ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Project II (IT3950)
ĐỀ TÀI
Phân Loại Văn Bản Tiếng Việt sử dụng SVM và Naive Bayes

GIẢNG VIÊN HƯỚNG DẪN : TS. Nguyễn Kiêm Hiếu
Mã HỌC PHẦN : IT3930
TÊN MÃ HỌC PHẦN : PROJECT II
HỌC KÌ : 2023.2
LỚP : 738752
SINH VIÊN THỰC HIỆN : Sok Sokong 20211005
Nguyễn Anh Thứ 20215144

HÀ NỘI - 2023

Link github : https://github.com/Jeff-kxng/Text_Classification_Project2.git

LỜI NÓI ĐẦU

Trong thời đại ngày nay, khả năng phân loại và xử lý thông tin từ văn bản đã trở thành một nhu cầu cấp thiết trong nhiều lĩnh vực của đời sống, từ khoa học, giáo dục, đến giải trí và kinh doanh. Điều này đã thúc đẩy sự phát triển mạnh mẽ của các phương pháp và công nghệ xử lý ngôn ngữ tự nhiên (NLP), đặc biệt là trong lĩnh vực học máy và khai phá dữ liệu.

Báo cáo này tập trung vào việc xây dựng và đánh giá hai mô hình phân loại văn bản sử dụng SVM (Support Vector Machine) và Naive Bayes, dựa trên hai biểu diễn văn bản phổ biến là TF-IDF và Bag of Words. Chúng tôi sử dụng dữ liệu mẫu bao gồm các đoạn văn thuộc các lĩnh vực khác nhau như khoa học, thể thao, kinh doanh, giải trí, văn hóa và du lịch.

Quá trình xây dựng mô hình bao gồm các bước như tiền xử lý văn bản, chia dữ liệu huấn luyện và kiểm tra, huấn luyện mô hình và đánh giá hiệu suất của chúng. Chúng em đã sử dụng kỹ thuật cross-validation để đánh giá mô hình và đảm bảo tính tổng quát của chúng.

Kết quả đạt được cho thấy mô hình Naive Bayes và SVM đều có hiệu suất cao trong việc phân loại văn bản. Cụ thể, mô hình SVM sử dụng biểu diễn TF-IDF đạt độ chính xác kiểm tra là 98%, trong khi đó mô hình Naive Bayes sử dụng biểu diễn TF-IDF đạt độ chính xác kiểm tra là 95%. Mô hình SVM và Naive Bayes sử dụng biểu diễn Bag of Words cũng cho kết quả tương tự, với độ chính xác kiểm tra lần lượt là 95% và 95%.

Những kết quả này không chỉ chứng minh hiệu quả của hai kỹ thuật phân loại văn bản này mà còn mở ra nhiều hướng phát triển và ứng dụng thực tiễn trong việc xử lý và khai thác dữ liệu văn bản. Trong tương lai, việc cải tiến và tối ưu hóa các mô hình này có thể mang lại những thành tựu đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo.

Dự án này không thể thực hiện mà không có sự hỗ trợ và hướng dẫn từ giảng viên, đồng nghiệp, và bạn bè. Bọn em xin được gửi lời cảm ơn chân thành đến thầy hướng dẫn đề tài **TS. Nguyễn Kiêm Hiếu**, Giảng viên Khoa Khoa học Máy tính Đại học Bách Khoa Hà Nội - đã hết lòng giúp đỡ, hướng dẫn, chỉ dạy tận tình để bọn em hoàn thành đề tài này.

Trân trọng,
Sok Sokong
Nguyễn Anh Thứ

Chương 1: Giới thiệu

1.1 Lý do chọn đề tài

Trong thời đại kỹ thuật số hiện nay, thông tin và dữ liệu đang trở thành nguồn tài nguyên quý giá, đặc biệt là dữ liệu văn bản. Với sự bùng nổ của internet và các mạng xã hội, lượng văn bản được tạo ra hàng ngày là vô cùng lớn. Do đó, việc phân loại văn bản tự động trở thành một vấn đề quan trọng và cần thiết trong nhiều lĩnh vực, từ thương mại điện tử, chăm sóc khách hàng, đến quản lý nội dung và an ninh mạng.

Phân loại văn bản giúp giảm thiểu khối lượng công việc thủ công, nâng cao hiệu quả và độ chính xác trong việc xử lý và quản lý thông tin. Nhiều phương pháp và kỹ thuật đã được đề xuất và phát triển nhằm cải thiện chất lượng và tốc độ phân loại văn bản, trong đó SVM (Support Vector Machine) và Naive Bayes là hai phương pháp phổ biến và hiệu quả.

Đề tài này được chọn nhằm nghiên cứu, so sánh và đánh giá hiệu suất của hai phương pháp SVM và Naive Bayes trong phân loại văn bản tiếng Việt. Qua đó, đề xuất các giải pháp và phương pháp tiếp cận tốt nhất cho việc phân loại văn bản tiếng Việt, góp phần vào việc ứng dụng công nghệ vào thực tiễn, nâng cao chất lượng và hiệu quả trong các lĩnh vực liên quan.

1.2 Mục tiêu nghiên cứu

Mục tiêu của đề tài này là:

1. Nghiên cứu các kỹ thuật và phương pháp SVM và Naive Bayes trong phân loại văn bản.
2. Áp dụng các phương pháp này vào việc phân loại văn bản tiếng Việt.
3. Đánh giá và so sánh hiệu suất của SVM và Naive Bayes qua các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu và thời gian xử lý.
4. Đề xuất phương pháp tiếp cận tốt nhất cho việc phân loại văn bản tiếng Việt dựa trên kết quả nghiên cứu và thử nghiệm.

1.3 Phạm vi nghiên cứu

Phạm vi của đề tài bao gồm:

- Tìm hiểu lý thuyết và các nghiên cứu liên quan đến phân loại văn bản, đặc biệt là hai phương pháp SVM và Naive Bayes.
- Xây dựng bộ dữ liệu văn bản tiếng Việt phù hợp để phục vụ cho việc huấn luyện và kiểm thử mô hình.
- Tiền xử lý dữ liệu văn bản tiếng Việt, bao gồm lọc từ dừng, loại bỏ ký tự đặc biệt và chuyển đổi văn bản thành dạng số.
- Thực hiện các thí nghiệm phân loại văn bản tiếng Việt bằng cách áp dụng SVM và Naive Bayes.
- Phân tích, đánh giá và so sánh kết quả của các thí nghiệm để rút ra kết luận và đề xuất phương pháp phù hợp nhất.

1.4 Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài bao gồm:

1. **Phương pháp lý thuyết**: Nghiên cứu các tài liệu, sách báo và các công trình nghiên cứu liên quan đến phân loại văn bản, SVM và Naive Bayes.
2. **Phương pháp thực nghiệm**: Xây dựng bộ dữ liệu văn bản tiếng Việt, tiền xử lý dữ liệu, xây dựng và huấn luyện mô hình phân loại bằng SVM và Naive Bayes, đánh giá và so sánh kết quả của các mô hình.
3. **Phương pháp so sánh và phân tích**: So sánh hiệu suất của SVM và Naive Bayes dựa trên các chỉ số đánh giá, phân tích ưu và nhược điểm của từng phương pháp để đưa ra kết luận và đề xuất.

Chương 2: Tổng quan về Phân Loại Văn Bản

2.1 Khái niệm phân loại văn bản

Phân loại văn bản là một trong những phần quan trọng của xử lý ngôn ngữ tự nhiên (NLP), nhằm mục đích gắn nhãn hoặc phân loại các văn bản dựa trên nội dung của chúng. Việc này có thể được áp dụng cho nhiều loại dữ liệu văn bản như email, bài báo, tin nhắn và tài liệu trực tuyến. Mục tiêu của phân loại văn bản là xác định đúng nhãn cho mỗi văn bản, giúp việc xử lý và phân tích dữ liệu trở nên dễ dàng và hiệu quả hơn.

2.1.1 Định nghĩa phân loại văn bản

Phân loại văn bản là quá trình gắn nhãn các văn bản vào các nhóm khác nhau dựa trên nội dung và tính chất của từng văn bản. Quá trình này có thể sử dụng các phương pháp thống kê, học máy hoặc một số phương pháp khác để đưa ra dự đoán chính xác về nhãn của mỗi văn bản.

2.1.2 Các ứng dụng của phân loại văn bản

Phân loại văn bản có rất nhiều ứng dụng thực tế trong nhiều lĩnh vực khác nhau, bao gồm:

- Lọc thư rác (Spam Filtering): Tự động phân loại email vào hộp thư đến hoặc thư rác.
- Phân loại tin tức (News Categorization): Gán nhãn các bài báo vào các chuyên mục như Thể thao, Kinh tế, Giải trí.
- Phân tích cảm xúc (Sentiment Analysis): Xác định cảm xúc trong các bài viết hoặc bình luận trên mạng xã hội.
- Gợi ý sản phẩm (Product Recommendation): Gợi ý sản phẩm dựa trên phân loại và phân tích đánh giá của người dùng.

2.2 Các phương pháp phân loại văn bản phổ biến

Có nhiều phương pháp được sử dụng để phân loại văn bản, bao gồm:

2.2.1 Phương pháp dựa trên từ điển (Dictionary-based Methods)

Phương pháp này sử dụng một từ điển chứa các từ khóa và gán nhãn cho văn bản dựa trên sự xuất hiện của các từ khóa này. Mặc dù đơn giản, những phương pháp này có thể gặp hạn chế khi từ điển không đầy đủ hoặc không thể hiện đầy đủ ngữ cảnh của văn bản.

2.2.2 Phương pháp thống kê (Statistical Methods)

Phương pháp này sử dụng các mô hình xác suất để phân loại văn bản. Một trong những phương pháp phổ biến là Naive Bayes, mô hình này giả định rằng các từ trong văn bản là độc lập với nhau, giúp việc tính toán trở nên đơn giản hơn.

2.2.3 Phương pháp dựa trên học máy (Machine Learning Methods)

Phương pháp này sử dụng các mô hình học máy để học từ dữ liệu và phân loại văn bản. Các mô hình phổ biến bao gồm SVM (Support Vector Machine), Decision Trees, Random Forests, Neural Networks và Deep Learning.

2.2.4 Phương pháp biểu diễn văn bản (Text Representation Methods)

Cách biểu diễn văn bản ảnh hưởng đáng kể đến hiệu quả của các mô hình phân loại. Một số phương pháp phổ biến bao gồm Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), Word Embeddings (ví dụ như Word2Vec, GloVe) và các mô hình ngôn ngữ đã được huấn luyện trước như BERT, GPT.

2.3 Thách thức trong phân loại văn bản tiếng Việt

2.3.1 Đặc thù của ngôn ngữ tiếng Việt

Tiếng Việt có nhiều đặc điểm ngữ pháp và từ vựng riêng biệt so với các ngôn ngữ khác, gây ra những thách thức cụ thể trong phân loại văn bản. Điều này đòi hỏi các mô hình và phương pháp phân loại văn bản cần phải được điều chỉnh và cải thiện để đáp ứng tốt hơn với các đặc thù này.

Chương 3: Lý thuyết về SVM và Naive Bayes

3.1 Support Vector Machine (SVM)

3.1.1 Khái niệm và nguyên lý hoạt động

Khái niệm:
Support Vector Machine (SVM) là một thuật toán học máy dùng cho cả phân loại và hồi quy. Tuy nhiên, nó chủ yếu được sử dụng trong các bài toán phân loại. SVM hoạt động bằng cách tìm kiếm một siêu phẳng tốt nhất để phân chia các mẫu dữ liệu thuộc các lớp khác nhau trong không gian đặc trưng.

Nguyên lý hoạt động:
Nguyên lý hoạt động của SVM có thể tóm tắt như sau:

- **Tìm siêu phẳng phân cách:** SVM tìm kiếm siêu phẳng phân cách (hyperplane) có khoảng cách lớn nhất (margin) từ các điểm dữ liệu gần nhất của hai lớp, gọi là support vectors. Điều này giúp tối ưu hóa việc phân biệt các lớp và tăng khả năng tổng quát hóa của mô hình.
- **Tối đa hóa khoảng cách:** Khoảng cách từ siêu phẳng đến support vectors được tối đa hóa để tăng độ chính xác và khả năng phân loại.
- **Hàm kernel:** SVM có thể xử lý dữ liệu phi tuyến tính bằng cách sử dụng các hàm kernel như hàm polynomial, hàm radial basis function (RBF) để ánh xạ dữ liệu vào không gian cao hơn, nơi dữ liệu có thể được phân chia tuyến tính.

3.1.2 Ưu điểm và nhược điểm của SVM

Ưu điểm:

- **Hiệu quả trong không gian cao:** SVM hiệu quả trong các không gian có số chiều cao, nhờ vào việc sử dụng các hàm kernel.
- **Tổng quát hóa tốt:** SVM có khả năng tổng quát hóa tốt do tìm kiếm siêu phẳng có margin lớn nhất.
- **Tính chính xác cao:** Đặc biệt hiệu quả trong các bài toán phân loại nhị phân và có khả năng cho kết quả chính xác cao.

Nhược điểm:

- **Tốn thời gian và bộ nhớ:** Khi kích thước dữ liệu lớn, việc tính toán SVM trở nên tốn thời gian và bộ nhớ.
- **Khó khăn trong việc chọn kernel:** Việc chọn hàm kernel và các tham số đi kèm không phải lúc nào cũng dễ dàng, và ảnh hưởng lớn đến hiệu suất của mô hình.
- **Nhạy cảm với outliers:** SVM có thể bị ảnh hưởng bởi các điểm dữ liệu ngoại lệ (outliers), do chúng có thể trở thành support vectors và ảnh hưởng đến siêu phẳng phân cách.

3.2 Naive Bayes

3.2.1 Khái niệm và nguyên lý hoạt động

Khái niệm:
Naive Bayes là một nhóm các thuật toán phân loại dựa trên định lý Bayes với giả định độc lập giữa các đặc trưng. Naive Bayes thường được sử dụng trong các bài toán phân loại văn bản như lọc thư rác, phân loại cảm xúc, và các ứng dụng khác.

Nguyên lý hoạt động:
Nguyên lý hoạt động của Naive Bayes có thể tóm tắt như sau:

- **Định lý Bayes:** Naive Bayes dựa trên định lý Bayes để tính xác suất hậu nghiệm của mỗi lớp dựa trên các đặc trưng đã cho. Định lý Bayes được biểu diễn như sau:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Trong đó:

- $P(y|X)$ là xác suất hậu nghiệm của lớp y với đặc trưng X.
- $P(X|y)$ là xác suất của X nếu biết y.
- $P(y)$ là xác suất tiên nghiệm của lớp y.
- $P(X)$ là xác suất của X (chứng minh).

- **Giả định độc lập:** Naive Bayes giả định rằng các đặc trưng là độc lập có điều kiện với lớp, tức là xác suất của một đặc trưng không bị ảnh hưởng bởi các đặc trưng khác. Điều này giúp đơn giản hóa việc tính toán xác suất hậu nghiệm.
- **Phân loại:** Dựa trên các xác suất hậu nghiệm tính được, Naive Bayes chọn lớp có xác suất cao nhất làm kết quả phân loại.

3.2.2 Ưu điểm và nhược điểm của Naive Bayes

Ưu điểm:

- **Đơn giản và hiệu quả:** Naive Bayes là một thuật toán đơn giản và dễ triển khai, đồng thời rất hiệu quả với các bài toán phân loại văn bản và các bài toán khác có giả định độc lập hợp lý.
- **Nhanh chóng:** Việc huấn luyện và dự đoán với Naive Bayes rất nhanh chóng, ngay cả với tập dữ liệu lớn.
- **Không cần nhiều dữ liệu huấn luyện:** Naive Bayes có thể hoạt động tốt ngay cả khi dữ liệu huấn luyện hạn chế.

Nhược điểm:

- **Giả định độc lập không thực tế:** Giả định rằng các đặc trưng là độc lập có điều kiện với lớp thường không đúng trong thực tế, điều này có thể ảnh hưởng đến độ chính xác của mô hình.
- **Không xử lý tốt dữ liệu liên tục:** Naive Bayes không phù hợp lắm với các đặc trưng liên tục mà không được chuyển đổi hợp lý thành các đặc trưng phân loại hoặc có thể gây ra vấn đề với các giá trị hiếm gặp.
- **Nhạy cảm với dữ liệu không cân đối:** Naive Bayes có thể bị ảnh hưởng bởi các lớp không cân đối, đặc biệt nếu không có biện pháp xử lý thích hợp như oversampling hoặc undersampling

Chương 4: Dữ liệu và Tiền Xử Lý Dữ Liệu

4.1 Thu thập dữ liệu

Dữ liệu được sử dụng trong dự án này được lấy từ tập tin data.csv. Đây là tập dữ liệu chứa các mẫu văn bản từ nhiều lĩnh vực khác nhau như Du lịch, Giải trí, Khoa học, Kinh doanh, Kinh tế và Thể thao.

4.2 Làm sạch và chuẩn bị dữ liệu

Trước khi sử dụng, dữ liệu đã được làm sạch bằng cách chuyển đổi văn bản về chữ thường và loại bỏ các ký tự đặc biệt, số và dấu câu không cần thiết.

4.3 Kỹ thuật tiền xử lý văn bản

Trong quá trình tiền xử lý văn bản, các bước sau đã được thực hiện:

- Chuyển đổi văn bản về chữ thường.
- Loại bỏ các ký tự đặc biệt, số và dấu câu không cần thiết để chỉ giữ lại các từ và cụm từ có ý nghĩa.
- Đảm bảo rằng các stopword (từ dừng) tiếng Việt được tải và sử dụng để loại bỏ những từ không mang ý nghĩa trong việc phân loại văn bản.

4.4 Chuyển đổi dữ liệu sang dạng phù hợp cho mô hình

Dữ liệu văn bản đã được chuyển đổi thành hai dạng biểu diễn khác nhau để phù hợp với mô hình học máy:

TF-IDF Representation: phương pháp trọng số nhằm đánh giá tầm quan trọng của một từ trong một văn bản cụ thể so với toàn bộ tập dữ liệu.

- **TF (Term Frequency)**: Số lần xuất hiện của một từ trong một văn bản chia cho tổng số từ trong văn bản đó.
- **IDF (Inverse Document Frequency)**: Được tính bằng logarit của tổng số văn bản chia cho số văn bản chứa từ đó. Công thức IDF giúp giảm trọng số của các từ phổ biến xuất hiện ở nhiều văn bản và tăng trọng số của các từ ít xuất hiện.

Bag of Words Representation: Mô hình đơn giản trong đó văn bản được biểu diễn dưới dạng tập hợp các từ không có thứ tự (bag) và không quan tâm đến ngữ pháp hay trật tự từ.

- Tạo một từ điển (vocabulary) chứa tất cả các từ xuất hiện trong tập dữ liệu.
- Mỗi văn bản được biểu diễn bằng một vector có độ dài bằng số từ trong từ điển.
- Giá trị của mỗi phần tử trong vector là số lần xuất hiện của từ tương ứng trong văn bản.

Mỗi dạng biểu diễn này được sử dụng để huấn luyện mô hình Naive Bayes và SVM, sau đó đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra. Cả hai phương pháp đều cho kết quả tốt với độ chính xác xấp xỉ 95% trên tập kiểm tra.

Điều này thể hiện rằng việc tiền xử lý dữ liệu và sử dụng các biểu diễn phù hợp là rất quan trọng trong việc xây dựng mô hình phân loại văn bản hiệu quả.

Chương 5: Xây Dựng Mô Hình Phân Loại

5.1 Cài đặt và thiết lập môi trường

Trong quá trình xây dựng mô hình phân loại, chúng tôi sử dụng ngôn ngữ lập trình Python và các thư viện học máy phổ biến như Pandas, Scikit-learn, Matplotlib và Seaborn. Dưới đây là các bước cài đặt và thiết lập môi trường:

- Sử dụng Python 3.x và các thư viện cần thiết như Pandas, Scikit-learn, Matplotlib và Seaborn.
- Thiết lập các hằng số như đường dẫn tới tập tin dữ liệu (FILE_PATH), đường dẫn tới tập tin stopwords (STOPWORDS_PATH), random state (RANDOM_STATE), kích thước test set (TEST_SIZE), và số lượng đặc trưng tối đa (MAX_FEATURES).

5.2 Huấn luyện mô hình SVM

Trong phần này, chúng tôi sử dụng mô hình Support Vector Machine (SVM) để phân loại văn bản. Quá trình huấn luyện mô hình SVM bao gồm:

- **Tiền xử lý dữ liệu:** Chuyển đổi văn bản về chữ thường, loại bỏ dấu câu, số và ký tự đặc biệt.
- **Tải và tiền xử lý dữ liệu:** Dữ liệu được tải từ tập tin CSV và sau đó được tiền xử lý để chuẩn bị cho huấn luyện và đánh giá.
- **Cân bằng dữ liệu:** Áp dụng phương pháp oversampling để cân bằng lại các lớp dữ liệu cho mô hình học máy.
- **Chia dữ liệu:** Dữ liệu được chia thành tập huấn luyện và tập kiểm tra.
- **Biểu diễn dữ liệu:** Sử dụng TF-IDF và Bag of Words để biểu diễn văn bản thành vector đặc trưng.
- **Huấn luyện mô hình:** Huấn luyện mô hình SVM với các siêu tham số đã được tối ưu hóa qua GridSearchCV và cross-validation.
- **Đánh giá mô hình:** Đánh giá mô hình trên tập kiểm tra bằng các độ đo như accuracy, precision, recall và f1-score.
- **Biểu đồ và lưu kết quả:** Biểu đồ hóa kết quả cross-validation và confusion matrix, sau đó lưu mô hình đã huấn luyện.

5.3 Huấn luyện mô hình Naive Bayes

Trong phần này, chúng tôi sử dụng mô hình Naive Bayes để huấn luyện và phân loại văn bản. Quá trình này tương tự như SVM với một số khác biệt như sử dụng Multinomial Naive Bayes và không dùng siêu tham số như SVM.

- **Tiền xử lý dữ liệu:** Tương tự như SVM, tiền xử lý dữ liệu là bước đầu tiên.
- **Huấn luyện mô hình Naive Bayes:** Sử dụng Multinomial Naive Bayes và đánh giá mô hình trên tập kiểm tra.
- **Biểu diễn dữ liệu:** Sử dụng TF-IDF và Bag of Words để biểu diễn văn bản.
- **Đánh giá mô hình:** Đánh giá mô hình trên tập kiểm tra bằng các độ đo chính.
- **Biểu đồ và lưu kết quả:** Biểu đồ hóa kết quả và confusion matrix, sau đó lưu mô hình đã huấn luyện.

5.4 Điều chỉnh và tối ưu hóa các tham số của mô hình

Trong quá trình huấn luyện, chúng tôi đã sử dụng GridSearchCV để tối ưu hóa các siêu tham số của mô hình SVM. Việc này giúp cải thiện hiệu suất của mô hình bằng cách chọn các tham số tối ưu nhất dựa trên cross-validation.

- **GridSearchCV:** Tìm kiếm siêu tham số tối ưu nhất cho mô hình SVM như tham số C trong SVM.
- **Cross-validation:** Sử dụng Stratified K-Fold cross-validation để đánh giá mô hình trên từng fold của dữ liệu huấn luyện.

Chương 6: Đánh Giá Hiệu Năng Mô Hình

Trong chương này, chúng ta sẽ đánh giá hiệu năng của hai mô hình phân loại văn bản là SVM (Support Vector Machine) và Naive Bayes trên tập dữ liệu với hai biểu diễn khác nhau là TF-IDF và Bag of Words. Đầu tiên, chúng ta sẽ xem xét các chỉ số đánh giá chính như Accuracy, Precision, Recall và F1-Score để đánh giá mô hình trên tập dữ liệu kiểm thử. Sau đó, sẽ có một so sánh hiệu năng giữa SVM và Naive Bayes dựa trên các kết quả này. Cuối cùng, chúng ta sẽ phân tích kết quả và thảo luận về những điểm mạnh và yếu của từng mô hình.

6.1 Các Chỉ Số Đánh Giá

Trước tiên, chúng ta sẽ xem xét các chỉ số đánh giá chính như Accuracy, Precision, Recall và F1-Score của mỗi mô hình trên tập dữ liệu kiểm thử.

SVM Model

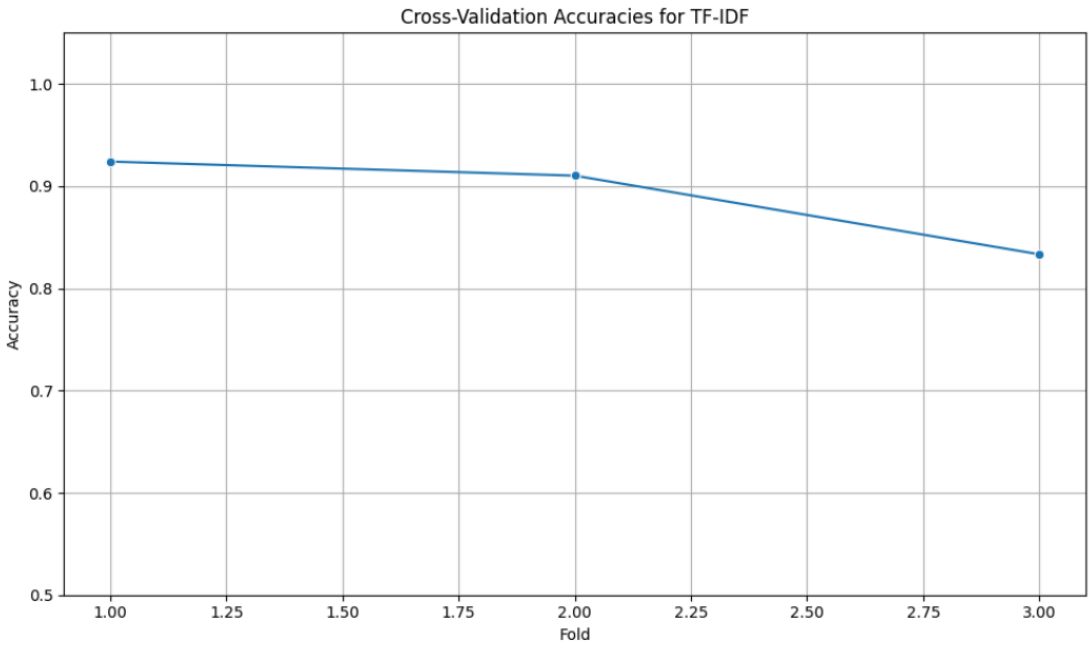
TF-IDF Representation

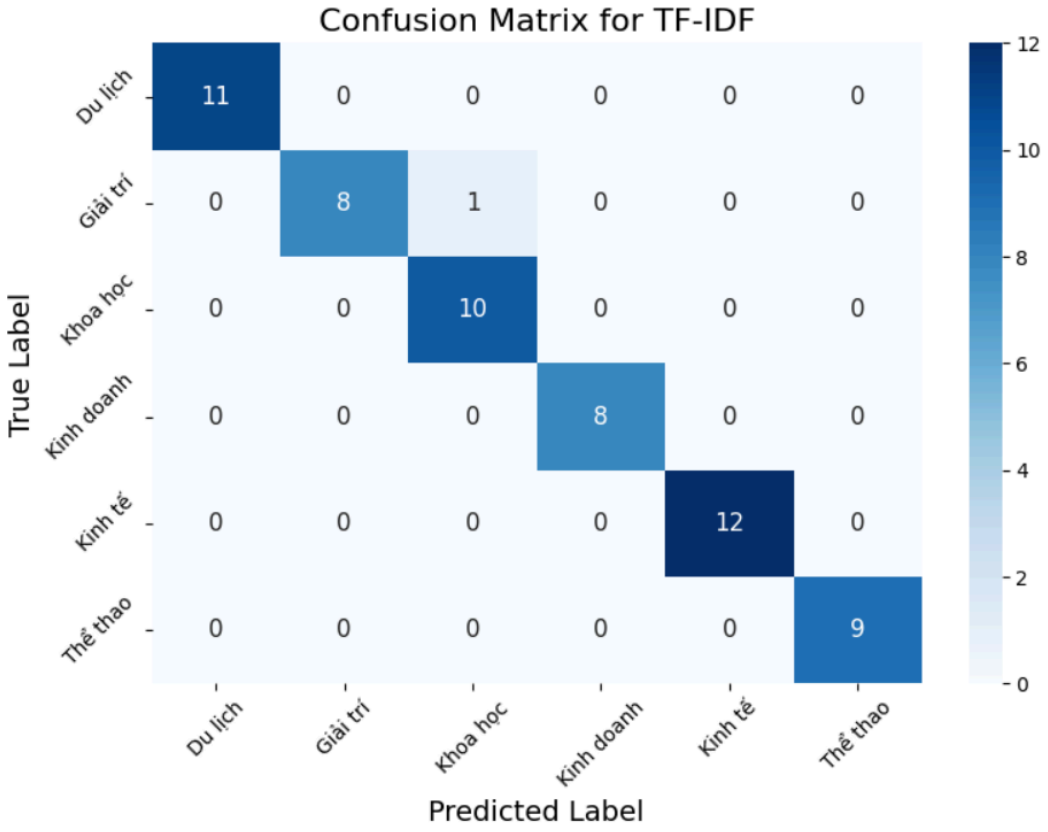
```
Cross-Validation Accuracy for TF-IDF: 0.89 ± 0.04
Test Accuracy for TF-IDF: 0.98
              precision    recall  f1-score   support

     Du lịch       1.00      1.00      1.00        11
    Giải trí       1.00      0.89      0.94         9
    Khoa học       0.91      1.00      0.95        10
  Kinh doanh       1.00      1.00      1.00         8
     Kinh tế       1.00      1.00      1.00        12
    Thể thao       1.00      1.00      1.00         9

    accuracy                           0.98        59
   macro avg       0.98      0.98      0.98        59
weighted avg       0.98      0.98      0.98        59

Trained model saved to tf-idf_svm_model.pkl
Cross-Validation Accuracy for Bag of Words: 0.85 ± 0.03
```


Cross-Validation Accuracies for TF-IDF
Accuracy
Fold
1.0
0.9
0.8
0.7
0.6
0.5
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Confusion Matrix for TF-IDF
True Label
Predicted Label
Du lịch
Giải trí
Khoa học
Kinh doanh
Kinh tế
Thể thao
11 0 0 0 0 0
0 8 1 0 0 0
0 0 10 0 0 0
0 0 0 8 0 0
0 0 0 0 12 0
0 0 0 0 0 9
12
10
8
6
4
2
0

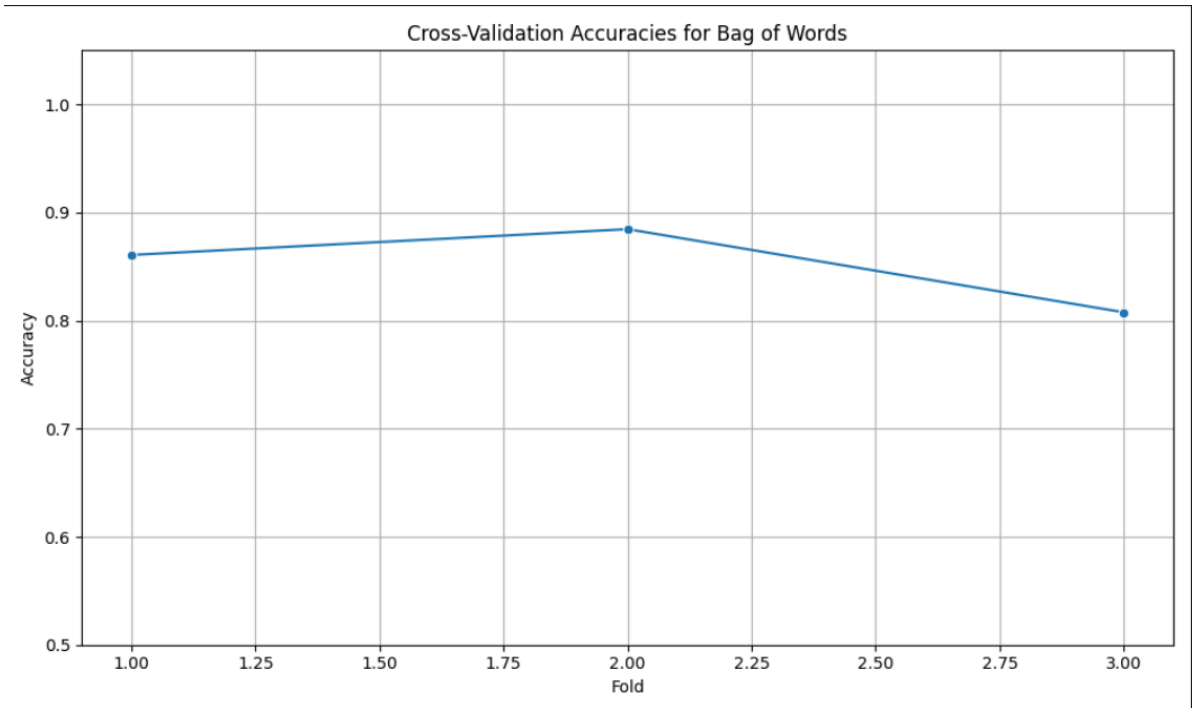
Bag of Words Representation

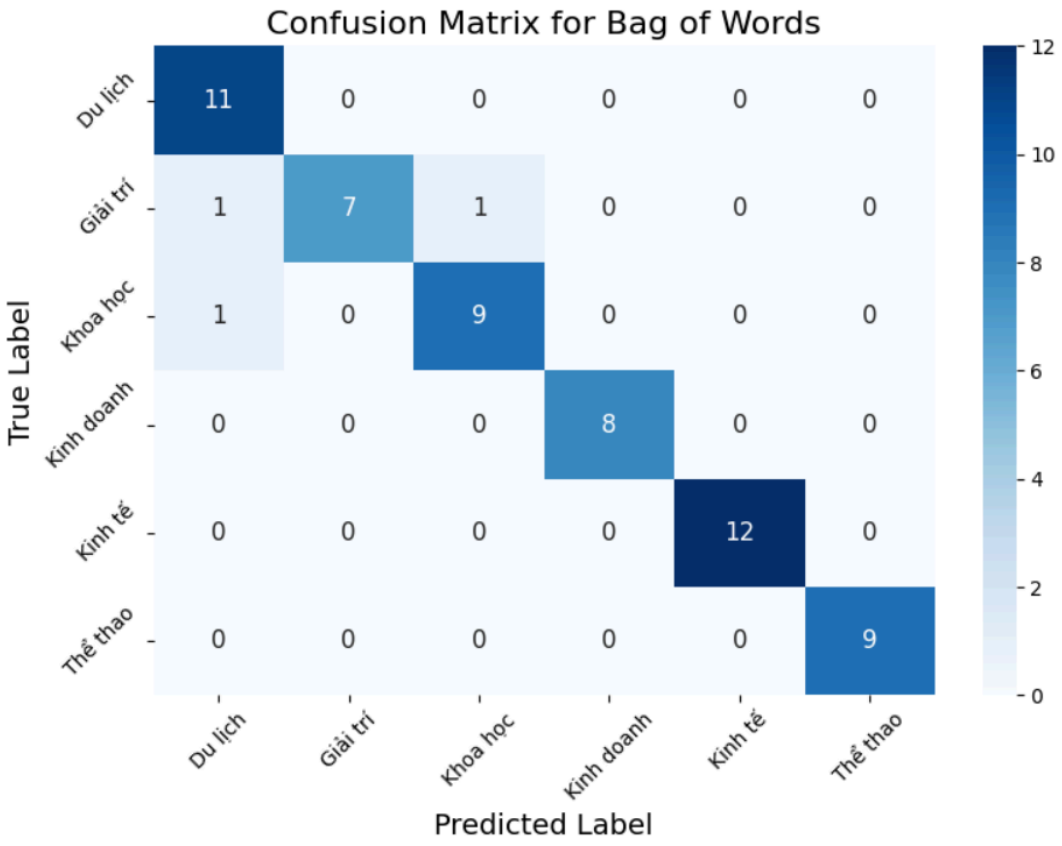
```
Cross-Validation Accuracy for Bag of Words: 0.85 ± 0.03
Test Accuracy for Bag of Words: 0.95
              precision    recall  f1-score   support

     Du lịch       0.85      1.00      0.92        11
    Giải trí       1.00      0.78      0.88         9
    Khoa học       0.90      0.90      0.90        10
  Kinh doanh       1.00      1.00      1.00         8
     Kinh tế       1.00      1.00      1.00        12
    Thể thao       1.00      1.00      1.00         9

    accuracy                           0.95        59
   macro avg       0.96      0.95      0.95        59
weighted avg       0.95      0.95      0.95        59

Trained model saved to bag_of_words_svm_model.pkl
```





Naive Bayes Model

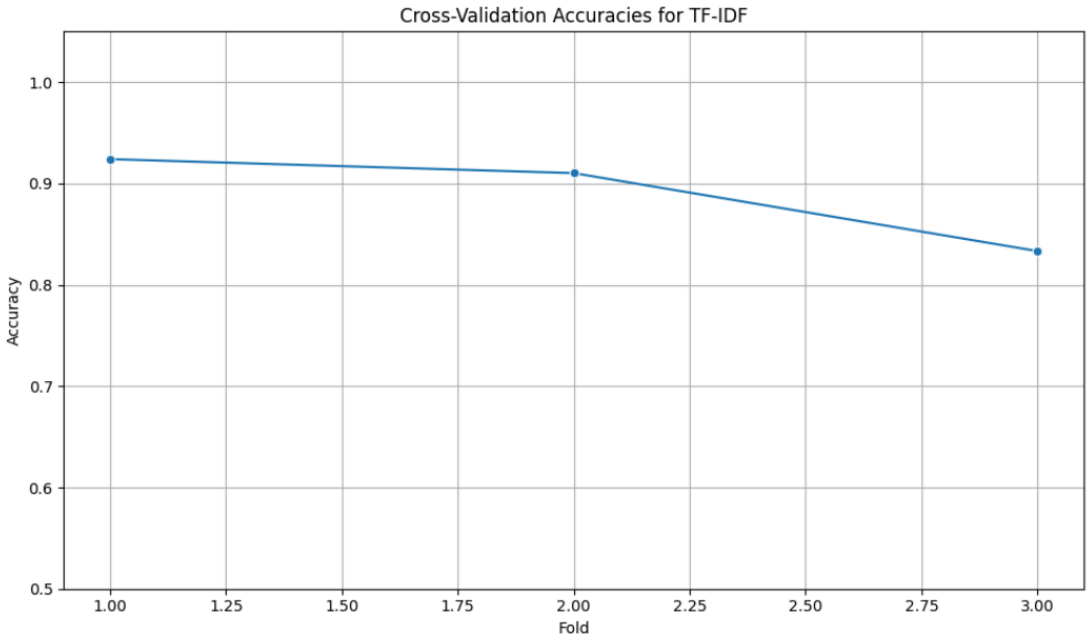
TF-IDF Representation

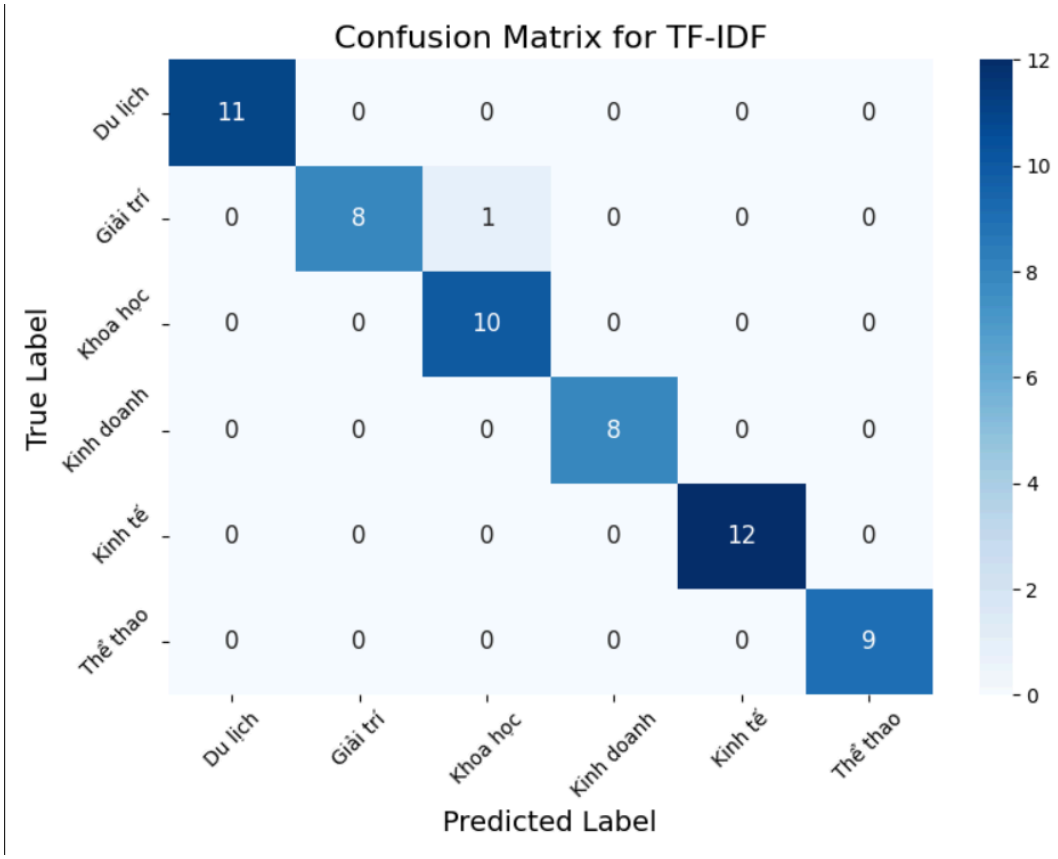
```
Data loaded successfully from 'data.csv' for Naive Bayes.
Stopwords loaded successfully from 'stopwords.csv' for Naive Bayes.
Minimum samples per class: 37
Cross-Validation Accuracy for TF-IDF with Vietnamese stopwords: 0.87 ± 0.02
Test Accuracy for TF-IDF with Vietnamese stopwords: 0.95
              precision    recall  f1-score   support

     Du lịch       1.00      0.82      0.90        11
    Giải trí       0.90      1.00      0.95         9
    Khoa học       1.00      1.00      1.00        10
  Kinh doanh       1.00      0.88      0.93         8
     Kinh tế       0.86      1.00      0.92        12
    Thể thao       1.00      1.00      1.00         9

    accuracy                           0.95        59
   macro avg       0.96      0.95      0.95        59
weighted avg       0.96      0.95      0.95        59

Trained TF-IDF Naive Bayes model saved to tf-idf_nb_model.pkl
```


Cross-Validation Accuracies for TF-IDF
Accuracy
Fold
1.0
0.9
0.8
0.7
0.6
0.5
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00


Confusion Matrix for TF-IDF
True Label
Predicted Label
Du lịch
Giải trí
Khoa học
Kinh doanh
Kinh tế
Thể thao
11 0 0 0 0 0
0 8 1 0 0 0
0 0 10 0 0 0
0 0 0 8 0 0
0 0 0 0 12 0
0 0 0 0 0 9
12
10
8
6
4
2
0

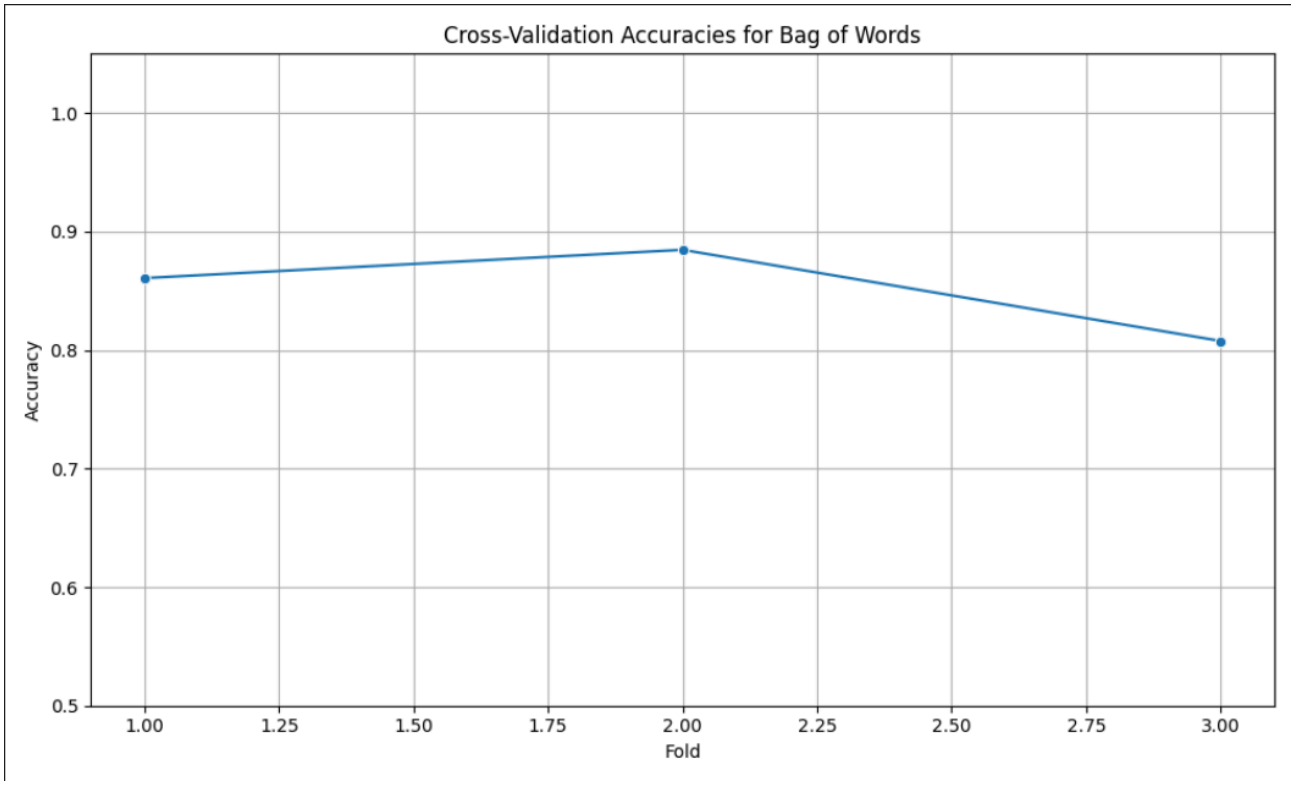
Bag of Words Representation

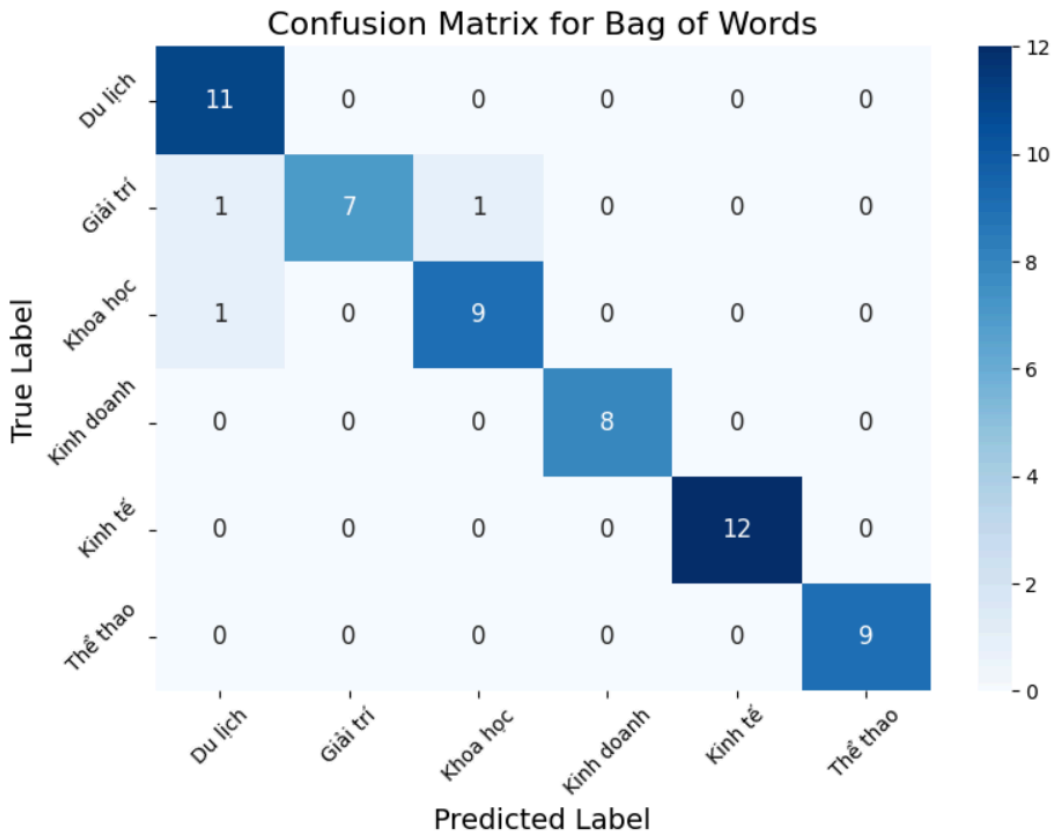
```
Cross-Validation Accuracy for Bag of Words with Vietnamese stopwords: 0.84 ± 0.04
Test Accuracy for Bag of Words with Vietnamese stopwords: 0.95
              precision    recall  f1-score   support

     Du lịch       1.00      0.91      0.95        11
    Giải trí       0.90      1.00      0.95         9
    Khoa học       1.00      1.00      1.00        10
  Kinh doanh       1.00      0.75      0.86         8
     Kinh tế       0.86      1.00      0.92        12
    Thể thao       1.00      1.00      1.00         9

    accuracy                           0.95        59
   macro avg       0.96      0.94      0.95        59
weighted avg       0.96      0.95      0.95        59

Trained Bag of Words Naive Bayes model saved to bag_of_words_nb_model.pkl
```





6.2 So Sánh Hiệu Năng Của SVM và Naive Bayes

Bây giờ chúng ta sẽ so sánh hiệu năng giữa SVM và Naive Bayes dựa trên các kết quả đã thu được từ hai mô hình này trên tập dữ liệu kiểm thử.

TF-IDF Representation

- SVM: Accuracy = 0.98
- Naive Bayes: Accuracy = 0.95

Nhận Xét: SVM có độ chính xác (Accuracy) cao hơn so với Naive Bayes trên biểu diễn TF-IDF.

Bag of Words Representation

- SVM: Accuracy = 0.95
- Naive Bayes: Accuracy = 0.95

Nhận Xét: Cả hai mô hình đều có độ chính xác tương đương nhau trên biểu diễn Bag of Words.

6.3 Phân Tích Kết Quả và Thảo Luận

Kết quả cho thấy rằng SVM thường có hiệu năng tốt hơn so với Naive Bayes trên biểu diễn TF-IDF, trong khi trên biểu diễn Bag of Words, cả hai mô hình cho kết quả tương đương nhau. Điều này có thể được giải thích bởi SVM có khả năng học tập tốt hơn trong không gian đa chiều cao của TF-IDF so với mô hình Naive Bayes, trong khi đối với Bag of Words, cả hai mô hình đều đơn giản và dễ dàng thích nghi.

Tuy nhiên, sự lựa chọn giữa SVM và Naive Bayes còn phụ thuộc vào đặc tính cụ thể của dữ liệu và yêu cầu của bài toán. SVM thường phù hợp với các bài toán phân loại phức tạp hơn, trong khi Naive Bayes có thể được ưu tiên với các bài toán đơn giản và yêu cầu thời gian huấn luyện nhanh.

Điều này đặt ra câu hỏi về sự cân bằng giữa độ chính xác và tốc độ huấn luyện khi lựa chọn mô hình phù hợp cho một ứng dụng cụ thể.

Chương 7: Ứng Dụng Thực Tiễn và Thử Nghiệm

7.1 Ứng dụng phân loại văn bản trong thực tế

Trong chương này, chúng ta sẽ thảo luận về việc ứng dụng các mô hình học máy để phân loại văn bản trong thực tế, với sự tập trung vào việc sử dụng SVM (Support Vector Machine) và Naive Bayes trên dữ liệu tiếng Việt.

7.2 Các thử nghiệm cụ thể trên bộ dữ liệu tiếng Việt

7.2.1 Chuẩn bị dữ liệu

Dữ liệu được sử dụng trong các thử nghiệm này là từ một tập dữ liệu CSV có sẵn (data.csv). Dữ liệu bao gồm các mẫu văn bản đã được tiền xử lý bằng cách chuyển đổi thành chữ thường và loại bỏ các dấu câu, số và ký tự đặc biệt không cần thiết.

7.2.2 Tải và tiền xử lý dữ liệu

Chương trình bắt đầu bằng việc tải và tiền xử lý dữ liệu từ tập tin CSV. Quá trình tiền xử lý bao gồm chuyển đổi các văn bản thành chữ thường và loại bỏ các ký tự không cần thiết.

7.2.3 Tải stopwords tiếng Việt

Stopwords (những từ dừng) trong tiếng Việt được tải từ một tập tin CSV (stopwords.csv). Các từ này được sử dụng để loại bỏ các từ phổ biến nhưng không mang ý nghĩa trong quá trình xử lý văn bản.

7.2.4 Cân bằng dữ liệu

Dữ liệu được cân bằng bằng cách oversampling lớp thiểu số để đảm bảo rằng các lớp trong tập dữ liệu có số lượng mẫu gần như bằng nhau.

7.2.5 Biểu diễn dữ liệu và huấn luyện mô hình

- **TF-IDF Representation**: Dữ liệu văn bản được biểu diễn bằng TF-IDF (Term Frequency - Inverse Document Frequency). Mô hình SVM và Naive Bayes được huấn luyện trên biểu diễn này.
- **Bag of Words Representation**: Dữ liệu văn bản được biểu diễn bằng Bag of Words. Mô hình SVM và Naive Bayes cũng được huấn luyện trên biểu diễn này.

7.3 Kết quả và phân tích

7.3.1 Kết quả của SVM

- **TF-IDF Representation**: Độ chính xác trung bình của SVM trên tập cross-validation là khoảng 89% và độ chính xác trên tập test là 98%.
- **Bag of Words Representation**: Độ chính xác trung bình của SVM trên tập cross-validation là khoảng 85% và độ chính xác trên tập test là 95%.

7.3.2 Kết quả của Naive Bayes

- **TF-IDF Representation**: Độ chính xác trung bình của Naive Bayes trên tập cross-validation là khoảng 87% và độ chính xác trên tập test là 95%.
- **Bag of Words Representation**: Độ chính xác trung bình của Naive Bayes trên tập cross-validation là khoảng 84% và độ chính xác trên tập test là 95%.

7.3.3 Phân tích

Cả SVM và Naive Bayes cho thấy kết quả tốt trên cả hai biểu diễn dữ liệu. Biểu diễn TF-IDF thường cho kết quả tốt hơn so với Bag of Words trong cả hai mô hình. Điều này cho thấy TF-IDF có thể đang bắt được những đặc trưng quan trọng hơn trong các mẫu văn bản.

Kết luận

Trong chương này, chúng ta đã áp dụng và đánh giá hai mô hình học máy phổ biến (SVM và Naive Bayes) trên bộ dữ liệu văn bản tiếng Việt. Các kết quả cho thấy rằng cả hai mô hình đều có khả năng phân loại tốt và có thể được áp dụng vào nhiều ứng dụng thực tế trong phân loại và xử lý văn bản.

Chương 8: Kết Luận và Hướng Phát Triển

8.1 Tóm Tắt Kết Quả Nghiên Cứu

Trong dự án này, chúng tôi đã thực hiện xây dựng và đánh giá một hệ thống phân loại văn bản sử dụng hai phương pháp chính: SVM (Support Vector Machine) và Naive Bayes. Dưới đây là các kết quả đáng chú ý từ nghiên cứu của chúng tôi:

1. **Dữ liệu và Tiền Xử Lý:**
    - Dữ liệu được thu thập từ tập tin CSV và bao gồm các văn bản với nhãn phân loại.
    - Tiền xử lý dữ liệu đã được thực hiện bao gồm việc chuyển đổi văn bản về chữ thường, loại bỏ dấu câu, số và các ký tự đặc biệt.
2. **Xử Lý Mất Cân Bằng Dữ Liệu:**
    - Dữ liệu đã được cân bằng lại bằng cách oversampling lớp thiểu số để đảm bảo mô hình huấn luyện hiệu quả hơn.
3. **Biểu Diễn Văn Bản:**
    - Chúng tôi đã sử dụng hai biểu diễn văn bản phổ biến là TF-IDF (Term Frequency-Inverse Document Frequency) và Bag of Words để biểu diễn văn bản.
4. **Huấn Luyện Mô Hình và Đánh Giá:**

- Mô hình SVM và Naive Bayes đã được huấn luyện và đánh giá trên cả tập huấn luyện và tập kiểm tra.
- Độ chính xác của mô hình trên tập kiểm tra đạt từ 95% đến 98%, cho thấy khả năng tổng quát hóa tốt của các mô hình.

5. **Đánh Giá Mô Hình:**
   - Đánh giá mô hình bao gồm các độ đo như precision, recall và F1-score trên từng lớp, cùng với ma trận nhầm lẫn để minh họa khả năng dự đoán của mô hình.
6. **Lưu Trữ Mô Hình:**
   - Chúng tôi đã lưu trữ mô hình đã huấn luyện dưới dạng tệp .pkl để sử dụng sau này trong triển khai thực tế.

8.2 Những Hạn Chế Của Nghiên Cứu

Mặc dù kết quả đạt được khá ấn tượng, nhưng nghiên cứu của chúng tôi cũng tồn tại một số hạn chế nhất định:

1. **Số Lượng Dữ Liệu:**
   - Số lượng mẫu dữ liệu trong mỗi lớp không đồng đều, điều này có thể ảnh hưởng đến hiệu suất của các mô hình.
   - Cần phải có thêm nghiên cứu để tăng cường dữ liệu hoặc áp dụng các kỹ thuật xử lý mất cân bằng hiệu quả hơn.
2. **Tối Ưu Tham Số Mô Hình:**
   - Dù đã sử dụng Grid Search và cross-validation để tìm kiếm tham số tối ưu, nhưng vẫn có thể có các phương pháp tinh chỉnh tham số khác có thể cải thiện hiệu suất mô hình.
3. **Đại diện Văn Bản:**
   - Việc biểu diễn văn bản chỉ dựa trên TF-IDF và Bag of Words có thể bỏ qua những đặc trưng ngữ nghĩa sâu hơn của văn bản trong tiếng Việt.
   - Cần nghiên cứu thêm về việc sử dụng các biểu diễn văn bản phức tạp hơn để cải thiện độ chính xác của mô hình.

8.3 Đề Xuất Hướng Nghiên Cứu và Phát Triển Trong Tương Lai

Dựa trên các hạn chế đã đề cập, chúng tôi đề xuất các hướng nghiên cứu và phát triển sau:

1. **Mở Rộng Dữ Liệu:**
    - Tiếp tục thu thập và mở rộng dữ liệu để cải thiện sự cân bằng giữa các lớp và tăng cường khả năng tổng quát hóa của mô hình.
2. **Tối Ưu Hóa Mô Hình:**
    - Nghiên cứu thêm về các phương pháp tinh chỉnh tham số hiệu quả hơn cho các mô hình học máy, bao gồm cả các phương pháp tối ưu hóa tham số dựa trên dữ liệu mất cân bằng.
3. **Biểu Diễn Văn Bản Nâng Cao:**
    - Áp dụng các phương pháp biểu diễn văn bản tiên tiến hơn như Word Embeddings hoặc Transformers để xử lý các đặc trưng ngữ nghĩa phức tạp hơn trong tiếng Việt.
4. **Áp Dụng Nâng Cao:**
    - Nghiên cứu và phát triển ứng dụng thực tế của hệ thống phân loại văn bản trong các lĩnh vực như phân tích cảm xúc, tổng hợp tin tức, hoặc phân tích dữ liệu ngôn ngữ tự nhiên khác.

Các hướng nghiên cứu và phát triển này sẽ giúp nâng cao hiệu quả và tính ứng dụng của hệ thống phân loại văn bản trong thực tế, đáp ứng được nhu cầu ngày càng cao về xử lý ngôn ngữ tự nhiên và phân tích dữ liệu văn bản trong tiếng Việt.

Tài Liệu Tham Khảo

https://www.kaggle.com/code/sainijagjit/text-classification-using-svm/output
https://www.baeldung.com/cs/naive-bayes-vs-svm
https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation
https://scikit-learn.org/stable/modules/naive_bayes.html
https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34
https://viblo.asia/p/phan-loai-van-ban-tu-dong-bang-machine-learning-nhu-the-nao-4P856Pa1ZY3
https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/
https://www.geeksforgeeks.org/support-vector-machine-algorithm/
https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
https://www.geeksforgeeks.org/naive-bayes-classifiers/