BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN TỐT NGHIỆP

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN NGÔN NGỮ DỰA TRÊN MULTINOMIAL NAÏVE BAYES

BÙI MINH THÀNH
thanh1546365@huce.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: TS. Phạm Hồng Phong _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Đại Học Xây Dựng Hà Nội

Hà Nội, ngày/09/2024

Lời cảm ơn

Trong khoảng thời gian làm đồ án tốt nghiệp, em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và sự dẫn dắt chỉ bảo nhiệt tình của thầy cô, gia đình và bạn bè.

Em xin gửi lời cảm ơn chân thành đến giảng viên hướng dẫn - TS. Phạm Hồng Phong - Trường Đại học Xây Dựng Hà Nội, người đã tận tình hướng dẫn, chỉ bảo em trong suốt quá trình thực hiện đồ án tốt nghiệp.

Em cũng xin gửi cảm ơn chân thành nhất tới các thầy cô giáo trong trường Đại học Xây Dựng Hà Nội nói chung, các thầy cô trong Bộ môn Khoa học máy tính nói riêng đã dạy dỗ cho em kiến thức về các môn đại cương cũng như các môn chuyên ngành, giúp em có được cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ em trong suốt quá trình em tham gia học tập.

Lời cuối cùng, em xin chân thành cảm ơn gia đình và bạn bè, những người luôn ở bên cạnh đã tạo điều kiện, quan tâm, giúp đỡ, động viên em trong suốt quá trình học tập và hoàn thành đồ án tốt nghiệp.

Với điều kiện về thời gian cũng như lượng kiến thức về đề tài rất rộng mà kinh nghiệm còn hạn chế của một sinh viên, đề án này không thể tránh được những thiếu sót. Em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các thầy cô để em có điều kiện bổ sung, nâng cao ý thức của mình, phục vụ tốt hơn công tác thực tế sau này.

Mục lục

Danh mục thuật ngữ và từ viết tắt

Thuật ngữ	**Giải thích**
NLP	Natural Language Processing (Xử lý ngôn ngữ tự nhiên)
SVM	Support Vector Machine (Máy vector hỗ trợ)
RNN	Recurrent Neural Network (Mạng nơ-ron hồi quy)
LSTM	Long Short-Term Memory (Mạng nơ-ron bộ nhớ ngắn dài)
Naïve Bayes	Thuật toán Naïve Bayes

Chương 1

Tổng quan

1.1 Tổng quan về xử lý ngôn ngữ tự nhiên (NLP):

1.1.1 Xử lý ngôn ngữ tự nhiên là gì:

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của trí tuệ nhân tạo (AI), là một công nghệ máy học, giúp cho máy tính hiểu, tạo và thao tác ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa của ngôn ngữ - công cụ hoàn hảo nhất của tuy duy và giao tiếp.

Về cơ bản, NLP là quá trình AI được dạy để hiểu quy tắc, cú pháp của ngôn ngữ, đồng thời máy móc được lập trình giúp phát triển các thuật toán phức tạp, nhằm biểu diễn những quy tắc đã học. Sau đó, chúng áp dụng thuật toán để thực hiện tác vụ cụ thể.

NLP kết hợp ngôn ngữ học tính toán, mô hình hóa ngôn ngữ con người dựa trên quy tắc với các mô hình thống kê, học máy (Machine Learning) và học sâu (Deep Learning).

Cùng với nhau, những công nghệ này cho phép máy tính xử lý ngôn ngữ của con người dưới dạng dữ liệu văn bản hoặc giọng nói và 'hiểu' ý nghĩa đầy đủ, hoàn chỉnh với ý định và tình cảm của người nói hoặc người viết.

NLP thúc đẩy chương trình máy tính dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác, phản hồi lệnh được yêu cầu và tóm tắt khối lượng lớn văn bản một cách nhanh chóng, ngay cả trong thời gian thực.

Các tổ chức ngày nay có khối lượng lớn dữ liệu thoại và văn bản từ nhiều kênh liên lạc khác nhau như email, tin nhắn văn bản, bảng tin trên mạng xã hội, tệp video, tệp âm thanh và nhiều hơn nữa. Họ sử dụng phần mềm NLP để tự động xử lý dữ liệu này, phân tích ý định hoặc cảm xúc trong tin nhắn và phản hồi bằng người thật theo thời gian thực.

1.1.2 Cách thức hoạt động của mô hình xử lý ngôn ngữ tự nhiên (Xác định ngôn ngữ):

Để hiểu rõ hơn về xử lý ngôn ngữ tự nhiên, ta có thể chia NLP thành hai nhánh lớn:

1. Xử lý tiếng nói (Speech Processing): tập trung nghiên cứu và phát triển các thuật toán, chương trình máy tính xử lý ngôn ngữ của con người ở dạng tiếng nói (dữ liệu âm thanh).
    - Các ứng dụng quan trọng của xử lý tiếng nói bao gồm:
        - Nhận dạng tiếng nói: chuyển ngôn ngữ từ dạng tiếng nói sang văn bản.
        - Tổng hợp tiếng nói: chuyển ngôn ngữ từ dạng văn bản thành tiếng nói.
2. Xử lý văn bản (Text Processing): tập trung vào phân tích dữ liệu văn bản. Các ứng dụng quan trọng của xử lý văn bản bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động hay kiểm lỗi chính tả tự động.
    - Xử lý văn bản đôi khi được chia tiếp thành hai nhánh nhỏ hơn bao gồm:
        - Hiểu văn bản: liên quan tới các bài toán phân tích văn bản.
        - Sinh văn bản: liên quan tới nhiệm vụ tạo ra văn bản mới như trong các ứng dụng về dịch máy hoặc tóm tắt văn bản tự động.
3. Các thao tác xử lý văn bản:
    (a) Tiền xử lý dữ liệu (Pre-processing): Bước đầu tiên trong quá trình NLP là tiền xử lý dữ liệu. Văn bản đầu vào thường chứa nhiều dữ liệu không cần thiết hoặc không quan trọng như dấu câu hoặc ký tự đặc biệt. Để làm cho dữ liệu phù hợp để xử lý, các bước tiền xử lý bao gồm loại bỏ dấu câu, chuyển đổi văn bản thành chữ thường để loại bỏ sự phân biệt chữ hoa/chữ thường, và loại bỏ các ký tự đặc biệt không cần thiết.

    (b) Tách từ (Tokenization): Sau khi tiền xử lý, dòng văn bản được chia thành các phần tử cơ bản gọi là "token". Mỗi token thường tương ứng với một từ, dấu câu, hoặc ký tự. Tách từ là quá trình quan trọng trong NLP vì nó tạo ra đơn vị cơ bản để máy tính có thể xử lý.

    (c) Phân tích từ loại (Part-of-speech Tagging): Mỗi token sau khi được tách từ được gán một nhãn để chỉ ra vai trò ngữ pháp của nó trong câu. Ví dụ, một từ có thể được gán nhãn là danh từ, động từ, tính từ, hoặc trạng từ. Phân tích từ loại giúp máy tính hiểu cấu trúc ngữ pháp của câu.

    (d) Phân tích cú pháp (Parsing): Câu được phân tích thành cấu trúc cây để hiểu mối quan hệ cú pháp giữa các từ. Cấu trúc cây này cho phép máy tính hiểu ý nghĩa và mối quan hệ cú pháp giữa các thành phần của câu. Ví dụ, trong câu "Con mèo đen nhảy qua cái bàn", cấu trúc cây có thể chỉ ra mối quan hệ giữa "mèo"và "đen"như một cụm danh từ.

    (e) Trích xuất thông tin (Information Extraction): Một trong những nhiệm vụ quan trọng của NLP là trích xuất thông tin từ văn bản. Điều này bao gồm việc nhận dạng và trích xuất các loại thông tin như tên riêng, địa chỉ, số điện thoại, ngày tháng, hoặc thông tin khác có ý nghĩa trong văn bản.

(f) Phân giải ngữ nghĩa (Semantic Parsing): Quá trình này liên quan đến việc hiểu ý nghĩa của câu bằng cách phân tích ý nghĩa của các thành phần ngôn ngữ và mối quan hệ giữa chúng. Điều này bao gồm việc phân tích ý nghĩa của từ ngữ, cấu trúc câu, và ngữ cảnh.

(g) Học máy (Machine Learning) và Học sâu (Deep Learning): Nhiều phương pháp trong NLP sử dụng các thuật toán học máy và deep learning để huấn luyện các mô hình dựa trên dữ liệu. Các mô hình này có thể học từ dữ liệu đầu vào và tự điều chỉnh để cải thiện hiệu suất của chúng trong các nhiệm vụ NLP.

(h) Áp dụng các ứng dụng cụ thể: Cho input và dự đoán thử.

1.1.3 Ưu điểm và nhược điểm của xử lý ngôn ngữ tự nhiên:

1. Ưu điểm của xử lý ngôn ngữ tự nhiên (NLP):

   (a) Tăng tốc độ xử lý và hiệu quả công việc:
   - Tự động hóa nhiều tác vụ liên quan đến ngôn ngữ, từ phân tích văn bản đến tạo văn bản tự động.
   - Tiết kiệm thời gian và công sức cho con người.
   - Có thể xử lý lượng dữ liệu lớn một cách nhanh chóng và hiệu quả.

   (b) Cải thiện trải nghiệm người dùng:
   - Đóng vai trò quan trọng trong phát triển các ứng dụng tương tác người-máy như trợ lý ảo và chatbot.
   - Hiểu và phản hồi yêu cầu của người dùng một cách tự nhiên và linh hoạt, cải thiện trải nghiệm người dùng..

   (c) Dễ dàng tiếp cận thông tin:
   - Trích xuất thông tin từ văn bản và ngôn ngữ tự nhiên.
   - Giúp người dùng dễ dàng tiếp cận và tìm kiếm thông tin từ các nguồn dữ liệu khác nhau như internet, email, báo cáo doanh nghiệp, v.v.

   (d) Dịch máy và giao tiếp đa ngôn ngữ:
   - Phát triển các công nghệ dịch máy, giúp giao tiếp và trao đổi thông tin giữa người dùng từ các ngôn ngữ khác nhau trở nên dễ dàng hơn.

   (e) Phân tích dữ liệu và phản hồi thị trường:
   - Giúp doanh nghiệp phân tích và hiểu dữ liệu từ phản hồi của khách hàng trên mạng xã hội, email, các bài đánh giá sản phẩm, v.v.
   - Cung cấp thông tin quan trọng để điều chỉnh chiến lược kinh doanh và cung cấp sản phẩm/dịch vụ phù hợp hơn với nhu cầu thị trường.

2. Nhược điểm của Xử lý Ngôn ngữ Tự nhiên (NLP):

   (a) Độ chính xác chưa đạt 100%:
   - Dù đã có tiến bộ đáng kể, NLP vẫn gặp hạn chế trong việc hiểu và xử lý ngôn ngữ tự nhiên.
   - Kết quả có thể không chính xác trong một số trường hợp, đặc biệt là trong ngữ cảnh phức tạp hoặc đa nghĩa.

(b) Phụ thuộc vào chất lượng dữ liệu:

- Hiệu suất của hệ thống NLP phụ thuộc nhiều vào chất lượng và đa dạng của dữ liệu huấn luyện.
- Dữ liệu không cân đối hoặc không đại diện có thể dẫn đến kết quả không chính xác hoặc thiếu độ tin cậy.

(c) Khả năng hiểu ngữ cảnh hạn chế:

- Khả năng hiểu ngữ cảnh và ngữ nghĩa của NLP vẫn còn hạn chế, đặc biệt trong các tình huống phức tạp hoặc đa nghĩa.

(d) Vấn đề về quyền riêng tư và bảo mật:

- Sử dụng NLP để phân tích dữ liệu cá nhân có thể gây ra các vấn đề liên quan đến quyền riêng tư và bảo mật, khi thông tin cá nhân của người dùng có thể bị tiết lộ hoặc sử dụng không đúng đắn.

1.1.4 Tầm quan trọng của xử lý ngôn ngữ tự nhiên:

1. Giao tiếp người-máy hiệu quả:

   - NLP đóng vai trò quan trọng trong việc tạo ra các giao diện người-máy hiệu quả.
     - Ví dụ: trợ lý ảo trên điện thoại di động và hệ thống hỗ trợ khách hàng tự động trên các trang web.
   - Giúp tạo ra trải nghiệm người dùng tốt hơn và giảm thời gian và chi phí cho doanh nghiệp.

2. Phụ thuộc vào chất lượng dữ liệu:

   - Hiệu suất của hệ thống NLP phụ thuộc nhiều vào chất lượng và đa dạng của dữ liệu huấn luyện.
   - Dữ liệu không cân đối hoặc không đại diện có thể dẫn đến kết quả không chính xác hoặc thiếu độ tin cậy.

3. Tích hợp ngôn ngữ tự nhiên vào ứng dụng công nghệ:

   - NLP cho phép tích hợp ngôn ngữ tự nhiên vào nhiều loại ứng dụng công nghệ khác nhau.
     - Ví dụ: hệ thống điều khiển bằng giọng nói trong ô tô thông minh giúp lái xe tương tác với hệ thống điều khiển một cách an toàn và thuận tiện.

4. Dịch máy và giao tiếp đa ngôn ngữ:

   - NLP đã đóng vai trò quan trọng trong việc phát triển các dịch vụ dịch máy như Google Translate.
   - Giúp giao tiếp trên toàn thế giới trở nên dễ dàng hơn.
   - Hỗ trợ giao tiếp kinh doanh quốc tế và tạo ra các ứng dụng học ngoại ngữ.

5. Phân tích dữ liệu và thông tin:
    - NLP giúp tổ chức và phân tích dữ liệu từ các nguồn như email, bài đăng trên mạng xã hội, và các văn bản trên internet.
    - Cung cấp thông tin quan trọng cho doanh nghiệp để hiểu ý kiến của khách hàng, phát triển sản phẩm, và định hình chiến lược kinh doanh.
6. Hỗ trợ trong lĩnh vực y tế và y tế cộng đồng:
    - Trong lĩnh vực y tế, NLP giúp tổ chức và phân tích thông tin trong hồ sơ bệnh án điện tử.
    - Cải thiện chẩn đoán, dự đoán bệnh, và tối ưu hóa quản lý bệnh nhân.
    - Giúp phát hiện và phòng chống dịch bệnh thông qua phân tích dữ liệu từ các nguồn khác nhau.
7. Nâng cao trải nghiệm người dùng trên internet:
    - NLP được sử dụng trong các công cụ tìm kiếm và phân loại nội dung trên internet.
    - Cung cấp kết quả tìm kiếm chính xác hơn và tùy chỉnh dựa trên ngữ cảnh và sở thích của người dùng.
    - Cải thiện trải nghiệm người dùng và tăng cơ hội tiếp cận thông tin hữu ích.
8. Hỗ trợ giáo dục và học tập:
    - Trong giáo dục, NLP có thể được sử dụng để tự động tạo nội dung giảng dạy, cung cấp phản hồi tức thì cho học viên, và phân tích hiệu suất học tập.
    - Giúp cá nhân học tập theo cách cá nhân hóa và hiệu quả hơn.

1.1.5 Ứng dụng của xử lý ngôn ngữ tự nhiên:

1. Nhận dạng chữ viết:
    (a) Nhận dạng chữ in:
        - Chuyển chữ trên sách giáo khoa thành văn bản điện tử.
        - Giúp số hóa hàng ngàn đầu sách trong thời gian ngắn.
    (b) Nhận dạng chữ viết tay:
        - Phức tạp hơn do không có khuôn dạng cố định.
        - Ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).
2. Nhận dạng tiếng nói:
    (a) Chuyển âm thanh thành văn bản:
        - Giúp thao tác trên thiết bị nhanh hơn, ví dụ thay vì gõ tài liệu, bạn đọc và trình soạn thảo tự ghi lại.
        - Hữu ích cho người khiếm thị và là bước đầu trong giao tiếp giữa con người và robot.

(b) Tổng hợp tiếng nói:

- Chuyển văn bản thành âm thanh tương ứng.
- Giúp đọc tự động sách và nội dung trang web.
- Trợ giúp tốt cho người khiếm thị và là bước cuối cùng trong giao tiếp giữa robot và con người.

3. Dịch tự động (Machine Translation):

(a) Chuyển ngôn ngữ: Chuyển ngôn ngữ này sang ngôn ngữ khác.

- Ví dụ: Evtrans của Softex dịch từ tiếng Anh sang tiếng Việt và ngược lại.
- Các công ty như Lạc Việt và Google cũng tham gia lĩnh vực này.

4. Tìm kiếm thông tin (Information Retrieval):

(a) Đặt câu hỏi và tìm nội dung phù hợp:

- Internet giúp tiếp cận thông tin dễ dàng nhưng tìm đúng thông tin cần thiết là thách thức.
- Các máy tìm kiếm chưa hiểu được ngôn ngữ tự nhiên của con người.

5. Tóm tắt văn bản:

(a) Tóm tắt văn bản dài thành ngắn.

(b) Giữ nguyên các nội dung thiết yếu.

6. Khai phá dữ liệu (Data Mining) và phát hiện tri thức:

(a) Phát hiện tri thức mới từ tài liệu:

- Công cụ tự tìm câu trả lời dựa trên thông tin web, dù trước đó có câu trả lời lưu trên web hay không.

7. Sửa lỗi chính tả:

(a) Phát hiện và sửa lỗi chính tả:

- Tích hợp trong các ứng dụng văn phòng như Microsoft Word, Google Docs.
- Hỗ trợ nhiều ngôn ngữ, bao gồm tiếng Việt.

8. Gán nhãn từ loại:

(a) Gắn nhãn từ dựa theo ngữ cảnh:

- Xác định từ loại như danh từ, động từ, tính từ, trạng từ.
- Giúp máy tính hiểu mối quan hệ nghĩa giữa các từ.

9. Xử lý nhập nhằng nghĩa của từ:

(a) Xác định ý nghĩa chủ đích của từ:

- Ví dụ: từ "bat"có thể nghĩa là dơi hoặc gậy bóng chày tùy ngữ cảnh.

10. Nhận dạng thực thể:

    (a) Xác định tên riêng cho thực thể:

        - Ví dụ: trong câu “Jane đã đi nghỉ ở Pháp và cô ấy say mê các món ăn địa phương,” xác định "Jane"và "Pháp"là các thực thể.

11. Phân tích cảm xúc:

    (a) Diễn giải cảm xúc qua văn bản:

        - Tìm từ hoặc cụm từ thể hiện cảm xúc như không hài lòng, hạnh phúc, nghi ngờ, hối hận và các cảm xúc khác.

12. Chatbot:

    (a) Chương trình hội thoại văn bản:

        - Có khả năng trò chuyện, hỏi đáp với con người.

1.2 Tổng quan về đề tài:

1.2.1 Xác định ngôn ngữ là gì:

Định dạng ngôn ngữ, hay xác định ngôn ngữ (Language Identification) là quá trình nhận biết và phân loại ngôn ngữ của một đoạn văn bản hoặc một chuỗi ký tự. Mục đích chính của việc định dạng ngôn ngữ là xác định ngôn ngữ mà đoạn văn bản được viết bằng, để từ đó có thể thực hiện các xử lý hoặc phân tích ngữ cảnh phù hợp.

1.2.2 Cách thức hoạt động của xác định ngôn ngữ:

1. Tiền xử lý dữ liệu: Trước hết, văn bản đầu vào cần được tiền xử lý để làm sạch và chuẩn hóa. Điều này bao gồm loại bỏ dấu câu, chuyển đổi văn bản thành chữ thường, và loại bỏ các ký tự đặc biệt không cần thiết.

2. Tách từ (Tokenization): Dòng văn bản sau khi được tiền xử lý được chia thành các phần tử cơ bản gọi là "token". Mỗi token thường tương ứng với một từ hoặc ký tự. Tách từ là quá trình quan trọng để máy tính có thể xử lý các phần tử riêng lẻ trong văn bản.

3. Phân tích từ loại (Part-of-speech tagging): Mỗi token sau khi được tách từ được gán một nhãn để chỉ ra vai trò ngữ pháp của nó trong câu. Điều này giúp máy tính hiểu cấu trúc ngữ pháp của văn bản.

4. Trích xuất đặc trưng (Feature extraction): Sau khi đã tách từ và phân tích từ loại, các đặc trưng ngôn ngữ được trích xuất từ văn bản. Các đặc trưng này có thể bao gồm tần suất xuất hiện của các từ, ký tự, hoặc các đặc điểm ngữ cảnh khác.

5. Sử dụng mô hình phân loại (Classification model): Các đặc trưng được trích xuất từ văn bản sau đó được đưa vào một mô hình phân loại để xác định ngôn ngữ của văn bản. Mô hình này thường được huấn luyện trên dữ liệu được gán nhãn trước để có thể dự đoán ngôn ngữ của văn bản mới.

6. Đưa ra dự đoán (Predict): Dựa trên đầu ra của mô hình phân loại, hệ thống NLP có thể xác định ngôn ngữ của đoạn văn bản đầu vào.

7. Đánh giá và điều chỉnh: Cuối cùng, kết quả dự đoán có thể được đánh giá và điều chỉnh để cải thiện hiệu suất của mô hình phân loại.

1.2.3 Các yếu tố ảnh hưởng đến xác định ngôn ngữ:

1. Tần suất xuất hiện của từng ngôn ngữ trong dữ liệu: Ngôn ngữ xuất hiện nhiều hơn sẽ dễ xác định hơn do sự quen thuộc và dữ liệu phong phú.

2. Độ dài văn bản: Đoạn văn dài cung cấp nhiều thông tin ngữ pháp và từ vựng, giúp xác định ngôn ngữ chính xác hơn so với văn bản ngắn.

3. Tần suất từ đặc trưng: Mỗi ngôn ngữ có các từ hoặc cụm từ đặc trưng. Phân tích sự xuất hiện của các từ này giúp nhận diện ngôn ngữ hiệu quả.

4. Đa dạng ngôn ngữ trong dữ liệu: Dữ liệu chứa nhiều ngôn ngữ hoặc các bảng chữ cái khác nhau có thể làm phức tạp việc xác định ngôn ngữ. Khả năng phân biệt giữa các ngôn ngữ có ký tự và cấu trúc tương tự là cần thiết.

5. Độ chính xác của mô hình xác định ngôn ngữ: Mô hình có thể không luôn chính xác, đặc biệt đối với ngôn ngữ có từ vựng hoặc cấu trúc ngữ pháp tương đồng. Hiệu quả của mô hình phụ thuộc vào dữ liệu huấn luyện và thuật toán.

6. Từ đồng âm hoặc đồng nghĩa: Một số từ tồn tại trong nhiều ngôn ngữ và có nghĩa khác nhau, gây nhầm lẫn trong quá trình xác định ngôn ngữ.

7. Ngữ cảnh: Xác định ngôn ngữ còn phụ thuộc vào ngữ cảnh sử dụng. Ví dụ, trong email, địa chỉ IP của người gửi hoặc danh sách nhận có thể gợi ý về ngôn ngữ được sử dụng.

1.2.4 Kỹ thuật và công nghệ trong xác định ngôn ngữ:

1. Mô hình dựa trên từ điển (Dictionary-based models):

   - Nguyên lý: Sử dụng từ điển chứa các từ hoặc cụm từ đặc trưng của mỗi ngôn ngữ.
   - Cách thức hoạt động: Khi một đoạn văn bản mới được đưa vào, các từ trong đoạn văn bản được so sánh với các từ trong từ điển để xác định ngôn ngữ.

2. Mô hình dựa trên thống kê (Statistical models):

   - Nguyên lý: Sử dụng các phương pháp thống kê để phân loại ngôn ngữ.
   - Cách thức hoạt động: Một phương pháp phổ biến là sử dụng tần suất xuất hiện của các từ hoặc cụm từ trong văn bản. Các đặc trưng thống kê như tần suất xuất hiện của các từ, ký tự, hoặc các đặc điểm ngữ cảnh khác được sử dụng để xây dựng mô hình.

3. Học máy (Machine learning):
    - Nguyên lý: Sử dụng các thuật toán học máy để xây dựng các mô hình phân loại ngôn ngữ.
    - Thuật toán phổ biến: Máy vector hỗ trợ (SVM), mạng nơ-ron, Naive Bayes.
    - Cách thức hoạt động: Các mô hình được huấn luyện trên dữ liệu gắn nhãn với ngôn ngữ tương ứng và sau đó có thể dự đoán ngôn ngữ của các đoạn văn bản mới
4. Mạng nơ-ron hồi quy đơn giản (Simple Recurrent Neural Network – RNN):
    - Nguyên lý: Sử dụng các lớp đầu vào và lớp ẩn để xử lý dữ liệu chuỗi.
    - Cách thức hoạt động: RNN học các mẫu ngôn ngữ và xu hướng xuất hiện từ dữ liệu huấn luyện.
5. Mô hình học sâu (Deep learning models):
    - Nguyên lý: Sử dụng các mô hình học sâu để phân tích dữ liệu chuỗi.
    - Thuật toán phổ biến: Mạng nơ-ron Hồi quy Dài ngắn (LSTM), Mạng nơ-ron Hồi quy Gated (GRU).
    - Cách thức hoạt động: Cung cấp khả năng học và hiểu các mẫu phức tạp trong dữ liệu chuỗi, cải thiện khả năng phân loại ngôn ngữ.
6. Học không giám sát (Unsupervised learning):
    - Nguyên lý: Phân loại văn bản mà không cần dữ liệu huấn luyện gắn nhãn.
    - Cách thức hoạt động: Các phương pháp học không giám sát như phân cụm (clustering) được sử dụng để phân loại văn bản thành các nhóm dựa trên đặc trưng ngôn ngữ.
7. Kết hợp các kỹ thuật:
    - Nguyên lý: Kết hợp nhiều kỹ thuật khác nhau để cải thiện hiệu suất mô hình.
    - Cách thức hoạt động: Kết hợp từ điển, học máy, và học sâu để đảm bảo tính chính xác và đáng tin cậy trong quá trình xác định ngôn ngữ, đặc biệt trong các tình huống phức tạp.

1.2.5 Tầm quan trọng của xác định ngôn ngữ:

1. Dịch máy tự động và thông dịch:
    - Vai trò: Xác định ngôn ngữ của văn bản nguồn là bước quan trọng để chọn mô hình dịch phù hợp.
    - Lợi ích: Hiểu đúng ngôn ngữ giúp dịch máy tái tạo văn bản chính xác.

2. Phân loại và lọc dữ liệu:

    - Vai trò: Xác định ngôn ngữ giúp tổ chức và phân loại nội dung một cách chính xác.
    - Lợi ích: Ví dụ, trong phân loại email, xác định ngôn ngữ giúp phân loại email vào các danh mục như "công việc", "thư cá nhân", "quảng cáo", giúp người dùng quản lý hòm thư hiệu quả hơn.

3. Tìm kiếm thông tin trên internet:

    - Vai trò: Xác định ngôn ngữ cải thiện kết quả tìm kiếm bằng cách hiểu và áp dụng các quy tắc ngữ cảnh của ngôn ngữ.
    - Lợi ích: Khi người dùng tìm kiếm bằng ngôn ngữ cụ thể, việc hiểu được ngôn ngữ giúp cung cấp kết quả tìm kiếm chính xác và liên quan hơn.

4. Phân tích dữ liệu và thông tin:

    - Vai trò: Xác định ngôn ngữ giúp trích xuất thông tin quan trọng từ các nguồn dữ liệu đa ngôn ngữ.
    - Lợi ích: Giúp tổ chức dữ liệu và hiểu rõ hơn về xu hướng và ý kiến của người dùng từ các nguồn thông tin khác nhau..

5. Giao tiếp người - máy:

    - Vai trò: Xác định ngôn ngữ trong các hệ thống trợ lý ảo và chatbot.
    - Lợi ích: Giúp cung cấp trải nghiệm tương tác người-máy tự nhiên và hiệu quả hơn.

6. Y tế và y tế cộng đồng:

    - Vai trò: Xác định ngôn ngữ giúp cải thiện quản lý thông tin bệnh án và phân tích dữ liệu về sức khỏe cộng đồng.
    - Lợi ích: Giúp tạo ra các dịch vụ y tế hiệu quả hơn, đặc biệt là trong việc phát hiện và quản lý các vấn đề sức khỏe cộng đồng.

1.2.6 Các nghiên cứu liên quan:

1. Nghiên cứu: "Algorithmic Programming Language Identification":
    - Tác giả: David Klein, Kyle Murray, Simon Weber.
    - Phương pháp: Sử dụng các phương pháp học máy và xử lý ngôn ngữ tự nhiên để phân tích đặc điểm cú pháp và từ vựng của các ngôn ngữ lập trình.
    - Kết quả: Đề xuất một mô hình phân loại ngôn ngữ lập trình với độ chính xác cao, dựa trên các đặc trưng trích xuất từ mã nguồn.
    - Ứng dụng: Cung cấp một công cụ hỗ trợ cho các nhà phát triển phần mềm và các hệ thống quản lý mã nguồn trong việc nhận diện và phân loại mã nguồn theo ngôn ngữ lập trình.
    - Mục tiêu: Nghiên cứu này nhằm mục đích phát triển và đánh giá các thuật toán để nhận diện ngôn ngữ lập trình từ mã nguồn một cách tự động.
2. Nghiên cứu: "Automatic Language Identification in Texts: A Survey":
    - Tác giả: Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, Krister Lindén.
    - Mục tiêu: Cung cấp một tổng quan toàn diện về các phương pháp nhận diện ngôn ngữ tự động trong văn bản, bao gồm các phương pháp truyền thống và hiện đại.
    - Phương pháp: Tổng hợp và phân tích các phương pháp từ học máy, thống kê, và học sâu, đồng thời đánh giá hiệu suất của các phương pháp này trên các bộ dữ liệu khác nhau.
    - Kết quả: Khảo sát các yếu tố ảnh hưởng đến hiệu suất của các phương pháp nhận diện ngôn ngữ, như kích thước tập dữ liệu, độ dài văn bản, và ngôn ngữ đích.
    - Ứng dụng: Đưa ra các khuyến nghị cho việc lựa chọn phương pháp nhận diện ngôn ngữ phù hợp trong các ứng dụng thực tế như dịch máy, phân tích ngôn ngữ, và quản lý nội dung.

Hai nghiên cứu này đều tập trung vào việc nhận diện ngôn ngữ, nhưng một nghiên cứu hướng đến ngôn ngữ lập trình trong khi nghiên cứu còn lại tập trung vào ngôn ngữ tự nhiên trong văn bản.

Chương 2

Cơ sở lý thuyết

2.1 Định lý Bayes:

2.1.1 Lịch sử nguồn gốc ra đời của thuật toán:

Định lý Bayes (Bayes theorem) được đặt tên theo nhà toán học Thomas Bayes (1701 – 1761).

Thomas Bayes là một nhà toán học người Scotland, sinh vào năm 1701. Ông không được biết đến rộng rãi trong thời đại của mình và không để lại nhiều thông tin về cuộc đời cá nhân. Bayes là một nhà thần học Presbytery tại Nonconformist Bunhill Fields ở London và không có bất kỳ bằng cấp học vị đại học nào. Tuy nhiên, ông có một niềm đam mê sâu sắc với toán học và khoa học tự nhiên.

Vào năm 1763, tại Hội Hoàng gia ở Luân Đôn, một người bạn của Thomas Bayes là Richard Price đã trình bày bài luận của Thomas Bayes, giải quyết một khó khăn trong lý thuyết xác suất. Bayes đã quan tâm đến việc làm thế nào để biến quan sát một sự kiện thành một ước tính về cơ hội của sự kiện xảy ra lần nữa.

Trong bài báo của mình, Bayes minh họa vấn đề với một câu hỏi bí truyền về vị trí của quả bóng bi-a lăn trên bàn. Ông đã đưa ra một công thức mà biến quan sát các địa điểm cuối cùng của bi-a thành một ước tính về cơ hội quả bóng lăn sau đó. Nói cách khác, công việc của ông cho phép quan sát được sử dụng để suy ra xác suất mà giả thuyết có thể đúng. Bayes do đó đã đặt nền tảng cho việc định lượng niềm tin.

2.1.2 Phát biểu định lý Bayes:

Định lý Bayes là một trong những công cụ quan trọng nhất trong lý thuyết xác suất và thống kê. Là một phương pháp tính xác suất có điều kiện, cho phép tính xác suất xảy ra của một sự kiện ngẫu nghiên A khi biết sự kiện liên quan đến B đã xảy ra.

- Xác xuất này được kí hiệu là:

$$P(A \mid B)$$

- Đọc là "xác xuất của A nếu có B". Đại lượng này được gọi là xác suất có điều kiện (hay xác suất hậu nghiệm) vì nó được rút ra từ giá trị được cho của X hoặc phụ thuộc vào giá trị đó.

2.1.3 Công thức của định lý Bayes:

- Nếu A và B là hai sự kiện trong một không gian xác suất và $P(B) \neq 0$, thì xác suất điều kiện của A khi biết B được tính theo công thức:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

  - $P(A \mid B)$ là xác suất của sự kiện A xảy ra khi biết rằng B đã xảy ra.
  - $P(B \mid A)$ là xác suất của sự kiện B xảy ra khi biết rằng A đã xảy ra.
  - $P(A)$ là xác suất tiên nghiệm của sự kiện A.
  - $P(B)$ là xác suất tiên nghiệm của sự kiện B.

- Giải thích chi tiết công thức:

  - Xác suất tiên nghiệm (Prior Probability) P(A): Đây là xác suất ban đầu của sự kiện A trước khi có bất kỳ thông tin nào về sự kiện B.
  - Xác suất có điều kiện (Conditional Probability) $P(B \mid A)$: Đây là xác suất của sự kiện B xảy ra khi biết rằng sự kiện A đã xảy ra.
  - Xác suất hậu nghiệm (Posterior Probability) $P(A \mid B)$: Đây là xác suất của sự kiện A xảy ra sau khi biết rằng sự kiện B đã xảy ra. Đây là xác suất mà chúng ta muốn tính toán dựa trên thông tin mới.
  - Xác suất biên (Marginal Probability) P(B): Đây là xác suất của sự kiện B xảy ra, không phụ thuộc vào sự kiện A. Nó có thể được tính như là tổng của các xác suất của B dưới mọi trường hợp có thể xảy ra của A:

    *
    $$P(B) = P(B \mid A) \cdot P(A) + P(B \mid \neg A) \cdot P(\neg A)$$

    * Trong đó ¬A là sự kiện đối ngẫu với A, tức là sự kiện A không xảy ra.

- Ngoài ra thì Định lý Bayes còn có các dạng khác được mở rộng từ định lý và được dùng trong các trường hợp cụ thể như:

  1. Định lý Bayes mở rộng (Extended Bayes' Theorem):
     - Mô tả: Định lý Bayes mở rộng áp dụng khi có nhiều hơn hai sự kiện được xem xét. Nó cung cấp một cách tiếp cận cho việc tính toán xác suất có điều kiện của một sự kiện trong một tình huống phức tạp hơn, khi có nhiều yếu tố ảnh hưởng.
     - Ứng dụng: Dùng trong các hệ thống phức tạp như phân tích thị trường chứng khoán, dự báo thời tiết, hoặc các mô hình thống kê đa biến.
  2. Định lý Bayes phi thường (Bayesian Theorem of Rare Events):
     - Mô tả: Định lý Bayes phi thường áp dụng khi sự kiện X là một sự kiện hiếm, có xác suất xảy ra rất nhỏ. Trong trường hợp này, công thức của Định lý Bayes có thể được đơn giản hóa để dễ dàng tính toán.
     - Ứng dụng: Phân tích rủi ro, dự báo thiên tai, xác định xác suất sự cố hiếm gặp trong công nghiệp.

3. Định lý Bayes phi tuyến tính (Nonlinear Bayesian Theorem):
   - Mô tả: Định lý Bayes phi tuyến tính áp dụng trong các tình huống mà mối quan hệ giữa các biến không phải là tuyến tính. Trong trường hợp này, các phép tính xác suất có điều kiện sẽ không tuân theo quy tắc nhân ma trận hay tích chập, mà thay vào đó sẽ sử dụng các phương pháp tính toán phức tạp hơn như phương pháp Monte Carlo.
   - Ứng dụng: Mô hình hóa các hệ thống sinh học phức tạp, phân tích dữ liệu phi tuyến tính trong kinh tế học, vật lý hạt nhân.
4. Định lý Bayes với ước lượng tiền nghiệm không chính xác (Bayes' Theorem with Inaccurate Prior Estimates):
   - Mô tả: Định lý Bayes này được áp dụng khi ước lượng tiền nghiệm (prior estimates) của các xác suất ban đầu không chắc chắn hoặc không chính xác. Trong trường hợp này, các phương pháp thống kê Bayesian có thể được sử dụng để cập nhật ước lượng của chúng ta dựa trên dữ liệu mới.
   - Ứng dụng: Phân tích các kết quả nghiên cứu khoa học với dữ liệu không hoàn hảo, cập nhật mô hình dự báo thị trường tài chính dựa trên dữ liệu mới.
5. Định lý Bayes trong mô hình học máy (Bayes' Theorem in Machine Learning Models):
   - Mô tả: Trong lĩnh vực học máy, Định lý Bayes được sử dụng trong các mô hình học máy Bayes, trong đó các xác suất được ước lượng dựa trên các điểm dữ liệu đào tạo và sử dụng để đưa ra dự đoán cho các điểm dữ liệu mới.
   - Ứng dụng: Phân loại văn bản, nhận dạng hình ảnh, dự báo chuỗi thời gian, các hệ thống khuyến nghị.

2.1.4 Chứng minh định lý Bayes:

1. Sử dụng định nghĩa xác suất có điều kiện:

   - Xác suất có điều kiện của một biến cố A khi đã biết biến cố B đã xảy ra được định nghĩa là tỷ lệ giữa các xác suất của sự kiện A và xác suất của sự kiện B đã xảy ra, khi biết rằng B đã xảy ra:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{với } P(B) \neq 0$$

   - Tương tự, xác suất có điều kiện của sự kiện B khi biết A đã xảy ra:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad \text{với } P(A) \neq 0$$

   - Suy diễn từ định nghĩa:

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

$$P(A \cap B) = P(B \mid A) \cdot P(A)$$

- Vì $P(A \cap B)$ là xác suất của A và B xảy ra đồng thời, hai công thức trên đều mô tả cùng một giá trị. Do đó, ta có:

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

- Thay thế vào công thức xác suất có điều kiện:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

2.1.5 Ví dụ minh họa cho định lý Bayes:

- Yêu Cầu: Xác định ngôn ngữ (Tiếng Anh, Tiếng Pháp, hoặc Tiếng Tây Ban Nha) của một tài liệu dựa trên các từ xuất hiện trong tài liệu đó.
- Đầu Vào:
  - Một tài liệu chứa một loạt các từ.
  - Tỉ lệ xuất hiện của từng từ trong tài liệu so với mỗi ngôn ngữ.
- Các bước thực hiện tính toán:

  Bước 1: Thu thập dữ liệu:

  - Giả sử ta có một tài liệu với các từ sau: "the", "chat", "gato".
  - Ta cũng có thông tin về tỉ lệ xuất hiện của các từ này trong ba ngôn ngữ:
    - Tiếng Anh:
      - "the": 0.07
      - "chat": 0.01
      - "gato": 0.001
    - Tiếng Pháp:
      - "the": 0.01
      - "chat": 0.05
      - "gato": 0.002
    - Tiếng Tây Ban Nha:
      - "the": 0.01
      - "chat": 0.001
      - "gato": 0.04
  - Giả sử xác suất tiên nghiệm của mỗi ngôn ngữ (trước khi biết bất kỳ từ nào) là như nhau:

$$P(\text{English}) = P(\text{French}) = P(\text{Spanish}) = \frac{1}{3}$$

Bước 2: Áp dụng Định lý Bayes:

Ta sẽ tính xác suất hậu nghiệm của mỗi ngôn ngữ dựa trên các từ trong tài liệu. Định lý Bayes:

$$P(\text{Language} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Language}) \cdot P(\text{Language})}{P(\text{Words})}$$

- Trong đó:
  - $P(\text{Words} \mid \text{Language})$ là xác suất các từ xuất hiện trong tài liệu khi biết ngôn ngữ.
  - $P(\text{Language})$ là xác suất tiên nghiệm của ngôn ngữ.
  - $P(\text{Words})$ là xác suất biên của các từ xuất hiện trong tài liệu.

Bước 3: tính toán xác suất các từ xuất hiện trong tài liệu đối với từng ngôn ngữ:

- Giả sử các từ xuất hiện độc lập, ta có:

$$P(\text{Words} \mid \text{English}) = P(\text{the} \mid \text{English}) \cdot P(\text{chat} \mid \text{English}) \cdot P(\text{gato} \mid \text{English})$$

$$P(\text{Words} \mid \text{English}) = 0.07 \cdot 0.01 \cdot 0.001 = 0.0000007$$

$$P(\text{Words} \mid \text{French}) = P(\text{the} \mid \text{French}) \cdot P(\text{chat} \mid \text{French}) \cdot P(\text{gato} \mid \text{French})$$

$$P(\text{Words} \mid \text{French}) = 0.01 \cdot 0.05 \cdot 0.002 = 0.000001$$

$$P(\text{Words} \mid \text{Spanish}) = P(\text{the} \mid \text{Spanish}) \cdot P(\text{chat} \mid \text{Spanish}) \cdot P(\text{gato} \mid \text{Spanish})$$

$$P(\text{Words} \mid \text{Spanish}) = 0.01 \cdot 0.001 \cdot 0.04 = 0.0000004$$

- Xác suất tiên nghiệm của mỗi ngôn ngữ:

$$P(\text{English}) = P(\text{French}) = P(\text{Spanish}) = \frac{1}{3}$$

- Xác suất biên của các từ xuất hiện trong tài liệu:

$$\begin{aligned} P(\text{Words}) = {} & P(\text{Words} \mid \text{English}) \cdot P(\text{English}) \\ & + P(\text{Words} \mid \text{French}) \cdot P(\text{French}) \\ & + P(\text{Words} \mid \text{Spanish}) \cdot P(\text{Spanish}) \end{aligned}$$

$$\begin{aligned} P(\text{Words}) &= 0.0000007 \cdot \frac{1}{3} + 0.000001 \cdot \frac{1}{3} + 0.0000004 \cdot \frac{1}{3} \\ &= 0.000000233 + 0.000000333 + 0.000000133 \\ &= 0.000000699 \end{aligned}$$

- Xác suất hậu nghiệm của từng ngôn ngữ:

$$P(\text{English} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{English}) \cdot P(\text{English})}{P(\text{Words})}$$

$$P(\text{English} \mid \text{Words}) = \frac{0.0000007 \cdot \frac{1}{3}}{0.000000699} = \frac{0.000000233}{0.000000699} \approx 0.33$$

$$P(\text{French} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{French}) \cdot P(\text{French})}{P(\text{Words})}$$

$$P(\text{French} \mid \text{Words}) = \frac{0.000001 \cdot \frac{1}{3}}{0.000000699} = \frac{0.000000333}{0.000000699} \approx 0.476$$

$$P(\text{Spanish} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Spanish}) \cdot P(\text{Spanish})}{P(\text{Words})}$$

$$P(\text{Spanish} \mid \text{Words}) = \frac{0.0000004 \cdot \frac{1}{3}}{0.000000699} = \frac{0.000000133}{0.000000699} \approx 0.190$$

Bước 4: Kết luận:

- Dựa trên các tính toán trên, xác suất hậu nghiệm cho từng ngôn ngữ là:
  - Tiếng Anh: 33.3%
  - Tiếng Pháp: 47.6%
  - Tiếng Tây Ban Nha: 19.0%
- Do đó, tài liệu có khả năng cao nhất là bằng Tiếng Pháp.

Tóm tắt các bước tính toán:

- Thu thập dữ liệu về tỉ lệ xuất hiện của từ trong từng ngôn ngữ và xác suất tiên nghiệm.
- Tính xác suất các từ xuất hiện trong tài liệu đối với từng ngôn ngữ.
- Tính xác suất biên của các từ xuất hiện trong tài liệu.
- Áp dụng Định lý Bayes để tính xác suất hậu nghiệm của từng ngôn ngữ.
- So sánh các xác suất hậu nghiệm để xác định ngôn ngữ có khả năng cao nhất.

2.2 Thuật toán phân loại Naïve Bayes (Naive Bayes Classifier (NBC)):

2.2.1 Lịch sử nguồn gốc ra đời của thuật toán:

Cùng với sự phát triển của định lý Bayes và sự phát triển của lĩnh vực máy học và trí tuệ nhân tạo, thuật toán Naïve Bayes đã được ra đời và áp dụng vào lý thuyết xác suất với vai trò là phân loại dữ liệu.

2.2.2 Nguồn gốc tên gọi "Naïve":

Thuật toán được gọi là "Naïve" vì giả định này thường không đúng trong thực tế, trong nhiều trường hợp bởi vì các đặc trưng thường có sự phụ thuộc lẫn nhau. Tuy nhiên, giả định Naïve thường không ảnh hưởng đến hiệu suất của thuật toán vì nó vẫn cho kết quả phân loại tốt trên nhiều tập dữ liệu thực tế.

2.2.3 Công thức của Phân loại Bayes:

Giả sử ta có một tập dữ liệu với các đặc trưng $X = \{x_1, x_2, \ldots, x_n\}$ và một tập các lớp $C = \{C_1, C_2, \ldots, C_n\}$. Ta muốn xác định lớp C_i của một mẫu mới dựa trên các đặc trưng của nó.

- Theo định lý Naïve Bayes, xác suất của lớp C_i khi biết các đặc trưng X được tính như sau:
$$P(C_i \mid X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)}$$
- Giả định độc lập:
  - Giả định rằng các đặc trưng x_j là độc lập với nhau khi biết lớp C_i, ta có:
  $$P(X \mid C_i) = P(x_1, x_2, \ldots, x_n \mid C_i) = P(x_1 \mid C_i) \cdot P(x_2 \mid C_i) \cdot \cdots \cdot P(x_n \mid C_i)$$
  - Do đó, công thức Naïve Bayes trở thành:
  $$P(C_i \mid X) = \frac{P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)}{P(X)}$$
  - Bỏ qua xác suất biên: Vì $P(X)$ là một hằng số không phụ thuộc vào lớp C_i, ta có thể bỏ qua nó khi so sánh xác suất giữa các lớp. Ta chỉ cần tính:
  $$P(C_i \mid X) \propto P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)$$
  - Phương pháp tối đa hóa: Để xác định lớp C_i cho một mẫu mới, ta chọn lớp có xác suất lớn nhất:
  $$\hat{C} = \arg\max_{C_i \in C} P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)$$

2.2.4 Chứng minh tính đúng của thuật toán phân loại Naïve Bayes:

- Từ định nghĩa xác suất có điều kiện của một sự kiện A khi biết sự kiện B được định nghĩa là:
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$
- Áp dụng định lý Bayes, xác suất có điều kiện của lớp C_i khi biết X chính là:
$$P(C_i \mid X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)}$$
- Ta có giả định độc lập:
  - Giả định rằng các đặc trưng x_j là độc lập với nhau khi biết lớp C_i, ta có:
  $$P(X \mid C_i) = P(x_1, x_2, \ldots, x_n \mid C_i) = P(x_1 \mid C_i) \cdot P(x_2 \mid C_i) \cdot \cdots \cdot P(x_n \mid C_i)$$

– Do đó, công thức Naïve Bayes trở thành:

$$P(C_i \mid X) = \frac{P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)}{P(X)}$$

– Bỏ qua xác suất biên: Vì $P(X)$ là một hằng số không phụ thuộc vào lớp C_i, ta có thể bỏ qua nó khi so sánh xác suất giữa các lớp. Ta chỉ cần tính:

$$P(C_i \mid X) \propto P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)$$

- Phương pháp tối đa hóa: Để xác định lớp C_i cho một mẫu mới, ta chọn lớp có xác suất lớn nhất:

$$\hat{C} = \arg\max_{C_i \in C} P(C_i) \cdot \prod_{j=1}^{n} P(x_j \mid C_i)$$

2.2.5 Ví dụ minh họa của thuật toán phân loại Naïve Bayes:

Giả sử chúng ta có một tập dữ liệu đơn giản với hai lớp: "spam"và "not spam", và các đặc trưng là các từ trong email.

Tập dữ liệu gồm 10 email, trong đó 5 email là "spam"và 5 email là "not spam". Các đặc trưng (từ) bao gồm "cheap", "offer", "winner", "click". Bảng dữ liệu như sau:

Email	**Cheap**	**Offer**	**Winner**	**Click**	**Class**
Email 1	1	1	0	1	Spam
Email 2	1	0	1	0	Spam
Email 3	0	1	0	1	Spam
Email 4	1	1	1	0	Spam
Email 5	1	0	0	1	Spam
Email 6	0	1	1	0	Not Spam
Email 7	0	0	1	0	Not Spam
Email 8	0	1	0	1	Not Spam
Email 9	0	0	1	0	Not Spam
Email 10	0	1	0	0	Not Spam

Bảng 2.1: Tập dữ liệu với email, từ và lớp.

Bước 1: Tính xác suất tiên nghiệm:

Xác suất tiên nghiệm của mỗi lớp:

$$P(\text{Spam}) = \frac{5}{10} = 0.5$$

$$P(\text{Not Spam}) = \frac{5}{10} = 0.5$$

Bước 2: Tính xác suất có điều kiện:

Xác suất có điều kiện của mỗi từ trong mỗi lớp:

- Cheap:
  - Trong 5 email spam, từ "cheap"xuất hiện 4 lần.
  - Trong 5 email not spam, từ "cheap"không xuất hiện.

$$P(\text{Cheap} \mid \text{Spam}) = \frac{4}{5} = 0.8$$

$$P(\text{Cheap} \mid \text{Not Spam}) = \frac{0}{5} = 0.0$$

- Offer:
  - Trong 5 email spam, từ "offer"xuất hiện 3 lần.
  - Trong 5 email not spam, từ "offer"cũng xuất hiện 3 lần.

$$P(\text{Offer} \mid \text{Spam}) = \frac{3}{5} = 0.6$$

$$P(\text{Offer} \mid \text{Not Spam}) = \frac{3}{5} = 0.6$$

- Winner:
  - Trong 5 email spam, từ "winner"xuất hiện 1 lần.
  - Trong 5 email not spam, từ "winner"xuất hiện 3 lần.

$$P(\text{Winner} \mid \text{Spam}) = \frac{1}{5} = 0.2$$

$$P(\text{Winner} \mid \text{Not Spam}) = \frac{3}{5} = 0.6$$

- Click:
  - Trong 5 email spam, từ "click"xuất hiện 3 lần.
  - Trong 5 email not spam, từ "click"xuất hiện 1 lần.

$$P(\text{Click} \mid \text{Spam}) = \frac{3}{5} = 0.6$$

$$P(\text{Click} \mid \text{Not Spam}) = \frac{1}{5} = 0.2$$

Bước 3: Tính xác suất tổng hợp:

Giả sử có một email mới với các từ: "cheap", "offer", "click". Email mới có thể được biểu diễn như sau:

Email	Cheap	Offer	Winner	Click
New Email	1	1	0	1

Bảng 2.2: Email mới với các từ và đặc trưng.

- Tính xác suất tổng hợp cho lớp "Spam":

$$P(\text{Spam} \mid \text{New Email}) \propto P(\text{Spam}) \cdot P(\text{Cheap} \mid \text{Spam}) \cdot P(\text{Offer} \mid \text{Spam}) \cdot P(\text{Click} \mid \text{Spam})$$

$$P(\text{Spam} \mid \text{New Email}) \propto 0.5 \cdot 0.8 \cdot 0.6 \cdot 0.6 = 0.5 \cdot 0.288 = 0.144$$

- Tính xác suất tổng hợp cho lớp "Not Spam":

$$P(\text{Not Spam} \mid \text{New Email}) \propto P(\text{Not Spam}) \cdot P(\text{Cheap} \mid \text{Not Spam}) \cdot P(\text{Offer} \mid \text{Not Spam}) \cdot P(\text{Click} \mid \text{Not Spam})$$

$$P(\text{Not Spam} \mid \text{New Email}) \propto 0.5 \cdot 0.0 \cdot 0.6 \cdot 0.2 = 0.5 \cdot 0 = 0$$

Bước 4: Phân lớp:

So sánh xác suất tổng hợp:

$$P(\text{Spam} \mid \text{New Email}) = 0.144$$

$$P(\text{Not Spam} \mid \text{New Email}) = 0$$

Vì $0.144 > 0$, email mới được phân loại là "Spam".

2.2.6 Một số kiểu mô hình Naive Bayes:

1. Multinomial Naïve Bayes (MNB):
   - Mô tả: Multinomial Naïve Bayes là một loại mô hình Naïve Bayes thường được sử dụng trong xử lý dữ liệu rời rạc, nơi mỗi đặc trưng là số lần xuất hiện của một từ hoặc một thuộc tính trong một mẫu. Mô hình này giả định rằng các đặc trưng theo phân phối đa thức (multinomial distribution).
   - Ứng dụng:
     - Phân loại văn bản: Ví dụ như lọc email spam, phân loại tài liệu, hoặc phân tích cảm xúc. Trong trường hợp này, mỗi tài liệu là một mẫu và mỗi từ trong từ vựng là một đặc trưng. Mô hình sẽ tính xác suất một tài liệu thuộc về một lớp dựa trên tần suất xuất hiện của các từ trong tài liệu đó.
     - Xử lý ngôn ngữ tự nhiên: Áp dụng trong các hệ thống gợi ý, tìm kiếm thông tin và các bài toán phân loại khác liên quan đến văn bản.

2. Gaussian Naïve Bayes:

   - Mô tả: Gaussian Naïve Bayes được sử dụng khi dữ liệu đầu vào là liên tục và giả định rằng các đặc trưng tuân theo phân phối Gaussian (hay phân phối chuẩn). Điều này có nghĩa là giá trị của mỗi đặc trưng được mô tả bằng một đường cong chuông.
   - Ứng dụng:
     - Dự đoán hành vi mua hàng: Ví dụ, nếu chúng ta đang xây dựng một mô hình để dự đoán nếu một người nào đó sẽ mua một sản phẩm dựa trên chiều cao và cân nặng của họ thì có thể sử dụng Gaussian Naïve Bayes.
     - Phân loại các biến liên tục: Như dự đoán điểm số học tập dựa trên các đặc trưng liên tục như thời gian học tập và số giờ ngủ.

3. Bernoulli Naïve Bayes:

   - Mô tả: Bernoulli Naïve Bayes tương tự như Multinomial Naive Bayes, nhưng thay vì đếm số lần xuất hiện của một từ trong một mẫu, nó chỉ xem xét xem một từ có xuất hiện trong một mẫu hay không. Loại mô hình này thích hợp khi dữ liệu đầu vào là nhị phân, có nghĩa là mỗi đặc trưng chỉ nhận một trong hai giá trị như "có" hoặc "không".
   - Ứng dụng:
     - Phát hiện thư rác (spam detection): Mỗi email được biểu thị bằng các đặc trưng nhị phân cho biết sự hiện diện hoặc vắng mặt của một từ cụ thể.
     - Phân loại nhị phân: Như dự đoán xem một người có mắc một bệnh cụ thể hay không dựa trên các triệu chứng có/không.

4. Complement Naïve Bayes:

   - Mô tả: Là một biến thể của Multinomial Naïve Bayes, Complement Naïve Bayes được thiết kế đặc biệt để xử lý các vấn đề mất cân bằng lớp (class imbalance). Thay vì tính xác suất cho mỗi lớp độc lập, mô hình này tính xác suất đối với các lớp còn lại và sau đó lấy nghịch đảo của nó.
   - Ứng dụng:
     - Phân loại văn bản với lớp mất cân bằng: Hiệu quả hơn khi phân loại các lớp với số lượng mẫu không đều nhau, ví dụ như khi lớp "spam" ít xuất hiện hơn so với lớp "non-spam".
     - Xử lý các bài toán mất cân bằng lớp: Trong các lĩnh vực như y tế (phát hiện bệnh hiếm gặp) hoặc an ninh (phát hiện gian lận).

5. Hybrid Naïve Bayes:

   - Mô tả: Hybrid Naïve Bayes kết hợp các loại hình Naïve Bayes khác nhau, thường kết hợp Multinomial Naive Bayes và Gaussian Naive Bayes để xử lý các dữ liệu có cả đặc trưng rời rạc và liên tục.

- Ứng dụng:
  - Xử lý dữ liệu hỗn hợp: Cải thiện hiệu suất của mô hình trên các loại dữ liệu phức tạp có cả đặc trưng rời rạc (như từ ngữ) và liên tục (như các thông số vật lý).
  - Các hệ thống gợi ý và dự báo: Trong các hệ thống đề xuất sản phẩm hoặc phân tích hành vi người dùng với dữ liệu hỗn hợp.

6. Tree Augmented Naïve Bayes (TAN):

- Mô tả: TAN là một biến thể của Naïve Bayes, thay vì giả định rằng các đặc trưng đầu vào là độc lập có điều kiện, nó sử dụng một cây Bayesian để mô hình hóa các mối quan hệ tương tác giữa các đặc trưng. TAN cho phép các đặc trưng tương tác với nhau, giúp giảm bớt sự ngây thơ của giả định Naïve Bayes.
- Ứng dụng:
  - Mô hình hóa mối quan hệ phức tạp giữa các đặc trưng: Thích hợp cho các bài toán mà các đặc trưng không độc lập, chẳng hạn như phân tích hành vi người tiêu dùng hoặc dự đoán y học.
  - Các hệ thống phức tạp: Như hệ thống chẩn đoán y tế hoặc các hệ thống dự đoán tài chính nơi các đặc trưng có mối quan hệ tương tác phức tạp.

2.3 Phân loại đa thức Naïve Bayes:

2.3.1 Đặc điểm của thuật toán phân loại đa thức Multinomial Naïve Bayes:

Giả định Naïve Bayes: giả định rằng các đặc trưng đầu vào là độc lập có điều kiện, tức là giá trị của một đặc trưng không phụ thuộc vào các đặc trưng khác khi đã biết lớp của mẫu. Mặc dù giả định này thường không được thực tế, nhưng mô hình vẫn hoạt động hiệu quả trong nhiều trường hợp thực tế.

Phân phối Multinomial: mô hình này thích hợp cho các tác vụ phân loại dựa trên các đặc trưng đếm, nơi mỗi đặc trưng có thể có nhiều hơn hai giá trị rời rạc.

- Ví dụ phổ biến là phân loại văn bản dựa trên tần suất xuất hiện các từ.

Sử dụng xác suất: Multinomial Naïve Bayes sử dụng nguyên tắc của xác suất để tính toán xác suất của một mẫu thuộc về mỗi lớp dựa trên các đặc trưng của nó.

Thuật ngữ " đa thức" (Multinomial) dùng để chỉ loại phân phối dữ liệu được mô hình giả định. Các tính năng trong phân loại văn bản thường là số từ hoặc tần số thuật ngữ. Phân phối đa thức được sử dụng để ước tính khả năng nhìn thấy một tập hợp số từ cụ thể trong tài liệu.

2.3.2 Công thức của định lý Multinomial Naïve Bayes:

- Để tính toán xác suất của một văn bản D thuộc về một lớp C_k, ta sử dụng công thức:

$$P(C_k \mid D) \propto P(C_k) \prod_{i=1}^{n} P(w_i \mid C_k)^{f_i}$$

- Trong đó:
  - $P(C_k \mid D)$ là xác suất của lớp C_k khi biết văn bản D.
  - $P(C_k)$ là xác suất tiên nghiệm của lớp C_k.
  - w_i là từ thứ i trong từ vựng.
  - f_i là số lần xuất hiện của từ w_i trong văn bản D.
  - $P(w_i \mid C_k)$ là xác suất có điều kiện của từ w_i xuất hiện trong lớp C_k.
- Xác suất tiên nghiệm $P(C_k)$: Xác suất tiên nghiệm của lớp C_k được tính bằng cách lấy tỷ lệ số văn bản trong lớp C_k so với tổng số văn bản:

$$P(C_k) = \frac{N_{C_k}}{N}$$

- Trong đó:
  - N_{C_k} là số văn bản thuộc lớp C_k.
  - N là tổng số văn bản.
- Xác suất có điều kiện $P(w_i \mid C_k)$: Để tính xác suất có điều kiện của một từ w_i xuất hiện trong lớp C_k, chúng ta sử dụng công thức:

$$P(w_i \mid C_k) = \frac{N_{w_i|C_k}}{N_{C_k}}$$

- Trong đó:
  - $N_{w_i|C_k}$ là số lần từ w_i xuất hiện trong tất cả các văn bản thuộc lớp C_k.
  - N_{C_k} là tổng số từ trong tất cả văn bản thuộc lớp C_k.

2.3.3 Laplace Smoothing:

Laplace Smoothing, còn được gọi là Additive Smoothing, là một kỹ thuật được sử dụng trong xác suất và thống kê để tránh vấn đề xác suất bằng 0 khi xử lý dữ liệu văn bản. Trong bài toán Multinomial Naive Bayes, Laplace Smoothing được áp dụng để đảm bảo rằng tất cả các từ trong từ vựng đều có một xác suất dương không bằng 0, ngay cả khi chúng không xuất hiện trong dữ liệu huấn luyện của một lớp cụ thể.

Công Thức Laplace Smoothing:

- Để tính xác suất có điều kiện $P(w_i \mid C_k)$ của một từ w_i trong lớp C_k khi sử dụng Laplace Smoothing, chúng ta sử dụng công thức sau:

$$P(w_i \mid C_k) = \frac{N_{w_i|C_k} + \alpha}{N_{C_k} + \alpha \cdot |V|}$$

- Trong đó:
  - $N_{w_i|C_k}$ là số lần từ w_i xuất hiện trong tất cả các văn bản thuộc lớp C_k.
  - N_{C_k} là tổng số từ trong tất cả các văn bản thuộc lớp C_k.
  - α là tham số làm trơn, thường được đặt là 1.
  - $|V|$ là kích thước từ vựng (tổng số từ duy nhất trong tập dữ liệu).

- Giải thích công thức:
  - Tham số làm trơn α: Giá trị α thường được đặt là 1 để đảm bảo rằng tất cả các từ trong từ vựng đều có xác suất dương. Giá trị này có thể được điều chỉnh dựa trên yêu cầu cụ thể của bài toán.
  - Số lần xuất hiện của từ $N_{w_i|C_k}$: Là số lần từ w_i xuất hiện trong tất cả các văn bản thuộc lớp C_k.
  - Tổng số từ trong lớp N_{C_k}: Là tổng số từ trong tất cả các văn bản thuộc lớp C_k.
  - Kích thước từ vựng $|V|$: Là tổng số từ duy nhất trong từ vựng của tập dữ liệu.

2.3.4 Chứng minh định lý Multinomial Naïve Bayes:

Để chứng minh tính đúng đắn của thuật toán Multinomial Naïve Bayes, ta cần trải qua nhiều bước tính toán:

1. Định lý Bayes:

   Theo định lý Bayes, xác suất của lớp C_k khi biết văn bản D được tính như sau:

$$P(C_k \mid D) = \frac{P(D \mid C_k) \cdot P(C_k)}{P(D)} \tag{1}$$

   - Trong đó:
     - $P(D \mid C_k)$ là xác suất có điều kiện của văn bản D khi biết lớp C_k.
     - $P(C_k)$ là xác suất tiên nghiệm của lớp C_k.
     - $P(D)$ là xác suất tiên nghiệm của văn bản D.

2. Xác suất tiên nghiệm:

   Xác suất tiên nghiệm của lớp C_k được tính như sau:

$$P(C_k) = \frac{N_{C_k}}{N} \tag{2}$$

   - Trong đó:
     - N_{C_k} là số văn bản thuộc lớp C_k.
     - N là tổng số văn bản.
     -

3. Xác suất có điều kiện của văn bản D khi biết lớp C_k:

   - Với giả định các từ trong văn bản là độc lập có điều kiện khi biết lớp C_k, xác suất có điều kiện của văn bản D khi biết lớp C_k là:

$$P(D \mid C_k) = P(w_1, w_2, \ldots, w_n \mid C_k)$$

   - Do giả định độc lập có điều kiện, ta có:

$$P(D \mid C_k) = \prod_{i=1}^{n} P(w_i \mid C_k)^{f_i} \tag{3}$$

   - Trong đó:
     - f_i là số lần xuất hiện của từ w_i trong văn bản D.

4. Xác suất có điều kiện của từ w_i xuất hiện trong lớp C_k:

$$P(w_i \mid C_k) = \frac{N_{w_i|C_k}}{N_{C_k}} \tag{4}$$

- Trong đó:
  - $N_{w_i|C_k}$ là số lần từ w_i xuất hiện trong tất cả các văn bản thuộc lớp C_k.
  - N_{C_k} là tổng số từ trong tất cả các văn bản thuộc lớp C_k.
- Để tránh vấn đề xác suất bằng 0, ta sử dụng Laplace smoothing:

$$P(w_i \mid C_k) = \frac{N_{w_i|C_k} + \alpha}{N_{C_k} + \alpha \cdot |V|} \tag{5}$$

- Trong đó:
  - α là tham số làm trơn, thường được đặt là 1.
  - $|V|$ là kích thước từ vựng.

5. Áp dụng vào Công thức định lý Bayes:

- Thay các công thức (1), (2), (3), (4), (5) vào định lý Bayes, ta có:

$$P(C_k \mid D) = \frac{P(D \mid C_k) \cdot P(C_k)}{P(D)}$$

- Do $P(D)$ là hằng số không phụ thuộc vào lớp C_k, ta có thể bỏ qua $P(D)$ khi so sánh giữa các lớp. Từ đó công thức trở thành:

$$P(C_k \mid D) \propto P(C_k) \cdot P(D \mid C_k)$$

- Thay $P(D \mid C_k)$ bằng $\prod_{i=1}^{n} P(w_i \mid C_k)^{f_i}$, ta được:

$$P(C_k \mid D) \propto P(C_k) \cdot \prod_{i=1}^{n} P(w_i \mid C_k)^{f_i}$$

- Tối đa hóa xác suất:
  Để xác định lớp C_k có xác suất cao nhất, ta tối đa hóa $P(C_k \mid D)$:

$$\hat{C} = \arg\max_{C_k} P(C_k) \cdot \prod_{i=1}^{n} P(w_i \mid C_k)^{f_i}$$

2.3.5 Ví dụ minh họa Multinomial Naïve Bayes:

Giả sử chúng ta có một tập dữ liệu văn bản với ba ngôn ngữ: tiếng Anh (English), tiếng Pháp (French), và tiếng Tây Ban Nha (Spanish).

Ngôn ngữ	**Câu**
English	I love you
English	I like apples
English	You are amazing
French	Je t'aime
French	J'aime les pommes
French	Tu es incroyable
Spanish	Te quiero
Spanish	Me gustan las manzanas
Spanish	Eres increíble

Bảng 2.3: Bảng các ngôn ngữ và ví dụ.

Bước 1: Xây dựng từ vựng:
Từ vựng (*vocabulary*) bao gồm tất cả các từ xuất hiện trong tập dữ liệu:

{I, love, you, like, apples, are, amazing, Je, t'aime, J'aime, les, pommes,
Tu, es, incroyable, Te, quiero, Me, gustan, las, manzanas, Eres, increíble}

Tổng cộng có 22 từ.
Bước 2: Tính xác suất tiên nghiệm của mỗi ngôn ngữ:

$$P(\text{English}) = \frac{3}{9} = \frac{1}{3} \approx 0.333$$

$$P(\text{French}) = \frac{3}{9} = \frac{1}{3} \approx 0.333$$

$$P(\text{Spanish}) = \frac{3}{9} = \frac{1}{3} \approx 0.333$$

Bước 3: Tính xác suất có điều kiện $P(w_i \mid C_k)$ = số lần xuất hiện của từ trong ngôn ngữ | ngôn ngữ:

1. Trường hợp không áp dụng Laplace smoothing:

   - English:

$$P(\text{I} \mid \text{English}) = \frac{2}{7} \approx 0.286$$

$$P(\text{love} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

$$P(\text{you} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

$$P(\text{like} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

$$P(\text{apples} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

$$P(\text{are} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

$$P(\text{amazing} \mid \text{English}) = \frac{1}{7} \approx 0.143$$

- French:

$$P(\text{Je} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{t'aime} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{J'aime} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{les} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{pommes} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{Tu} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{es} \mid \text{French}) = \frac{1}{8} = 0.125$$

$$P(\text{incroyable} \mid \text{French}) = \frac{1}{8} = 0.125$$

- Spanish:

$$P(\text{Te} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{quiero} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{Me} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{gustan} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{las} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{manzanas} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{Eres} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

$$P(\text{increíble} \mid \text{Spanish}) = \frac{1}{8} = 0.125$$

2. Trường hợp áp dụng Laplace smoothing: Sử dụng Laplace smoothing với $\alpha = 1$.

- English:

$$P(\text{I}|\text{English}) = \frac{2+1}{7+22} = \frac{3}{29} \approx 0.103$$
$$P(\text{love}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{you}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{like}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{apples}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{are}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{amazing}|\text{English}) = \frac{1+1}{7+22} = \frac{2}{29} \approx 0.069$$
$$P(\text{other}|\text{English}) = \frac{0+1}{7+22} = \frac{1}{29} \approx 0.034$$

- French:

$$P(\text{Je}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{t'aime}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{J'aime}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{les}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{pommes}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{Tu}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{es}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{incroyable}|\text{French}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{other}|\text{French}) = \frac{0+1}{8+22} = \frac{1}{30} \approx 0.033$$

- Spanish:

$$P(\text{Te}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{quiero}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$
$$P(\text{Me}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{gustan}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{las}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{manzanas}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{Eres}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{increíble}|\text{Spanish}) = \frac{1+1}{8+22} = \frac{2}{30} \approx 0.067$$

$$P(\text{other}|\text{Spanish}) = \frac{0+1}{8+22} = \frac{1}{30} \approx 0.033$$

Bước 4: Tính xác suất tổng hợp:
Giả sử câu cần phân loại là: "I love you".
Câu này có thể biểu diễn như sau:

Sentence	**I**	**Love**	**You**
"I love 1	1	1	1

Bảng 2.4: Bảng biểu diễn dữ liệu đầu vào mới.

- Trường hợp không áp dụng Laplace smoothing:
  - English:

$$P(\text{English} \mid \text{I love you}) \propto P(\text{English}) \times P(\text{I} \mid \text{English}) \times P(\text{love} \mid \text{English}) \times P(\text{you} \mid \text{English})$$

$$P(\text{English} \mid \text{I love you}) \propto 0.333 \times 0.286 \times 0.143 \times 0.143 \approx 0.333 \times 0.0059 \approx 0.002$$

  - French:

$$P(\text{French} \mid \text{I love you}) \propto P(\text{French}) \times P(\text{I} \mid \text{French}) \times P(\text{love} \mid \text{French}) \times P(\text{you} \mid \text{French})$$

$$P(\text{French} \mid \text{I love you}) = 0$$

  - Spanish:

$$P(\text{Spanish} \mid \text{I love you}) \propto P(\text{Spanish}) \times P(\text{I} \mid \text{Spanish}) \times P(\text{love} \mid \text{Spanish}) \times P(\text{you} \mid \text{Spanish})$$

$$P(\text{Spanish} \mid \text{I love you}) = 0$$

- Trường hợp áp dụng Laplace smoothing:
  - English:

$$P(\text{English} \mid \text{I love you}) \propto P(\text{English}) \times P(\text{I} \mid \text{English}) \times P(\text{love} \mid \text{English}) \times P(\text{you} \mid \text{English})$$

$$P(\text{English} \mid \text{I love you}) \propto 0.333 \times 0.103 \times 0.069 \times 0.069 \approx 0.333 \times 0.0004941 \approx 0.0001649$$

  - French:

$$P(\text{French} \mid \text{I love you}) \propto P(\text{French}) \times P(\text{I} \mid \text{French}) \times P(\text{love} \mid \text{French}) \times P(\text{you} \mid \text{French})$$

$$P(\text{French} \mid \text{I love you}) \propto 0.333 \times 0.033 \times 0.033 \times 0.033 \approx 0.333 \times 0.000035937 \approx 0.00001198$$

  - Spanish:

$$P(\text{Spanish} \mid \text{I love you}) \propto P(\text{Spanish}) \times P(\text{I} \mid \text{Spanish}) \times P(\text{love} \mid \text{Spanish}) \times P(\text{you} \mid \text{Spanish})$$

$$P(\text{Spanish} \mid \text{I love you}) \propto 0.333 \times 0.033 \times 0.033 \times 0.033 \approx 0.333 \times 0.000035937 \approx 0.00001198$$

Bước 5: Phân lớp:
So sánh xác suất tổng hợp:

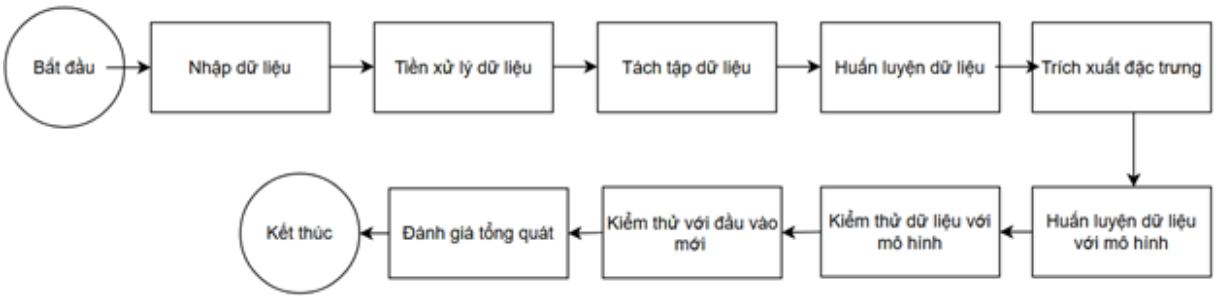
- Không Áp Dụng Laplace Smoothing:
  - $P(\text{English}|\text{I love you}) \approx 0.002$
  - $P(\text{French}|\text{I love you}) = 0$
  - $P(\text{Spanish}|\text{I love you}) = 0$
    Câu "I love you"được phân loại là tiếng Anh (English).
- Áp Dụng Laplace Smoothing:
  - $P(\text{English}|\text{I love you}) \approx 0.0001649$
  - $P(\text{French}|\text{I love you}) \approx 0.00001198$
  - $P(\text{Spanish}|\text{I love you}) \approx 0.00001198$
    Câu "I love you"được phân loại là tiếng Anh (English).

Chương 3

Thực nghiệm và đánh giá

3.1 Các bước thực hiện:

3.1.1 Lưu đồ thể hiện các bước trong bài code:



3.1.2 Các thư viện được sử dụng trong code:

1. NumPy: là thư viện cơ bản cho tính toán khoa học trong Python, cung cấp cấu trúc mảng đa chiều hiệu quả và các hàm toán học mạnh mẽ.

2. Pandas: là thư viện phân tích dữ liệu mạnh mẽ, cung cấp cấu trúc DataFrame cho phép lưu trữ và thao tác dữ liệu dạng bảng dễ dàng.

3. Scikit-learn (sklearn): là thư viện hàng đầu cho học máy, cung cấp các thuật toán phân loại, hồi quy, phân cụm và các công cụ đánh giá mô hình.

4. Seaborn: là thư viện trực quan hóa dữ liệu dựa trên matplotlib, giúp tạo ra các biểu đồ thống kê đẹp mắt và dễ hiểu.

5. Matplotlib: là thư viện cơ bản để tạo biểu đồ trong Python, hỗ trợ nhiều loại biểu đồ và cho phép tùy chỉnh cao.

6. SciPy là thư viện cho tính toán khoa học và kỹ thuật, cung cấp các công cụ cho đại số tuyến tính, tối ưu hóa, tích phân và vi phân.

7. Langdetect: là thư viện xác định ngôn ngữ của đoạn văn bản, hỗ trợ nhiều ngôn ngữ và dễ sử dụng.

3.1.3 Giải thích chi tiết từng bước (character):

Bước 1: Nhập dữ liệu:

	Text	language
0	klement gottwaldi surnukeha palsameeriti ning ...	Estonian
1	sebes joseph pereira thomas på eng the jesuit...	Swedish
2	ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...	Thai
3	விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...	Tamil
4	de spons behoort tot het geslacht haliclona en...	Dutch
...	...	...
21995	hors du terrain les années et sont des année...	French
21996	ใน พศ หลักจากที่เสด็จประพาสแหลมมลายู ชวา อินเ...	Thai
21997	con motivo de la celebración del septuagésimoq...	Spanish
21998	年月，當時還只有歲的她在美國出道，以mai-k名義推出首張英文《baby i like》，由...	Chinese
21999	aprilie sonda spațială messenger a nasa și-a ...	Romanian

22000 rows × 2 columns

Bước 2: Tiền xử lý dữ liệu:

```
Before cleaning:
0    klement gottwaldi surnukeha palsameeriti ning ...
1    sebes joseph pereira thomas  på eng the jesuit...
2    ถนนเจริญกรุง อักษรโรมัน thanon charoen krung เ...
3    விசாகப்பட்டினம் தமிழ்ச்சங்கத்தை இந்துப் பத்திர...
4    de spons behoort tot het geslacht haliclona en...
Name: Text, dtype: object

After cleaning:
0    klementgottwaldisurnukehapalsameeritiningpaigu...
1    sebesjosephpereirathomaspåengthejesuitsandthes...
2    ถนนเจรญกรงอกษรโรมนthanoncharoenkrungเรมดงแตถนน...
3    வசகபபடடனமதமழசசஙகததஇநதபபததரகவசகபபடடனஆசரயரசமபததட...
4    desponsbehoorttothetgeslachthaliclonaenbehoort...
Name: Text, dtype: object
```

Tách dữ liệu và kiểm tra dữ liệu sau khi đã tách

Bước 3: Tách dữ liệu:

```
X train: 17600
X test: 4400
Y train: 17600
Y test: 4400
```

Bước 4: Trích xuất đặc trưng:

- Uni-grams:

```
Number of unigrams in the training set: 6572
```

- Bi-grams:

```
Number of bigrams: 159652
```

- Cả Uni-grams và Bi-grams (1% Mixture):

```
Length of features: 2849
```

Bước 5: Xây dựng từ điển đặc trưng (tần xuất xuất hiện):

```
1: a
2: aa
3: ab
4: ac
5: ad
```

Bước 6: Huấn luyện dữ liệu:

- Uni-Grams:

```
Thai 119
Swedish 68
Tamil 110
Russian 68
Urdu 107
Chinese 3190
Spanish 49
English 49
Persian 76
Pushto 175
Romanian 91
Arabic 71
Portugese 56
Turkish 100
Estonian 77
Hindi 115
Dutch 48
Latin 118
French 58
Korean 1681
Indonesian 77
Japanese 2009
```

- Bi-Grams (1%):

```
Spanish: ['ad', 'al', 'an', 'ar', 'as', 'ci', 'co', 'de', 'el', 'en', 'er', 'es', 'la', 'nt', 'on', 'or', 'os', 'ra', 're', 'se', 'te']
Italian (Latin): ['ae', 'am', 'ar', 'at', 'en', 'er', 'es', 'ia', 'ic', 'in', 'is', 'it', 'li', 'ni', 'nt', 'on', 'ra', 'ri', 'st', 'ta', 'te', 'ti', 't
u', 'um', 'us']
English: ['an', 'ar', 'at', 'ed', 'en', 'er', 'es', 'he', 'in', 'nd', 'nt', 'on', 'or', 're', 'st', 'te', 'th', 'ti']
Dutch: ['aa', 'an', 'de', 'ee', 'el', 'en', 'er', 'et', 'ge', 'he', 'ie', 'in', 'nd', 'or', 'st', 'te']
Chinese: []
Japanese: []
```

- Mixture (Top 1%):

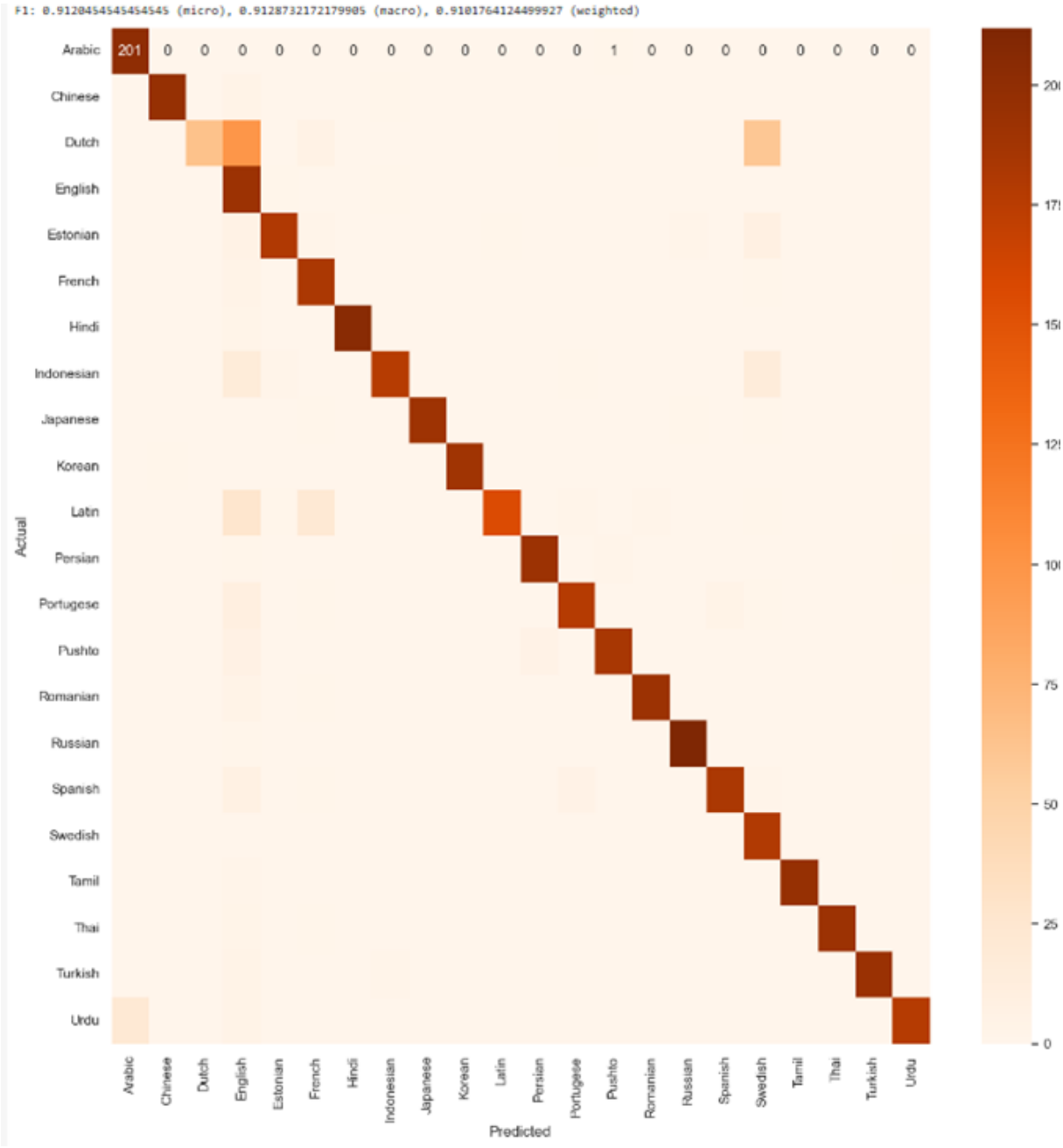
```
Thai: 658
Swedish: 556
Tamil: 530
Russian: 531
Urdu: 960
Chinese: 799
Spanish: 507
English: 490
Persian: 638
Pushto: 1050
Romanian: 530
Arabic: 661
Portugese: 531
Turkish: 585
Estonian: 581
Hindi: 590
Dutch: 518
Latin: 534
French: 517
Korean: 702
Indonesian: 491
Japanese: 830
```

- Mixture (Top 50):

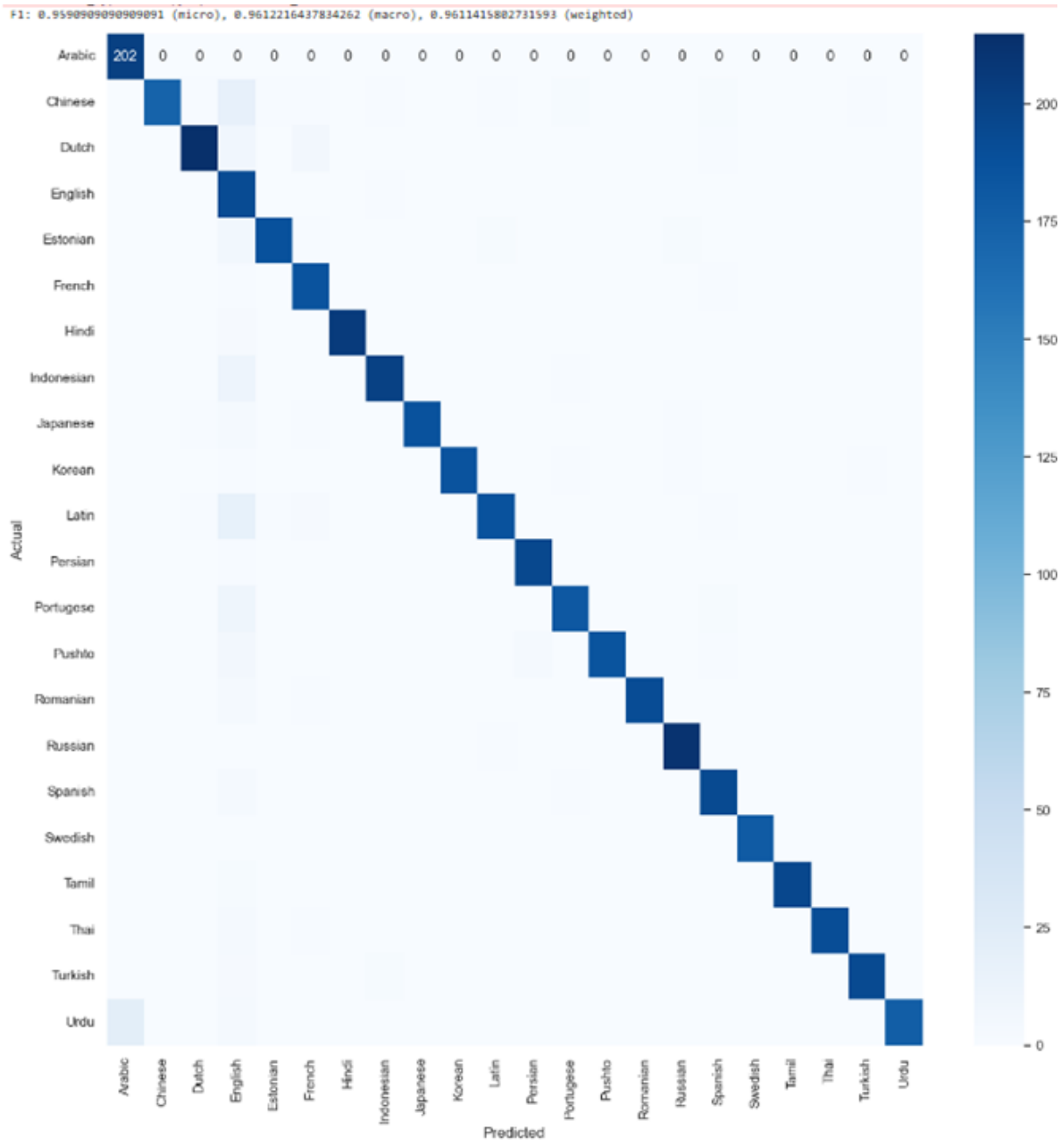
```
Top 5 n-grams for Swedish:
1: e
2: r
3: a
4: n
5: t
```

Bước 7: Kiểm thử dữ liệu với mô hình:

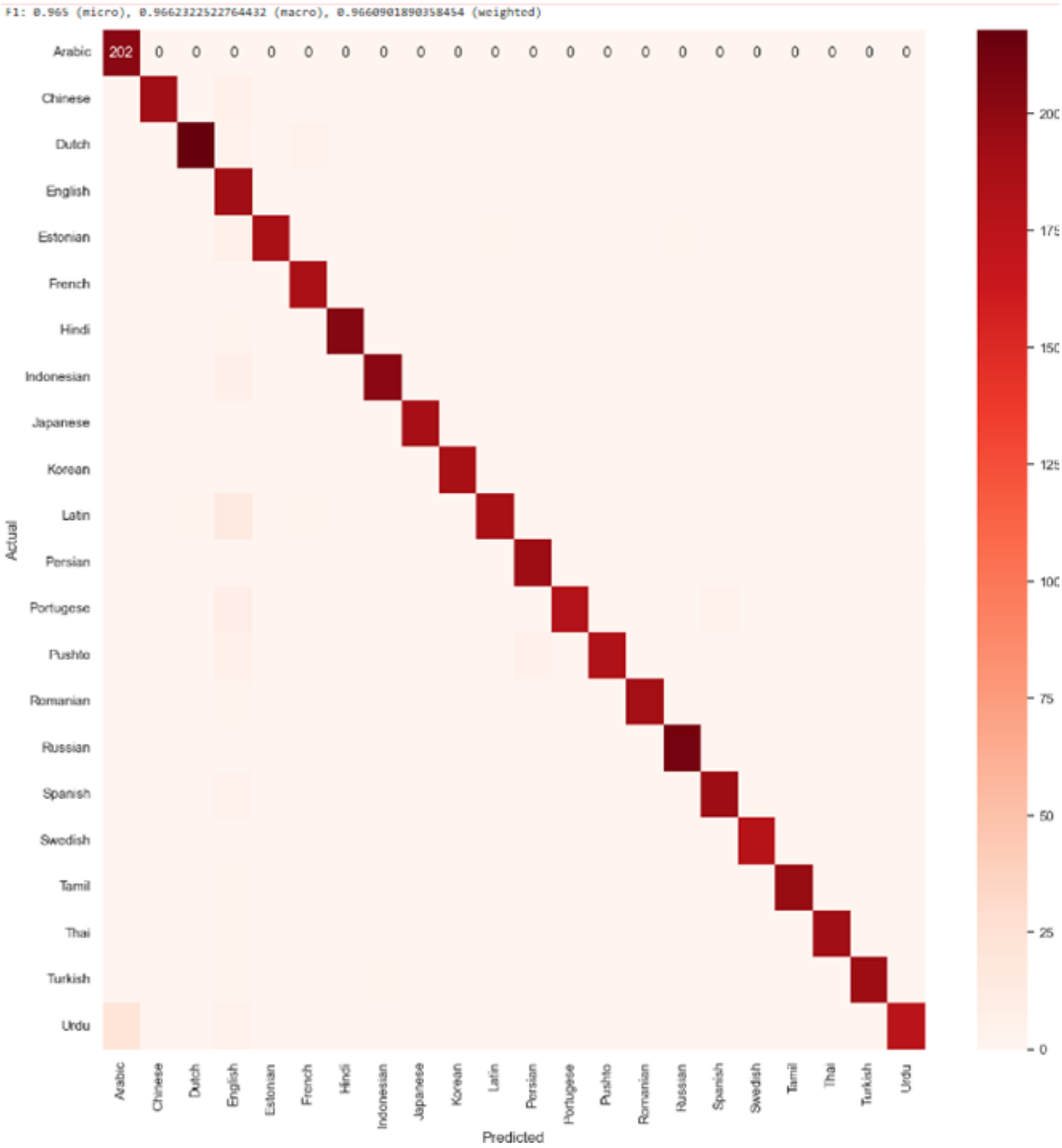
- Ma trận có được:
  - Uni-Grams:



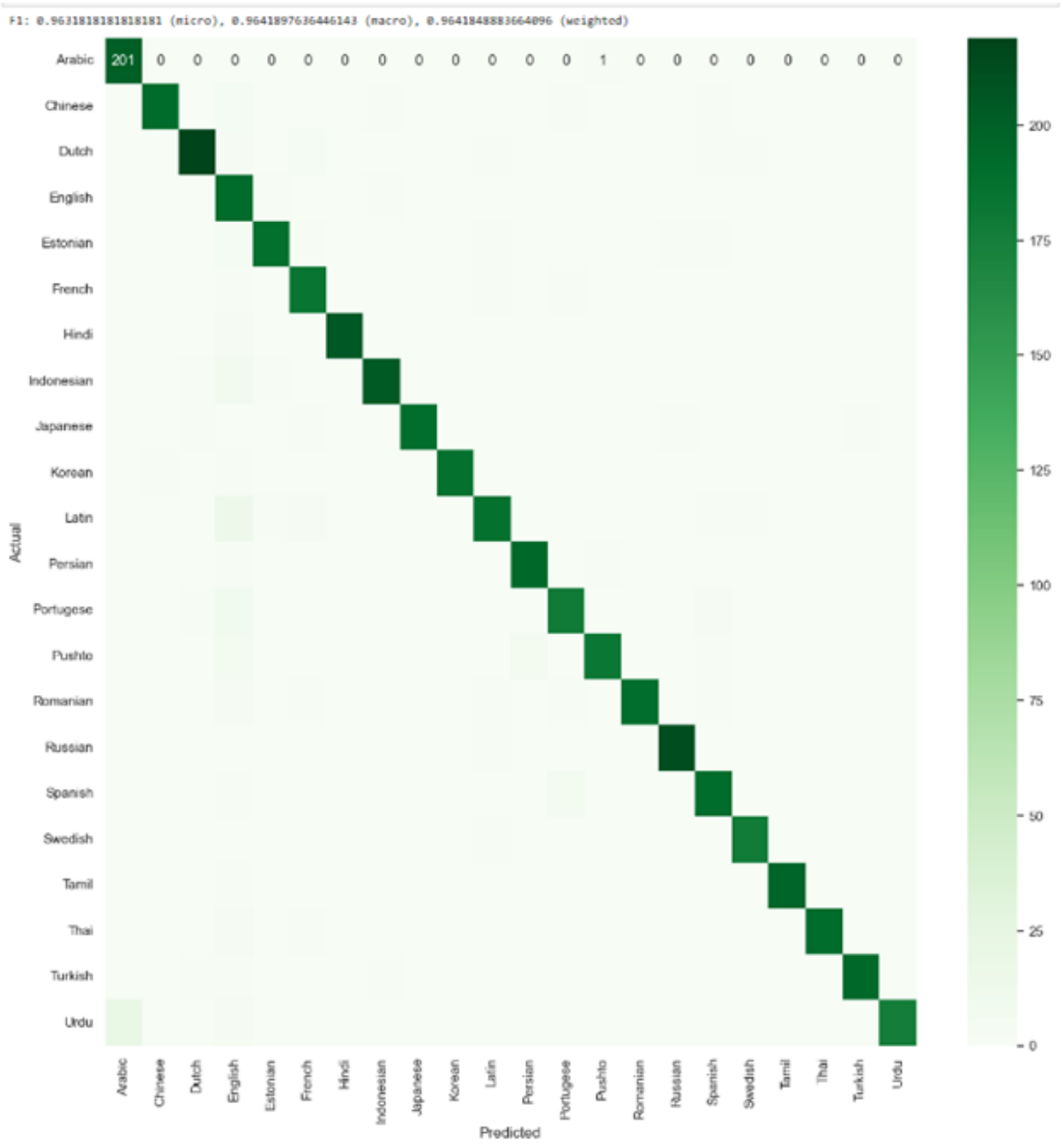
– Bi-Grams:


F1: 0.9590909090909091 (micro), 0.9612216437834262 (macro), 0.9611415882731593 (weighted)
Actual
Predicted
Arabic
Chinese
Dutch
English
Estonian
French
Hindi
Indonesian
Japanese
Korean
Latin
Persian
Portugese
Pushto
Romanian
Russian
Spanish
Swedish
Tamil
Thai
Turkish
Urdu
202
200
175
150
125
100
75
50
25
0

– Mixture (Top 1%):



– Mixture (Top 50):



Bước 8: Kiểm thử mô hình đã huấn luyện với dữ liệu đầu vào mới:
Bước 9: Đánh giá tổng quát:

- Các trường hợp dự đoán sai của máy:

```
#1: Expected: Urdu, Predicted: English
Text: belowwasthelistofthetypesofsportsplayedintheseagfromthebulletmarkindicatesthatthesportwasplayedintherespectiveyear
==================================================
#2: Expected: Portugese, Predicted: Dutch
Text: stephenmidgleypeterstevensbenrichardsonpaulgiolanicholaslanderworldwidewattlewebseiteüberdieakzienmiteinemschwerpunktaufdieaustralischenarten
==================================================
#3: Expected: Korean, Predicted: English
Text: 죽생도■生콥산스크리트어tiyagyonitiryañctiryaṁargа물리어tiracchānayoni또는죽생■生산스크리트어tiryañc물리어tiracchāna은별레군중이류조류짐승을포함한온갖동
물생명체를통칭하는대죽생에는사람의몸속에사는미세생명체도포함되며유용한가루다와같은신화적이거나신적인생명체도포함된다
==================================================
#4: Expected: Spanish, Predicted: English
Text: comoactrizhatrabajadoparaproductorascomoevilangelgirlfriendsfilmswickedbrazzerssweetheartvideonaughtyamericamilehighpureplaymediaforbiddenfruitsfil
msmilehighdigitalsinohardx
==================================================
#5: Expected: Thai, Predicted: English
Text: moralesfjarribaslorenzogtheformationofpotentiallyharmfulcompoundsinchurrosaspanishfrieddoughpastryasinfluencedbydeepfryingconditionsfoodchemistry
```

- Các trường hợp dự đoán đúng của máy:

- Kiểm tra độ chính xác của từng bộ huấn luyện:
  - Uni-Grams:

Naive Bayes on Unigrams:
Accuracy: 0.9120454545454545

	precision	recall	f1-score	support
Arabic	0.91	1.00	0.95	202
Chinese	0.99	0.98	0.98	201
Dutch	1.00	0.28	0.44	230
English	0.48	0.99	0.65	194
Estonian	0.98	0.90	0.94	200
French	0.84	0.97	0.90	188
Hindi	1.00	0.99	0.99	208
Indonesian	0.98	0.83	0.90	213
Japanese	1.00	0.98	0.99	194
Korean	1.00	0.99	1.00	190
Latin	0.99	0.74	0.85	210
Persian	0.97	0.98	0.98	196
Portugese	0.95	0.92	0.93	194
Pushto	0.98	0.94	0.96	196
Romanian	0.99	0.97	0.98	197
Russian	0.99	1.00	0.99	213
Spanish	0.98	0.92	0.95	199
Swedish	0.66	1.00	0.80	179
Tamil	1.00	0.99	0.99	198
Thai	1.00	0.98	0.99	196
Turkish	1.00	0.97	0.98	199
Urdu	0.99	0.88	0.93	203
accuracy			0.91	4400
macro avg	0.94	0.92	0.91	4400
weighted avg	0.94	0.91	0.91	4400

– Bi-Grams:

Naive Bayes on Bigrams:
Accuracy: 0.9590909090909091

	precision	recall	f1-score	support
Arabic	0.90	1.00	0.95	202
Chinese	1.00	0.86	0.93	201
Dutch	0.99	0.93	0.96	230
English	0.64	0.99	0.78	194
Estonian	0.99	0.94	0.96	200
French	0.93	0.99	0.96	188
Hindi	1.00	0.99	0.99	208
Indonesian	0.98	0.94	0.96	213
Japanese	1.00	0.96	0.98	194
Korean	1.00	0.98	0.99	190
Latin	0.98	0.89	0.93	210
Persian	0.98	0.99	0.99	196
Portugese	0.97	0.94	0.96	194
Pushto	1.00	0.94	0.97	196
Romanian	1.00	0.97	0.99	197
Russian	0.98	1.00	0.99	213
Spanish	0.96	0.97	0.97	199
Swedish	1.00	1.00	1.00	179
Tamil	1.00	0.99	0.99	198
Thai	1.00	0.97	0.99	196
Turkish	0.99	0.97	0.98	199
Urdu	1.00	0.87	0.93	203
accuracy			0.96	4400
macro avg	0.97	0.96	0.96	4400
weighted avg	0.97	0.96	0.96	4400

– Mixture (Top 50):

Naive Bayes on Mix Top 50:
Accuracy: 0.9631818181818181

	precision	recall	f1-score	support
Arabic	0.90	1.00	0.95	202
Chinese	0.99	0.96	0.97	201
Dutch	0.98	0.95	0.97	230
English	0.72	0.99	0.83	194
Estonian	0.98	0.94	0.96	200
French	0.93	0.98	0.96	188
Hindi	1.00	0.99	0.99	208
Indonesian	0.98	0.96	0.97	213
Japanese	1.00	0.98	0.99	194
Korean	1.00	0.99	0.99	190
Latin	0.96	0.90	0.93	210
Persian	0.97	0.99	0.98	196
Portugese	0.95	0.92	0.93	194
Pushto	0.99	0.93	0.96	196
Romanian	1.00	0.96	0.98	197
Russian	0.99	1.00	0.99	213
Spanish	0.95	0.96	0.96	199
Swedish	0.99	0.99	0.99	179
Tamil	1.00	0.99	0.99	198
Thai	1.00	0.98	0.99	196
Turkish	0.99	0.97	0.98	199
Urdu	1.00	0.87	0.93	203
accuracy			0.96	4400
macro avg	0.97	0.96	0.96	4400
weighted avg	0.97	0.96	0.96	4400

– Mixture (Top 1%):

Naive Bayes on Mix Top 1%:
Accuracy: 0.965

	precision	recall	f1-score	support
Arabic	0.90	1.00	0.95	202
Chinese	0.99	0.96	0.97	201
Dutch	0.97	0.95	0.96	230
English	0.71	0.99	0.83	194
Estonian	0.99	0.94	0.97	200
French	0.94	0.99	0.96	188
Hindi	1.00	0.99	0.99	208
Indonesian	0.98	0.95	0.96	213
Japanese	1.00	0.98	0.99	194
Korean	1.00	0.99	1.00	190
Latin	0.96	0.90	0.93	210
Persian	0.97	0.99	0.98	196
Portugese	0.98	0.93	0.95	194
Pushto	1.00	0.93	0.97	196
Romanian	1.00	0.97	0.98	197
Russian	0.99	1.00	0.99	213
Spanish	0.96	0.97	0.97	199
Swedish	0.99	1.00	1.00	179
Tamil	1.00	0.99	0.99	198
Thai	1.00	0.98	0.99	196
Turkish	1.00	0.97	0.99	199
Urdu	0.99	0.87	0.93	203
accuracy			0.96	4400
macro avg	0.97	0.97	0.97	4400
weighted avg	0.97	0.96	0.97	4400

3.1.4 Giải thích chi tiết từng bước (word):

Bước 1: Nhập dữ liệu:

	Text	language
0	هذه بغداد لم تكن دار كفر قط وجرى عليها هذا الذ...	Arabic
1	الصوف الذهبي هو صوف خيالي لكبش طائر خيالي تناق...	Arabic
2	دهرانة محلة تابعة لقرية عنقب التابعة لعزلة بني...	Arabic
3	أدان البابا "الاستغلال الجائر" للموارد الطبيعي...	Arabic
4	قامت داعش بتدمير مرقدي النبي يونس والنبي شيت إ...	Arabic
...	...	...
10995	تم نے اپنا عقیدہ بیان کر دیا اب میری باری ہے ک...	Urdu
10996	تاریخ الاسلام یہ علامہ ذہبی کی تاریخ ہے جو بیس...	Urdu
10997	آپ صورت و سیرت میں سلف صالحین کی جیتی جاگتی تص...	Urdu
10998	بعد ازاں جوزف سٹالن کے دور میں سوویت ثقافت کو ...	Urdu
10999	عدد یعنی ایک عدد اور ایک علامت ہے۔ کچھ عددی ...	Urdu

11000 rows × 2 columns

Bước 2: Tiền xử lý dữ liệu:

```
Before cleaning:
0    هذه بغداد لم تكن دار كفر قط وجرى عليها هذا الذ...
1    الصوف الذهبي هو صوف خيالي لكبش طائر خيالي تناق...
2    دهرانة محلة تابعة لقرية عنقب التابعة لعزلة بني...
3    أدان البابا "الاستغلال الجائر" للموارد الطبيعي...
4    قامت داعش بتدمير مرقدي النبي يونس والنبي شيت إ...
Name: Text, dtype: object

After cleaning:
0    هذه بغداد لم تكن دار كفر قط وجرى عليها هذا الذ...
1    الصوف الذهبي هو صوف خيالي لكبش طائر خيالي تناق...
2    دهرانة محلة تابعة لقرية عنقب التابعة لعزلة بني...
3    أدان البابا الاستغلال الجائر للموارد الطبيعية ...
4    قامت داعش بتدمير مرقدي النبي يونس والنبي شيت إ...
Name: Text, dtype: object
```

Bước 3: Tách dữ liệu:

```
X train: 8800
X test: 2200
Y train: 8800
Y test: 2200
```

Bước 4: Trích xuất đặc trưng:

- Uni-grams:

```
Number of unigrams in the training set: 138317
```

- Bi-grams:

```
Number of bigrams: 329240
```

- Cả Uni-grams và Bi-grams (1% Mixture):

```
Length of features: 210
```

Bước 5: Xây dựng từ điển đặc trưng (tần xuất xuất hiện):

```
1: aan
2: ad
3: adalah
4: al
5: anno
```

Bước 6: Huấn luyện dữ liệu:

- Uni-Grams:

```
Urdu 5865
Persian 6644
Russian 9092
Hindi 5423
Swedish 4229
Arabic 10662
Japanese 1007
Spanish 6401
Estonian 8026
Dutch 5737
Turkish 8972
Tamil 7397
Latin 7041
French 6866
Korean 14193
Pushto 9878
Indonesian 5702
Chinese 1061
Thai 5372
Portugese 6863
Romanian 7022
English 6467
```

- Bi-Grams (1%):

```
Spanish: []
Italian (Latin): []
English: []
Dutch: []
Chinese: []
Japanese: []
```

- Mixture (Top 1%):

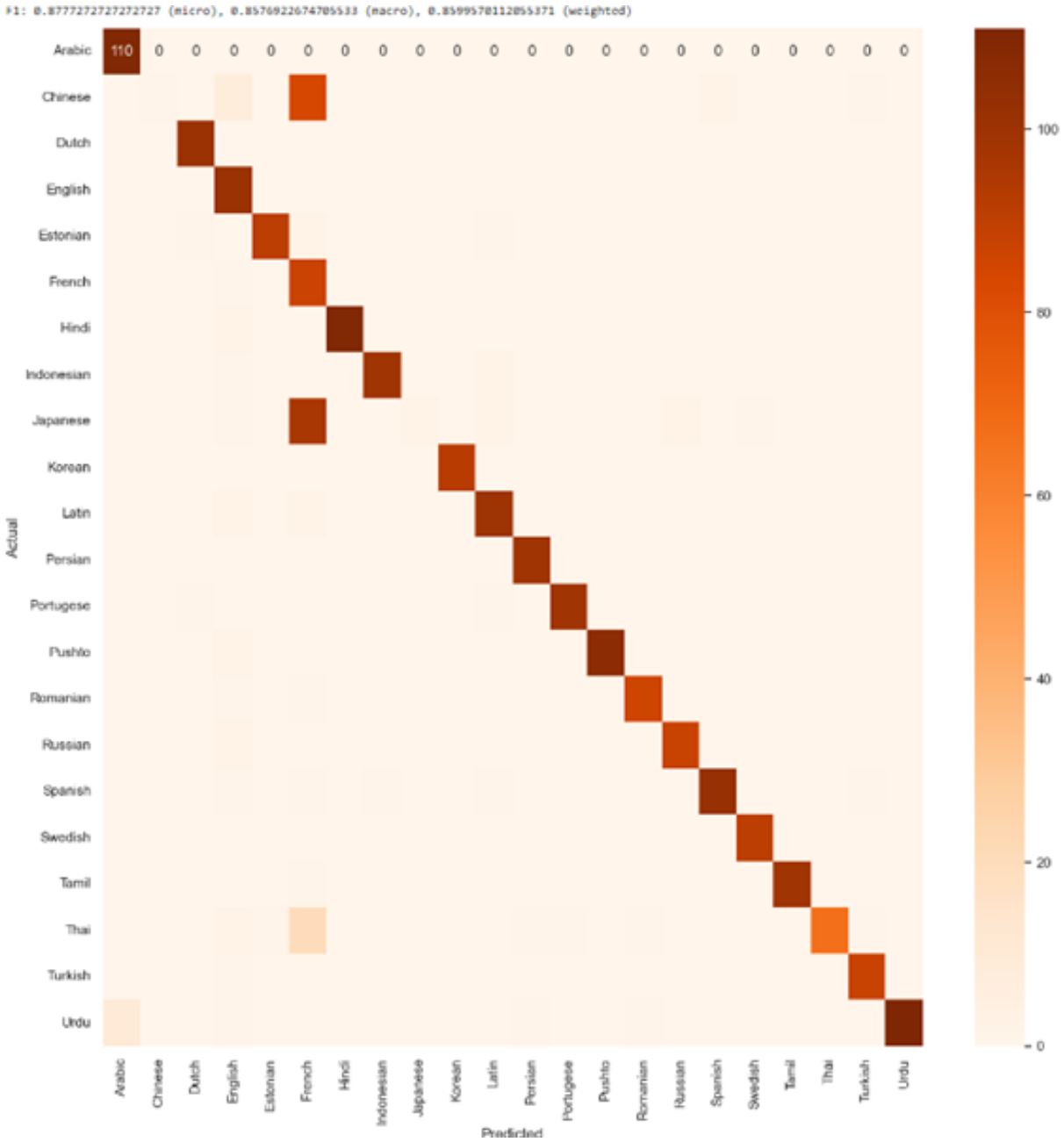
```
Urdu: 42
Persian: 43
Russian: 7
Hindi: 21
Swedish: 37
Arabic: 23
Japanese: 10
Spanish: 37
Estonian: 15
Dutch: 41
Turkish: 34
Tamil: 4
Latin: 30
French: 45
Korean: 5
Pushto: 48
Indonesian: 38
Chinese: 3
Thai: 7
Portugese: 40
Romanian: 39
English: 25
```

- Mixture (Top 50):

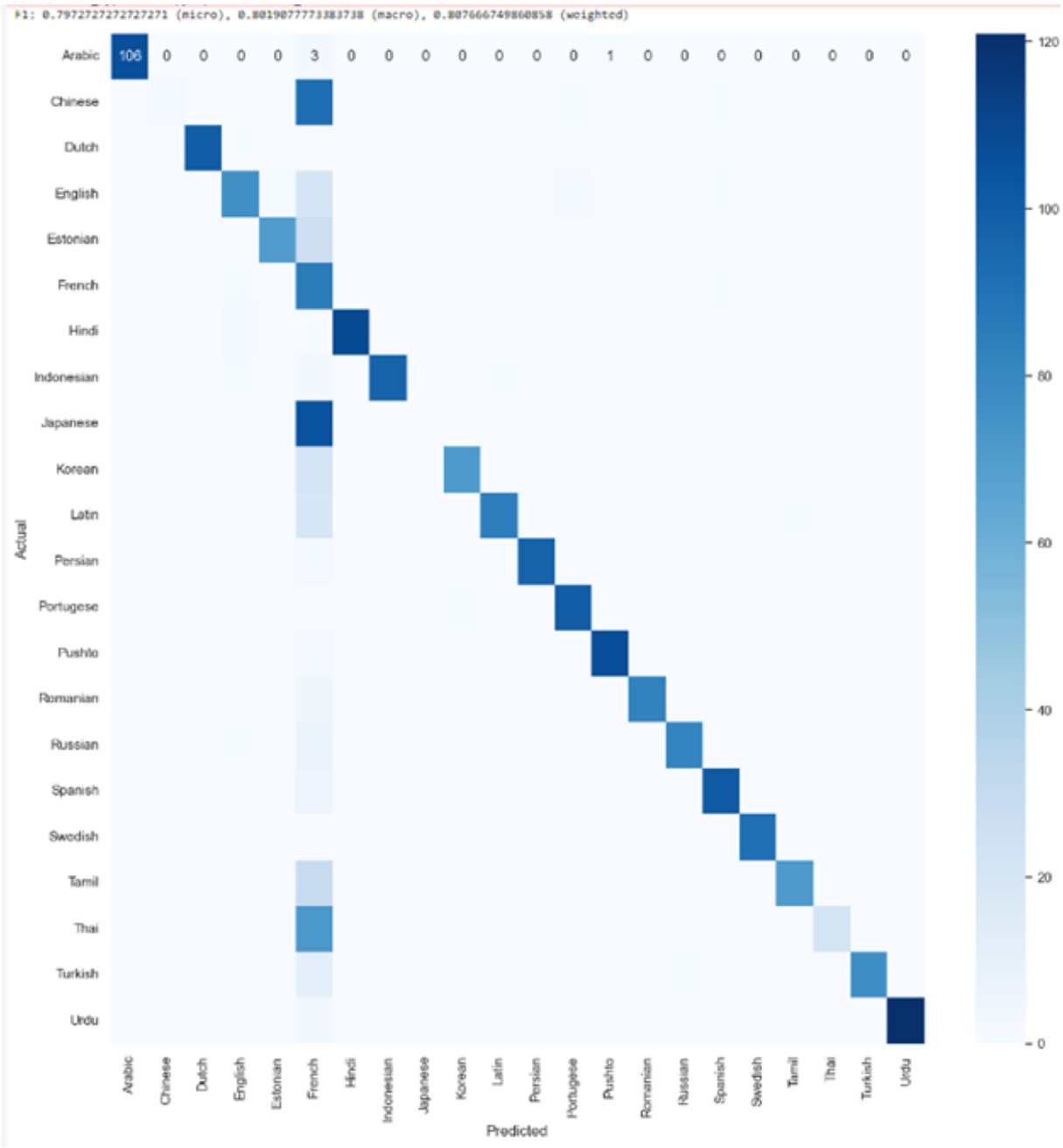
```
Top 5 n-grams for Dutch:
1: van
2: het
3: een
4: en
5: op
```

Bước 7: Kiểm thử dữ liệu với mô hình:

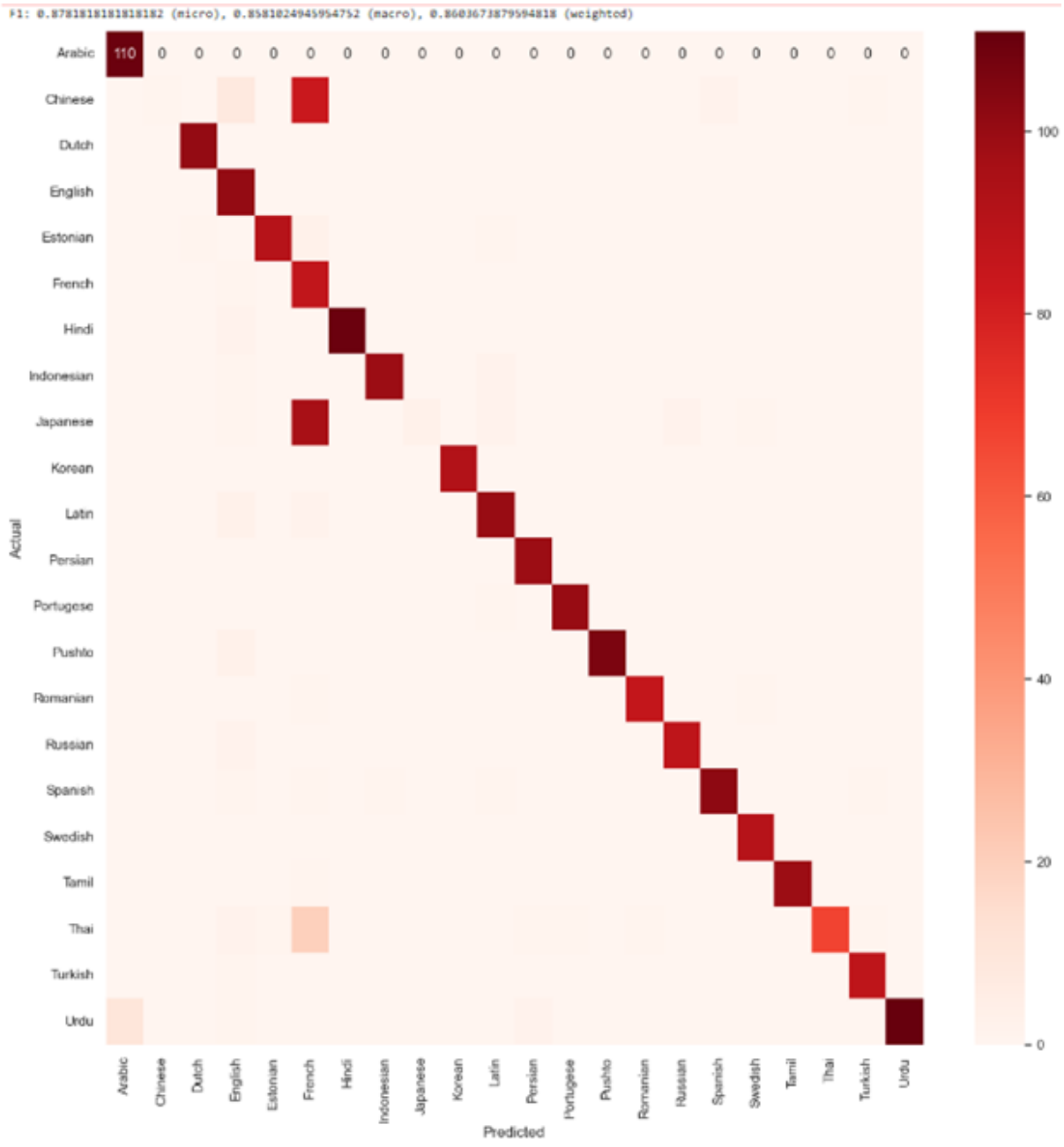
- Ma trận có được:
  - Uni-Grams:



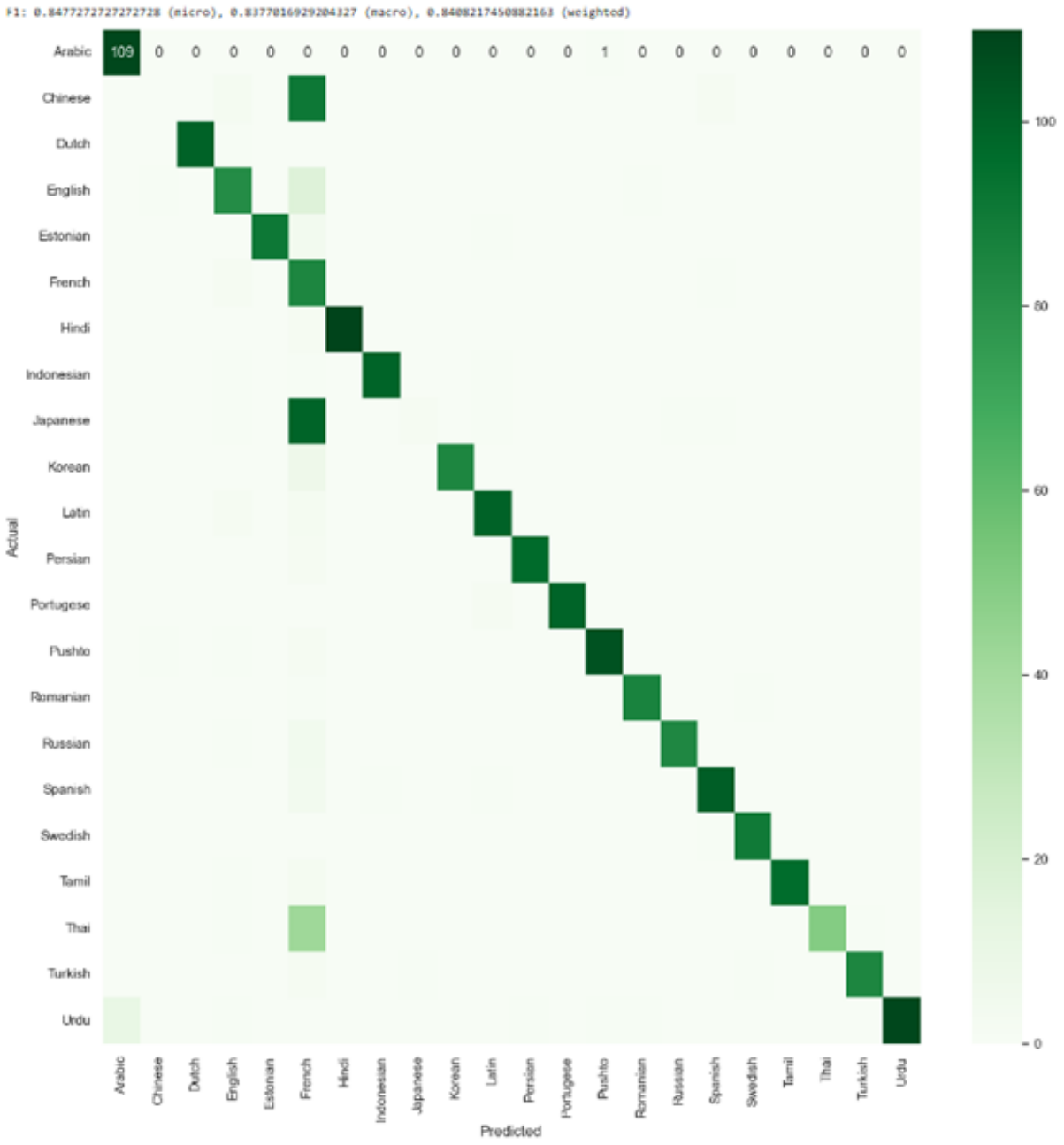
– Bi-Grams:


F1: 0.7972727272727271 (micro), 0.8019077773383738 (macro), 0.807666749860858 (weighted)
Arabic 106 0 0 0 0 3 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
Chinese
Dutch
English
Estonian
French
Hindi
Indonesian
Japanese
Korean
Latin
Persian
Portugese
Pushto
Romanian
Russian
Spanish
Swedish
Tamil
Thai
Turkish
Urdu
Actual
Predicted
120
100
80
60
40
20
0

– Mixture (Top 1%):

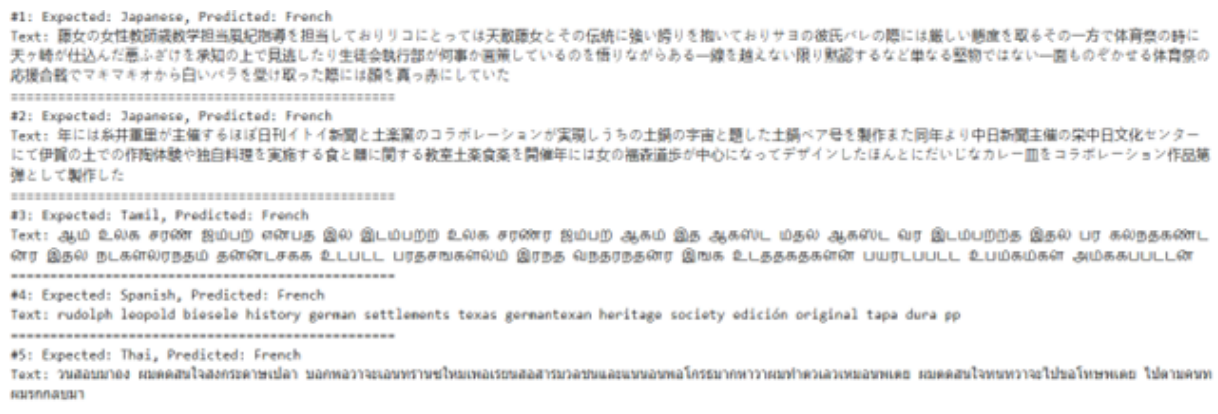


– Mixture (Top 50):



Bước 8: Kiểm thử mô hình đã huấn luyện với dữ liệu đầu vào mới:
Bước 9: Đánh giá tổng quát:

- Các trường hợp dự đoán sai của máy:



- Các trường hợp dự đoán đúng của máy:

- Kiểm tra độ chính xác của từng bộ huấn luyện:
  - Uni-Grams:

Naive Bayes on Unigrams:
Accuracy: 0.8777272727272727

	precision	recall	f1-score	support
Arabic	0.92	1.00	0.96	110
Chinese	1.00	0.01	0.02	96
Dutch	0.98	1.00	0.99	101
English	0.79	1.00	0.88	101
Estonian	0.99	0.95	0.97	96
French	0.29	0.99	0.45	88
Hindi	1.00	0.98	0.99	112
Indonesian	0.99	0.97	0.98	102
Japanese	1.00	0.03	0.06	105
Korean	1.00	1.00	1.00	92
Latin	0.93	0.95	0.94	105
Persian	0.98	1.00	0.99	99
Portugese	0.99	0.98	0.99	101
Pushto	1.00	0.97	0.99	109
Romanian	0.98	0.98	0.98	88
Russian	0.98	0.98	0.98	90
Spanish	0.98	0.95	0.97	107
Swedish	0.99	1.00	0.99	91
Tamil	1.00	0.99	0.99	100
Thai	1.00	0.71	0.83	94
Turkish	0.97	0.99	0.98	89
Urdu	1.00	0.90	0.94	124
accuracy			0.88	2200
macro avg	0.94	0.88	0.86	2200
weighted avg	0.95	0.88	0.86	2200

– Bi-Grams:

Naive Bayes on Bigrams:
Accuracy: 0.7972727272727272

	precision	recall	f1-score	support
Arabic	1.00	0.96	0.98	110
Chinese	1.00	0.02	0.04	96
Dutch	1.00	0.99	1.00	101
English	0.93	0.75	0.83	101
Estonian	0.99	0.73	0.84	96
French	0.17	0.98	0.29	88
Hindi	1.00	0.97	0.99	112
Indonesian	1.00	0.95	0.97	102
Japanese	0.00	0.00	0.00	105
Korean	0.99	0.77	0.87	92
Latin	0.99	0.81	0.89	105
Persian	1.00	0.98	0.99	99
Portugese	0.97	0.99	0.98	101
Pushto	0.99	0.98	0.99	109
Romanian	1.00	0.94	0.97	88
Russian	0.99	0.91	0.95	90
Spanish	0.97	0.94	0.96	107
Swedish	1.00	1.00	1.00	91
Tamil	0.99	0.71	0.83	100
Thai	1.00	0.23	0.38	94
Turkish	1.00	0.87	0.93	89
Urdu	1.00	0.98	0.99	124
accuracy			0.80	2200
macro avg	0.91	0.79	0.80	2200
weighted avg	0.91	0.80	0.81	2200

– Mixture (Top 50):

Naive Bayes on Mix Top 50:
Accuracy: 0.725909090909091

	precision	recall	f1-score	support
Arabic	0.91	0.98	0.94	110
Chinese	0.00	0.00	0.00	96
Dutch	0.97	0.99	0.98	101
English	0.73	0.16	0.26	101
Estonian	1.00	0.74	0.85	96
French	0.13	0.98	0.24	88
Hindi	1.00	0.98	0.99	112
Indonesian	1.00	0.97	0.99	102
Japanese	1.00	0.01	0.02	105
Korean	1.00	0.30	0.47	92
Latin	0.92	0.92	0.92	105
Persian	0.98	0.98	0.98	99
Portugese	1.00	0.98	0.99	101
Pushto	1.00	0.96	0.98	109
Romanian	0.99	0.98	0.98	88
Russian	1.00	0.79	0.88	90
Spanish	0.92	0.94	0.93	107
Swedish	0.93	0.99	0.96	91
Tamil	1.00	0.23	0.37	100
Thai	1.00	0.21	0.35	94
Turkish	0.99	0.90	0.94	89
Urdu	1.00	0.88	0.94	124
accuracy			0.73	2200
macro avg	0.88	0.72	0.73	2200
weighted avg	0.89	0.73	0.73	2200

– Mixture (Top 1%):

Naive Bayes on Mix Top 1%:
Accuracy: 0.725909090909091

	precision	recall	f1-score	support
Arabic	0.91	0.98	0.94	110
Chinese	0.00	0.00	0.00	96
Dutch	0.97	0.99	0.98	101
English	0.73	0.16	0.26	101
Estonian	1.00	0.74	0.85	96
French	0.13	0.98	0.24	88
Hindi	1.00	0.98	0.99	112
Indonesian	1.00	0.97	0.99	102
Japanese	1.00	0.01	0.02	105
Korean	1.00	0.30	0.47	92
Latin	0.92	0.92	0.92	105
Persian	0.98	0.98	0.98	99
Portugese	1.00	0.98	0.99	101
Pushto	1.00	0.96	0.98	109
Romanian	0.99	0.98	0.98	88
Russian	1.00	0.79	0.88	90
Spanish	0.92	0.94	0.93	107
Swedish	0.93	0.99	0.96	91
Tamil	1.00	0.23	0.37	100
Thai	1.00	0.21	0.35	94
Turkish	0.99	0.90	0.94	89
Urdu	1.00	0.88	0.94	124
accuracy			0.73	2200
macro avg	0.88	0.72	0.73	2200
weighted avg	0.89	0.73	0.73	2200

3.1.5 Giới thiệu về các thuật toán khác được sử dụng trong bài:

1. K Nearest Neighbour:

   - K Nearest Neighbour (KNN) là một thuật toán đơn giản và hiệu quả trong học máy, được sử dụng cho cả phân loại và hồi quy. Thuật toán hoạt động bằng cách tìm k điểm dữ liệu gần nhất trong không gian đặc trưng và dự đoán giá trị dựa trên các điểm đó.
   - KNN không yêu cầu giai đoạn huấn luyện. Thay vào đó, nó lưu trữ toàn bộ tập dữ liệu huấn luyện và thực hiện tính toán tại thời điểm dự đoán. Điều này làm cho KNN đơn giản nhưng có thể chậm nếu tập dữ liệu lớn.

2. (Ordinary) Least Squares:

   - Ordinary Least Squares (OLS) là một phương pháp thống kê để ước lượng các tham số trong mô hình hồi quy tuyến tính. Nó tìm các tham số sao cho tổng bình phương sai số giữa giá trị dự đoán và giá trị thực là nhỏ nhất.
   - OLS đơn giản và dễ triển khai. Nó cung cấp các ước lượng hiệu quả và không chệch trong điều kiện nhiễu có phân phối chuẩn.

3. Kolmogorov Smirnov Test:

   - Kolmogorov-Smirnov Test (KS Test) là một kiểm định phi tham số dùng để so sánh hai phân phối xác suất hoặc một phân phối với một phân phối tham chiếu. Nó dựa trên sự khác biệt lớn nhất giữa hai hàm phân phối tích lũy (CDF).
   - KS Test không phụ thuộc vào dạng phân phối của dữ liệu, làm cho nó rất linh hoạt. Nó có thể được sử dụng cho cả dữ liệu liên tục và dữ liệu rời rạc.

4. Lang Detect:

   - Lang Detect là một thư viện dùng để phát hiện ngôn ngữ của một đoạn văn bản dựa trên mô hình học máy. Nó sử dụng các kỹ thuật như Naive Bayes hoặc n-gram để xác định ngôn ngữ.
   - Lang Detect rất hiệu quả và hỗ trợ nhiều ngôn ngữ khác nhau. Nó dễ sử dụng và tích hợp vào các ứng dụng xử lý ngôn ngữ tự nhiên.

3.2 Kết quả thực nghiệm:

3.2.1 Thực nghiệm trên 10 từ (character):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages: ['English']
```

3. NB trên Mix Top 50:

```
Predicted languages: ['English']
```

4. NB trên Mix Top 1%:

```
Predicted languages: ['English']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): ['English']
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): Dutch
```

7. Langdetect:

```
Predicted Language: English
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['Latin']
```

3.2.2 Thực nghiệm trên 50 từ (character):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages: ['English']
```

3. NB trên Mix Top 50:

```
Predicted languages: ['English']
```

4. NB trên Mix Top 1%:

```
Predicted languages: ['English']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): English
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): English
```

7. Langdetect:

```
Predicted languages: ['English']
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['English']
```

3.2.3 Thực nghiệm trên 100 từ (character):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages: ['English']
```

3. NB trên Mix Top 50:

```
Predicted languages: ['English']
```

4. NB trên Mix Top 11%:

```
Predicted languages: ['English']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): English
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): English
```

7. Langdetect:

```
Predicted languages: ['English']
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['English']
```

3.2.4 Thực nghiệm trên 10 từ (word):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages for bigrams: ['French']
```

3. NB trên Mix Top 50:

```
Predicted Language (Top 50 Features): ['French']
```

4. NB trên Mix Top 1%:

```
Predicted languages: ['French']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): Japanese
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): Urdu
```

7. Langdetect:

```
Predicted Language: English
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['Urdu']
```

3.2.5 Thực nghiệm trên 50 từ (word):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages for bigrams: ['French']
```

3. NB trên Mix Top 50:

```
Predicted Language (Top 50 Features): ['French']
```

4. NB trên Mix Top 1%:

```
Predicted languages: ['French']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): Japanese
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): Urdu
```

7. Langdetect:

```
Predicted Language: English
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['Urdu']
```

3.2.6 Thực nghiệm trên 100 từ (word):

1. NB trên Unigrams:

```
Predicted languages: ['English']
```

2. NB trên Bigrams:

```
Predicted languages for bigrams: ['English']
```

3. NB trên Mix Top 50:

```
Predicted Language (Top 50 Features): ['English']
```

4. NB trên Mix Top 0.7%:

```
Predicted languages: ['English']
```

5. KNN trên Mix Top 50:

```
Predicted Language (Top 50 Features): English
```

6. OLS trên Mix Top 50:

```
Predicted Language (OLS, Top 50 Features): Swedish
```

7. Langdetect:

```
Predicted Language: English
```

8. KS trên Mix Top 50:

```
Predicted Language for Input Paragraph (Top 50 Features): ['_N/A_']
```

3.3 Đánh giá tổng quát về thuật toán:

3.3.1 Đánh giá riêng về thuật toán chính được sử dụng:

- Dễ hiểu: Thuật toán Naive Bayes (MNB) có cơ sở lý thuyết đơn giản và dễ dàng giải thích, giúp người dùng mới có thể nhanh chóng nắm bắt và áp dụng.
- Dễ triển khai: Với các thư viện hỗ trợ mạnh mẽ trong Python, việc triển khai MNB trở nên dễ dàng và không đòi hỏi nhiều công sức.
- Tập dữ liệu nhỏ: MNB thể hiện hiệu quả tính toán cao khi làm việc với các tập dữ liệu nhỏ, đảm bảo tốc độ xử lý nhanh và kết quả chính xác.
- Tập dữ liệu lớn: Khi áp dụng cho các tập dữ liệu lớn, hiệu suất của MNB không những không giảm mà còn được cải thiện đáng kể, nhờ vào khả năng xử lý và học hỏi từ nhiều dữ liệu.

3.3.2 So sánh với các thuật toán khác:

- Độ chính xác:
  - Kí tự(Character): có độ chính xác cao nhất trong cả 4 thuật toán: 96,5
  - Từ (Word): có độ chính xác không vượt trội lắm so với cả 4 thuật toán.
- Khả năng tiếp cận:
  - Dễ dàng tiếp cận cả mặt lý thuyết lẫn thực hành hơn so với 3 thuật toán còn lại.
  - Có nhiều thư viện hỗ trợ tính toán.

3.3.3 Định hướng trong tương lai:

Qua những ưu điểm và nhược điểm của thuật toán đã được trình bày trong bài đã cho thấy thuật toán xác định ngôn ngữ trong xử lý ngôn ngữ tự nhiên (NLP) là một quá trình không hề dễ dàng và đòi hỏi có sự đầu tư cả thời gian lẫn kiến thức để có thể ngày một phát triển. Trong thời gian làm đồ án em đã thực hiện được ở mức Uni-gram và Bi-gram.

- Những hướng phát triển trong lương lai:
  - Xử lý được bộ dữ liệu dày hơn, đa dạng hơn.
  - Áp dụng vào thực tiễn như các trang web đa quốc gia hoặc dùng để hộ trợ phiên dịch như Google Dịch.

Tài liệu tham khảo

[1] Daniel Jurafsky, James H. Martin. *Speech and Language Processing.* Stanford University, 2023. https://web.stanford.edu/~jurafsky/slp3/4.pdf

[2] A.Aylin Tokuc, Michal Aibin. *How to improve Naïve Bayes Classification Performance.* Baeldung, 2024. https://www.baeldung.com/cs/naive-bayes-classification-performance

[3] Jake VanderPlas. *Python Data Science Handbook.* O'Reilly Media, 2017. https://jakevdp.github.io/PythonDataScienceHandbook/

[4] Wikipedia. *Xử lý ngôn ngữ tự nhiên.* Wikipedia, 2024. https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn

[5] Scikit-learn. *Multinomial Naive Bayes Classifier.* Scikit-learn, 2024. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[6] David Klein, Kyle Murray, Simon Weber. *Algorithmic Programming Language Identification.* arXiv, 2011. https://arxiv.org/abs/1106.4064

[7] Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, Krister Lindén. *Automatic Language Identification in Texts: A Survey.* arXiv, 2018. https://arxiv.org/abs/1804.08186

[8] Sebastian Raschka. *Naive Bayes and Text Classification.* 2014. https://sebastianraschka.com/Articles/2014_naive_bayes_1.html

[9] JavaPoint. *Machine Learning - Naive Bayes Classifier.* JavaPoint, 2024. https://www.javatpoint.com/machine-learning-naive-bayes-classifier

[10] Sriram. *Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2024.* Upgrad, 2022. https://www.upgrad.com/blog/multinomial-naive-bayes-explained/