn hình: "con mèo" : Đây chính là output. Tóm lại: Testing set là tập các giá trị input và được dùng để kiểm thử độ chính xác của những mô hình máy học sau khi được huấn luyện.

1.2.4. Thu thập dữ liệu.

Thu thập dữ liệu hay cào dữ liệu là kỹ thuật thu thập dữ liệu từ các website trên mạng theo đường dẫn cho trước. Các công cụ thu thập sẽ truy cập vào đường dẫn và tải toàn bộ dữ liệu cũng như tìm kiếm thêm các đường dẫn bên trong để tiếp tục công việc khai thác dữ liệu. Kỹ thuật cào dữ liệu có thể thực hiện bởi hầu hết các ngôn ngữ lập trình hiện đại hỗ trợ HTTP, XML và DOM như: PHP , Python, Java, Javascript... Trong đề tài này sẽ sử dụng thư viện BeautifulSoup trên môi trường Python để thực hiện thu thập.



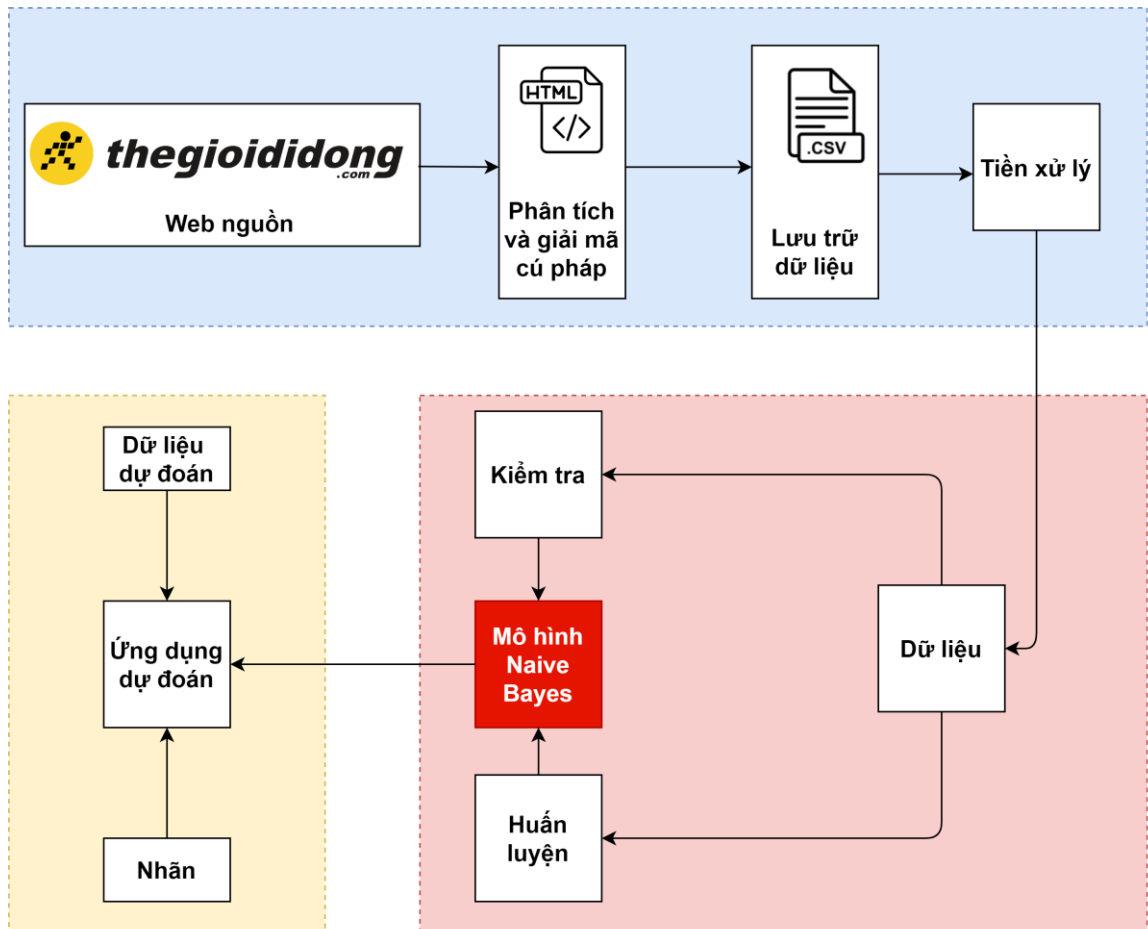
Hình 2 Mô tả hệ thống thu thập dữ liệu

Thu thập dữ liệu, ở dạng chung nhất, đề cập đến một kỹ thuật trong đó một chương trình máy tính trích xuất dữ liệu từ đầu ra được tạo ra từ một chương trình khác. Thu thập dữ liệu thường được biểu hiện trong việc thu thập dữ liệu trên web, quá trình sử dụng một ứng dụng hay một công cụ để trích xuất thông tin có giá trị từ một trang web.

CHƯƠNG 2

THIẾT KẾ VÀ CÀI ĐẶT

2.1. Thiết kế hệ thống.



Hình 3 Mô tả hoạt động của hệ thống

Hệ thống được chia thành 3 khối với 3 nhóm chức năng riêng biệt:

- Thu thập dữ liệu
- Xây dựng mô hình
- Xây dựng ứng dụng

2.2. Thu thập và tiền xử lý dữ liệu.

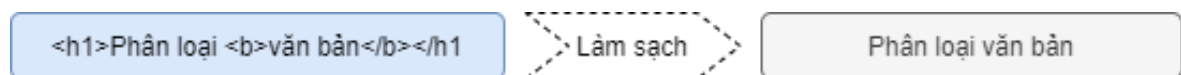
2.2.1. Thu thập dữ liệu

Quá trình thu thập dữ liệu web khá đơn giản, mặc dù việc thực hiện có thể phức tạp. Thu thập dữ liệu diễn ra trong 3 bước:

- Đầu tiên, đoạn mã được sử dụng để lấy thông tin mà chúng ta lập trình, gửi yêu cầu HTTP GET đến một trang web cụ thể.
- Khi trang web phản hồi, trình thu thập phân tích cú pháp tài liệu HTML cho một mẫu dữ liệu cụ thể.
- Khi dữ liệu được trích xuất, nó được chuyển đổi thành bất kỳ định dạng cụ thể nào mà ta cần thiết kể cho phần tiếp theo của bài toán. Một số định dạng lưu trữ: csv,json,txt,xls,...

2.2.2. Làm sạch dữ liệu

Đây là một bước để khử thành phần không cần cho bài toán. Đa phần các thành phần gây nhiễu là các thẻ HTML, JavaScript, và đương nhiên nếu cứ để thành phần gây nhiễu để tiến hành xử lý sẽ dẫn đến kết quả xử lý không tốt.



Hình 4 làm sạch văn bản

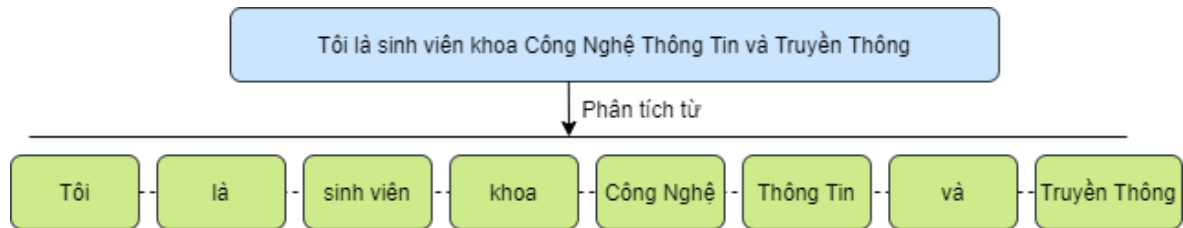
Thông thường chúng ta hay loại bỏ thành phần gây nhiễu là các thẻ HTML và JS như trên tuy nhiên thực tế thành phần nhiễu có thể không chỉ là HTML, JS, cũng có thể là những cụm từ không cần thiết, hay ký tự không có ý nghĩa (\$%&*&#u).

2.2.3. Tách từ

So với tiếng Anh mỗi từ được tạo sẽ có ngữ nghĩa cố định, nhưng trong tiếng Việt từ vựng có phần phức tạp hơn, bỏ qua mặt loại từ chỉ xét riêng cấu tạo từ có 2 loại từ cần phân biệt chính trong bài toán để máy học có thể hiểu đúng nghĩa:

- Từ đơn là từ chỉ gồm có một tiếng, có nghĩa, có thể đứng độc lập một mình.
Ví dụ: Ăn, ngủ, cây, truyện, kể, viết, đẹp,....
- Từ ghép là từ gồm hai hay nhiều tiếng, có nghĩa.
Ví dụ: Ăn uống, ăn nói, nhỏ nhẹ, con cháu, cha mẹ, anh chị, học sinh, giai cấp.

Dấu cách trong tiếng Việt không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ là 1 trong những bài toán cơ bản và quan trọng bậc nhất.



Hình 5 Mô tả hành động tách từ

Ví dụ : từ “đất nước” được tạo ra từ 2 âm tiết “đất” và “nước”, cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành 1 bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động. Tách từ chính xác hay không là công việc rất quan trọng, nếu không chính xác rất có thể dẫn đến việc ý nghĩa của câu sai, ảnh hưởng đến tính chính xác của chương trình.

2.2.4. Chuẩn hoá từ

Mục đích là đưa văn bản từ các dạng không đồng nhất về cùng một dạng. Dưới góc độ tối ưu bộ nhớ lưu trữ và tính chính xác cũng rất quan trọng. Chẳng hạn như: “Máy học”, ”đường”.

Ví dụ trong từ điển, dữ liệu huấn luyện của chúng ta không có “Máy học” và “đường” đây là các lỗi do bộ gõ và bộ mã unicode sinh ra do đặc thù các thiết bị và môi trường khác nhau, việc chuyển đổi những từ như “Máy học” về “máy học” và “đường” về “đường” là điều cần thiết để các bước xử lý sau như số hoá dữ liệu.



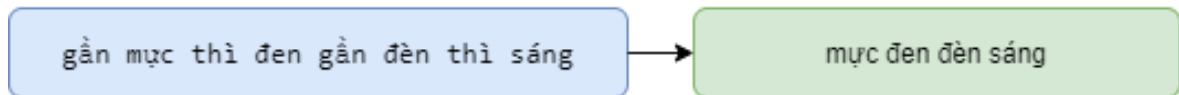
Hình 6 Mô tả hành động chuẩn hóa từ

2.2.5. Xóa từ dừng

Từ dừng là các từ xuất hiện nhiều nhưng lại vô nghĩa. Các từ dừng là những từ đặc biệt phổ biến trong ngữ liệu văn bản và do đó được coi là khá thiếu thông tin (ví dụ: các từ *như vậy, và, hoặc,...*).

- Một cách tiếp cận để loại bỏ từ dừng là tìm kiếm theo từ điển từ dành riêng cho ngôn ngữ cụ thể.

- Một cách tiếp cận khác là tạo một danh sách các từ dừng bằng cách sắp xếp tất cả các từ trong toàn bộ ngữ liệu văn bản theo tần suất. Danh sách từ dừng sau khi chuyển đổi thành một tập hợp các từ không thừa sau đó được sử dụng để loại bỏ tất cả các từ đó khỏi tài liệu đầu vào được xếp hạng trong số n từ hàng đầu trong danh sách dừng này.



Hình 7 Mô tả hành động xóa từ dừng

2.2.6. Số hoá dữ liệu

Số hoá dữ liệu hay véc tơ hóa là một bước quan trọng trong bất kỳ một bài toán nào của xử lý ngôn ngữ tự nhiên.

Tại sao lại cần vector hóa văn bản? Thông thường, máy tính không thể hiểu được ý nghĩa các từ. Như vậy, để xử lý được ngôn ngữ tự nhiên, ta cần có một phương pháp để biểu diễn văn bản dưới dạng mà máy tính có thể hiểu được.

Phương pháp tiêu chuẩn để biểu diễn văn bản đó là biểu diễn các văn bản theo véc tơ. Trong đó, các từ/cụm từ thuộc kho tài liệu ngôn ngữ được ánh xạ thành những véc tơ trên hệ không gian số thực.

Một số phương pháp phổ biến hiện nay:

- Bag of Word
- TF-IDF

2.3. Cài đặt mô hình.

2.3.1. Mô hình Naive Bayes.

Bài toán sẽ sử dụng mô hình Multinomial Naive Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản. Đặc trưng đầu vào ở đây chính là tần suất xuất hiện của từ trong văn bản đó. Ta tính xác suất từ xuất hiện trong văn bản $P(x_i/y)$ như sau:

$$P(x_i/y) = \frac{N_i}{N_c}$$

Trong đó:

N_i là tổng số lần từ x_i xuất hiện trong văn bản.

N_c là tổng số lần từ của tất cả các từ x_1, x_2, \dots, x_n xuất hiện trong văn bản.

Laplace Smoothing: Công thức trên có hạn chế là khi từ x_i không xuất hiện lần nào trong văn bản, ta sẽ có $N_i = 0$. Điều này làm cho $P(x_i/y) = 0$. Về phải của công thức bằng 0 nếu bất kì một giá trị nào bằng 0 (mặc dù có thể các giá trị khác rất lớn). Để khắc phục vấn đề này, người ta sử dụng kỹ thuật gọi là Laplace Smoothing bằng cách cộng thêm vào cả tử và mẫu để giá trị luôn khác 0.

$$P(x_i/y) = \frac{N_i + \alpha}{N_c + d\alpha}$$

Trong đó:

α thường là số dương, bằng 1.

$d\alpha$ được cộng vào mẫu để đảm bảo $\sum_{i=1}^d P(x_i/y) = 1$

2.3.2. Huấn luyện.

Quá trình huấn luyện một mô hình máy học bao gồm việc cung cấp một thuật toán máy học (tức là thuật toán học tập) cùng bộ dữ liệu để học tập. Thuật ngữ mô hình máy học đề cập đến mô hình tạo tác được tạo ra bởi quá trình đào tạo. Dữ liệu huấn luyện phải chứa câu trả lời đúng, được gọi là trạng thái mục tiêu hoặc thuộc tính mục tiêu. Thuật toán học tìm các mẫu trong dữ liệu đào tạo ánh xạ các thuộc tính dữ liệu đầu vào cho mục tiêu (câu trả lời muốn dự đoán) và nó xuất ra một mô hình máy học chứa các thuộc tính này. Một trong những trọng tâm khác của học máy là đạt được tính phổ quát (generalization), nói cách khác là mô hình sau khi được xây dựng bởi thuật toán máy học dùng để dự đoán hay đưa ra những quyết định khi gặp dữ liệu mà nó chưa gặp bao giờ. Một chương trình chỉ hiệu quả với dữ liệu đã gặp nhìn chung không có nhiều tính hữu dụng.

Ví dụ: giả sử bạn muốn đào tạo một mô hình máy học để dự đoán xem một người bất kì có phải là người miền Nam hay Bắc. Bạn sẽ cung cấp cho mô hình máy học dữ liệu đào tạo chứa các sở thích, món ăn, thông tin, ... mà bạn biết về đối tượng (nghĩa là một nhãn cho biết các thông tin về người Nam và Bắc). Mô hình máy học sẽ đào

tạo và sử dụng dữ liệu này, dẫn đến một mô hình cố gắng dự đoán liệu là người Bắc hay Nam.

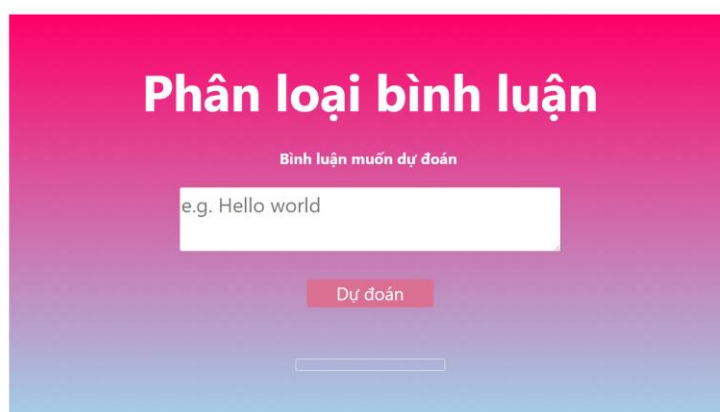
2.3.3 Đánh giá.

Kiểm thử hay đánh giá mô hình là công việc được thực hiện sau khi huấn luyện dữ liệu. Việc đánh giá mô hình chắc chắn cần thiết. Khi đánh giá mô hình, ta mới có thể kiểm soát, điều chỉnh thuật toán đạt hiệu quả tối ưu, so sánh giữa nhiều mô hình và chọn ra mô hình chính xác nhất, và xác nhận điều này trên một dữ liệu độc lập. Yếu tố thời gian cũng quyết định tính khả thi của một mô hình máy học. Chi tiết về quá trình đánh giá mô hình Naïve Bayes sẽ được trình bày ở Chương 3

2.4. Cài đặt giao diện người dùng.

Kiến trúc Client/ Server là kiến trúc nổi tiếng trong mạng máy tính, hầu hết các website hoạt động dựa trên kiến trúc này. Trong đó, Client là máy khách gửi yêu cầu đến máy Server. Tại đây thì server lắng nghe các yêu cầu từ máy Client, nhận thông tin từ Client sau đó xử lý, trả kết quả về máy Client. Khi người dùng gửi một yêu cầu lên hệ thống, yêu cầu có dạng dữ liệu là hình ảnh, hệ thống tiến hành tiền xử lý dữ liệu, đưa về dữ liệu dạng chuẩn của mô hình, sau đó sẽ trích xuất đặc trưng và đưa vào mô hình đã được xây dựng ở phần trước.

Mô hình sẽ trả ra kết quả là tên lớp văn bản, từ dữ liệu đầu vào mô hình sẽ đánh giá dữ liệu này và hiển thị kết quả cho người dùng. Kết quả phản hồi là trang thông tin dữ liệu phân loại như hình sau:



Hình 8 Giao diện người dùng

CHƯƠNG 3

ĐÁNH GIÁ MÔ HÌNH

3.1. Môi trường cài đặt.

- Scikit-learn (Sklearn) thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning .
- PyVi thư viện xử lý ngôn ngữ tiếng Việt trong Python .
- Gensim thư viện xử lý ngôn ngữ tự nhiên trong Python.
- BeautifulSoup thư viện thu thập dữ liệu từ các trang web
- Cùng một số thư viện phụ trợ khác như tqdm, docx2txt

3.2. Kết quả kiểm tra.

Khi xây dựng một mô hình máy học, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình. Hiệu năng của một mô hình thường được đánh giá dựa trên tập dữ liệu kiểm thử.

Cụ thể, giả sử đầu ra của mô hình khi đầu vào là tập kiểm thử được mô tả bởi véc tơ y_{pred} - là vector dự đoán đầu ra với mỗi phần tử là lớp được dự đoán của một điểm dữ liệu trong tập kiểm thử. Ta cần so sánh giữa véc tơ dự đoán y_{pred} này với vector lớp *thật* của dữ liệu, được mô tả bởi véc tơ y_{data} . Nghi thức đánh giá: sử dụng nghi thức hold-out : lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 tập dữ liệu còn lại dùng cho kiểm tra, có thể lặp lại quá bước này k lần rồi tính giá trị trung bình

Phương pháp Confusion Matrix: Là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp.

		Nhãn trị thực	
		Tích cực	Tiêu cực
Nhãn dự đoán	Tích cực	528	43
	Tiêu cực	134	222

Bảng 1 Kết quả thực nghiệm dự trên mô hình

Accuracy: tỷ lệ các trường hợp được dự báo đúng trên tổng số các trường hợp là bao nhiêu. Độ chính xác giúp ta đánh giá hiệu quả dự báo của mô hình trên một bộ dữ liệu. Độ chính xác càng cao thì mô hình của chúng ta càng chuẩn xác.

Bảng giá trị kiểm tra

Lần	Độ chính xác
1	0.795
2	0.795
3	0.792
4	0.812
5	0.790
6	0.814
7	0.795
8	0.790
9	0.807
10	0.799
Trung bình	0.798

Bảng 2 Độ chính xác của kết quả thực nghiệm

PHẦN KẾT LUẬN

1. Kết quả đạt được

- Cơ bản xây dựng được các bước xây dựng một website phân loại dữ liệu, thực hiện được các quy trình cơ bản của một website động.
- Cơ bản xây dựng được mô hình máy học cho dự phân loại văn bản.
- Xây dựng bố cục trang web hợp lý, dễ nhìn. Bước đầu hoàn thành được các nghiệp vụ cơ bản của hệ thống.

2. Ưu điểm.

- Nâng cao khả năng tiếp cận internet của người dùng thông qua ứng dụng.
- Tạo ra nền tảng phân loại tự động khi có một lượng lớn dữ liệu thay vì phải phân tích thủ công.

3. Nhược điểm.

- Hệ thống mới được triển khai và phát triển ở quy mô nhỏ.
- Hệ thống còn nhiều thiếu sót về chức năng.
- Độ chính xác chưa cao.
- Dữ liệu thu thập chưa đủ lớn

4. Hướng phát triển.

- Tối ưu hóa thuật toán và mô hình máy học.
- Cải thiện thời gian cũng như kích thước tập dữ liệu.

TÀI LIỆU THAM KHẢO

- [1] Steven Bird, Ewan Klei & Edwan Lober, —Natural Language Processing with Python, 2009.
- [2] Sebastian Raschka, —Naive Bayes and Text Classification – Introduction and Theory, 2014.
- [3] Nguyễn Thị Thùy Dương, Nghiên cứu lý thuyết Naive Bayes và ứng dụng trong phân loại văn bản tiếng Việt, Luận văn Thạc Sĩ Khoa Học Máy Tính– Đại học Thái Nguyên, 2015.