TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

NGÀNH HỆ THỐNG THÔNG TIN

Đề tài

XÂY DỰNG HỆ THỐNG PHÂN LOẠI VĂN BẢN TỰ ĐỘNG

Sinh viên: Nguyễn Thanh Liêm

Mã số: B2003790

Khóa: K46

Cần Thơ, 12/2024

TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC
NGÀNH HỆ THỐNG THÔNG TIN

Đề tài

XÂY DỰNG HỆ THỐNG PHÂN LOẠI VĂN BẢN TỰ ĐỘNG

Người hướng dẫn
TS. Nguyễn Minh Khiêm

Sinh viên: Nguyễn Thanh Liêm
Mã số: B2003790
Khóa: K46

Cần Thơ, 12/2024

TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

XÁC NHẬN CHỈNH SỬA LUẬN VĂN THEO YÊU CẦU CỦA HỘI ĐỒNG

Tên luận văn (tiếng Việt và tiếng Anh): Xây dựng hệ thống phân loại văn bản tự động (Automated Text Classification Application)

Họ tên sinh viên: Nguyễn Thanh Liêm MASV: B2003790

Mã lớp: DI2095A2

Đã báo cáo tại hội đồng ngành: Hệ thống thông tin

Ngày báo cáo: 11/12/2024

Luận văn đã được chỉnh sửa theo góp ý của Hội đồng.

Cần Thơ, ngày 20 tháng 12 năm 2024

Giáo viên hướng dẫn

(Ký và ghi họ tên)

Nguyễn Minh Khiêm

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành đến quý Thầy/Cô Trường Đại học Cần Thơ, quý Thầy/Cô thuộc Trường Công nghệ Thông tin và Truyền thông, và đặc biệt là quý Thầy/Cô thuộc Khoa Hệ Thống Thông Tin đã tận tình chỉ dạy và truyền đạt cho em những kiến thức bổ ích trong suốt quãng thời gian còn là sinh viên. Những kiến thức này sẽ là hành trang quý báu, giúp em tiếp bước trên con đường tương lai phía trước.

Em xin dành lời cảm ơn chân thành nhất đến Thầy TS. Nguyễn Minh Khiêm, người đã tận tình chỉ dạy, hướng dẫn, tạo điều kiện tốt nhất và luôn hỗ trợ sát sao để em có thể hoàn thành tốt đề tài luận văn tốt nghiệp.

Em xin cảm ơn Cô ThS. Bùi Đăng Hà Phương và Thầy ThS. Phạm Ngọc Quyền trong Hội đồng phản biện đã nhận lời và dành thời gian góp ý để đề tài luận văn của em được hoàn chỉnh hơn.

Em xin cảm ơn gia đình, bạn bè xung quanh đã luôn hỗ trợ, ủng hộ, động viên và tạo điều kiện tốt nhất để em có thể thực hiện tốt đề tài luận văn này.

Kiến thức và năng lực của bản thân em còn hạn chế, nên không tránh khỏi những sai sót trong quá trình thực hiện đề tài. Kính mong nhận được sự thông cảm và những đóng góp chân thành, quý báu từ quý Thầy/Cô và các bạn để em có thể phát triển đề tài tốt hơn trong tương lai.

Cuối lời, em xin kính chúc quý Thầy/Cô và các bạn có nhiều sức khỏe và nhiều thành công hơn trong tương lai.

Cần Thơ, ngày 11 tháng 12 năm 2024

Sinh viên thực hiện

Nguyễn Thanh Liêm

TÓM TẮT

Sự phát triển ngày càng lớn và bùng nổ của mạng internet đã kéo theo sự xuất hiện của các trang mạng xã hội, các bài báo, văn bản làm cho số lượng người sử dụng trao đổi thông tin trở nên rất lớn và không ngừng phát triển. Phần lớn những người sử dụng mạng internet thường chia sẻ cảm xúc, cuộc sống, kiến thức, ý kiến, quan điểm, … của chính mình. Việc phân tích và đưa ra chủ đề cho những trao đổi đó nhằm nắm bắt, dễ dàng quản lý, trích xuất thông tin và vô cùng quan trọng, có ý nghĩa lớn cho ngành giáo dục, kinh tế, chính trị, pháp luật, xã hội, … Một giải pháp hiệu quả cho công việc trên là phát triển một ứng dụng có khả năng phân loại văn bản tiếng Việt tự động.

Phân loại văn bản là một trong những vấn đề quan trọng của việc xử lý ngôn ngữ tự nhiên thuộc nhóm học có giám sát và trí tuệ nhân tạo. Nhiệm vụ chính của bài toán là đưa ra chủ đề cho văn bản vào một nhóm chủ đề cho trước. Để có thể giải quyết bài toán này cần xử lý hai vấn đề phức tạp và quan trọng ở hai giai đoạn: thu thập dữ liệu từ các trang mạng xã hội và phân tích đưa ra chủ đề thích hợp cho dữ liệu. Phương pháp được đề xuất sử dụng cho đề tài là biểu diễn văn bản bằng TF-IDF để chuyển đổi văn bản thành các vector đặc trưng số học, từ đó áp dụng mô hình phân loại Naïve Bayes cho việc phân loại văn bản. Các bước xử lý dữ liệu bao gồm: tiền xử lý văn bản như loại bỏ các ký tự không cần thiết, chuyển đổi chữ viết thường, loại bỏ từ dừng, chuẩn hóa bảng mã Unicode, … và mã hóa các nhãn chủ đề. Tiến hành đánh giá mô hình thông qua các chỉ số như độ chính xác, độ nhạy và F1-Score nhằm đảm bảo tính toàn diện của kết quả.

Đề tài đã có đóng góp vào việc xây dựng một quy trình phân loại văn bản đơn giản, hiệu quả, dễ dàng sử dụng và triển khai trên các hệ thống thông tin thực tế. Kết quả thực tế cho được độ chính xác cao và ổn định khi áp dụng trên các tập dữ liệu khác nhau. Bên cạnh đó, ứng dụng đã giúp quản lý và xử lý lượng lớn thông tin trong nhiều lĩnh vực như quản lý tài liệu, phân loại email, phân tích nội dung trên các nền tảng xã hội, …

Từ khóa: Phân loại văn bản, tiếng Việt, học có giám sát, TF-IDF, Naïve Bayes.

ABSTRACT

The growing and explosive development of the Internet has led to the emergence of social networking sites, articles, and documents, making the number of users exchanging information very large and constantly growing. Most Internet users often share their own feelings, lives, knowledge, opinions, views, etc. Analyzing and giving topics for those exchanges to capture, easily manage, extract information is extremely important, and has great significance for the education, economics, politics, law, society, etc. sectors. An effective solution for the above work is to develop an application that can automatically classify Vietnamese text.

Text classification is one of the important problems of natural language processing in the group of supervised learning and artificial intelligence. The main task of the problem is to give topics for the text into a given group of topics. To solve this problem, two complex and important issues need to be addressed in two stages: collecting data from social networking sites and analyzing to find appropriate topics for the data. The proposed method for the topic is to represent text using TF-IDF to convert text into numerical feature vectors, from which to apply the Naïve Bayes classification model for text classification. The data processing steps include: text preprocessing such as removing unnecessary characters, converting to lowercase, removing stop words, normalizing Unicode, ... and encoding topic labels. Evaluate the model through indicators such as accuracy, sensitivity and F1-Score to ensure the comprehensiveness of the results.

The topic has contributed to the construction of a simple, effective, easy-to-use and deployable text classification process on real information systems. The actual results show high accuracy and stability when applied on different data sets. In addition, the application has helped manage and process large amounts of information in many fields such as document management, email classification, content analysis on social platforms, etc.

Keywords: Text classification, Vietnamese, supervised learning, TF-IDF, Naïve Bayes.

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ CHUYÊN NGÀNH

Viết tắt	Giải thích
SVM	Máy vector hỗ trợ (Support Vector Machine)
KNN	Thuật toán K láng giềng gần nhất (K-Nearest Neighbors)
NLTK	Bộ công cụ xử lý ngôn ngữ tự nhiên (Natural Language Toolkit)
MLE	Ước lượng khả năng tối đa (Maximum Likelihood Estimation)
MAP	Ước lượng xác suất hậu nghiệm tối đa (Maximum A Posteriori)
LM	Mô hình ngôn ngữ (Language Model)
CRFs	Trường ngẫu nhiên có điều kiện (Conditional Random Fields)
HTML	Ngôn ngữ đánh dấu siêu văn bản (HyperText Markup Language)
IDE	Môi trường phát triển tích hợp (Integrated Development Environment)

CHƯƠNG 1. GIỚI THIỆU

1.1. Đặt vấn đề

Với sự phát triển ngày càng vượt bậc của nền công nghệ thông tin đặc biệt là mạng internet, sự bùng nổ về lượng thông tin được số hóa và đưa lên mạng ngày càng nhiều. Mạng internet đã trở thành một kho kiến thức khổng lồ về mọi lĩnh vực. Do đó, số lượng các văn bản xuất hiện ngày càng nhiều đặc biệt là các bài báo điện tử. Việc phân loại các bài báo theo các chủ đề giúp người đọc dễ theo dõi và thuận tiện hơn khi tìm đọc các bài báo có liên quan cùng chủ đề đó. Các chủ đề trong cuộc sống thường bao gồm: Chính trị, kinh tế, giáo dục, khoa học, pháp luật, chuyện lạ, …

Hiện nay, việc tìm kiếm các bài đọc hoặc bài báo liên quan đến một lĩnh vực cụ thể sẽ khá khó khăn và tốn nhiều thời gian. Vì vậy, phát triển một mô hình có khả năng tự động phân loại các văn bản, bài viết vào các chủ đề danh mục đã được xác định trước sẽ rất có ý nghĩa. Mô hình này giúp tối ưu hóa quá trình xử lý và phân tích dữ liệu văn bản, hỗ trợ đưa ra quyết định, tự động hóa các quy trình phân loại văn bản vào một lĩnh vực cụ thể.

1.2. Các nghiên cứu liên quan

Bài toán phân loại tài liệu văn bản được đặt ra với mục đích giúp con người có thể tiết kiệm được thời gian trong việc tìm kiếm, tổng hợp thông tin và quản lý dữ liệu. Có nhiều phương pháp phân loại văn bản như Naïve Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbor, … Các bài báo liên quan đến vấn đề phân loại văn bản được công bố như bài báo "Text Categorization Based on Regularized Linear Classification Methods" [1] của nhóm tác giả Tong Zhang & Frank J.Oles (Mathematical Sciences Department, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598). Bài báo này trình bày phương pháp phân loại văn bản tuyến tính dựa vào các kỹ thuật Linear Least Squares Fit, Logistic Regression và SVM. Bài báo "Text Categorization with Support Vector Machines" [2] của tác giả Thorsten Joachims. Bài báo trình bày về việc sử dụng và cải tiến kỹ thuật Support Vector Machines (SVM) cho việc học máy có hiệu quả trong việc phân loại văn bản.

Hầu hết các bài báo trên đều đạt được kết quả khá tốt, tuy nhiên khó có thể so sánh chúng vì tập dữ liệu thực nghiệm cho mỗi phương pháp là khác nhau. Các nghiên cứu này dựa trên các tiếp cận máy học, mô hình xác suất và thống kê, thường tập trung vào bài toán được phân làm hai lớp và gặp khó khăn với dữ liệu lớn. Mặc khác, các bài báo trên dành cho vấn đề xử lý ngôn ngữ nước ngoài, cụ thể là tiếng Anh. Để áp dụng cho các văn bản là tiếng Việt thì sẽ không đạt được kết quả mong muốn.

Ở Việt Nam có một số bài báo nghiên cứu về phân loại văn bản như "Phân lớp văn bản tiếng Việt tự động theo chủ đề" [3] của nhóm tác giả Lê Văn Nguyên. Bài báo trình bày các thuật toán phân loại văn bản như Naïve Bayes, SVM và KNN để thực nghiệm phân

lớp văn bản tiếng Việt trên 5 bộ dữ liệu thuộc 4 chủ đề khác nhau. Bài báo cho kết quả khá tốt với thuật toán SVM có độ chính xác cao nhất và thời gian thực nghiệm mô hình là thấp nhất. Bài báo "Khảo sát các mô hình phân loại văn bản tiếng Việt" [4] của tác giả Nguyễn Chí Hiếu đã trình bày rõ từng ưu, khuyết điểm đối với các phương pháp phân loại văn bản cho văn bản tiếng Việt với các thuật toán như Naïve Bayes, SVM, Học sâu (Deep Learning) và K-NN. Bên cạnh đó bài báo còn đề xuất một số giải pháp cho việc phân loại văn bản trên các tập dữ liệu khác nhau.

1.3. Mục tiêu đề tài

Mục tiêu chính của đề tài là xây dựng mô hình ứng dụng phân loại văn bản, bài báo tiếng Việt theo các chủ đề cụ thể giúp người dùng có thể tiết kiệm được thời gian, dễ dàng tìm kiếm các bài báo liên quan với nhau trong cùng một chủ đề.

1.4. Đối tượng và phạm vi đề tài

1.4.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu chính của đề tài:

- Định lý Bayes và các mô hình phân loại Naive Bayes.
- Cấu trúc cấu tạo của từ, phân loại từ, tách từ trong văn bản tiếng Việt.
- Cấu trúc ngữ pháp của văn bản tiếng Việt.
- Quy trình phân loại văn bản với mô hình học máy.
- Các bước tiền xử lý dữ liệu cho các văn bản tiếng Việt.

1.4.2. Phạm vi đề tài

Xây dựng mô hình phân loại các bài báo, văn bản tiếng việt với quy mô vừa và nhỏ, hỗ trợ cho mọi người dung có thể tìm kiếm những chủ đề liên quan đến nội dung trong cùng một lĩnh vực một cách nhanh chóng.

1.5. Nội dung đề tài

Những công việc đã thực hiện, các giai đoạn và khoảng thời gian thực hiện của mỗi công việc để hoàn thành đề tài được trình bày chi tiết trong ***Bảng 1.1*** bên dưới:

Bảng 1.1: Kế hoạch thực hiện

STT	CÔNG VIỆC THỰC HIỆN	TUẦN																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	Khảo sát vấn đề	x	x																	
2	Phân tích yêu cầu			x	x															
3	Thiết kế dữ liệu					x	x													
4	Cài đặt chương trình							x	x	x	x	x								
5	Kiểm thử											x	x							
6	Sửa lỗi chương trình												x	x	x	x				
7	Viết báo cáo												x	x	x	x	x	x	x	x

1.6. Những đóng góp chính của đề tài

- Xây dựng được các chức năng tiền xử lý cho dữ liệu văn bản tiếng Việt.

- Xây dựng được hệ thống phân loại văn bản tiếng Việt tự động có thể áp dụng vào ứng dụng cụ thể trong đời sống, góp phần giảm thiểu sự tiêu tốn về thời gian và công sức của con người cho việc phân loại thủ công.

1.7. Bố cục của luận văn

Bố cục luận văn gồm các phần như sau:

- Chương 1. Giới thiệu:
  - Tính cấp thiết của đề tài.
  - Các nghiên cứu tương tự có liên quan đến đề tài.
  - Mục tiêu, vấn đề mà đề tài muốn giải quyết.
  - Các đối tượng có liên quan và phạm vi của đề tài.
  - Những công việc cần thực hiện để đạt được mục tiêu đề tài.
  - Các kết quả đề tài đã đạt được qua các quá trình nghiên cứu.
- Chương 2. Mô tả bài toán:
  - Mô tả chi tiết bài toán
  - Các hướng tiếp cận giải quyết của đề tài.
- Chương 3. Thiết kế và cài đặt giải pháp:
  - Kiến trúc tổng quát của hệ thống.
  - Chi tiết các mô hình liên quan.
  - Các giải pháp và quá trình cài đặt hệ thống.
- Chương 4. Kiểm thử và đánh giá:
  - Các kịch bản và tiến hành kiểm thử phần mềm
  - Kết quả của quá trình kiểm thử phần mềm.
- Chương 5. Kết luận và hướng phát triển:
  - Kết quả mà đề tài đã đạt được sau khi đã kiểm thử phần mềm.
  - Những hạn chế mà đề tài chưa thực hiện được.
  - Các chức năng và mục tiêu mà đề tài hướng đến trong tương lai.

1.8. Tổng kết chương

Chương 1 nhằm giới thiệu cho người đọc hiểu tổng quan về đề tài, tính cấp thiết và các giải pháp giải quyết vấn đề được đặt ra. Giới thiệu những nghiên cứu đã được ra đời trước có liên quan đến đề tài. Giới thiệu đối tượng và phạm vi mà đề tài hướng đến bên cạnh những kết quả đã được thực hiện thành công qua quá trình nghiên cứu. Bài toán sẽ được mô tả chi tiết ở chương tiếp theo.

CHƯƠNG 2. MÔ TẢ BÀI TOÁN

2.1. Mô tả chi tiết bài toán

Phân loại văn bản là một trong những bài toán cổ điển trong khai thác dữ liệu, thuộc nhóm học có giám sát trong học máy. Nội dung chính của bài toán này là đi tìm chủ đề thích hợp trong tập hữu hạn các chủ đề đã được xác định trước. Tiêu chí cho việc chọn chủ đề phù hợp cho văn bản là dựa trên sự tương đồng về mặt ngữ nghĩa giữa chúng với các văn bản trong tập dữ liệu huấn luyện. Bài toán yêu cầu dữ liệu cần phải có các nhãn chủ đề, mô hình sẽ học từ các dữ liệu có nhãn đó, sau đó nó sẽ được dùng để dự đoán các nhãn cho các dữ liệu mới mà mô hình chưa từng gặp.

Dữ liệu sử dụng cho bài toán phân loại văn bản có thể được lấy từ các văn bản, các bài báo, các văn kiện có các chủ đề cụ thể về một lĩnh vực như pháp luật, kinh tế, chính trị, giáo dục, … Sau đó dữ liệu sẽ được tiến hành qua các bước tiền xử lý văn bản trước khi phân loại các nhãn cho các chủ đề.

Dữ liệu là các văn bản sẽ được tiền xử lý dữ liệu để tách từ, loại bỏ các từ stopwords, dấu câu, các ký tự đặc biệt, đưa về văn bản thường, chuẩn hóa lại kiểu gõ dấu câu tiếng việt, … Quá trình tiền xử lý dữ liệu là một bước quan trọng giúp chuẩn hóa dữ liệu đầu vào, loại bỏ các thành phần dư thừa và tạo một cơ sở dữ liệu sạch cho việc huấn luyện mô hình phân loại văn bản. Mục tiêu chính của quá trình này là biến văn bản thô thành văn bản dữ liệu có cấu trúc và dễ xử lý hơn. Dưới đây là các bước để tiến hành tiền xử lý cho văn bản tiếng Việt:

- Xóa các mã code HTML: Dữ liệu được thu thập từ các trang website đôi khi vẫn sẽ xót lại các đoạn mã HTML. Các mã này không có tác dụng cho việc phân loại mà còn có thể làm cho kết quả phân loại văn bản bị kém đi. Vì vậy cần loại bỏ các thẻ HTML hoặc các thành phần không cần thiết có trong văn bản.
- Chuẩn hóa bảng mã Unicode: Đảm bảo rằng tất cả các ký tự trong văn bản được chuyển đổi về dạng Unicode tiêu chuẩn.
- Chuẩn hóa kiểu gõ dấu tiếng Việt: Loại bỏ sự không nhất quán trong cách gõ dấu cho văn bản tiếng Việt, như dấu sai vị trí hoặc sử dụng các ký tự không đúng đắn.
- Thực hiện tách từ cho văn bản: Đơn vị từ trong tiếng Việt có từ đơn và từ ghép. Nên cần cung cấp cho mô hình học thuật biết rằng đâu là từ đơn và đâu là từ ghép, nhằm tránh sự nhầm lẫn rằng tất cả các từ mà máy học nhận được đều là từ đơn.
- Đưa về văn bản thường: Chuyển đổi toàn bộ chữ viết hoa thành chữ thường để tránh sự phân biệt không cần thiết giữa các từ. Việc này có thể giúp giảm số lượng đặc trưng và tăng độ chính xác cao hơn cho mô hình.

- Xóa các ký tự đặc biệt, khoảng trắng dư: Các ký tự @, #, $, &, … hoặc các khoảng trắng thừa giữa các từ trong câu không mang nhiều ý nghĩa cho bài toán. Việc loại bỏ này tránh làm ảnh hưởng đến kết quả mô hình và tăng tốc độ học và xử lý, giảm số chiều đặc trưng.
- Loại bỏ các từ stopwords: Loại bỏ các từ không mang ý nghĩa quan trọng trong ngữ cảnh của văn bản như và, của, là, đó, … nhằm tăng độ chính xác cho mô hình, giảm nhiễu dữ liệu, tối ưu hóa được hiệu suất.

Sau đó dữ liệu sẽ được huấn luyện dựa vào mô hình học máy để học từ các dữ liệu được huấn luyện và phân loại văn bản ra các nhãn chủ đề cụ thể cho từng lĩnh vực. Hiệu suất hệ thống của bài toán sẽ được đánh giá dựa trên các độ đo như độ đo chính xác (accuracy), đồng nhất (precision), phủ định (recall) và F1-score.

Hệ thống sẽ cung cấp cho người dùng một giao diện đơn giản. Tại đây, người dùng có thể nhập văn bản đầu vào hoặc lựa chọn đầu vào khác là một video có chứa nội dung cần phân loại. Nếu người dùng chọn đầu vào là một video thì phần mềm sẽ tiến hành trích xuất, chuyển đổi nội dung của video về dạng văn bản và tiến hành xử lý văn bản và phân loại chủ đề cho video đó.

2.2. Hướng tiếp cận giải quyết của đề tài

2.2.1. Hướng tiếp cận thứ 1

Đối với đề tài này cần được tiến hành tiến xử lý dữ liệu qua các bước sau: xóa các đoạn mã code HTML, chuẩn hóa kiểu gõ dấu tiếng Việt, chuẩn hóa bảng mã Unicode, tách từ, đưa về văn bản thường, xóa các ký tự đặc biệt, loại bỏ stopwords (từ dừng). Mục đích chính của công việc này là biến đổi văn bản thô thành một văn bản có cấu trúc dễ dàng xử lý và giảm bớt việc nhiễu thông tin dữ liệu. Dưới đây là các phương pháp có thể thực hiện để áp dụng cho các các bước tiền xử lý dữ liệu:

- Xóa các mã code HTML: Sử dụng biểu thức chính quy r'<[^>]*>' để tìm các đoạn văn bản nào nằm trong các dấu < > và thay thế chúng bằng các đoạn văn bản rỗng.
- Chuẩn hóa kiểu gõ dấu tiếng Việt: Xác định vị trí đặt câu và các dấu câu sau đó tiến hành chuẩn hóa dấu câu.
- Chuẩn hóa bảng mã Unicode: Áp dụng phương pháp ánh xạ để chuyển đổi các ký tự từ dạng Unicode dựng sẵn sang bản Unicode tổ hợp. Sau đó dùng biểu thức chính quy để thay thế các ký tự không chuẩn bằng các ký tự tương ứng trong bảng mã từ điển được thiết lập sẵn.
- Tách từ: Sử dụng thư viện xử lý ngôn ngữ tự nhiên là Underthesea để thực hiện tách câu thành các từ hợp với ngữ cảnh tiếng Việt.
- Đưa về văn bản thường: Sử dụng phương thức lower() trong python.
- Xóa các ký tự đặc biệt và khoảng trắng dư: Sử dụng biểu thức chính quy để tìm và loại bỏ các ký tự không cần thiết.
- Loại bỏ từ dừng (stopwords): Sử dụng các thư viện counter và NLTK để tiến hành xóa các từ không có ý nghĩa ra khỏi văn bản.

2.2.2. Hướng tiếp cận thứ 2

Đối với đề tài này có thể sử dụng thuật toán Naive Bayes làm giải pháp tiếp cận. Mô hình được lựa chọn là Multinomial Naive Bayes là một trong những mô hình phù hợp với phát triển trong phạm vi nghiên cứu đề tài.

2.3. Tổng kết chương

Chương 2 giới thiệu chi tiết bài toán mà đề tài cần phải thực hiện. Trình bày các phương pháp mà đề tài hướng đến để có thể giải quyết được bài toán một cách tốt nhất. Ở chương tiếp theo sẽ mô tả rõ hơn về cách thiết kế các mô hình và hướng dẫn cài đặt hệ thống.

CHƯƠNG 3. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP

3.1. Kiến trúc tổng quát hệ thống

Hình 3.1 bên dưới thể hiện rõ cái nhìn tổng quát về cách hoạt động của bài toán phân loại văn bản.

Ban đầu sẽ chia các bộ tập tin dữ liệu (dataset) thành hai thành phần là dữ liệu để huấn luyện (data training) và dữ liệu để kiểm thử (data testing). Tổng thể cho bài toán sẽ chia làm hai công việc như sau:

- Huấn luyện dữ liệu: Dùng bộ Training Text (data train) tiến hành qua các bước tiền xử lý văn bản, sau đó trích xuất thành các bộ vector đặc trưng (Feature Vectors) để đưa vào mô hình huấn luyện (Machine Learning Algorithm), từ đó tạo thành một mô hình đã huấn luyện (Predictive Model) dùng để đưa vào phân loại sau này. Kết quả có được cho quá trình thực hiện công việc này là mô hình đã được huấn luyện và thu được nhãn (Expected Label) của văn bản cần được phân loại.

- Hiện thực kết quả: Dùng văn bản mới (New Document/ Text) đưa vào tiền xử lý văn bản và trích xuất đặc trưng thành các bộ vectors đặc trưng (Feature Vectors) đưa vào mô hình đã được tạo ra từ quá trình huấn luyện (Predictive Model) để tiến hành phân tích và cho ra kết quả là tên nhãn (Label) cho văn bản.

Mô hình cho bài toán phân loại văn bản bao gồm các thành phần sau:

- Training Text/ Document: là các văn bản đầu vào cho mô hình máy học để tiến hành huấn luyện dữ liệu và học từ các văn bản này sau khi qua các bước xử lý văn bản.

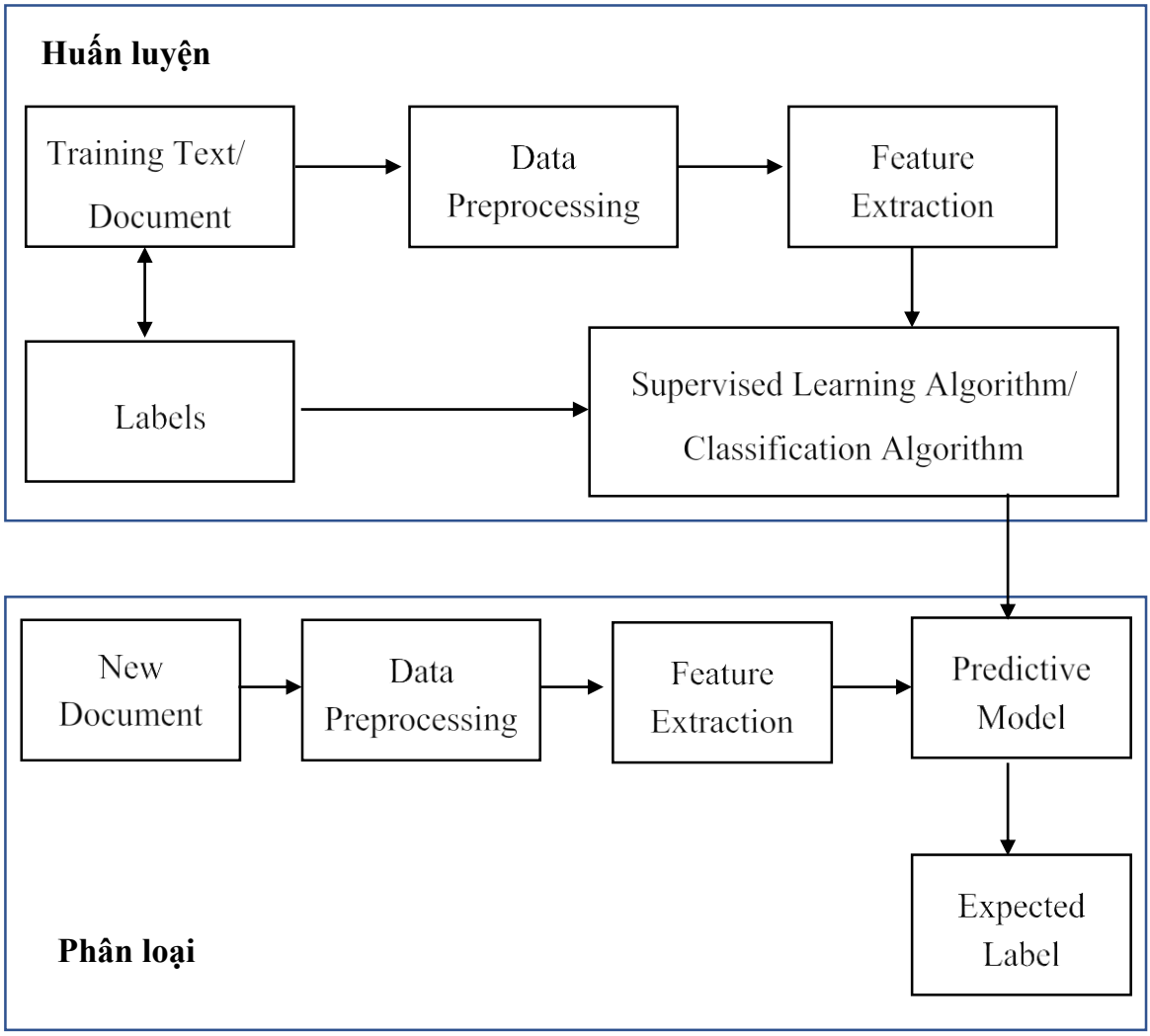
- Text Preprocessing: mô tả quá trình tiến hành tiền xử lý dữ liệu văn bản để biến một văn bản thô thành một văn bản có cấu trúc.

- Feature Vectors: chuyển đổi văn bản thành các dữ liệu số có cấu trúc để tiến hành đưa vào mô hình huấn luyện.

- Machine Learning Algorithm: thuật toán phân loại cho bài toán, thông qua đó mô hình có thể xử lý văn bản và phân loại văn bản tự động.

- Predictive Model: mô hình đã được huấn luyện dữ liệu dựa trên các bộ dữ liệu văn bản đầu vào đã được học trước đó và mô hình này được dùng để dự đoán nhãn cho các dữ liệu đầu vào là các văn bản mà người dùng nhập vào khi cần phân loại chủ đề cho chúng.

- Label: là các nhãn/ chủ đề cho các dữ liệu được xác định trước mà mô hình sẽ học và dự đoán.



Hình 3.1: Mô hình tổng quát bài toán phân loại văn bản

3.2. Xây dựng các mô hình

3.2.1. Tổng quan về từ và các phương pháp tách từ

❖ **Tổng quan về từ**

Trong quá trình học tập và sử dụng ngôn ngữ trong đời sống hằng ngày, mọi người đều quen thuộc với khái niệm về "từ". Nhưng để định nghĩa được chính xác từ là gì thì hoàn toàn không phải là một vấn đề đơn giản. Trong ngành ngôn ngữ học đã có nhiều định nghĩa về từ được đưa ra, nhưng hầu hết chưa có định nghĩa nào có thể bao quát được mọi vấn đề liên quan đến khái niệm "từ". Theo bài báo cáo [5] của tác giả Đinh Điền, có một số khái niệm tiêu biểu sau đây về từ:

- Theo L.Bloomfield thì: "*từ là một hình thái tự do nhỏ nhất".*

- Solncev quan niệm: "*từ là đơn vị ngôn ngữ có tính hai mặt: âm và nghĩa. Từ có khả năng độc lập về cú pháp khi sử dụng trong lời".*

- Còn B.Golovin quan niệm: "*từ là đơn vị nhỏ nhất có nghĩa của ngôn ngữ, được vận dụng độc lập, tái hiện tự do trong lời nói để xây dựng nên câu".*

Trong tiếng Việt, cũng có nhiều khái niệm về từ được nêu ra như:

- Theo Trương Văn Trình và Nguyễn Hiến Lê thì: "*từ là âm có nghĩa, dùng trong ngôn ngữ để diễn tả một ý đơn giản nhất, nghĩa là ý không thể phân tích ra được".*

- Theo Nguyễn Kim Thản định nghĩa thì: *"từ là đơn vị cơ bản của ngôn ngữ, có thể tách khỏi các đơn vị khác của lời nói để vận dụng một cách độc lập và là một khối hoàn chỉnh về ý nghĩa (từ vựng hay ngữ pháp) và cấu tạo".*

❖ **Hình vị tiếng Việt**

Theo nghiên cứu của bài báo [5] thì tiếng là đơn vị cơ bản trong tiếng Việt vì nó có thể nhận diện tương đối dễ dàng bởi người bản ngữ cũng như nhận diện một cách tự động bởi máy tính. Xét về mặt kỹ thuật trên máy tính, có thể sắp xếp các tiếng một cách dễ dàng do số lượng cũng như chiều dài của các tiếng là nhỏ.[1]

Ngoài ra, tiếng còn được xem là từ chính tả. Tuy nhiên, nếu xét trên các tiêu chí của ngôn ngữ học thì tiếng không thể được xem là một từ thực sự vì chưa thỏa tiêu chí về nội dung (phải có ý nghĩa hoàn chỉnh). Vì vậy, dựa theo quan điểm của tác giả Đinh Điền của bài nghiên cứu [6] thì xem tiếng chỉ là hình vị tiếng Việt".

Hình vị tiếng Việt ở đây được hiểu là hình vị như trong ngôn ngữ học đại cương và còn phải xét thêm yếu tố hình tố, là yếu tố thuần túy hình thức biểu diễn những kiểu quan hệ bên trong giữa các thành tố trong từ được gọi là những "thanh hình vị" hay "á hình vị". Như vậy, trong tiếng Việt sẽ có ba loại hình vị như sau:

- *Hình vị gốc*: là những nguyên tố, đơn vị nhỏ nhất, có nghĩa, chúng có thể là hình vị thực (là những từ vựng) hay hình vị hư (ngữ pháp), chúng có thể đứng độc lập hay bị ràng buộc.

- Tha hình vị: vốn cũng là hình vị gốc, nhưng vì mối tương quan với các thành tố khác trong từ mà chúng biến đổi đi về âm, nghĩa, …Tha hình vị bao gồm: tha hình vị láy nghĩa (*giá cả, hỏi han, tuổi tác, nhà cửa, yêu thương, …*), tha hình vị láy âm (*chúm chím, đo đỏ, lé đé, đủng đỉnh, …*), tha hình vị định tính (*xanh lè, tối ôm, cười khẩy, …*), tha hình vị tựa phụ tố(*giáo viên, hiện đại hóa, tân tổng thống, ….*).

- Á hình vị: là những âm chiết đoạn ngữ âm được phần xuất một cách tiêu cực, thuần túy dựa vào hình thức, không rõ nghĩa, song có giá trị khu biệt làm các chức năng cấu tạo từ. (*dưa hấu, dưa gang, bí ử, đậu nành, bồ nông, …)*

❖ **Từ tiếng Việt**

Theo nghiên cứu [6] thì "*từ tiếng Việt được cấu tạo bởi những hình vị tiếng Việt".*

Từ trong tiếng Việt bao gồm: từ đơn (một âm tiết), từ ghép (đa âm tiết) và từ láy (từ có sự lặp lại về tiếng âm thanh).

Xuất phát từ nhu cầu xử lý tự động ngữ liệu tiếng Việt bằng máy tính, tác giả [6] đã đề nghị cách thức hình thức hóa các quan niệm về hình vị tiếng Việt và từ tiếng Việt nói trên như sau:

[1] Trong tiếng Việt, có khoảng 10000 tiếng các loại và chiều dài của mỗi tiếng cũng được giới hạn là 7 ký tự (nghiêng là tiếng dài nhất với 7 ký tự).

- Do hình vị tiếng Việt cũng chính là từ chính tả nên việc hình thức hóa rất đơn giản. Trong ngữ liệu tiếng Việt thì đơn vị cơ bản được lưu trữ cũng chính là từ chính tả này. Tuy nhiên, nếu chỉ lưu trong kho ngữ liệu thì sẽ rất hạn chế và không thể khai thác hiệu quả vốn có của nó.

- Để lưu trữ thông tin về ranh giới từ tiếng Việt sẽ sử dụng khái niệm từ từ điển học được nêu trong nghiên cứu [6]. Từ từ điển học ở đây được định nghĩa là "*những đơn vị mà căn cứ vào đặc điểm ý nghĩa của nó phải xếp riêng trong từ điển và có đánh dấu đây là đơn vị từ của ngôn ngữ".*

❖ **Các phương pháp tách từ cho bài toán phân loại văn bản**

Đối với các văn bản ngôn ngữ tiếng Anh việc tách từ được thực hiện khá đơn giản dựa vào các ký tự phân cách như khoảng trắng, ký tự tab, các dấu câu, dấu ngoặc, … Ngược lại, đối với văn bản tiếng Việt thì khoảng trắng ngoài việc có thể dùng để ngăn cách các từ với nhau thì còn được dùng để ngăn cách các âm tiết của một từ ghép. Ví dụ với câu "Học sinh đi học" phải được tách thành "Học_sinh/ đi_học". Khoảng trắng thứ nhất và thứ ba được dùng để ngăn cách các âm tiết của một từ và khoảng trắng thứ hai được dùng để ngăn cách hai từ với nhau. Điều này gây khó khăn cho quá trình tách từ.

Từ trong tiếng Việt, ngoài từ đơn (một âm tiết), còn có từ ghép (đa âm tiết), chính vì không thể sử dụng khoảng trắng để xác định được ranh giới của các từ. Những âm tiết được kết hợp để tạo thành các từ khác nhau tùy thuộc vào ngữ cảnh của văn bản. Để nhận dạng đúng ranh giới của các từ (tách từ) phục vụ cho các bài toán phân tích dữ liệu văn bản, các nhà khoa học đã đề xuất nhiều phương pháp tách từ. Dựa trên đặc điểm của "từ" kết hợp với các cách tiếp cận khác nhau, các phương pháp tách từ có thể được chia thành ba nhóm chính: dựa trên từ điển (dictionary-based), dựa trên thống kê (statistic-based) và phương pháp lai (hybrid).

- ***Tách từ dựa trên từ điển (dictionary-based)***

Ý tưởng chính cho phương pháp tách từ dựa trên từ điển là từ một cơ sở dữ liệu từ điển sẵn có, thực hiện so khớp từng âm tiết trong văn bản với các từ có trong từ điển. Tùy vào cách so khớp mà ta có được các phương pháp khác nhau như: so khớp từ dài nhất, so khớp từ ngắn nhất, so khớp chồng lắp và so khớp cực đại. Độ chính xác của phương pháp tách từ dựa trên từ điển phụ thuộc rất lớn vào kích thước và độ bao phủ của từ điển được xây dựng, đảm bảo được độ chính xác cao với các từ quen thuộc. Với đặc điểm là không cần các bước huấn luyện vì vậy thời gian xử lý của phương pháp này tương đối nhanh, đơn giản, dễ hiểu và hiệu quả với các văn bản có cấu trúc tốt. Tuy nhiên phương pháp này sẽ khó có thể xử lý được các tình huống nhập nhằng cũng như xử lý tình huống từ mới xuất hiện nhưng không tồn tại trong từ điển. Hai phương pháp thường được sử dụng của tách

từ dựa trên từ điển là phương pháp so khớp cực đại và phương pháp so khớp từ dài nhất.

- Phương pháp so khớp từ dài nhất (Longest Matching) được đề xuất trong bài báo [7] bởi Surapant Meknavin và cộng sự vào năm 1997. Đây là kỹ thuật cơ bản trong các phương pháp tách từ dựa trên từ điển. Ý tưởng chính là tìm kiếm chuỗi ký tự dài nhất có thể trong văn bản mà khớp với từ điển, ưu tiên từ dài nhất trước khi xét các từ ngắn hơn. Với mỗi câu, duyệt từ trái qua phải các âm tiết trong câu, kiểm tra xem có nhóm các âm tiết nào có tồn tại trong từ điển hay không. Chuỗi dài nhất các âm tiết được xác định là từ sẽ được chọn ra. Tiếp tục thực hiện lặp lại việc so khớp cho đến hết câu. Ví dụ "Học sinh học sinh vật học" xét từ trái qua phải, âm tiết đầu tiên là từ "học", "học" cũng có thể là một từ đơn nhưng với câu trên từ "học" có thể kết hợp với âm tiết "sinh" để tạo nên từ ghép "học sinh", ta được từ đầu tiên là "học sinh". Tiếp tục xét các âm tiết còn lại cho đến khi hết câu ta có thu được các từ sau: "học sinh", "học sinh", "vật", "học". Với ví dụ này, phương pháp so khớp từ dài nhất không đem lại được kết quả như mong muốn.
- Phương pháp so khớp cực đại (Maximum Matching Method) được nêu ra trong bài báo [8] do Chih-Hao Tsai đề xuất vào năm 1996. Đây là một phương pháp tương tự Longest Matching, nhưng tập trung vào việc tìm kiếm một phân đoạn tối ưu nhất dựa trên số lượng từ tối thiểu hoặc tối đa từ trong câu. Ứng với mỗi câu dữ liệu đầu vào, tìm tất cả các trường hợp mà các âm tiết có thể kết hợp lại để tạo nên các từ có nghĩa. Ứng với mỗi loại ngôn ngữ khác nhau thì sự lựa chọn các nhóm âm tiết này có thể khác nhau. Phương pháp này là so khớp toàn diện cho một câu thay vì so khớp cục bộ âm tiết đang được xét. Với ví dụ "Học sinh học sinh vật học" thì các trường hợp kết hợp của các âm tiết có thể có "học sinh", "học", "sinh vật học", từ được tách trong câu sẽ chính xác hơn so với phương pháp so khớp từ dài nhất.

- ***Tách từ dựa trên thống kê (statistic-based)***

Mô hình ngôn ngữ (Language Model – LM) là một phương pháp dựa trên thống kê, được sử dụng rộng rãi trong bài toán xử lý ngôn ngữ tự nhiên, bao gồm cả tách từ. Mô hình này đánh giá xác suất xuất hiện của một từ hoặc ký tự trong văn bản dựa trên các mẫu thống kê thu thập từ dữ liệu huấn luyện. Ý tưởng chính cho phương pháp này là tận dụng đồng xuất hiện của các từ hoặc ký tự để dự đoán ranh giới từ, thường được xây dựng dựa trên việc thu thập thống kê số lần xuất hiện hoặc đồng thời xuất hiện của các từ trong một tập hợp văn bản.

Mô hình LM được đề xuất và phát triển bởi Jelinek và cộng sự (1991) trong [9] trong đó mô hình ngôn ngữ chủ yếu tập trung vào việc tối ưu hóa xác suất chuỗi từ dựa trên dữ liệu huấn luyện lớn.

Mô hình n-grams được áp dụng để biểu diễn xác suất của một chuỗi từ hoặc ký tự. Công thức tính xác suất cho chuỗi từ $W = w1, w2, w3, ...$ trong văn bản được tính toán như sau:

$$P(W) = P(w_1, w_2, w_3, \dots, w_n) \approx \prod_{i=1}^{n} P(w_i \,|w_{i-n+1}, \dots, w_{i-1}) \quad (1)$$

Với n là độ dài của ngữ cảnh được xem xét:

$n = 1\ (unigram): P(W) \approx \prod_{i=1}^{n} P(w_i)$

$n = 2\ (bigram): P(W) \approx \prod_{i=1}^{n} P(w_i \,|w_{i-1})$

$n = 3\ (trigram): P(W) \approx \prod_{i=1}^{n} P(w_i \,|w_{i-2}, w_{i-1})$

Xác suất có điều kiện của từ w_i xảy ra khi biết rằng từ trước đó là w_{i-1} được tính bằng công thức sau:

$$P(w_i|w_{i-1}) = \frac{freq(w_{i-1}, w_i)}{freq(w_{i-1})} \quad (2)$$

Trong đó,

$freq(w_{i-1}, w_i)$ là số lần cặp từ w_{i-1} và w_i xuất hiện liên tiếp trong tập huấn luyện.

$freq(w_{i-1})$ là số lần từ w_{i-1} xuất hiện trong dữ liệu huấn luyện.

Trong một số trường hợp khi các cặp từ hoặc từ không xuất hiện trong dữ liệu huấn luyện, xác suất của chúng sẽ bằng 0. Để khắc phục, các phương pháp làm trơn như Add-One, Good Turing được sử dụng.

Công thức xác định cách phân đoạn g^* sao cho xác suất của chuỗi từ được phân đoạn đạt giá trị cao nhất:

$$g^* = \arg max_g P(g|w) \quad (3)$$

Trong đó:

g là một cách phân đoạn cụ thể.

$P(g|w)$ là xác suất của cách phân đoạn g cho chuỗi ký tự w.

Áp dụng định lý Bayes ta thu được công thức sau:

$$P(g|w) = \frac{P(w|g)P(g)}{P(w)} \quad (4)$$

Vì $P(w)$ không phụ thuộc vào cách phân đoạn g, nên có thể bỏ qua quá trình tối ưu. Vì vậy giá trị g^* được tính công thức sau:

$$g^* = \arg max_g P(w|g)P(g) \quad (5)$$

Giả định $P(w|g)$ được ước lượng bằng mô hình ngôn ngữ n-grams và $P(g)$ có thể được coi là đồng đều hoặc ước lượng dựa trên dữ liệu huấn luyện.

- ***Tách từ dựa trên phương pháp lai (hybrid)***

Phương pháp tách từ dựa trên tiếp cận lai là sự kết hợp các phương pháp khác nhau, đặc biệt là phương pháp từ điển và phương pháp thống kê. Mục tiêu của phương pháp này là tận dụng các điểm mạnh của từng phương pháp để đạt được hiệu quả cao hơn trong việc tách từ, đặc biệt là trong các ngôn ngữ có cấu trúc phức tạp như tiếng Việt. Tách từ dựa trên tiếp cận lai cho kết quả chính xác cao hơn và linh hoạt hơn so với việc chỉ sử dụng một phương pháp duy nhất.

Một số phương pháp kết hợp giữa phương pháp tiếp cận từ điển và phương pháp tiếp cận thống kê được nêu ra như: kết hợp giữa mô hình so khớp cực đại và máy học véc-tơ hỗ trợ (SVMs) được nêu trong tài liệu [10], kết hợp giữa mô hình so khớp cực đại và ngôn ngữ mô hình n-grams được trình bày trong bài báo [11]. Nhóm tác giả bài báo [11] cũng đã đề xuất phương pháp tách từ tiếng Việt dựa trên sự kết hợp giữa phương pháp tiếp cận dựa trên từ điển và phương pháp thống kê với thư viện vnTokenizer cung cấp phương pháp tách từ tiếng Việt dựa trên kỹ thuật lai (từ điển, automat hữu hạn trạng thái, biểu thức chính quy và so khớp từ dài nhất). Bên cạnh đó, trong bài báo [12] cũng nêu ra thư viện underthesea[2] sử dụng mô hình học máy (CRF) kết hợp với từ điển để tách từ. Phương pháp này giúp xử lý chính xác các từ có sẵn trong từ điển và đồng thời giúp phân đoạn các từ không có trong từ điển bằng cách sử dụng các kỹ thuật thống kê.

3.2.2. Định lý Bayes và bộ phân lớp Naïve Bayes

Định lý Bayes là một định lý của toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan đến B xảy ra.

Ý tưởng cơ bản cho của cách tiếp cận Naive Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại.

Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện B khi biết A như sau:

$$P(B|A) = \frac{P(A|B)\ P(B)}{P(A)} \quad (6)$$

Trong đó:

P(B|A): xác suất xảy ra B với điều kiện A đã xảy ra (xác suất hậu nghiệm).

P(A|B): xác suất xảy ra A với điều kiện B đã xảy ra (xác suất có điều kiện).

P(B): xác suất tiên nghiệm của B, tức là xác suất xảy ra B trước khi có thông tin về A.

P(A): xác suất xảy ra sự kiện A, được tính bằng tổng xác suất của A với tất cả các khả năng của B

$$P(A) = \sum_i P(A|B_i).P(B_i) \quad (7)$$

[2] https://underthesea.readthedocs.io/en/latest/readme.html

Giả sử phân chia 1 sự kiện X thành n thành phần độc lập khác nhau X_1, X_2, X_3, … X_n. Từ đó, có thể tính được:

$$\mathrm{P(x,y)} = \mathrm{P}(x_1 \cap x_2 \cap x_3 \ldots \cap x_n|\mathrm{y}) = \mathrm{P}(x_1|\mathrm{y})\mathrm{P}(x_2|\mathrm{y})\mathrm{P}(x_3|\mathrm{y}) \ldots \mathrm{P}(x_n|\mathrm{y}) \quad (8)$$

Do đó ta có: $P(y|x) \propto P(y) \sum_{i=1}^{n} P(x_i|y), với \propto là phép tỉ lệ thuận.$

Trên thực tế thì rất ít khi tìm được dữ liệu mà các thành phần hoàn toàn độc lập với nhau. Tuy nhiên với giả thiết trên có thể giúp tính toán trở nên đơn giản, huấn luyện dữ liệu nhanh và đem lại hiệu quả bất ngờ với các lớp bài toán nhất định.

3.2.3. Thuật toán Naïve Bayes

Phương pháp Naïve Bayes là một tập hợp các mô hình học có giám sát dựa trên việc áp dụng định lý Bayes với giả định ngây thơ (Naïve) về sự độc lập có điều kiện giữa mọi cặp đặc trưng cho giá trị của biến lớp. Định lý Bayes đã phát biểu mối quan hệ giữa biến lớp cho trước *y* với vector đặc trưng x_1 đến x_n theo công thức (9):

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)} \quad (9)$$

Với giả định rằng, các đặc trưng là hoàn toàn độc lập có điều kiện với nhau khi biết giá trị của *y*:

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y) \quad (10)$$

Khi đó, với tất cả *i*, công thức (9) được đơn giản hóa thành

$$P(\, y \mid x_1, x_2, \ldots, x_n\,) = \frac{\sum_{i=1}^{n} P(x_i|y)P(y)}{P(x_1\,, x_2, \ldots, x_n)} \quad (11)$$

Vì $P(x_1\,, x_2, \ldots, x_n)$ là hằng số cho đầu vào và độc lập vì vậy có thể sử dụng quy tắc phân loại sau

$$P(\, y \mid x_1, x_2, \ldots, x_n\,) \propto \prod_{i=1}^{n} P(x_i|y)P(y) \quad (12)$$

$$=> \hat{y} = \arg\max P(y). \prod_{i=1}^{n} P(x_i \mid y) \quad (13)$$

Từ công thức (13), 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 giá trị của $P(x_i|y)$, sau đó là tần suất tương đối của lớp *y* trong huấn luyện. Các bộ phân loại Naïve Bayes khác nhau chủ yếu bởi các giả định được đưa ra liên quan đến phân phối $P(x_i|y)$.

Naïve Bayes yêu cầu một lượng nhỏ dữ liệu huấn luyện để có thể ước tính các thông số cần thiết. Naïve Bayes học và phân loại dữ liệu rất nhanh so với các phương pháp phức tạp khác. Việc tách và phân bổ 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. Mặc khác, Naïve Bayes được biết đến như một bộ phân loại tốt, nhưng nó lại là một bộ ước lượng kém, vì vậy kết quả xác suất dự đoán không được coi trọng. Việc tính $P(x_i|y)$ phụ thuộc vào loại dữ liệu huấn luyện. [5] [6]

Có ba mô hình phân loại được sử dụng phổ biến hiện nay là: *Gaussian Naïve Bayes*, *Multinomial Naïve Bayes* và *Bernoulli Naïve Bayes.*

Mô hình Gaussian Naïve Bayes được sử dụng phổ biến trong loại dữ liệu mà các thành phần là các biến liên tục. Với mỗi chiều dữ liệu *i* và lớp *y*, lớp 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 (14):

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{14}$$

Trong đó các tham số σ_y và μ_y được ước tính bằng cách sử dụng maximum likelihood.

Mô hình Multinomial Naïve Bayes khai triển thuật toán Naïve Bayes cho dữ liệu theo phân bố đa thức và là một trong hai biến thể của Naïve Bayes cổ điển được sử dụng trong phân loại văn bản, trong đó dữ liệu thường được biểu diễn dưới dạng đếm số lượng vector đặc trưng với các phần tử nguyên có giá trị là tần suất xuất hiện của từ đó trong tài liệu. Phân phối được tham số hóa bởi vector $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ cho mỗi lớp y, trong đó *n* là số đặc trưng *(feature)* 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 θ_y được ước tính bằng một phiên bản làm mịn của xác suất tối đa, tức là đếm tần số tương đối theo công thức (15):

$$\widehat{\theta_{yı}} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{15}$$

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 $\alpha \geq 0$ 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 $\alpha = 1$ được gọi là làm mịn *Laplace,* trong khi $\alpha < 1$ được gọi là làm mịn *Lidstone.*

Mô hình Bernoulli Naïve Bayes khai triển các thuật toán huấn luyện và phân loại Naïve Bayes cho dữ liệu được phân bố theo các phân bố Bernoulli đa biến. Tức là, có thể có nhiều chiều đặc trưng nhưng mỗi 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 từ tương ứng có trong tài liệu và p nếu từ không có trong tài liệu). Do đó, lớp này yêu cầu các mẫu phải được biểu diễn dưới dạng vector đặc trưng có giá trị nhị phân. Luật quyết định cho mô hình *Bernoulli Naïve Bayes* dựa trên công thức (16):

$$P(x_i \mid y) = P(i \mid y)\, x_i + (1 - P(i \mid y))\,(1 - x_i) \tag{16}$$

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

3.2.4. Phân loại văn bản tiếng Việt với thuật toán Naïve Bayes

Thuật toán Naïve Bayes là kỹ thuật phổ biến trong học máy có giám sát. Ý tưởng chính của kỹ thuật này dựa vào xác suất có điều kiện giữa từ và cụm từ và nhãn phân loại để dự đoán văn bản mới cần phân loại thuộc vào chủ đề, lớp nào. Naïve Bayes được ứng dụng nhiều trong giải quyết các bài toán phân loại văn bản, xây dựng các bộ lọc thư rác tự động, hay trong các bài toán khai phá quan điểm bởi tính dễ hiểu, dễ triển khai cũng như cho độ chính xác tốt.

Ý tưởng cơ bản cho việc tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa các từ đặc trưng và nhãn để dự đoán xác suất nhãn của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các đặc trưng trong văn bản đều hoàn toàn độc lập với nhau. Giả định đó làm cho việc tính toán Naïve Bayes hiểu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các đặc trưng để đưa ra phán đoán nhãn. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng, …

❖ **Áp dụng cài đặt cho bài toán phân loại văn bản:**

Thuật toán gồm hai giai đoạn huấn luyện và phân lớp:

Huấn luyện: tính $P(C_i)$ và $P(x_k|C_i)$

Đầu vào:

- Ma trận TF-IDF: Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
- Tập nhãn/ lớp/ chủ đề $C = \{C_1, C_2, \dots, C_m\}$ cho từng vector đặc trưng của tập huấn luyện.

Đầu ra:

- Các giá trị xác suất tiên nghiệm $P(C_i)$ cho mỗi lớp C_i và Xác suất có điều kiện $P(x_k|C_i)$ cho mỗi từ x_k trong mỗi lớp C_i.

Công thức tính $P(C_i)$ đã làm mịn Laplace

$$P(C_i) = \frac{|docs_i| + 1}{|total\ docs| + m} \quad (17)$$

Trong đó

$|docs_i|$: số văn bản của tập huấn luyện thuộc phân lớp C_i.

$|total\ docs|$: tổng số văn bản trong tập huấn luyện.

m: số phân lớp.

Khởi tại mảng A, B có kích thước m.

Duyệt qua các văn bản trong tập dữ liệu, đếm số văn bản trong mỗi phân lớp lưu vào A.

Tính xác suất cho từng phân lớp theo công thức trên và lưu vào mảng B.

Công thức tính $P(x_k|C_i)$ đã làm mịn Laplace:

$$P(x_k \mid C_i) = \frac{\sum_{j \in docs_i} tfidf(x_k, j) + 1}{\sum_{j \in docs_i} TF - IDF\ tổng(j) + \ |V|} \quad (18)$$

Trong đó,

$P(x_k \mid C_i)$: xác suất có điều kiện cửa từ x_k thuộc vào lớp C_i.

$\sum_{j \in docs_i} tfidf(x_k, j)$: tổng giá trị TF-IDF của từ x_k trong tất cả các văn bản thuộc lớp C_i.

$\sum_{j \in docs_i} TF - IDF\ tổng(j)$: tổng tất cả giá trị TF-IDF của mọi từ trong các văn bản thuộc lớp C_i. Biểu thị tổng trọng số của từ vựng trong lớp C_i, được sử dụng để chuẩn hóa.

$|V|$: Kích thước từ vựng (tổng số từ duy nhất trong từ điển).

Khởi tạo mảng 2 chiều C, với chiều 1 có kích thước là m (số lượng phân nhãn), chiều 2 có kích thước là $|V|$ (số từ trong từ điển).

Duyệt qua các văn bản trong tập dữ liệu, tiến hành thống kê các chỉ số cần thiết để tính xác suất $P(x_k \mid C_i)$ theo công thức trên và lưu vào mảng $C[m][|V|]$. Trong đó, $C[i][k]$ là giá trị $P(x_k \mid C_i)$, tức là xác suất của từ x_k thuộc vào lớp C_i.

Phân lớp:

Đầu vào:

- Vector đặc trưng TF-IDF của văn bản cần phân lớp.
- Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

Đầu ra:

- Nhãn/lớp của văn bản cần phân loại.

Công thức tính xác suất thuộc phân lớp *i* khi biết trước mẫu X (xác suất hậu nghiệm).

$$P(C_i|X) = P(C_i) \prod_{k=1}^{n} P(x_k \ |C_i) \quad (19)$$

Công thức (19) là công thức chuẩn của Naïve Bayes. Tuy nhiên, để tránh underflow (xảy ra khi nhân nhiều xác suất nhỏ) thì nên sử dụng log xác suất:

$$\log P(C_i|X) = \log P(C_i) + \sum_{k=1}^{N} \log P(x_k|C_i) \quad (20)$$

Trong đó,

$X = [x_1, x_2, \dots, x_N]$: vector TF-IDF của văn bản cần phân lớp.

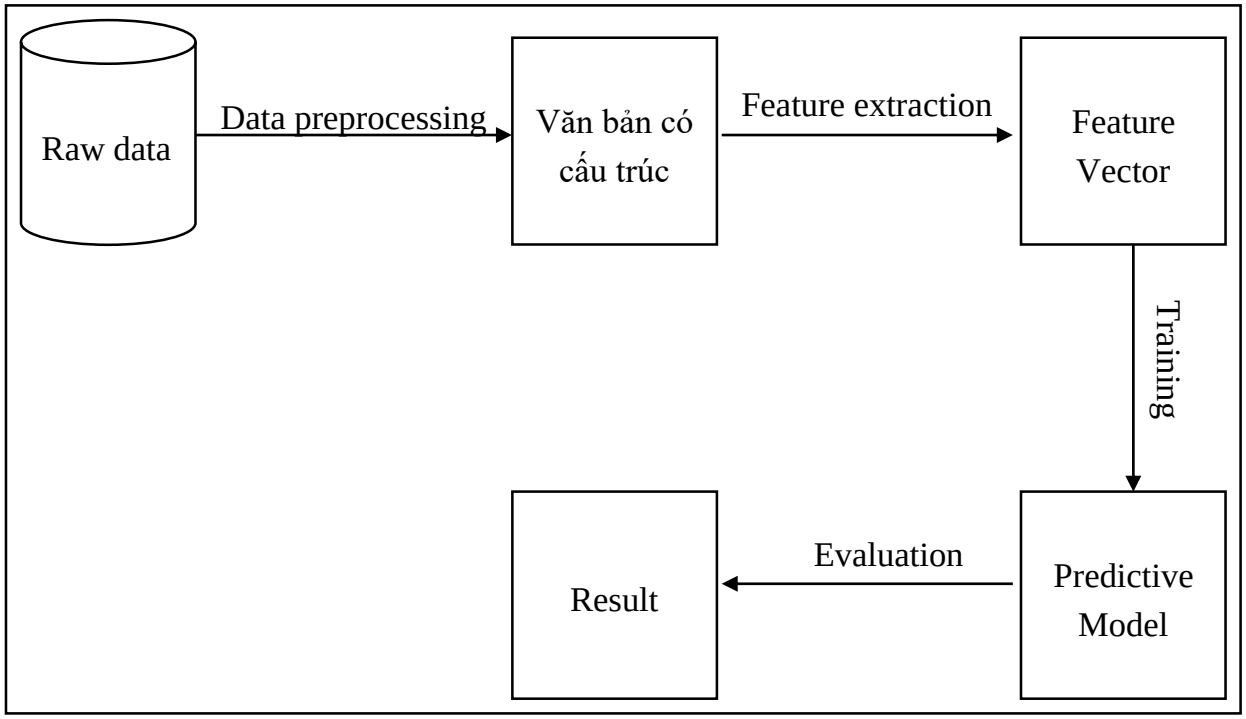
$P(C_i)$: xác suất tiên nghiệm của lớp C_i.

$P(x_k \mid C_i)$: xác suất có điều kiện của từ x_k thuộc vào lớp C_i.

Dựa vào vector đặc trưng của văn bản cần phân lớp, áp dụng công thức (20) tính xác suất thuộc từng phân lớp cho văn bản và chọn ra lớp C_i có xác suất hậu nghiệm cao nhất với công thức $\hat{C} = arg \max_{C_i} P(C_i|X)$.

3.2.5. Quá trình thực hiện bài toán phân loại văn bản

Quy trình cài đặt hệ thống phân loại tự động cho văn bản tiếng Việt cần thực hiện qua các bước được mô tả tổng quan qua mô như ***Hình 3.2*** bên dưới:



Hình 3.2: Mô hình thực hiện bài toán phân loại văn bản

❖ **Bước 1: Raw data (Thu thập dữ liệu)**

Tập tin dữ liệu là yếu tố quan trọng nhất cho tất cả các bài toán máy học đặc biệt đối với bài toán phân loại văn bản tiếng Việt thì công đoạn chuẩn bị tập tin dữ liệu dataset là càng được chú trọng nhiều hơn bởi cấu trúc phức tạp của các văn bản tiếng Việt.

Dữ liệu này được sưu tầm từ các bài báo, các tập tin văn bản trên mạng internet. Khi thu thập các tập tin dữ liệu cần xác định rõ chủ đề của tập tin văn bản đó. Đối với đề tài

này thì bộ tập tin dữ liệu dataset được tham khảo từ [15] với 10 chủ đề (Chính trị Xã hội, Đời sống, Khoa học, Kinh doanh, Pháp luật, Sức khỏe, Thế giới, Thể thao, Văn hóa, Vi tính). Dữ liệu đã được tác giả [15] thu thập là nội dung các bài báo có đã có sẵn chủ đề từ các website của các trang báo điện tử như trang báo VnExpress[3], báo Thanh niên[4], báo Tuổi trẻ[5] và báo Người lao động[6] được mô tả chi tiết số lượng tập tin cho dành cho tập huấn luyện và tập kiểm thử như ***Bảng 3.1*** bên dưới:

Bảng 3.1: Bảng mô tả số lượng tập dữ liệu

STT	Tên chủ đề/ nhãn	Số lượng tập tin train	Số lượng tập tin test
1	Chính trị Xã hội	5.219	7.567
2	Đời sống	3.159	2.036
3	Khoa học	1.820	2.096
4	Kinh doanh	2.552	5.276
5	Pháp luật	3.868	3.788
6	Sức khỏe	3.384	5.417
7	Thế giới	2.898	6.716
8	Thể thao	5.298	6.667
9	Văn hóa	3.080	6.250
10	Vi tính	2.481	4.560
TỔNG		33.759	50.373

❖ **Bước 2: Data preprocessing (Tiền xử lý dữ liệu)**

Tiền xử lý dữ liệu là một bước quan trọng và không thể thiếu khi làm việc với dữ liệu cho việc huấn luyện mô hình học máy. Công việc này là quá trình chuẩn hóa và làm sạch dữ liệu, loại bỏ các thành phần không có ý nghĩa cho việc phân loại văn bản sau này, biến đổi một văn bản thô không có cấu trúc thành một văn bản có cấu trúc. Tiền xử lý dữ liệu giúp giảm số chiều đặc trưng cho các vector khi đưa vào huấn luyện và tăng thời gian xử lý hơn cho mô hình.

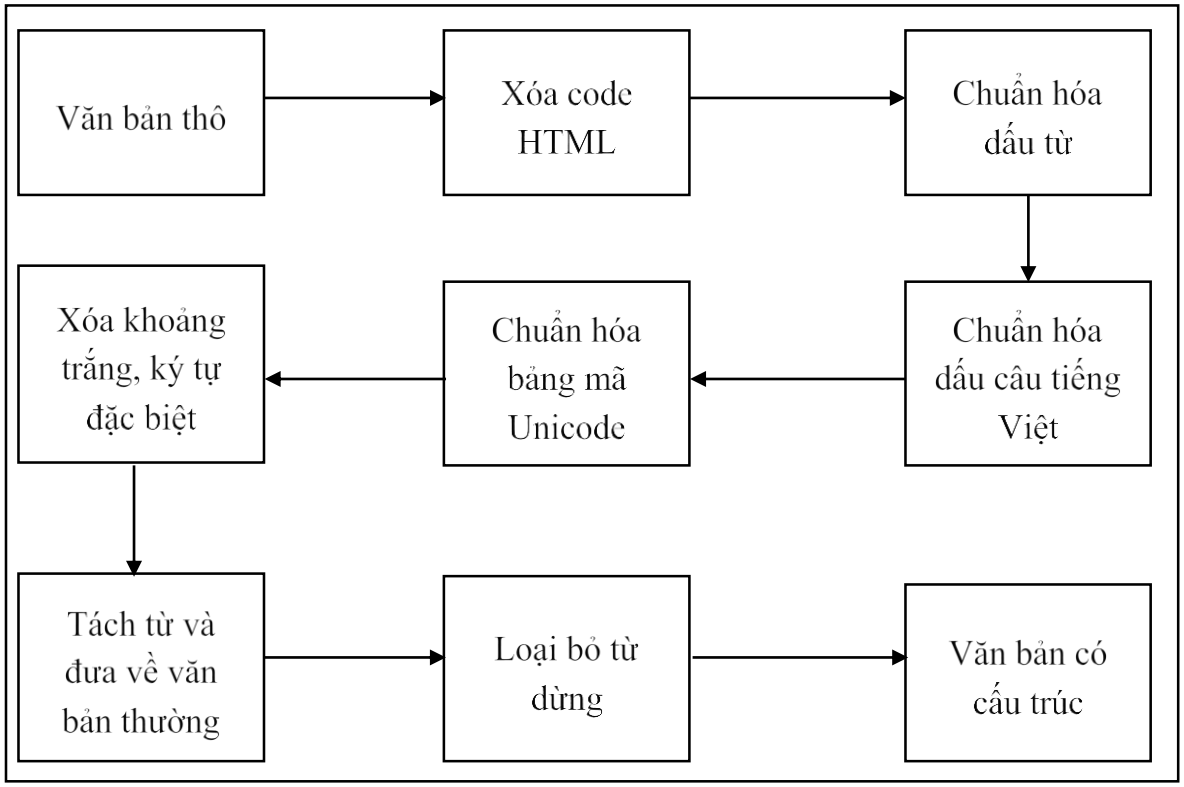
Quá trình tiền xử lý dữ liệu tiếng Việt cho bài toán phân loại văn bản thường bao gồm các công việc như ***Hình 3.3*** bên dưới:

[3] https://vnexpress.net/

[4] https://thanhnien.vn/

[5] https://tuoitre.vn/

[6] https://nld.com.vn/



Hình 3.3: Hình mô hình tiền xử lý văn bản tiếng Việt

- Xóa code HTML: Dữ liệu thu thập từ các website đôi khi vẫn sẽ có các đoạn mã code HTML lẫn lộn trong các văn bản. Các đoạn mã này không có tác dụng cho việc phân loại tìm nhãn chủ đề mà còn làm cho kết quả phân loại bị ảnh hưởng và không đạt được kết quả tối ưu.

Cách làm: sử dụng biểu thức chính quy `re.sub(r'<[^>]*>', '', txt)` để loại bỏ các thẻ HTML ra khỏi văn bản.

- Chuẩn hóa dấu từ: Chuẩn hóa cách đặt dấu thanh trong một từ tiếng Việt.

Cách làm:

- Tách từ thành danh sách các ký tự.
- Xác định vị trí dấu thanh và các nguyên âm.
- Điều chỉnh dấu thanh về đúng vị trí theo quy tắc tiếng Việt.
- Nếu từ không hợp lệ, trả về từ gốc.

- Chuẩn hóa dấu câu tiếng Việt: Tiến hành chuẩn hóa dấu thanh cho toàn bộ câu tiếng Việt.

Cách làm:

- Chuyển câu thành văn bản thường.
- Tách câu thành các từ.
- Chuẩn hóa dấu từ dựa vào hàm chuẩn hóa dấu từ phía trên.
- Kết hợp các từ thành câu đã được chuẩn hóa.

- Chuẩn hóa bảng mã Unicode: Hiện nay có hai bộ mã Unicode là bộ mã Unicode được dựng sẵn và bộ mã Unicode tổ hợp. Về mặt hình thức thì hai bộ mã này

không có gì khác nhau nhưng là nó lại không giống nhau. Vì vậy cần đưa chúng về cùng một bộ mã Unicode dựng sẵn.

Cách làm:

- Tìm các ký tự trong bảng ánh xạ được xây dựng sẵn.
- Dựa trên từ điển ánh xạ, thay thế ký tự khớp trong chuỗi đầu vào.
- Trả về chuỗi văn bản đã chuyển đổi.

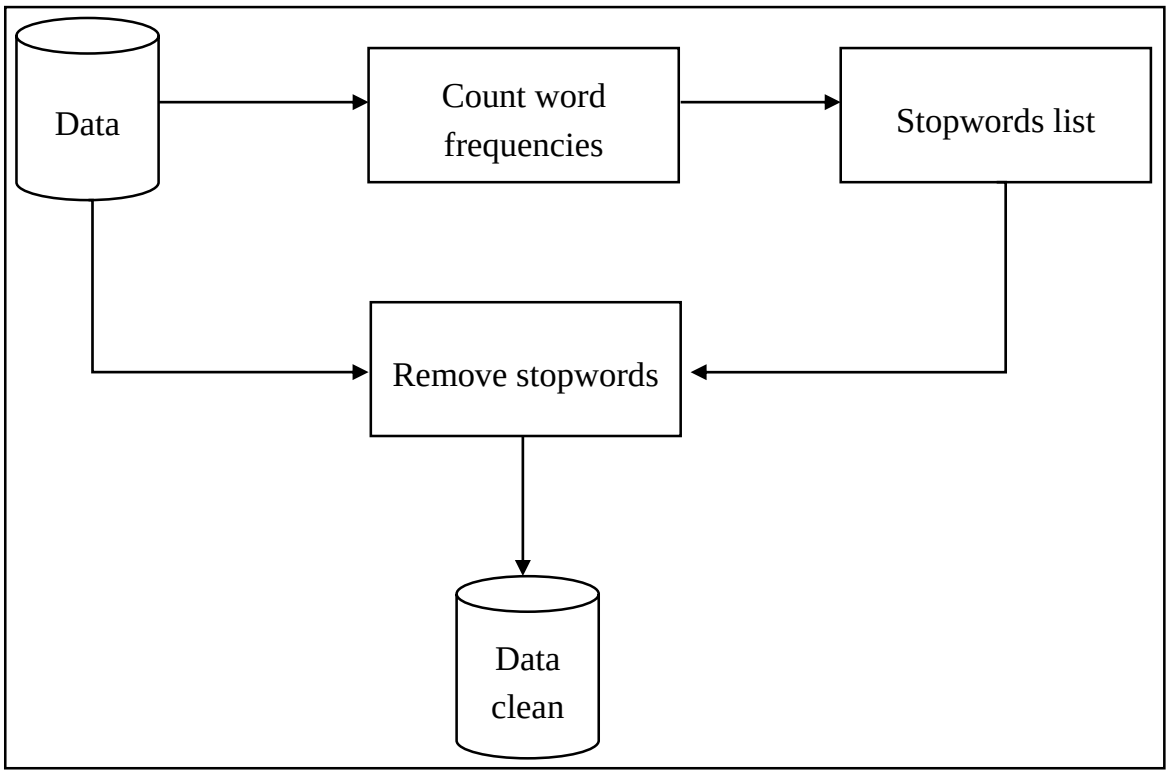
- Xóa khoảng trắng và các ký tự đặc biệt: Các dấu ngắt câu, số đếm, các ký tự đặc biệt khách không giúp cho công việc phân loại và huấn luyện mô hình. Ngoài ra, công việc này còn giúp tăng tốc độ học và xử lý.

Cách làm:

- Sử dụng các biểu thức chính quy để thực hiện xóa và loại bỏ các khoảng trắng thừa.

- Tách từ và đưa về văn bản thường: Sử dụng thư viện underthesea của phương pháp lai để tiến hành tách từ như được trình bày phía trên.
- Loại các bỏ từ dừng (stopwords): Là một bước quan trọng trong xử lý ngôn ngữ tự nhiên. Mục đích chính cho công việc này là loại bỏ các từ không có ý nghĩa để tập trung vào các từ quan trọng như được trình bày phía trên. Bên cạnh đó, việc xóa các từ dừng còn giảm được kích thước dữ liệu và tăng tốc độ xử lý cho các quá trình phân tích TF-IDF phía sau, giúp máy học nhanh hơn và tiết kiệm được tài nguyên, giảm tính dư thừa và tăng độ chính xác cho máy học.

Quy trình thực hiện xóa các từ dừng được mô tả chi tiết như ***Hình 3.4*** bên dưới:



Hình 3.4: Mô hình quy trình thực hiện xóa các từ dừng

- Data: dữ liệu đầu vào cần được loại bỏ các từ dừng.

- Count word frequencies: đại diện cho hàm đếm số lần xuất hiện của từ trong văn bản sau đó tiến hành sắp xếp các từ đó theo thứ tự giảm dần về độ xuất hiện trong toàn bộ dữ liệu.
- Stopwords list: danh sách các từ dừng được lọc ra từ dữ liệu thông qua hàm count_word_frequencies.
- Remove stopwords: là hàm xóa các từ dừng trong tài liệu. Tiến hành lặp qua tất cả các từ trong dữ liệu so sánh với các từ có trong danh sách từ dừng (stopwords list) tiến hành xóa bỏ khỏi dữ liệu và ghép các từ còn lại trong câu lại với nhau.
- Data clean: dữ liệu sau khi được loại bỏ từ dừng.

❖ **Bước 3: Feature extraction (Trích xuất đặc trưng)**: Là bước quan trọng trong xử lý ngôn ngữ tự nhiên, nơi các đặc trưng của văn bản được chuyển thành dạng số để sử dụng các mô hình máy học.

Tập dữ liệu từ thu được từ tập dữ liệu sau khi được tiến hành tiền xử lý đang ở dạng không đúng với cấu trúc cho học máy, do đó để xử lý cho phân loại bằng các phương pháp máy học cần vector hóa chúng. Trong đề tài này lựa chọn sử dụng kỹ thuật TF-IDF để tiến hành biểu diễn văn bản sang dạng các vector đặc trưng.

TF (Term Frequency – Tần suất từ): biểu diễn tần suất xuất hiện của một từ trong tài liệu được tính theo công thức (20):

$$TF(t_i, d_j) = \frac{Số\ lần\ từ\ t_i\ xuất\ hiện\ trong\ tài\ liệu\ d_j}{Tổng\ số\ từ\ trong\ tài\ liệu\ d_j} = \frac{f_i}{n_j} \quad (20)$$

IDF (Inverse Document Frequency – Tần suất ngược của tài liệu): là tần suất nghịch của một từ trong tập tài liệu. Đo lường tầm quan trọng của từ trên toàn bộ tập tài liệu. Các từ xuất hiện phổ biến trong tài liệu (“và”, “là”, “của”, …) sẽ có giá trị IDF thấp. IDF được tính theo công thức (21):

$$IDF(t) = \log \frac{N}{1 + DF(t)} = \log \frac{N}{f(t_i)} = IDF_{ij} \quad (21)$$

Trong đó,

N: Tổng số tài liệu trong tập ngữ liệu.
$DF(t)$: Số tài liệu chứa từ t.
$f(t_i)$: Số lượng các tài liệu chứa từ t_i.

Công thức tính TF-IDF (Terms Frequency Inverse Document Frequency) được tính theo công thức (22) và (23):

$$TF - IDF(t, d) = TF(t, d) * IDF(t) \quad (22)$$

$$w_{ij} = TF_{ij} * IDF_{ij} \quad (23)$$

Bên cạnh việc sử dụng các công thức tính TF-IDF được nêu phía trên, thư viện scikit-learn hỗ trợ việc tính toán TF-IDF thông qua lớp TfidfVectorizer[7] và TfidfTransformer[8]. Đây là một công cụ mạnh mẽ để chuyển đổi văn bản thô thành biểu diễn TF-IDF. Đối với đề tài đã áp dụng công cụ TfidfVectorizer để thuận tiện cho quá trình cài đặt.

❖ **Bước 4: Training (Huấn luyện)**

Tiến hành huấn luyện mô hình dựa trên thuật toán Naïve Bayes với mô hình Multinomial Naïve Bayes.

Xây dựng các tập dữ liệu huấn luyện/ kiểm thử (train/ test) với bộ dữ liệu dataset qua các nhãn dữ liệu đã thu được ***Bảng 3.2*** bên dưới về các mẫu tập tin và tỉ lệ huấn luyện là 80% dành cho huấn luyện và 20% dành cho kiểm thử trên từng nhãn dữ liệu của bài toán.

Bảng 3.2 Bảng phân bố tập huấn luyện và kiểm thử dùng mô hình Naive Bayes

Nhãn	#Train	#Test	%Train	%Test	#Tổng
Văn hóa	2.427	653	78.79%	21.20%	3.080
Pháp luật	3.110	758	80.40%	19.59%	3.868
Thể thao	4.261	1.037	80.42%	19.57%	5.298
Chính trị xã hội	4.173	1.046	79.95%	20.04%	5.219
Khoa học	1.463	357	80.38%	19.61%	1.820
Sức khỏe	2.708	676	80.02%	19.97%	3.384
Đời sống	2.520	639	79.77%	20.22%	3.159
Thế giới	2.317	581	79.95%	20.04%	2.898
Vi tính	1.999	482	80.57%	19.42%	2.481
Kinh doanh	2.029	523	79.50%	20.49%	2.552

❖ Bước 5: Evaluation (Đánh giá)

Tiến hành đánh giá mô hình thông qua các thông số accuracy, precision, recall và f1-score.

3.3. Giải pháp cài đặt

Ngôn ngữ được lựa chọn sử dụng để cài đặt cho hệ thống là Python. Cần cài đặt Python vào máy tính để hỗ trợ cho ngôn ngữ này. Có thể tùy chọn các phiên bản, tuy nhiên

[7] https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[8] https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

thời điểm cài đặt và hoạt động hệ thống ổn định nhất được hoạt động trên phiên bản Python 3.11.9.

Để lập trình và phát triển hệ thống cần có IDE (môi trường tích hợp dùng để viết code và phát triển ứng dụng). IDE được lựa chọn là Visual Studio Code.

Sau khi cài đặt Python và Visual Studio Code hoàn tất, tiến hành mở thư mục dự án trên Visual Studio Code, mở terminal và cài đặt môi trường ảo để lưu trữ các thư viện cần thiết cho hệ thống bằng lệnh "python3 –m venv myenv", sau đó tiến hành kích hoạt môi trường ảo bằng lệnh "myenv\Scripts\activate". Tại đây, người dùng có thể cài đặt các gói thư viện cần thiết cho hệ thống hoạt động tốt:

- pip[9]==24.2
- regex[10]==2024.7.24
- underthesea[11]==6.8.4
- tqdm[12]==4.66.5
- nltk[13]==3.9.1
- numpy[14]==1.26.4
- pytz[15]==2024.1
- scikit-learn[16]==1.5.1
- sklearn-crfsuite[17]==0.5.0
- pandas[18]==2.2.2

Trong đó:

- Pip: trình quản lý gói tiêu chuẩn cho Python, dùng để cài đặt và quản lý các gói thư viện hoặc gói Python từ PyPI.

- Regex: thư viện này dùng để xử lý các biểu thức chính quy một cách hiệu quả và mạnh mẽ.

[9] https://pypi.org/project/pip/

[10] https://pypi.org/project/regex/

[11] https://pypi.org/project/underthesea/

[12] https://pypi.org/project/tqdm/

[13] https://pypi.org/project/nltk/

[14] https://pypi.org/project/numpy/

[15] https://pypi.org/project/pytz/

[16] https://pypi.org/project/scikit-learn/

[17] https://pypi.org/project/sklearn-crfsuite/

[18] https://pypi.org/project/pandas/

- Underthesea: thư viện này dùng để xử lý ngôn ngữ tự nhiên cho tiếng Việt, hỗ trợ các tính năng như tách từ, tách câu, gán nhãn từ loại và sinh câu tự động.

- Tqdm: thư viện này dùng để hiển thị các thanh tiến trình trực quan khi xử lý dữ liệu lớn.

- NLTK: thư viện này được dùng để xử lý ngôn ngữ tự nhiên mạnh mẽ, cung cấp các công cụ để làm việc với văn bản như phân tách câu, gán nhãn từ loại, tạo cây cú pháp và phân tích ngữ nghĩa, ...

- Numpy: thư viện này dùng để xử lý các mảng đa chiều, cung cấp các chức năng tính toán cho đại số tuyến tính, thống kê, ...

- Pytz: thư viện này dùng để cung cấp hỗ trợ cho múi giờ.

- Scikit-learn: thư viện này dùng để cung cấp các thuật toán học có giám sát và không giám sát, công cụ xử lý dữ liệu và các công cụ đánh giá mô hình.

- Sklearn-crfsuite: thư viện để tích hợp Conditional Random Fields (CRFs) với Scikit-learn dùng trong các bài toán học chuỗi như gán nhãn thực thể, phân loại chuỗi, ...

- Pandas: thư viện này giúp phân tích, xử lý và trực quan hóa dữ liệu dạng bảng hoặc chuỗi thời gian.

Sử dụng câu lệnh với cú pháp "pip install + *tên thư viện==phiên bản*" để tiến hành cài đặt danh sách các thư viện của môi trường ảo trên Visual Studio Code. Sau khi đã cài đặt đầy đủ các thư viện cần thiết thì tiến hành cài đặt hệ thống.

Các chức năng của hệ thống hoạt động tốt nhất và ổn định khi đã cài đặt đủ và đúng các thư viện cùng với các thông số phiên bản được cung cấp phía trên.

3.4. Tổng kết chương

Chương 3 giới thiệu kiến trúc tổng quát cho mô hình hệ thống lựa chọn sử dụng, mô tả các tập dữ liệu cần thiết, các mô hình của xử lý dữ liệu cho hệ thống trước khi cài đặt và mô tả cách cài đặt các mô hình cho hệ thống. Sau khi cài đặt thành công sẽ đến bước kiểm thử hệ thống ở chương 4 để đưa ra các đánh giá khách quan và chính xác nhất về các chức năng hệ thống.

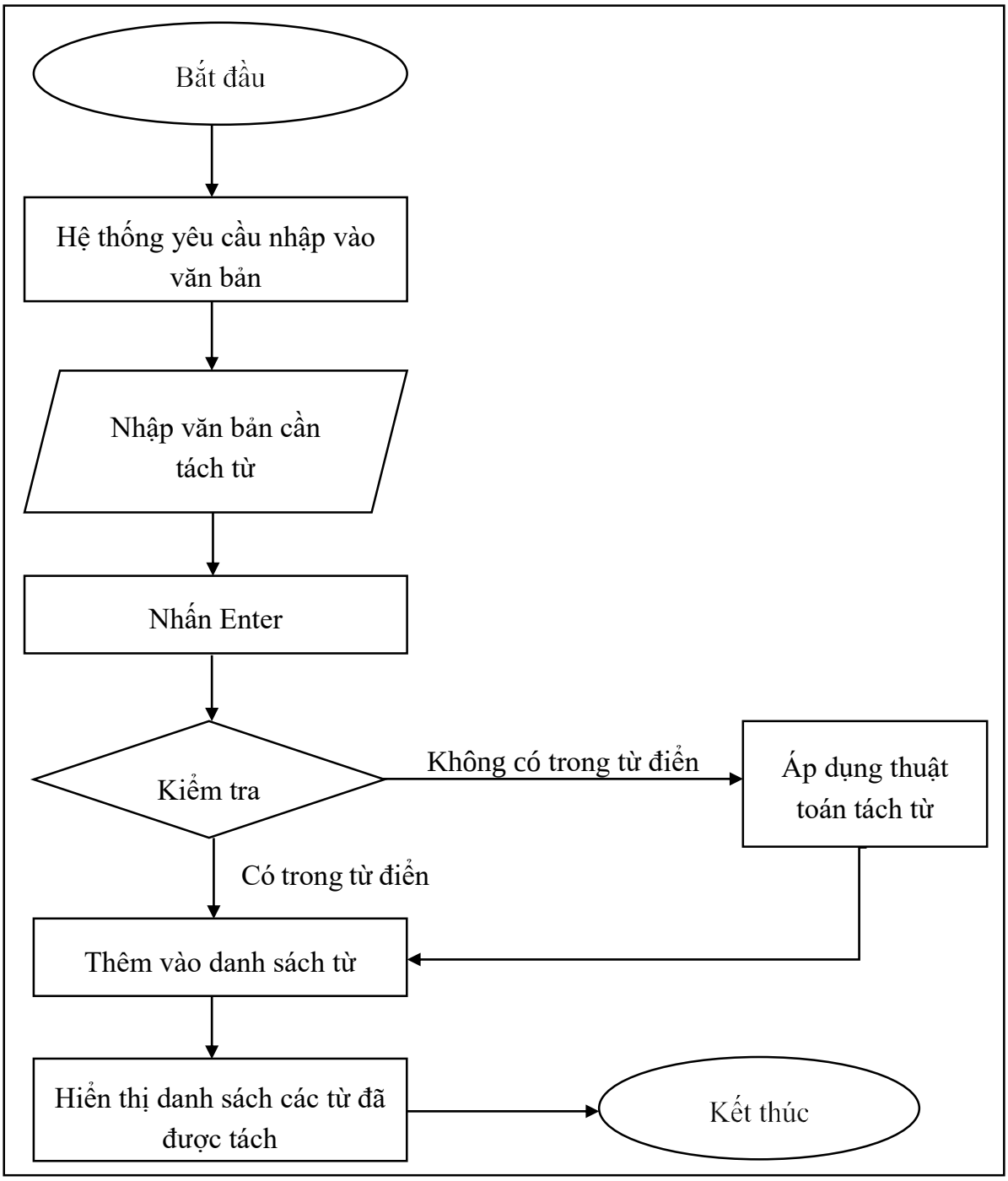
CHƯƠNG 4. KIỂM THỬ VÀ ĐÁNH GIÁ

4.1. Kịch bản kiểm thử

Để đảm bảo hệ thống hoạt động được chính xác cần phải qua các quá trình kiểm thử nhiều lần nghiêm ngặt.

4.1.1. Chức năng tách từ: CN01

- Lưu đồ giải thuật ***Hình 4.1*** mô tả quy trình tách từ cho một văn bản.



Hình 4.1: Lưu đồ giải thuật quy trình tách từ

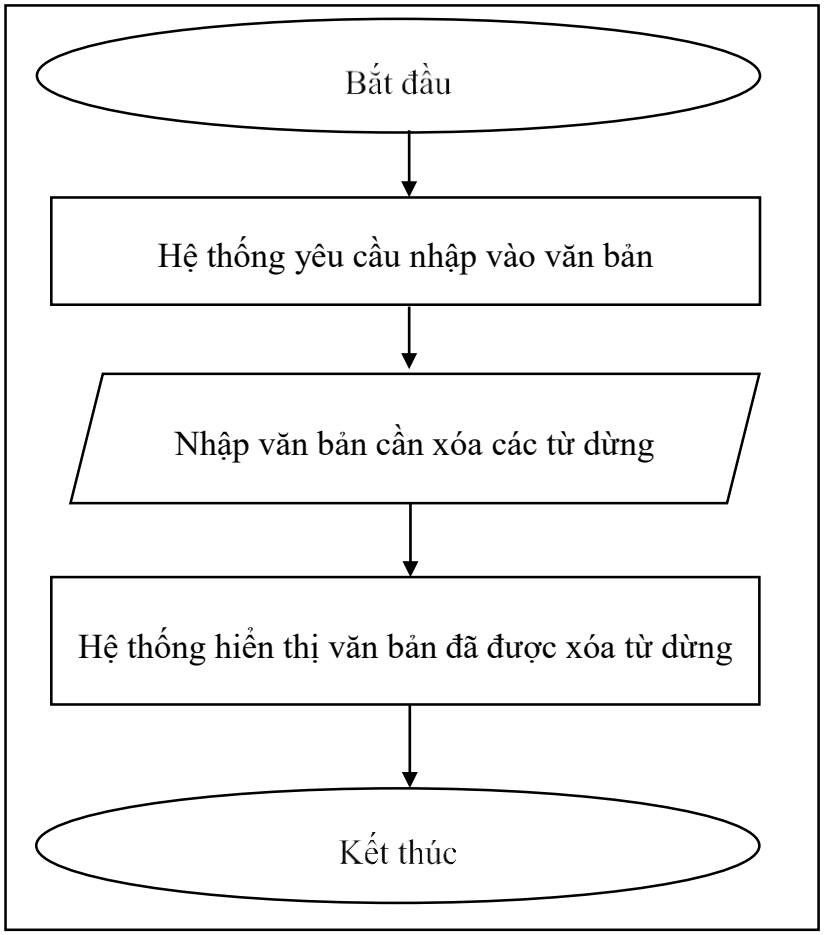
Bảng 4.1 mô tả các trường hợp kiểm thử Input – Output mong đợi của quy trình tách từ của một câu văn bản.

Bảng 4.1: Các trường hợp kiểm thử cho quy trình tách từ

Trường hợp kiểm thử	Input		Output
	Từ có trong từ điển	Từ không có trong từ điển	
1	X		Thêm vào danh sách từ và hiển thị danh sách các từ.
2		X	Áp dụng thuật toán tách từ để thêm từ vào danh sách từ và hiển thị danh sách các từ.

4.1.2. Chức năng xóa các từ dừng: CN02

- Lưu đồ giải thuật ***Hình 4.2*** mô tả quy trình xóa các từ dừng (stopwords) cho một văn bản.



Hình 4.2: Lưu đồ giải thuật cho quy trình xóa các từ dừng

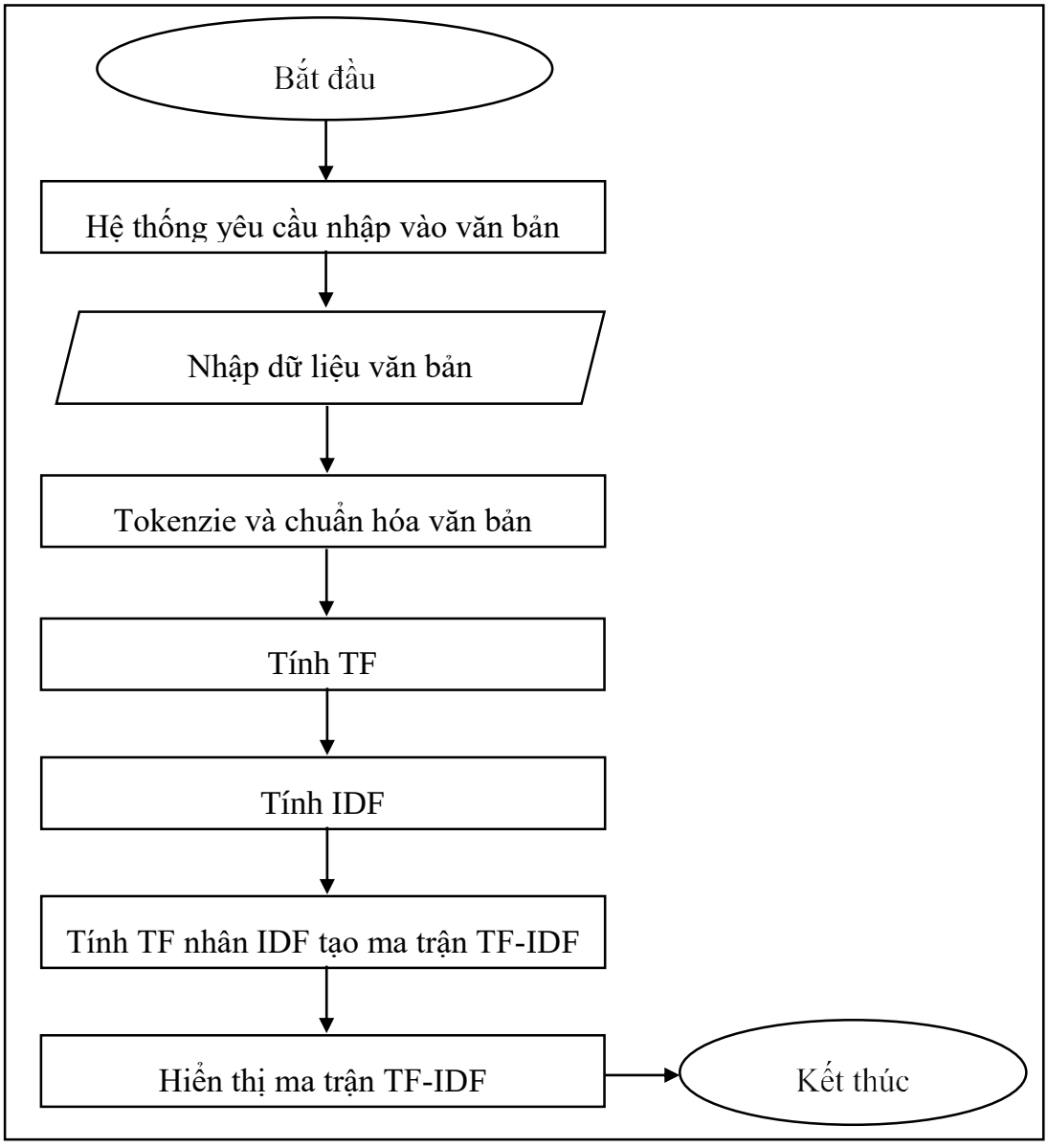
Bảng 4.2 mô tả các trường hợp kiểm thử Input – Output mong đợi của quy trình tách từ của một câu văn bản.

Bảng 4.2: Các trường hợp kiểm thử cho chức năng xóa từ dừng

Trường hợp kiểm thử	Input	Output
	Văn bản chưa được xóa các từ dừng	
1	X	Văn bản đã được xóa các từ dừng

4.1.3. Chức năng tính TF-IDF: CN03

- Lưu đồ giải thuật ***Hình 4.2*** mô tả quy trình tính các giá trị TF-IDF cho một văn bản.



Hình 4.3: Lưu đồ giải thuật quy trình tính giá trị TF-IDF

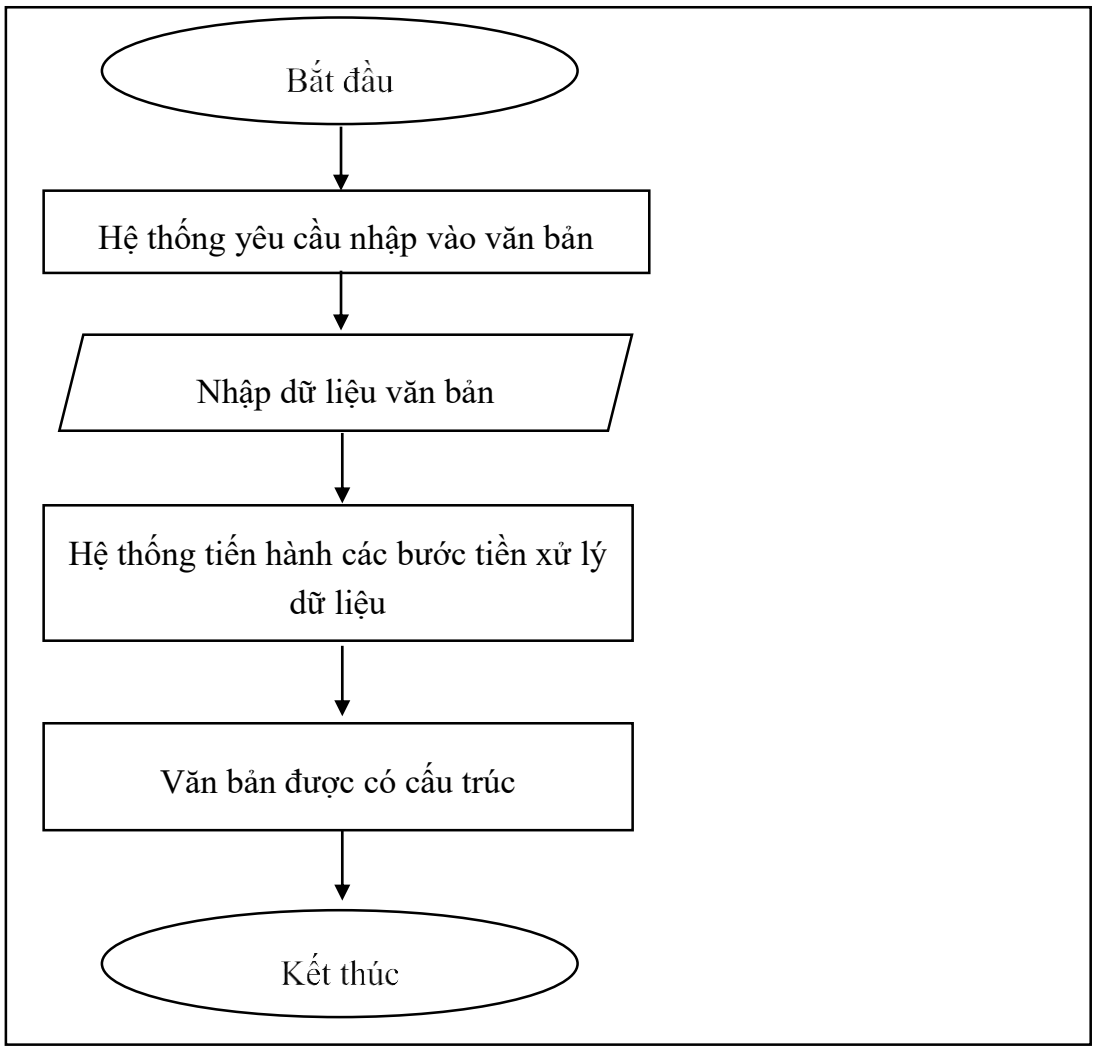
Bảng 4.3 mô tả các trường hợp kiểm thử Input – Output mong đợi của quy trình tính các giá trị TF-IDF của văn bản.

Bảng 4.3: Các trường hợp kiểm thử cho chức năng tính TF-IDF

Trường hợp kiểm thử	Input	Output
	Văn bản cần tính TF-IDF	
1	X	Ma trận TF-IDF

4.1.4. Chức năng tiền xử lý dữ liệu: CN04

- Lưu đồ giải thuật ***Hình 4.4*** mô tả quy trình tiền xử lý dữ liệu cho một văn bản.



Hình 4.4: Lưu đồ giải thuật cho quy trình tiền xử lý dữ liệu

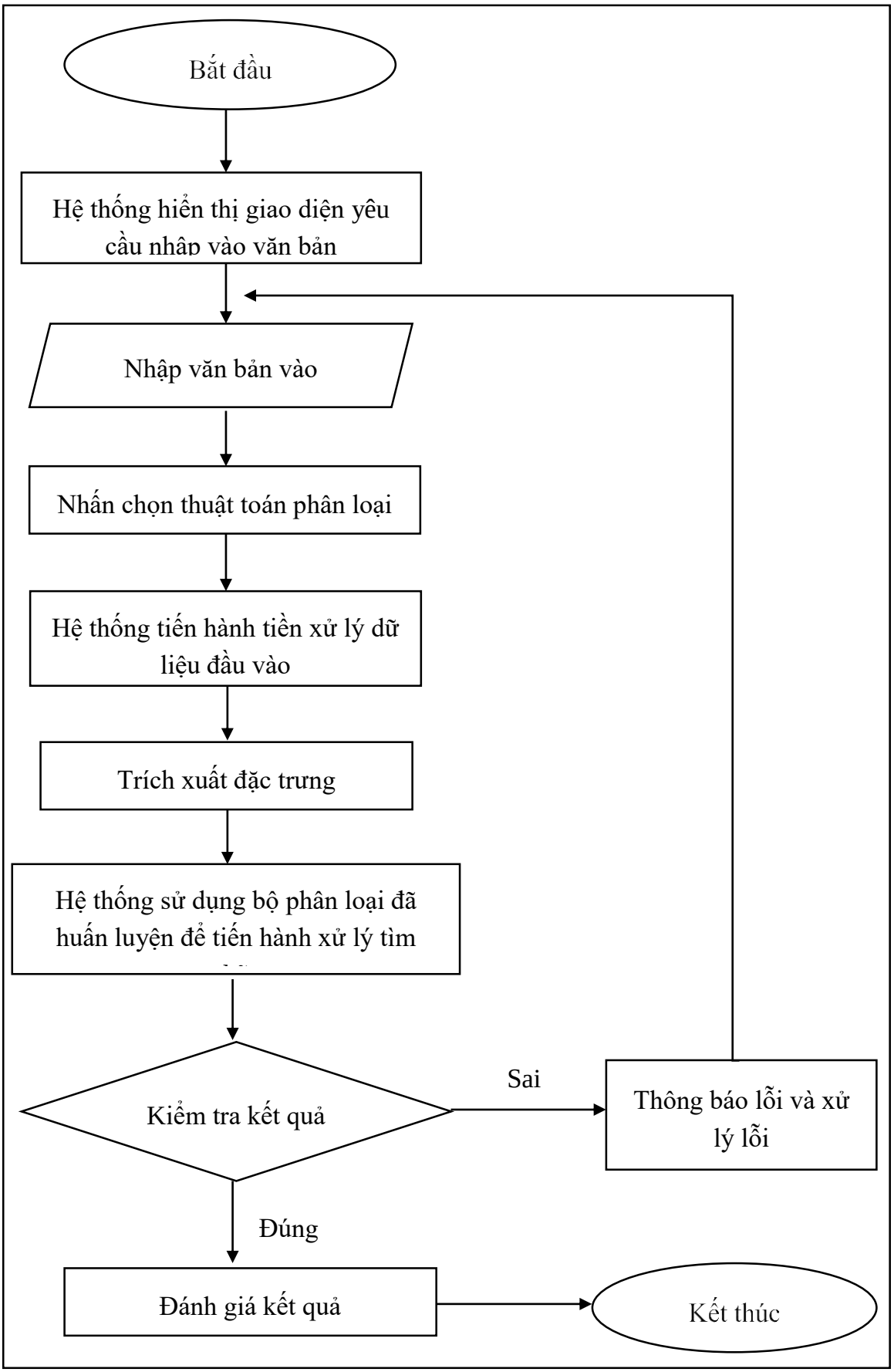
Bảng 4.3 mô tả các trường hợp kiểm thử Input – Output mong đợi của quy trình tiền xử lý dữ liệu đầu vào cho văn bản.

Bảng 4.4: Các trường hợp kiểm thử cho chức năng tiền xử lý dữ liệu

Trường hợp kiểm thử	Input				Output
	Văn bản có các các dấu thanh của từ đặt sai vị trí	Văn bản chứa các mã code HTML	Văn bản chứa nhiều từ dừng và chưa tách từ	Văn bản chứa nhiều ký tự đặc biệt	
1	X				Văn bản có cấu trúc
2		X			Văn bản có cấu trúc
3			X		Văn bản có cấu trúc
4				X	Văn bản có cấu trúc

4.1.5. Chức năng phân loại văn bản: CN05

- Lưu đồ giải thuật ***Hình 4.5*** mô tả quy trình phân loại văn bản và tìm nhãn cho văn bản.



Hình 4.5: Lưu đồ giải thuật mô tả quy trình phân loại văn bản

Bảng 4.5 mô tả các trường hợp kiểm thử Input – Output mong đợi của quá trình phân loại tìm nhãn cho văn bản.

Bảng 4.5: Các trường hợp kiểm thử cho chức năng phân loại văn bản

Trường hợp kiểm thử	Input			Output
	Văn bản bình thường	Văn bản chứa nhiều thông tin gây nhiễu	Văn bản lẫn lộn ngôn ngữ	
1	X			Phân loại đúng nhãn
2		X		Xử lý văn bản và phân loại đúng nhãn
3			X	Xử lý văn bản và phân loại đúng nhãn

4.2. Kết quả kiểm thử

4.2.1. Chức năng tách từ: CN01

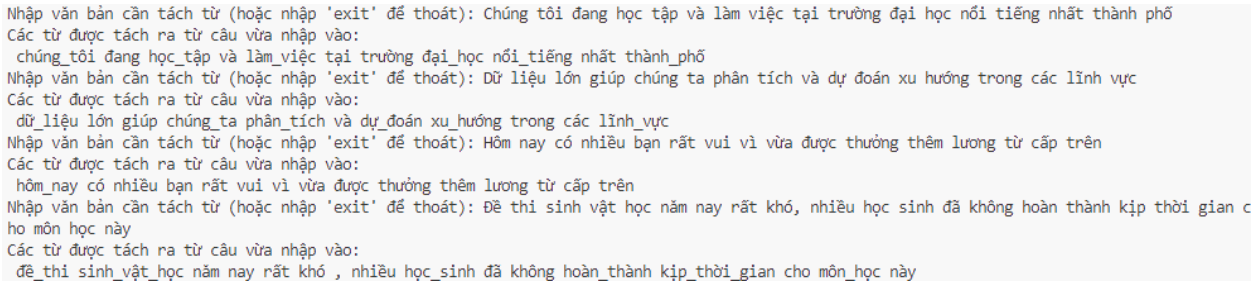
- Trường hợp kiểm thử 1: Từ có trong từ điển

Bảng 4.6 trình bày kết quả của quá trình tách từ của những từ có trong từ điển thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.6: Kết quả tách từ của văn bản có chứa từ có trong từ điển

STT	Input	Output	Đúng với output mong đợi
1	Chúng tôi đang học tập và làm việc tại trường đại học nổi tiếng nhất thành phố	chúng_tôi đang học_tập và làm_việc tại trường đại_học nổi_tiếng nhất thành_phố	X
2	Dữ liệu lớn giúp chúng ta phân tích và dự đoán xu hướng trong các lĩnh vực	dữ_liệu lớn giúp chúng_ta phân_tích và dự_đoán xu_hướng trong các lĩnh_vực	X
3	Hôm nay có nhiều bạn rất vui vì vừa được thưởng thêm lương từ cấp trên	hôm_nay có nhiều bạn rất vui vì vừa được thưởng thêm lương từ cấp trên	X
4	Đề thi sinh vật học năm nay rất khó, nhiều học sinh đã không hoàn thành kịp thời gian cho môn học này	đề_thi sinh_vật_học năm nay rất khó , nhiều học_sinh đã không hoàn_thành kịp_thời_gian cho môn_học này	X

Hình 4.6 minh họa cho kết quả kiểm thử của quy trình tách từ của các câu có các từ có trong từ điển



Hình 4.6: Kết quả kiểm thử của quy trình tách từ từ có trong từ điển

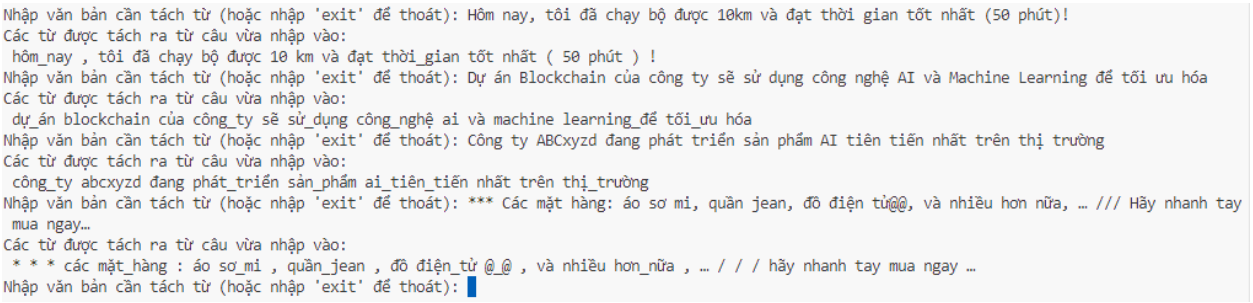
- Trường hợp kiểm thử 2: Từ không có trong từ điển

Bảng 4.7 trình bày kết quả của quá trình tách từ của những từ không có trong từ điển thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.7: Kết quả tách từ của văn bản có chứa từ không có trong từ điển

STT	Input	Output	Đúng với outut mong đợi
1	Hôm nay, tôi đã chạy bộ được 10km và đạt thời gian tốt nhất (50 phút)!	hôm_nay , tôi đã chạy bộ được 10 km và đạt thời_gian tốt nhất (50 phút) !	X
2	Dự án Blockchain của công ty sẽ sử dụng công nghệ AI và Machine Learning để tối ưu hóa	dự_án blockchain của công_ty sẽ sử_dụng công_nghệ ai và machine learning_để tối_ưu hóa	X
3	Công ty ABCxyzd đang phát triển sản phẩm AI tiên tiến nhất trên thị trường	công_ty abcxyzd đang phát_triển sản_phẩm ai_tiên_tiến nhất trên thị_trường	X
4	*** Các mặt hàng: áo sơ mi, quần jean, đồ điện tử@@, và nhiều hơn nữa, … /// Hãy nhanh tay mua ngay…	* * * các mặt_hàng : áo sơ_mi , quần_jean , đồ điện_tử @_@ , và nhiều hơn_nữa , … / / / hãy nhanh tay mua ngay …	X

Hình 4.7 minh họa cho kết quả kiểm thử của quy trình tách từ của các câu có các từ có trong từ điển.



Hình 4.7: Kết quả kiểm thử của quy trình tách từ có chứa từ không có trong từ điển

4.2.2. Chức năng xóa các từ dừng: CN02

- Trường hợp kiểm thử 1: Văn bản chưa được xóa các từ dừng

Bảng 4.8 trình bày kết quả của quá trình loại bỏ các từ dừng ra khỏi văn bản thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.8 Kết quả quá trình loại bỏ từ dừng

STT	Input	Output	Đúng với output mong đợi
1	TPHCM: Khai trương dịch vụ lặn biển săn cá mập Ngày 29/4, công ty TNHH TM - DV Hải Thanh đã khánh thành và đưa vào hoạt động khu tham quan Thủy cung Đại Dương tại số 491 Nguyễn Thị Thập, phường Tân Quy, quận 7. Thủy cung Đại Dương được coi là "điểm nhấn" ấn tượng cho khách tham quan, với hàng ngàn cá cảnh, san hô cùng nhiều loài sinh vật biển quý hiếm được trưng bày.	tphcm khai_trương dịch_vụ lặn biển săn cá_mập 294 tnhh tm dv hải_thanh khánh_thành hoạt_động khu tham_quan thủy_cung đại_dương 491 nguyễn_thị_thập phường tân_quy quận 7 thủy cung đại_dương coi nhấn ấn_tượng khách tham_quan ngàn cá_cảnh san_hô loài sinh_vật biển quý_hiếm trưng_bày	X
2	Thành lập dự án POLICY phòng chống HIV/AIDS ở VN (NLĐ)-Quỹ hỗ trợ khẩn cấp về AIDS của Hoa Kỳ vừa thành lập dự án POLICY tại VN với cam kết hỗ trợ Chính phủ và nhân dân VN đối phó HIV/AIDS.	thành_lập dự_án policy phòng_chống hivaids nlđ quỹ hỗ_trợ khẩn_cấp aids hoa_kỳ vừa thành_lập dự_án policy cam_kết hỗ_trợ chính_phủ nhân_dân đối_phó hivaids	X

Hình 4.8 minh họa cho kết quả kiểm thử của quy trình tách từ của các câu có các từ có trong từ điển.

```
Nhập văn bản cần xóa từ dừng (hoặc nhập 'e' để thoát): TPHCM: Khai trương dịch vụ lặn biển săn cá mập Ngày 29/4, công ty TNHH TM - DV Hải Thanh
đã khánh thành và đưa vào hoạt động khu tham quan Thủy cung Đại Dương tại số 491 Nguyễn Thị Thập, phường Tân Quy, quận 7. Thủy cung Đại Dương đư
ợc coi là "điểm nhấn" ấn tượng cho khách tham quan, với hàng ngàn cá cảnh, san hô cùng nhiều loài sinh vật biển quý hiếm được trưng bày.
Kết quả văn bản sau khi xóa các từ dừng:
 tphcm khai_trương dịch_vụ lặn biển săn cá_mập 294 tnhh tm dv hải_thanh khánh_thành hoạt_động khu tham_quan thủy_cung đại_dương 491 nguyễn_thị_t
hập phường tân_quy quận 7 thủy cung đại_dương coi nhấn ấn_tượng khách tham_quan ngàn cá_cảnh san_hô loài sinh_vật biển quý_hiếm trưng_bày
Nhập văn bản cần xóa từ dừng (hoặc nhập 'e' để thoát): Thành lập dự án POLICY phòng chống HIV/AIDS ở VN (NLĐ)- Quỹ hỗ trợ khẩn cấp về AIDS của H
oa Kỳ vừa thành lập dự án POLICY tại VN với cam kết hỗ trợ Chính phủ và nhân dân VN đối phó HIV/AIDS.
Kết quả văn bản sau khi xóa các từ dừng:
 thành_lập dự_án policy phòng_chống hivaids nlđ quỹ hỗ_trợ khẩn_cấp aids hoa_kỳ vừa thành_lập dự_án policy cam_kết hỗ_trợ chính_phủ nhân_dân đối
_phó hivaids
```

Hình 4.8: Kết quả kiểm thử của quy trình loại bỏ các từ dừng

4.2.3. Chức năng tính TF-IDF: CN03

- Trường hợp kiểm thử 1: Văn bản cần tính TF-IDF

Bảng 4.9 trình bày kết quả tính TF-IDF của một văn bản thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.9: Kết quả tính TF-IDF của văn bản

STT	Input	Output	Đúng với output mong đợi
1	Pháo hoa sáng rực trên bầu trời đêm. Màn pháo hoa chào đón năm mới rất đẹp.	Thu được ma trận TF-IDF như hình	X
2	Học lập trình rất vui. Tôi yêu thích lập trình máy tính từ lớp 6. Tôi đang sử dụng máy tính hp.	Thu được ma trận TF-IDF như hình	X

Hình 4.9 và ***Hình 4.10*** minh họa cho kết quả kiểm thử của quy trình tính giá trị TF-IDF của một văn bản.

```
Tổng số từ có trong tài liệu: 16
     bầu   chào    hoa    màn    mới    năm   pháo    rất    rực   sáng   trên   trời    đêm    đón    đẹp
0 0.3776 0.0000 0.2687 0.0000 0.0000 0.0000 0.2687 0.0000 0.3776 0.3776 0.3776 0.3776 0.3776 0.0000 0.0000
1 0.0000 0.3533 0.2514 0.3533 0.3533 0.3533 0.2514 0.3533 0.0000 0.0000 0.0000 0.0000 0.0000 0.3533 0.3533
```

Hình 4.9: Kết quả kiểm thử tính TF-IDF cho văn bản có giá trị input thứ 1

```
Tổng số từ có trong tài liệu: 17
    dụng     hp    học    lập    lớp    máy    rất     sử  thích  trình   tính    tôi     từ    vui    yêu   đang
0 0.0000 0.0000 0.4905 0.3730 0.0000 0.0000 0.4905 0.0000 0.0000 0.3730 0.0000 0.0000 0.0000 0.4905 0.0000 0.0000
1 0.0000 0.0000 0.0000 0.2897 0.3809 0.2897 0.0000 0.0000 0.3809 0.2897 0.2897 0.2897 0.3809 0.0000 0.3809 0.0000
2 0.4176 0.4176 0.0000 0.0000 0.0000 0.3176 0.0000 0.4176 0.0000 0.0000 0.3176 0.3176 0.0000 0.0000 0.0000 0.4176
```

Hình 4.10: Kết quả kiểm thử tính TF-IDF cho văn bản có giá trị input thứ 2

4.2.4. Chức năng tiền xử lý dữ liệu: CN04

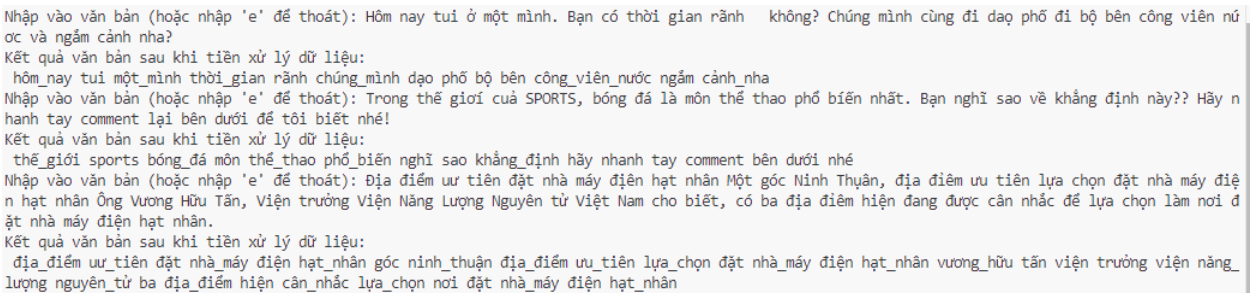
- Trường hợp kiểm thử 1: Văn bản có dấu thanh của từ đặt sai vị trí

Bảng 4.10 bên dưới trình bày kết quả tiền xử lý dữ liệu cho văn bản có dấu thanh của từ đặt sai vị trí trong câu, thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.10: Kết quả tiền xử lý dữ liệu cho văn bản có dấu thanh của từ đặt sai vị trí

STT	Input	Output	Đúng với output mong đợi
1	Hôm nay tui ở một mình. Bạn có thời gian rãnh không? Chúng mình cùng đi dạo phố đi bộ bên công viên nước và ngắm cảnh nha?	hôm_nay tui một_mình thời_gian rãnh chúng_mình dạo phố bộ bên công_viên_nước ngắm cảnh_nha	X
2	Trong thế giơi cuả SPORTS, bóng đá là môn thể thao phổ biến nhất. Bạn nghĩ sao về khẳng định này?? Hãy nhanh tay comment lại bên dưới để tôi biết nhé!	thế_giới sports bóng_đá môn thể_thao phổ_biến nghĩ sao khẳng_định hãy nhanh tay comment bên dưới nhé	X
3	Địa điểm ưu tiên đặt nhà máy điện hạt nhân Một góc Ninh Thuận, địa điểm ưu tiên lựa chọn đặt nhà máy điện hạt nhân Ông Vương Hữu Tấn, Viện trưởng Viện Năng Lượng Nguyên tử Việt Nam cho biết, có ba địa điểm hiện đang được cân nhắc để lựa chọn làm nơi đặt nhà máy điện hạt nhân.	địa_điểm ưu_tiên đặt nhà_máy điện hạt_nhân góc ninh_thuận địa_điểm ưu_tiên lựa_chọn đặt nhà_máy điện hạt_nhân vương_hữu tấn viện trưởng viện năng_lượng nguyên_tử ba địa_điểm hiện cân_nhắc lựa_chọn nơi đặt nhà_máy điện hạt_nhân	X

Hình 4.11 bên dưới minh họa cho kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản có dấu thanh của từ đặt sai vị trí trong câu.



Hình 4.11: Kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản có dấu thanh của từ đặt sai vị trí trong câu

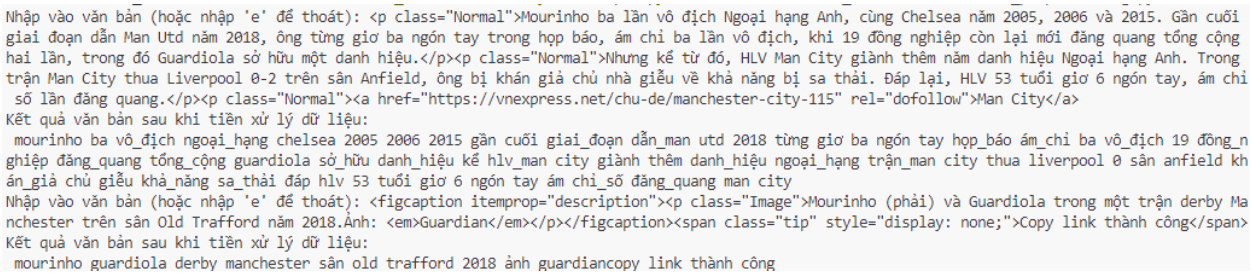
- Trường hợp kiểm thử 2: Văn bản có chứa các đoạn mã code HTML

Bảng 4.11 bên dưới trình bày kết quả tiền xử lý dữ liệu cho văn bản có chứa các đoạn mã code HTML trong câu, thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.11: Kết quả tiền xử lý dữ liệu cho văn bản chứa các đoạn mã code HTML

STT	Input	Output	Đúng với output mong đợi
1	Mourinho ba lần vô địch Ngoại hạng Anh, cùng Chelsea năm 2005, 2006 và 2015. Gần cuối giai đoạn dẫn Man Utd năm 2018, ông từng giơ ba ngón tay trong họp báo, ám chỉ ba lần vô địch, khi 19 đồng nghiệp còn lại mới đăng quang tổng cộng hai lần, trong đó Guardiola sở hữu một danh hiệu. Nhưng kể từ đó, HLV Man City giành thêm năm danh hiệu Ngoại hạng Anh. Trong trận Man City thua Liverpool 0-2 trên sân Anfield, ông bị khán giả chủ nhà giễu về khả năng bị sa thải. Đáp lại, HLV 53 tuổi giơ 6 ngón tay, ám chỉ số lần đăng quang. Man City	mourinho ba vô_địch ngoại_hạng chelsea 2005 2006 2015 gần cuối giai_đoạn dẫn_man utd 2018 từng giơ ba ngón tay họp_báo ám_chỉ ba vô_địch 19 đồng_nghiệp đăng_quang tổng_cộng guardiola sở_hữu danh_hiệu kể hlv_man city giành thêm danh_hiệu ngoại_hạng trận_man city thua liverpool 0 sân anfield khán_giả chủ giễu khả_năng sa_thải đáp hlv 53 tuổi giơ 6 ngón tay ám chỉ_số đăng_quang man city	X
2	<figcaption itemprop="description"> Mourinho (phải) và Guardiola trong một trận derby Manchester trên sân Old Trafford năm 2018.Ảnh: Guardian </figcaption>Copy link thành công	mourinho guardiola derby manchester sân old trafford 2018 ảnh guardiancopy link thành_công	X

Hình 4.12 bên dưới minh họa cho kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản chứa các đoạn mã code HTML trong câu.



Hình 4.12: Kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản chứa các đoạn mã code HTML

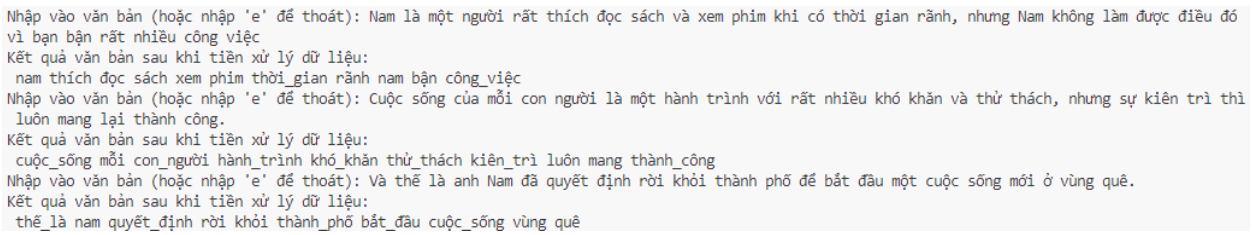
- Trường hợp kiểm thử 3: Văn bản chứa nhiều từ dừng và chưa được tách câu

Bảng 4.12 bên dưới trình bày kết quả tiền xử lý dữ liệu cho văn bản có chứa nhiều từ dừng và chưa được tách từ, thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.12: Kết quả tiền xử lý dữ liệu cho văn bản chứa nhiều từ dừng và chưa tách từ

STT	Input	Output	Đúng với output mong đợi
1	Nam là một người rất thích đọc sách và xem phim khi có thời gian rãnh, nhưng Nam không làm được điều đó vì bạn bận rất nhiều công việc	nam thích đọc sách xem phim thời_gian rãnh nam bận công_việc	X
2	Cuộc sống của mỗi con người là một hành trình với rất nhiều khó khăn và thử thách, nhưng sự kiên trì thì luôn mang lại thành công.	cuộc_sống mỗi con_người hành_trình khó_khăn thử_thách kiên_trì luôn mang thành_công	X
3	Và thế là anh Nam đã quyết định rời khỏi thành phố để bắt đầu một cuộc sống mới ở vùng quê.	thế_là nam quyết_định rời khỏi thành_phố bắt_đầu cuộc_sống vùng quê	X

Hình 4.13 bên dưới minh họa cho kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản chứa nhiều từ dừng và chưa tách câu.



Hình 4.13: Kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản chứa nhiều từ dừng và chưa tách câu

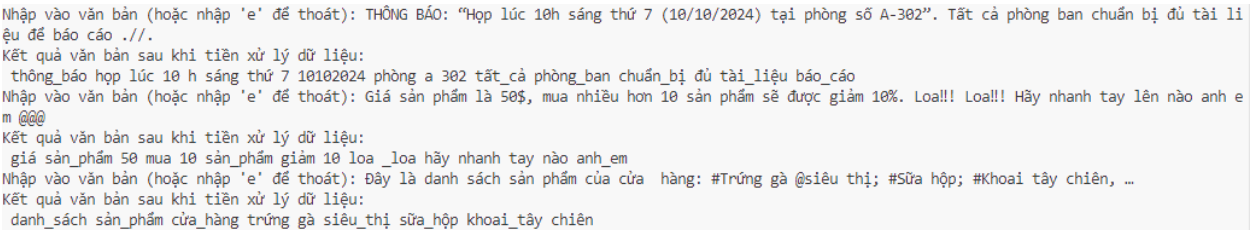
- Trường hợp kiểm thử 4: Văn bản chứa nhiều ký tự đặc biệt

Bảng 4.13 bên dưới trình bày kết quả tiền xử lý dữ liệu cho văn bản có chứa các ký tự đặc biệt, thông qua quá trình kiểm thử với các giá trị input khác nhau.

Bảng 4.13: Kết quả tiền xử lý dữ liệu cho văn bản chứa các ký tự đặc biệt

STT	Input	Output	Đúng với output mong đợi
1	THÔNG BÁO: "Họp lúc 10h sáng thứ 7 (10/10/2024) tại phòng số A-302". Tất cả phòng ban chuẩn bị đủ tài liệu để báo cáo .//.	thông_báo họp lúc 10 h sáng thứ 7 10102024 phòng a 302 tất_cả phòng_ban chuẩn_bị đủ tài_liệu báo_cáo	X
2	Giá sản phẩm là 50$, mua nhiều hơn 10 sản phẩm sẽ được giảm 10%. Loa!!! Loa!!! Hãy nhanh tay lên nào anh em @@@	giá sản_phẩm 50 mua 10 sản_phẩm giảm 10 loa _loa hãy nhanh tay nào anh_em	X
3	Đây là danh sách sản phẩm của cửa hàng: #Trứng gà @siêu thị; #Sữa hộp; #Khoai tây chiên, …	danh_sách sản_phẩm cửa_hàng trứng gà siêu_thị sữa_hộp khoai_tây chiên	X

Hình 4.14 bên dưới minh họa cho kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản có các ký tự đặc biệt trong câu.



Hình 4.14: Kết quả kiểm thử của quá trình tiền xử lý dữ liệu cho văn bản chứa các ký tự đặc biệt

4.2.5. Chức năng phân loại văn bản: CN05

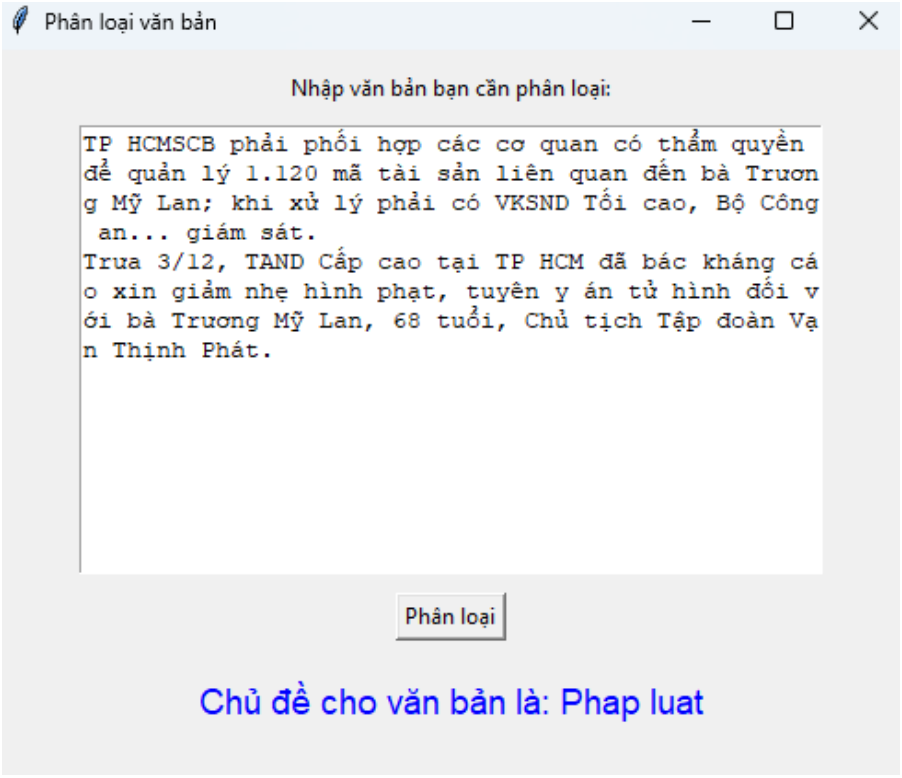
- Trường hợp kiểm thử 1: Văn bản bình thường

Bảng 4.14 bên dưới mô tả chi tiết các kết quả của quá trình phân loại tìm nhãn cho văn bản bình thường thông qua quá trình kiểm thử với các giá trị input khác nhau cho các chủ đề khác nhau đúng như mong đợi.

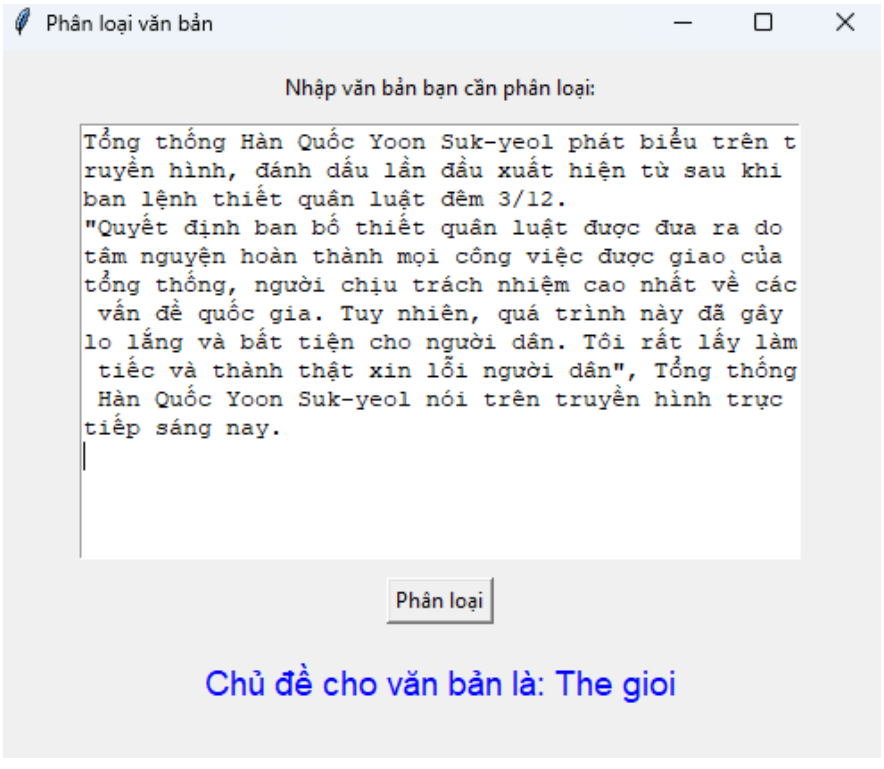
Bảng 4.14: Kết quả phân loại cho văn bản bình thường

STT	Input	Output	Đúng với output mong đợi
1	TP HCMSCB phải phối hợp các cơ quan có thẩm quyền để quản lý 1.120 mã tài sản liên quan đến bà Trương Mỹ Lan; khi xử lý phải có VKSND Tối cao, Bộ Công an... giám sát. Trưa 3/12, TAND Cấp cao tại TP HCM đã bác kháng cáo xin giảm nhẹ hình phạt, tuyên y án tử hình đối với bà Trương Mỹ Lan, 68 tuổi, Chủ tịch Tập đoàn Vạn Thịnh Phát.	Pháp luật	X
2	Tổng thống Hàn Quốc Yoon Suk-yeol phát biểu trên truyền hình, đánh dấu lần đầu xuất hiện từ sau khi ban lệnh thiết quân luật đêm 3/12. "Quyết định ban bố thiết quân luật được đưa ra do tâm nguyện hoàn thành mọi công việc được giao của tổng thống, người chịu trách nhiệm cao nhất về các vấn đề quốc gia. Tuy nhiên, quá trình này đã gây lo lắng và bất tiện cho người dân. Tôi rất lấy làm tiếc và thành thật xin lỗi người dân", Tổng thống Hàn Quốc Yoon Suk-yeol nói trên truyền hình trực tiếp sáng nay.	Thế giới	X
3	Ngày 6/12, Hoàng Duy Hưng, 34 tuổi, bị TAND Hà Nội tuyên y án 8 năm tù về tội Môi giới mại dâm; đồng phạm Đoàn Văn Trinh, 30 tuổi, án 5 năm tù về tội Chứa mại dâm. Tòa đánh giá tại phiên phúc thẩm Hưng và Trinh không đưa ra được tình tiết mới, do đó không có căn cứ chấp nhận kháng cáo xin giảm nhẹ hình phạt. Mức án tòa sơ thẩm đã tuyên là phù hợp.	Pháp luật	X
4	Messi nhận 38,43% phiếu bầu, vượt qua Cucho Hernandez của Columbus Crew (33,7%), Evander của Portland Timbers (9,24%), Christian Benteke của D.C. United (7,1%) và đồng đội Luis Suarez ở Inter Miami (2,17%), để giành danh hiệu cá nhân cao quý nhất giải. Các cầu thủ MLS, giới truyền thông và giám đốc điều hành CLB bỏ phiếu cho giải thưởng này. Messi trở thành cầu thủ Nam Mỹ thứ 10 và là người Argentina thứ năm giành Landon Donovan MVP, sau Luciano Acosta (FC Cincinnati, 2023), Diego Valeri (Portland Timbers, 2017), Guillermo Barros Schelotto (Columbus Crew, 2008), Christian Gomez (D.C. United, 2006). Anh cũng là cầu thủ đầu tiên của Inter Miami đoạt danh hiệu cá nhân này.	Thể thao	X

Người dùng tiến hành nhập hoặc sao chép văn bản vào phần mềm của hệ thống. Kết quả kiểm thử của quá trình phân loại tìm nhãn cho một văn bản bình thường được minh họa chi tiết như các ***Hình 4.15***, ***Hình 4.16***, ***Hình 4.17*** và ***Hình 4.18*** bên dưới:



Hình 4.15: Kết quả kiểm thử phân loại tìm nhãn cho văn bản bình thường 1



Hình 4.16: Kết quả kiểm thử phân loại tìm nhãn cho văn bản bình thường 2

Hình 4.17: Kết quả kiểm thử phân loại tìm nhãn cho văn bản bình thường 3

Hình 4.18: Kết quả kiểm thử phân loại tìm nhãn cho văn bản bình thường 4

- Trường hợp kiểm thử 2: Văn bản chứa nhiều thông tin gây nhiễu

Bảng 4.15 bên dưới trình bày kết quả của phân loại tìm nhãn cho văn bản chứa nhiều thông tin gây nhiễu giữa các văn bản với nhau, thông qua quá trình kiểm thử với các giá trị input khác nhau cho các chủ đề.

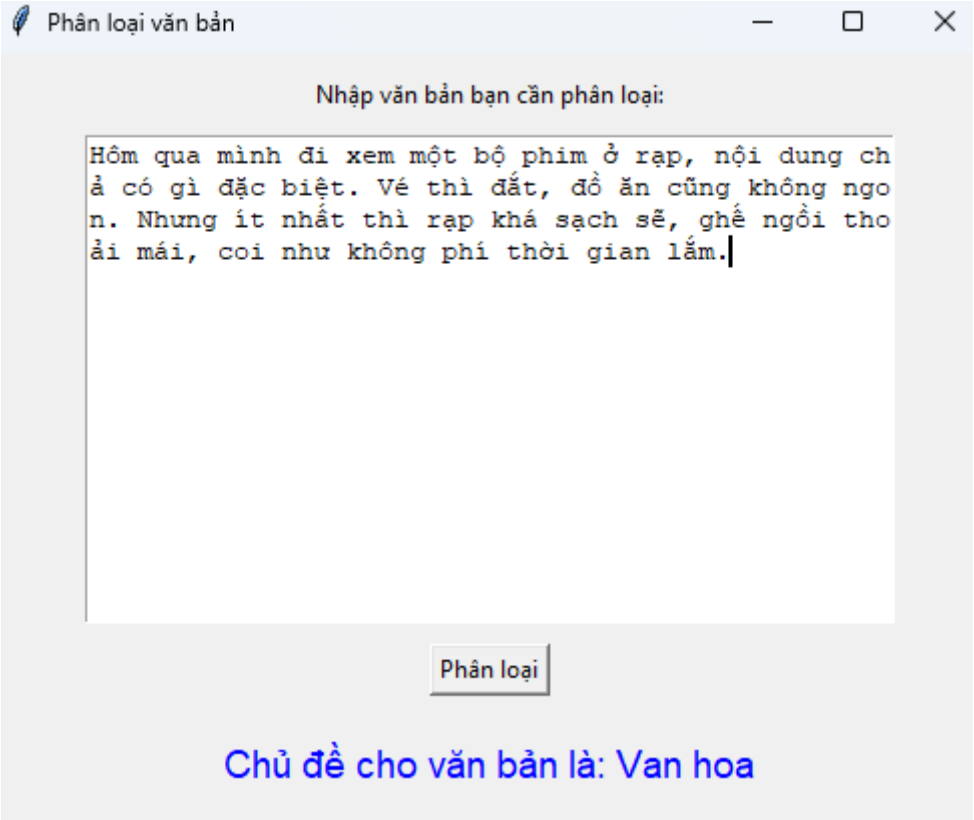
Bảng 4.15: Kết quả phân loại cho văn bản chứa nhiều thông tin gây nhiễu

STT	Input	Output	Đúng với output mong đợi
1	Tôi đi ngang qua tòa án hôm nay và thấy một đám đông tụ tập. &&& Có vẻ như một vụ án lớn đang được xét xử. Nhưng trời thì nóng nực, mọi người đứng chờ dưới cái nắng gay gắt. ☼☼☼ Chả biết bao giờ mới đến lượt xử lý. Nghe thoang thoáng đâu đó là các xét xử vụ tham ô của ông gì gì đó liên quan đến xây dựng cầu đường. Không biết đã xét xử như thế nào rồi. @@Copyright #Xử án #xây dựng #đời sống ****	Pháp luật	X
2	Hôm qua mình đi xem một bộ phim ở rạp, nội dung chả có gì đặc biệt. Vé thì đắt, đồ ăn cũng không ngon. Nhưng ít nhất thì rạp khá sạch sẽ, ghế ngồi thoải mái, coi như không phí thời gian lắm.	Văn hóa	X
3	Cụ thể, tại Singapore, đất mua đi, bán lại trong năm đầu tiên bị đánh thuế 100% trên giá trị chênh lệch mua, bán. 💰 Sau 2 năm, mức thuế suất giảm còn 50% và sau 3 năm là 25%. Tại Đài Loan, giao dịch bất động sản thực hiện trong 2 năm đầu sau khi mua áp dụng thuế suất là 45%. 🏦 Trong 2-5 năm, thuế suất là 35%, trong 5-10 năm thuế suất 20% và sau 10 năm là 15%. Xem thêm chi tiết tại đây.	Kinh doanh	X
4	"Cái máy tính của tôi hôm nay chạy chậm kinh khủng, chắc là lại bị virus ☹. Phải mang đi sửa hoặc cài lại hệ điều hành 📄. Đúng lúc đang cần làm gấp thì lại hỏng, bực thật 😤. Tôi ước có thể có một máy tính tốt hơn cho bản thân mình! 💻 Hy vọng máy tính mới sẽ không gặp phải vấn đề này nữa. Việc phải sửa chữa máy tính liên tục thật sự rất phiền phức, nó ảnh hưởng đến công việc và thời gian của tôi rất nhiều. Đã đến lúc cần đầu tư vào một chiếc máy tính mới hoặc tìm một giải pháp thay thế. Tìm hiểu thêm về các lựa chọn máy tính mới tại đây."	Vi tính	X

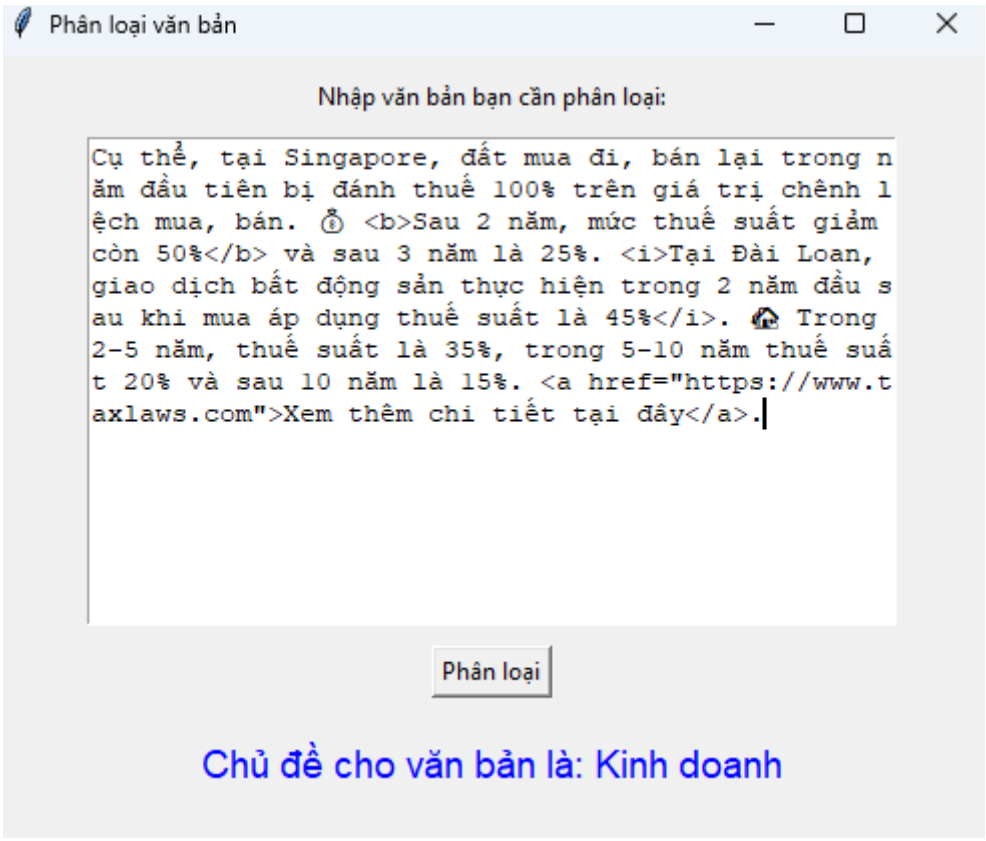
STT	Input	Output	Đúng với output mong đợi
5	báo điện tử VnExpress</br> Báo tiếng Việt nhiều người xem nhất Sáng nay mình ghé qua làng ĐH để ăn sáng. Có rất nhiều quán bán đồ ăn sáng như hủ tiếu, bánh mì, cơm tấm… Quán nào cũng khá đông khách, đặc biệt là vào giờ cao điểm. Tuy nhiên, thời tiết hôm nay nóng nực khiến mình cảm thấy không muốn ngồi lâu. À, nhân tiện ghé quán nước gần đó, mình thấy họ bán cà phê sữa đá cũng khá ngon và giá cả phải chăng. "Hôm nay là một ngày thật tuyệt vời!! :)) Mình và bạn bè đã cùng nhau đi picnic ở công viên, ăn uống và chơi đùa thật vui vẻ (^"^). ♥♥♥♥ #ĐờiSống #Picnic #NgàyHè"	Đời sống	X

Người dùng tiến hành nhập hoặc sao chép văn bản vào phần mềm của hệ thống. Kết quả kiểm thử của quá trình phân loại tìm nhãn cho một văn bản có chứa nhiều thông tin gẫy nhiễu được minh họa chi tiết như các ***Hình 4.19***, ***Hình 4.20***, ***Hình 4.21***, ***Hình 4.22*** và ***Hình 4.23*** bên dưới:

Hình 4.19: Kết quả kiểm thử phân loại tìm nhãn cho văn bản chứa thông tin gây nhiễu 1

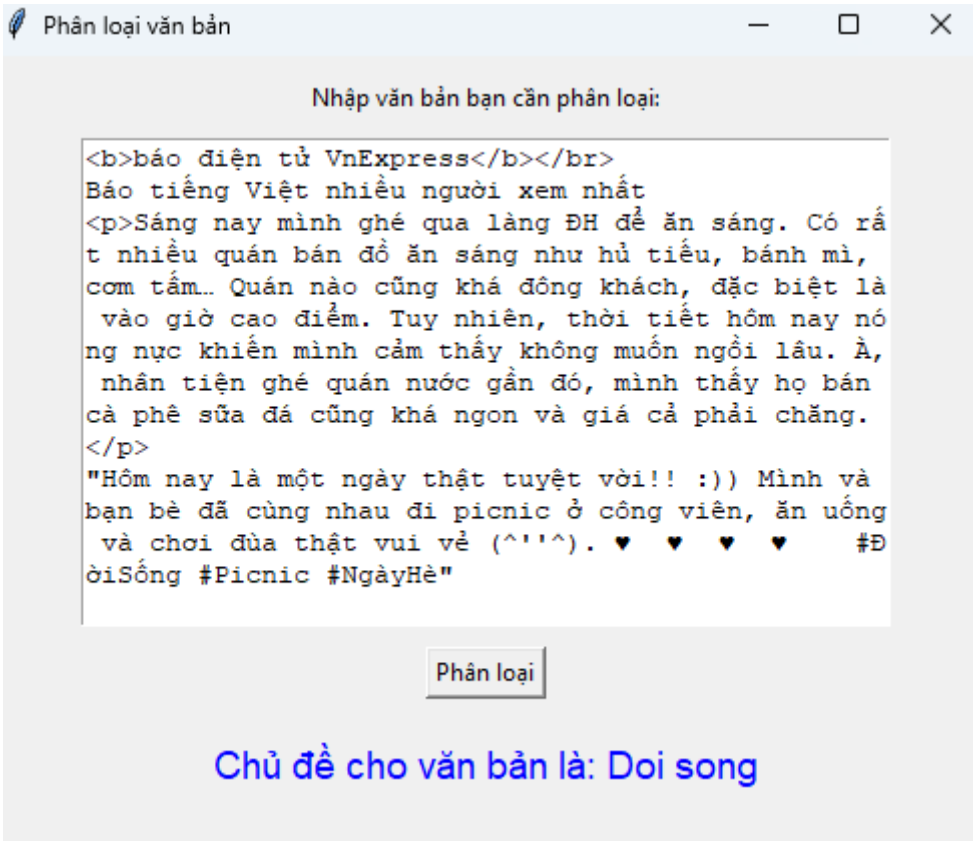


Hình 4.20: Kết quả kiểm thử phân loại tìm nhãn cho văn bản chứa thông tin gây nhiễu 2



Hình 4.21: Kết quả kiểm thử phân loại tìm nhãn cho văn bản chứa thông tin gây nhiễu 3

Hình 4.22: Kết quả kiểm thử phân loại tìm nhãn cho văn bản chứa thông tin gây nhiễu 4



Hình 4.23: Kết quả kiểm thử phân loại tìm nhãn cho văn bản chứa thông tin gây nhiễu 5

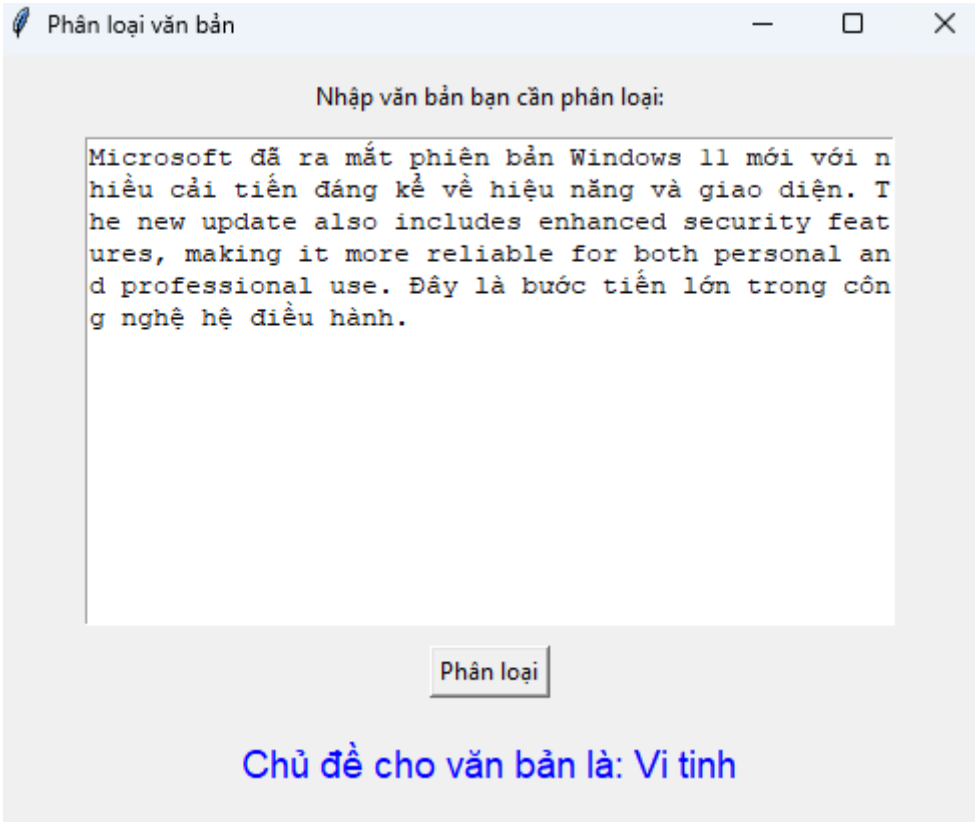
- Trường hợp kiểm thử 3: Văn bản lẫn lộn ngôn ngữ

Bảng 4.16 bên dưới trình bày kết quả của phân loại tìm nhãn cho văn bản lẫn lộn nhiều ngôn ngữ, thông qua quá trình kiểm thử với các giá trị input khác nhau cho các chủ đề.

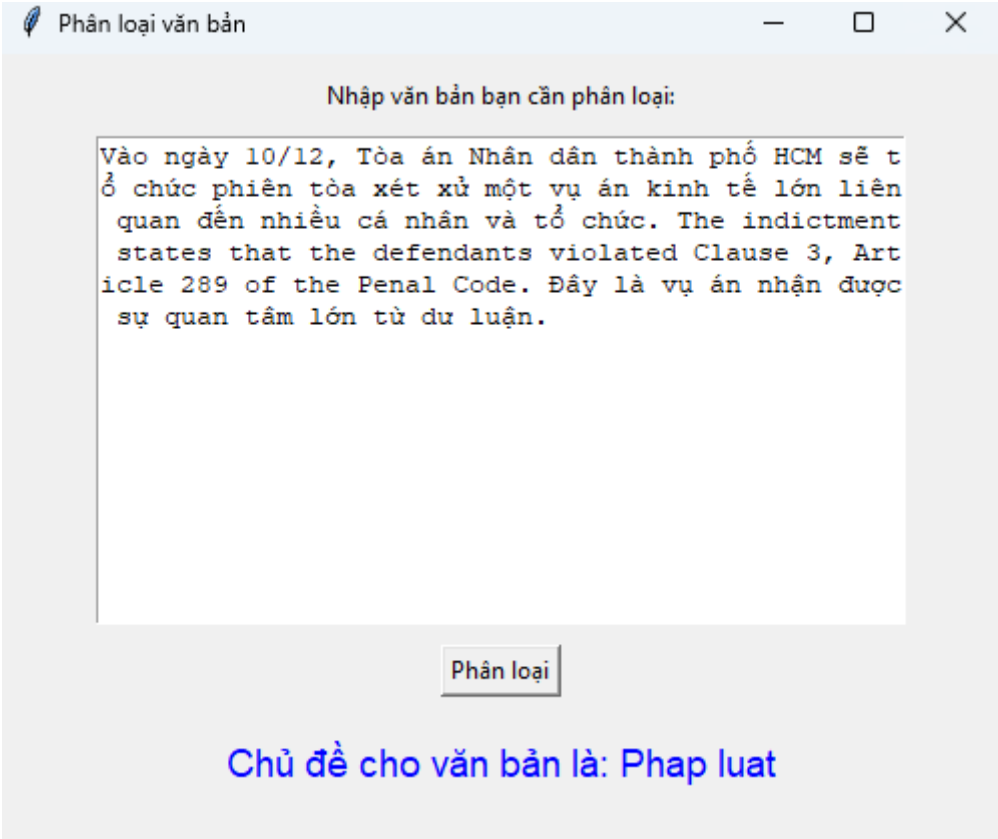
Bảng 4.16: Kết quả phân loại cho văn bản lẫn lộn ngôn ngữ

STT	Input	Output	Đúng với output mong đợi
1	Microsoft đã ra mắt phiên bản Windows 11 mới với nhiều cải tiến đáng kể về hiệu năng và giao diện. The new update also includes enhanced security features, making it more reliable for both personal and professional use. Đây là bước tiến lớn trong công nghệ hệ điều hành.	Vi tính	X
2	Vào ngày 10/12, Tòa án Nhân dân thành phố HCM sẽ tổ chức phiên tòa xét xử một vụ án kinh tế lớn liên quan đến nhiều cá nhân và tổ chức. The indictment states that the defendants violated Clause 3, Article 289 of the Penal Code. Đây là vụ án nhận được sự quan tâm lớn từ dư luận.	Pháp luật	X
3	Bộ Y tế vừa ban hành khuyến cáo mới về việc tiêm phòng cúm mùa. Vaccination is highly recommended for people over 60 and those with chronic illnesses. Đây là một bước quan trọng để giảm nguy cơ mắc các biến chứng nặng do bệnh cúm.	Sức khỏe	X

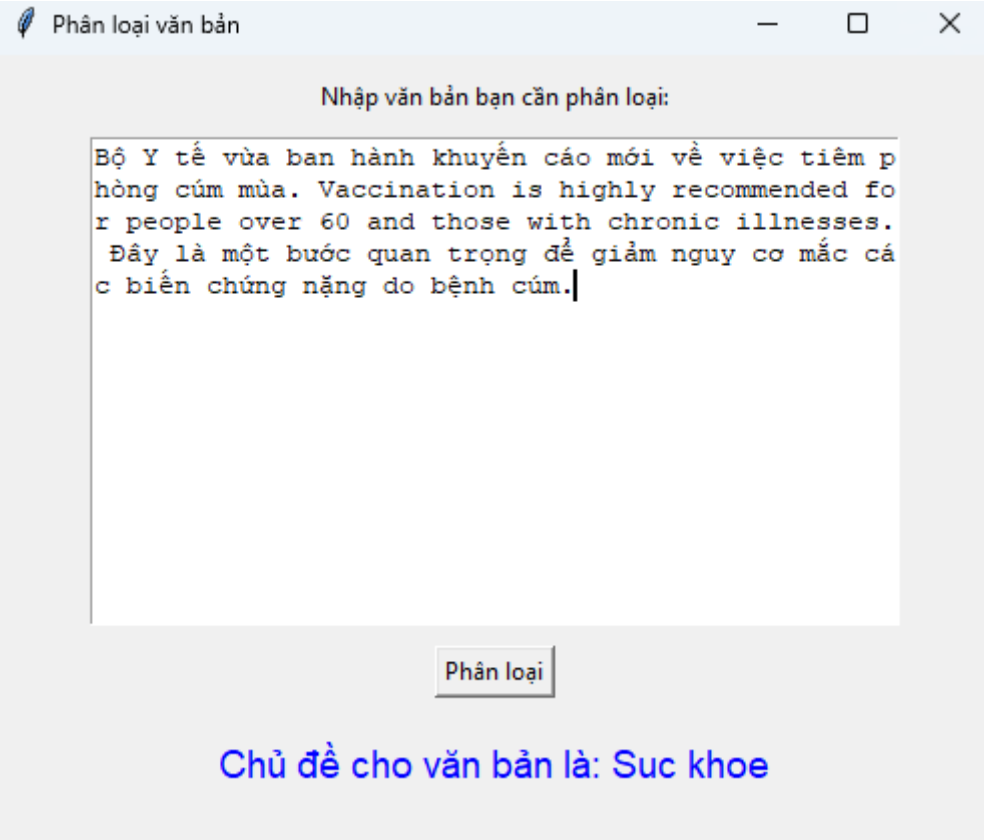
Người dùng tiến hành nhập hoặc sao chép văn bản vào phần mềm của hệ thống. Kết quả kiểm thử của quá trình phân loại tìm nhãn cho một văn bản lẫn lộn nhiều ngôn ngữ được minh họa chi tiết như các ***Hình 4.24***, ***Hình 4.25***, ***Hình 4.26*** bên dưới:



Hình 4.24: Kết quả kiểm thử phân loại tìm nhãn cho văn bản lẫn lộn ngôn ngữ 1



Hình 4.25: Kết quả kiểm thử phân loại tìm nhãn cho văn bản lẫn lộn ngôn ngữ 2



Hình 4.26: Kết quả kiểm thử phân loại tìm nhãn cho văn bản lẫn lộn ngôn ngữ 3

4.3. Đánh giá mô hình

4.3.1. Accuracy – Sự chính xác

Chia thành hai thành phần training và testing áp dụng một mô hình để train từ tập dữ liệu training. Tiếp theo sử dụng mô hình dự đoán trên tập testing và cuối cùng là tìm ra tỉ lệ số dữ liệu dự đoán đúng / tổng số dữ liệu kiểm thử.

$$Accuracy = \frac{Số\ dự\ đoán\ đúng}{Tổng\ số\ mẫu}$$

4.3.2. Precision – Sự đồng nhất

Precision là tỷ lệ dự đoán đúng trong số mẫu được dự đoán là nhãn đó.

Trong những bài toán phân loại, người ta thường định nghĩa lớp dữ liệu quan trọng hơn cần được xác định là lớp Positive, lớp còn lại là Negative.

Bảng 4.17 Dự đoán precision

		Model dự đoán	
		Positive	**Negative**
Thực tế	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Dựa vào ***Bảng 4.17*** công thức tính Precision cho một lớp cụ thể và Recal được tính như sau:

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

Xem xét trên tập dữ liệu kiểm thử (data testing) xem có bao nhiêu dữ liệu được mô hình dự đoán đúng. Tức là, số phát hiện đúng chia cho số đem đi kiểm thử. Đây chính là chỉ số accurary – độ chính xác như bên trên. Giá trị này càng cao cao tốt.

$$Precision = \frac{\gamma_{true} \cap \gamma_{selected}}{\gamma_{selected}}$$

4.3.3. Recall – Phủ định

Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive. Recall thể hiện tỉ lệ dự đoán chính xác trên tổng số mẫu thực sự thuộc nhãn đó.

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

4.3.4. F1-Score

Giá trị F1-Score được tính theo công thức sau:

$$F1 - Score = 2.\frac{Precision.Recall}{Precision + Recall}$$

F1-Score là trung bình điều hòa của các tiêu chí Precison và Recall. Nó có xu hướng lấy giá trị gần với giá trị nào nhỏ hơn giữa hai giá trị Precision và Recall và đồng thời nó có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn. Chính vì thế F1-Score thể hiện được khách quan hơn hiệu suất của một mô hình học máy.

4.3.5. Đánh giá mô hình

Đối với bài toán trong đề tài này, thuật toán được đánh giá qua các cách đánh giá Precision, Recall và F1-Score để đánh giá mô hình, thông tin chi tiết giá trị từng thông số của các loại nhãn được thể hiện chi tiết như ***Bảng 4.18*** bên dưới:

Bảng 4.18: Bảng kết quả đánh giá các thông số của mô hình Naive Bayes

Nhãn	Precision	Recall	F1-Score
Chính trị xã hội	0.67	0.92	0.77
Đời sống	0.82	0.87	0.84
Khoa học	0.97	0.50	0.66
Kinh doanh	0.93	0.74	0.82
Pháp luật	0.90	0.87	0.88
Sức khỏe	0.86	0.90	0.88

Nhãn	Precision	Recall	F1-Score
Thế giới	0.92	0.83	0.87
Thể thao	0.98	0.96	0.97
Văn hóa	0.93	0.88	0.90
Vi tính	0.96	0.87	0.92

Các nhãn "Thể thao", "Văn hóa", "Vi tính", "Thế giới" cho kết quả khá tốt, mô hình hoạt động hiệu quả trên các này. Các nhãn "Khoa học" và "Chính trị Xã hội" cho kết quả thấp không như mong đợi. Mô hình có thể nhầm lần các trường hợp không thuộc nhãn này.

4.4. Tổng kết chương

Nhìn chung, hệ thống đã được xây dựng với một số chức năng cơ bản cho quá trình phân loại văn bản tiếng Việt tự động như: chức năng tách từ, xóa các từ dừng, tính TF-IDF, tiền xử lý dữ liệu, phân loại tìm nhãn chủ đề cho một văn bản tiếng Việt. Tuy nhiên, vẫn còn một số thiếu sót cần được khắc phục: văn bản chưa được xử lý một cách tối ưu hóa, một số chủ đề có độ chính xác kết quả không được cao.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

5.1.1. Kết quả đạt được

- Tìm hiểu rõ được cấu trúc cấu tạo từ, phân loại từ và xây dựng được các chức năng cơ bản cho các quá trình của việc phân loại văn bản với tiếng Việt.
- Nêu rõ quy trình tách từ, các phương pháp tách từ, kết quả tách từ cho kết quả chính xác đúng với mong đợi.
- Hệ thống phân loại có thể xử lý các văn bản chứa nhiều thông tin gây nhiễu một cách chính xác tương tối cao.
- Hệ thống tiền xử lý dữ liệu đầu vào cho văn bản cho kết quả nhanh và hiệu quả, đạt kết quả chính xác cao.
- Xây dựng được hệ thống phân loại văn bản tiếng Việt giúp phân loại được chính xác các bài báo giúp người đọc có thể tiết kiệm thời gian khi tìm kiếm các bài báo, văn bản liên quan với nhau có cùng chủ đề.
- Phân loại các văn bản với thời gian nhanh chóng và cho kết quả chính xác cao với thuật toán Naive Bayes.

5.1.2. Hạn chế

Bên cạnh những chức năng cơ bản đạt được, hệ thống còn một số hạn chế cần được khắc phục:

- Mô hình có thể không đạt được tối ưu hóa khi mở rộng với tập dữ liệu quá lớn.
- Mô hình hiện tại chỉ dựa trên các đặc trưng của từ vựng mà không phân tích sâu về mặt ngữ cảnh và ngữ nghĩa nên đã dẫn đến những thiếu xót khi phân loại văn bản có ngữ nghĩa phức tạp.
- Chức năng xóa các từ dừng chưa được tối ưu hóa và chỉ được trích lọc từ dữ liệu, chưa phân tích rõ ngữ cảnh từ trong câu.

5.2. Hướng phát triển

- Thêm các mô hình học sâu để phân tích ngữ cảnh của văn bản một cách chính xác hơn.
- Áp dụng các mô hình học sâu vào chức năng xóa từ dừng để hệ thống hiểu rõ ngữ cảnh và ngữ nghĩa của văn bản, từ đó tối ưu hóa chức năng xóa từ dừng cho hệ thống.
- Tối ưu hóa các bước tiền xử lý dữ liệu, tăng cường bộ huấn luyện dữ liệu để cải thiện độ chính xác khi mô hình tiếp xúc với các tình huống thực tế đa dạng.
- Mở rộng mô hình với các tập dữ liệu lớn đúng với thực tế.
- Xây dựng hệ thống tích hợp khả năng học liên tục để cải thiện kết quả khi có dữ liệu mới.

TÀI LIỆU THAM KHẢO

[1] T. O. F. J. Zhang, "Text Categorization Based on Regularized Linear Classification Methods," *Information Retrieval*, 2001/04/01.

[2] T. Joachims, "Text Categorization with Support Vector Machines," *Proc. European Conf. Machine Learning (ECML'98)*, 1998.

[3] L. V. Nguyên, "PHÂN LỚP VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG THEO CHỦ ĐỀ," 2020.

[4] N. C. Hiếu, "KHẢO SÁT CÁC MÔ HÌNH PHÂN LOẠI VĂN BẢN TIẾNG VIỆT," *Journal of Science and Technology - IUH. 57. 10.46242/jstiuh.v57i03.4395,* 2022.

[5] H. B. Q. Đinh Điền, "VẤN ĐỀ VỀ RANH GIỚI TỪ TRONG NGỮ LIỆU SONG NGỮ ANH-VIỆT," Báo cáo Hội thảo Khoa học "Bảo vệ và Phát triển tiếng Việt". Viện Ngôn ngữ học, Khoa CNTT, ĐH Khoa học Tự nhiên – ĐHQG Tp.HCM, 2002.

[6] Đ. Điền, "Xây dựng và khai thác ngữ liệu song ngữ Anh-Việt điện tử," luận án tiến sĩ ngôn ngữ học so sánh, ĐH Khoa học Xã hội & Nhân văn, ĐHQG Tp.HCM, 3/2005.

[7] C. W. v. T. C. Surapant Meknavin, "Automatic Thai Word Segmentation Using Thai Dictionary-Based Approach," 1997.

[8] C.-H. Tsai, "A Maximum Matching Algorithm for Chinese Word Segmentation," 1996.

[9] B. M. S. R. a. M. S. F. Jelinek, A Dynamic Language Model for Speech Recognition, Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, {F}ebruary 19-22, 1991, 1991.

[10] K. H. V.-T. N. D. Dinh, Application of Maximum matching and SVMs for Vietnamese word segmentation, 2006.

[11] L. T. M. H. N. R. A. V. H. (. Hông Phuong, "A Hybrid Approach to Word Segmentation of Vietnamese Texts," In: Martín-Vide, C., Otto, F., Fernau, H. (eds) Language and Automata Theory and Applications. LATA 2008. Lecture Notes in Computer Science, vol 5196. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88282-4_23, 2008.

[12] M. &. N. T. (. Le, "Underthesea: A Python library for Vietnamese NLP tasks," ACL Workshop on NLP for Social Media.

[13] H. Zhang, "The Optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004. 2*, 2004.

[14] "scikit-learn," [Trực tuyến]. Available: https://scikit-learn.org/1.5/modules/naive_bayes.html. [Đã truy cập 21 11 2024].

[15] duyvuleo. [Trực tuyến]. Available: https://github.com/duyvuleo/VNTC/tree/master/Data/10Topics/Ver1.1. [Đã truy cập 10 10 2024].

[16] P. T. P. Đ. T. N. T. N. M. T. Phạm Nguyên Khang, "SỰ ẢNH HƯỞNG CỦA PHƯƠNG PHÁP TÁCH TỪ TRONG BÀI TOÁN PHÂN LỚP VĂN BẢN TIẾNG VIỆT," Tạp chí Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin FAIR 2016, ĐH. Cần Thơ 8/2016, 2016.

PHỤ LỤC

Bảng phụ lục 1: Danh sách từ các từ dừng được trích ra từ dữ liệu

STT	Từ dừng	STT	Từ dừng	STT	Từ dừng	STT	Từ dừng
1	Của	26	Phải	51	Biết	76	Đây
2	Và	27	Anh	52	Trước	77	Tới
3	Là	28	Ra	53	Việc	78	Cả
4	Có	29	Nhiều	54	Hai	79	Qua
5	Các	30	Từ	55	Sự	80	Lần
6	trong	31	Năm	56	Bạn	81	Chưa
7	Được	32	Nhưng	57	Đang	82	Tiền
8	Đã	33	Trên	58	Đi	83	Bằng
9	Cho	34	Ông	59	Mới	84	Số
10	Một	35	Tại	60	Họ	85	Thấy
11	Không	36	Sao	61	Nước	86	Tháng
12	Với	37	Lại	62	Nhất	87	Cùng
13	Người	38	Bị	63	Vì	88	Hay
14	Những	39	Còn	64	Vẫn	89	Chị
15	ở	40	Làm	65	Nhà	90	Cao
16	Khi	41	Như	66	Khác	91	Cô
17	Này	42	Theo	67	Lên	92	Cần
18	Để	43	Thì	68	Nên	93	Gì
19	Sẽ	44	Hơn	69	Nói	94	Điều
20	Cũng	45	Chỉ	70	Rằng	95	Việt_nam
21	Đến	46	Có_thể	71	Nếu	96	1
22	Về	47	Ngày	72	Do	97	2
23	Tôi	48	Rất	73	Trận	98	3
24	Vào	49	Mình	74	Con	99	4
25	Đó	50	Mà	75	Đội	100	5