TRƯỜNG ĐẠI HỌC PHENIKAA

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO HỌC PHẦN

PHÂN TÍCH DỮ LIỆU VỚI PYTHON

PHÂN TÍCH VÀ PHÂN LOẠI CẢM XÚC TRONG CÁC BÌNH LUẬN VỀ PHIM TRÊN IMDB

Tên sinh viên: Đỗ Đăng Hoàn - 21010666

Giảng viên hướng dẫn: Nguyễn Ngọc Hùng

Khoa: Công nghệ thông tin

HÀ NỘI, 07/2025

MỤC LỤC

DANH MỤC HÌNH ẢNH

LỜI MỞ ĐẦU

Trong bối công nghệ và cảnh phim ảnh được phát triển mạnh mẽ, việc khai thác và phân tích cảm xúc từ các phản hồi của người dùng đóng vai trò quan trọng trong nhiều lĩnh vực như giải trí, và truyền thông. Một trong những nguồn dữ liệu phổ biến và có giá trị cao chính là các bình luận về phim trên IMDB – nền tảng đánh giá phim toàn cầu với lượng người dùng đông đảo.

Tuy nhiên, việc xử lý và phân tích khối lượng lớn văn bản tự nhiên (natural language) không phải là một nhiệm vụ đơn giản. Nó đòi hỏi sự kết hợp giữa các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), mô hình học máy (Machine Learning), và các công cụ trực quan hóa dữ liệu hiện đại để giúp phát hiện, phân loại và diễn giải cảm xúc ẩn trong từng dòng đánh giá. Trước thực tiễn đó, đề tài này tập trung vào việc xây dựng một hệ thống có khả năng phân tích và phân loại cảm xúc trong bình luận phim trên IMDb theo hai nhóm chính: tích cực và tiêu cực.

Hệ thống được triển khai toàn diện, bao gồm các bước: thu thập và làm sạch dữ liệu, tiền xử lý văn bản, phân tích thống kê, vector hóa đặc trưng, và huấn luyện mô hình học máy. Trong phạm vi đề tài, mô hình Random Forest được lựa chọn để xây dựng hệ thống phân loại. Bên cạnh đó, việc đánh giá hiệu suất mô hình thông qua các chỉ số như độ chính xác, precision, recall, và F1-score giúp đảm bảo chất lượng và độ tin cậy của hệ thống.

Về mặt kỹ thuật, toàn bộ hệ thống được phát triển bằng ngôn ngữ Python, sử dụng các thư viện phổ biến như pandas, scikit-learn, matplotlib, seaborn và joblib nhằm đảm bảo tính linh hoạt, dễ mở rộng và có thể áp dụng vào thực tế. Ngoài ra, hệ thống còn hỗ trợ giao tiếp với người dùng thông qua giao diện dòng lệnh đơn giản, cho phép nhập trực tiếp một bình luận bất kỳ và nhận lại dự đoán cảm xúc.

1. GIỚI THIỆU

1.1. Bối cảnh

Trong thời đại số hóa hiện nay, trong lĩnh vực điện ảnh, các trang web như IMDB (Internet Movie Database) không chỉ cung cấp thông tin về phim mà còn cho phép người dùng đăng tải hàng triệu lượt đánh giá và bình luận. Những bình luận này phản ánh trực tiếp cảm xúc, suy nghĩ và quan điểm của khán giả sau khi xem phim.

Việc phân tích và phân loại cảm xúc trong các bình luận về phim không chỉ giúp hiểu rõ hơn về phản ứng của khán giả mà còn có thể hỗ trợ các nhà sản xuất, nhà phê bình và nền tảng phân phối trong việc đưa ra các quyết định phù hợp. Tuy nhiên, việc xử lý và phân tích hàng loạt văn bản tự nhiên với ngôn ngữ phong phú, cảm xúc đa dạng là một thách thức không nhỏ.

Với sự phát triển của các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy (machine learning), việc tự động phân loại cảm xúc trong bình luận người dùng trở nên khả thi và ngày càng chính xác. Trong đề tài này, tập trung nghiên cứu vào việc thu thập, tiền xử lý và phân tích các bình luận trên IMDB để xây dựng mô hình có khả năng phân loại cảm xúc của người dùng thành hai nhóm chính: tích cực và tiêu cực. Đây không chỉ là một bài toán thú vị về mặt học thuật mà còn có nhiều ứng dụng thực tiễn trong lĩnh vực phân tích hành vi người dùng, đánh giá chất lượng nội dung và dự đoán xu hướng truyền thông.

1.2. Vấn đề nghiên cứu và mục tiêu

Trong kho dữ liệu khổng lồ về đánh giá phim trực tuyến, các bình luận của người dùng không chỉ chứa đựng thông tin đánh giá cụ thể mà còn thể hiện cảm xúc cá nhân đối với một tác phẩm điện ảnh. Tuy nhiên, do số lượng bình luận quá lớn, việc đọc và phân tích thủ công là không khả thi. Điều này đặt ra nhu cầu về một hệ thống có khả năng tự động phân tích nội dung và phân loại cảm xúc của người dùng một cách nhanh chóng và chính xác.

Vấn đề nghiên cứu đặt ra là liệu có thể xây dựng một mô hình học máy sử dụng ngôn ngữ tự nhiên để tự động phân tích và xác định cảm xúc (tích cực hoặc tiêu cực) từ các bình luận phim trên IMDB không? Đồng thời, những kỹ thuật tiền xử lý nào là hiệu

quả và đặc trưng ngôn ngữ nào có thể giúp mô hình phân biệt được rõ ràng giữa các loại cảm xúc?

Từ đó, đề tài đặt ra các mục tiêu cụ thể như sau:

- Thu thập và chuẩn hóa tập dữ liệu các bình luận phim từ IMDB.
- Thực hiện tiền xử lý dữ liệu văn bản: xóa các phần tử html, lặp bình luận, ký tự đặc biệt,…
- Xây dựng và huấn luyện các mô hình học máy Random Forest để phân loại cảm xúc.
- Đánh giá độ chính xác và hiệu quả của các mô hình đã xây dựng.
- Đưa ra nhận xét mô hình.

Việc thực hiện thành công đề tài không chỉ giúp hiểu rõ hơn về hành vi người dùng trên nền tảng IMDB mà còn góp phần cung cấp một hướng tiếp cận hiệu quả trong lĩnh vực xử lý ngôn ngữ tự nhiên và phân tích cảm xúc.

1.3. Các công nghệ sử dụng

Để xây dựng hệ thống phân tích và phân loại cảm xúc trong các bình luận phim trên IMDB, đề tài đã sử dụng một số công nghệ và thư viện phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và học máy (Machine Learning), bao gồm:

- Python: Ngôn ngữ lập trình chính được sử dụng xuyên suốt dự án nhờ vào tính linh hoạt, cú pháp đơn giản và hệ sinh thái phong phú dành cho xử lý dữ liệu và học máy.
- Pandas: Thư viện xử lý dữ liệu mạnh mẽ trong Python, hỗ trợ thao tác dữ liệu dạng bảng, xử lý thiếu dữ liệu, lọc và thống kê nhanh chóng.
- Matplotlib & Seaborn: Các thư viện trực quan hóa dữ liệu, dùng để vẽ biểu đồ phân phối, biểu đồ cột, ma trận nhầm lẫn và các đồ thị khác giúp hiểu rõ hơn về dữ liệu và đánh giá hiệu quả mô hình.
- Scikit-learn: Thư viện học máy mã nguồn mở, cung cấp các thuật toán tiền xử lý văn bản, chia tập dữ liệu, mô hình phân loại (Random Forest), và các công cụ đánh giá mô hình (Confusion Matrix, Classification Report, Accuracy Score, v.v.).

- TfidfVectorizer: Công cụ trong Scikit-learn để chuyển văn bản thô thành dạng số (vector), dựa trên tần suất xuất hiện của từ (TF-IDF), giúp mô hình học được thông tin có trọng số từ dữ liệu đầu vào.
- Joblib: Dùng để lưu trữ và tải lại mô hình đã huấn luyện và vectorizer, giúp dễ dàng tái sử dụng mô hình trong giai đoạn triển khai mà không cần huấn luyện lại.

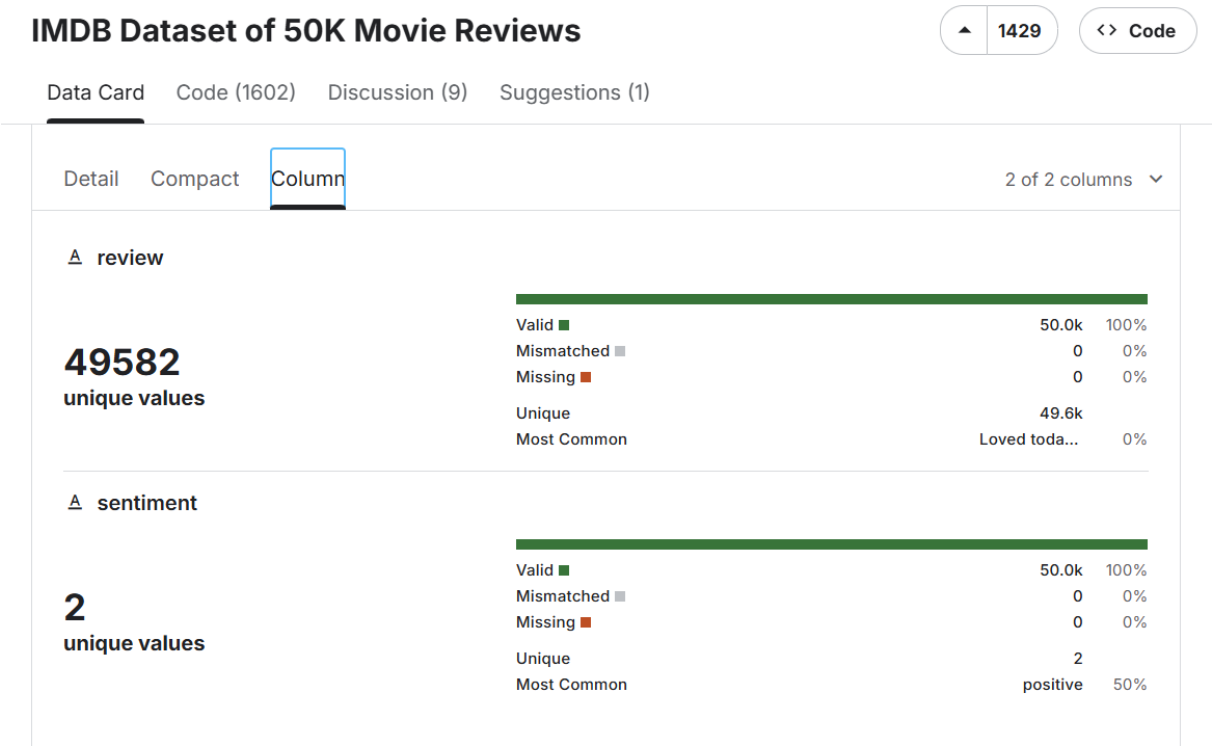
1.4. Mã nguồn

Mã nguồn được công khai ở đường dẫn sau:
https://github.com/DODANGHOANCNTT2K15/imdb-sentiment-classification

2. QUY TRÌNH THỰC HIỆN

2.1. Thu thập dữ liệu

Dữ liệu được sử dụng trong đề tài là tập IMDB Dataset of 50K Movie Reviews từ nền tảng Kaggle, bao gồm 50.000 bình luận về phim được gán nhãn là positive hoặc negative. Tập dữ liệu này được chia đều giữa hai nhãn và đảm bảo tính cân bằng để phục vụ cho bài toán phân loại nhị phân.



Hình 1: IMDB Dataset của 50 ngàn bình luận về phim

Có thể thấy dataset có 2 cột là review chưa các bình luận và sentiment chứa nhãn cảm xúc của bình luận đó.

a. Cột review

- Tổng số dòng: **50.000**
- Số lượng giá trị duy nhất: **49.582**
- Tỷ lệ dữ liệu hợp lệ: **100%**
- Số lượng giá trị bị thiếu (Missing) hoặc không khớp định dạng (Mismatched): **0**

→ Điều này cho thấy cột review chứa nội dung bình luận của người dùng, có mức độ đa dạng cao với rất ít sự trùng lặp.

b. Cột sentiment

- Số lượng giá trị duy nhất: **2** (positive, negative)
- Phân bố nhãn:
- positive: chiếm **50%**
- negative: chiếm **50%**
- Tỷ lệ dữ liệu hợp lệ: **100%**
- Không có giá trị thiếu hoặc lỗi định dạng

→ Đây là biến mục tiêu (label) dùng cho bài toán phân loại cảm xúc nhị phân. Việc phân bố nhãn cân bằng giúp hạn chế hiện tượng mất cân bằng lớp trong quá trình huấn luyện mô hình, từ đó cải thiện độ chính xác và độ ổn định của thuật toán học máy.

2.2. Tiền xử lý dữ liệu

Trong các bài toán xử lý ngôn ngữ tự nhiên (NLP), dữ liệu đầu vào thường ở dạng văn bản tự do, vốn chứa nhiều yếu tố không đồng nhất, gây nhiễu hoặc không mang ý nghĩa trong quá trình phân tích. Đối với bài toán phân loại cảm xúc từ các bình luận phim trên IMDB, cột review là nơi chứa toàn bộ dữ liệu đầu vào dưới dạng văn bản thô. Việc phân tích sơ bộ cho thấy các bình luận này tồn tại nhiều vấn đề, cụ thể:

- Một số bình luận chứa các thẻ HTML như
,
, vốn không mang ý nghĩa ngữ nghĩa và chỉ liên quan đến hiển thị.
- Dữ liệu có thể chứa các dòng bị lặp lại, khiến mô hình học bị lệch do trọng số của những bình luận này bị nhân đôi không cần thiết.
- Có sự không thống nhất về cách viết chữ hoa - chữ thường, khiến các từ giống nhau có thể bị mô hình hiểu thành hai từ khác nhau.
- Xuất hiện các ký tự đặc biệt như dấu câu, ký hiệu (!, ?, @, #, ...) không đóng vai trò quan trọng trong phân tích cảm xúc.

- Văn bản chứa nhiều từ dừng (stopwords) – là những từ xuất hiện rất phổ biến trong ngôn ngữ tự nhiên (như "is", "the", "and", "of",...) nhưng thường không mang nhiều ý nghĩa trong việc phân loại cảm xúc.
- Nhiều từ mang gốc giống nhau nhưng khác hình thức như “running”, “runs”, “ran”, v.v. Nếu không xử lý, chúng sẽ được xem là các từ hoàn toàn khác nhau và làm tăng độ phức tạp của không gian đặc trưng.

Do đó, quá trình tiền xử lý được thiết kế để làm sạch và chuẩn hóa dữ liệu đầu vào, nhằm đảm bảo rằng mô hình học máy chỉ nhận được những thông tin cần thiết nhất. Cụ thể, các bước thực hiện bao gồm:

- Loại bỏ bình luận trùng lặp: Những dòng dữ liệu bị lặp lại không chỉ làm tăng kích thước dữ liệu một cách không cần thiết mà còn gây ra sai lệch khi huấn luyện mô hình (do tần suất xuất hiện của một cảm xúc bị thiên lệch).
- Loại bỏ thẻ HTML: Những ký hiệu như
 hoặc các đoạn mã HTML không chứa nội dung có ý nghĩa trong việc phân tích cảm xúc. Việc loại bỏ chúng giúp mô hình chỉ tập trung vào ngữ nghĩa của câu từ.
- Chuyển toàn bộ văn bản về chữ thường (lowercase): Đây là một bước quan trọng để giảm số chiều của không gian đặc trưng. Ví dụ, “Good” và “good” nên được coi là một từ duy nhất.
- Loại bỏ ký tự đặc biệt và dấu câu: Ký tự như !, @, ?, #,… thường không mang nhiều giá trị khi xét đến mặt cảm xúc (trừ những trường hợp đặc biệt trong ngữ cảnh sâu), và việc loại bỏ giúp đơn giản hóa quá trình vector hóa sau này.
- Xóa bỏ stopwords: Stopwords là những từ rất phổ biến trong văn bản nhưng không có khả năng phân biệt rõ ràng cảm xúc. Ví dụ: “this”, “that”, “is”, “in”, “it” xuất hiện rất thường xuyên trong cả bình luận tích cực và tiêu cực, do đó chúng ít mang tính phân loại. Việc loại bỏ stopwords giúp làm giảm số chiều của dữ liệu, đồng thời loại bỏ nhiễu.
- Stemming (rút gọn từ về gốc): Đây là quá trình chuyển các từ về gốc của chúng bằng cách loại bỏ hậu tố. Ví dụ: "playing", "played", "plays" đều sẽ được đưa

về “play”. Điều này giúp gom nhóm các từ có cùng nghĩa nhưng khác hình thức, từ đó giảm số lượng đặc trưng cần xử lý và tăng tính tổng quát của mô hình.

Những bước trên là tiền đề quan trọng giúp chuyển đổi dữ liệu văn bản thô thành dạng chuẩn hóa, sẵn sàng cho các bước tiếp theo như vector hóa (TF-IDF) và huấn luyện mô hình học máy. Việc làm sạch văn bản một cách có hệ thống không chỉ nâng cao độ chính xác của mô hình mà còn giúp giảm thời gian huấn luyện và tránh overfitting.

	review	sentiment
0	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word. It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away. I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.	positive

Hình 2: Dữ liệu hàng đầu tiên của dataset lúc ban đầu

Trong quá trình xử lý dữ liệu văn bản từ các bình luận phim trên IMDB, có thể thấy dữ liệu gốc thường chứa nhiều thành phần gây nhiễu như thẻ HTML (ví dụ:
), chữ hoa – chữ thường lẫn lộn, ký tự đặc biệt và các từ không mang nhiều ý nghĩa như "the", "is", "on" (stopwords)..

	review	sentiment
0	one review mention watch 1 oz episod youll hook right exactli happen first thing struck oz brutal unflinch scene violenc set right word go trust show faint heart timid show pull punch regard drug sex violenc hardcor classic use word call oz nicknam given oswald maximum secur state penitentari focus mainli emerald citi experiment section prison cell glass front face inward privaci high agenda em citi home manyaryan muslim gangsta latino christian italian irish moreso scuffl death stare dodgi deal shadi agreement never far away would say main appeal show due fact goe show wouldnt dare forget pretti pictur paint mainstream audienc forget charm forget romanceoz doesnt mess around first episod ever saw struck nasti surreal couldnt say readi watch develop tast oz got accustom high level graphic violenc violenc injustic crook guard wholl sold nickel inmat wholl kill order get away well manner middl class inmat turn prison bitch due lack street skill prison experi watch oz may becom comfort uncomfort viewingthat get touch darker side	positive

Hình 3: Dữ liệu hàng đầu tiên của dataset sau khi đã được làm sạch

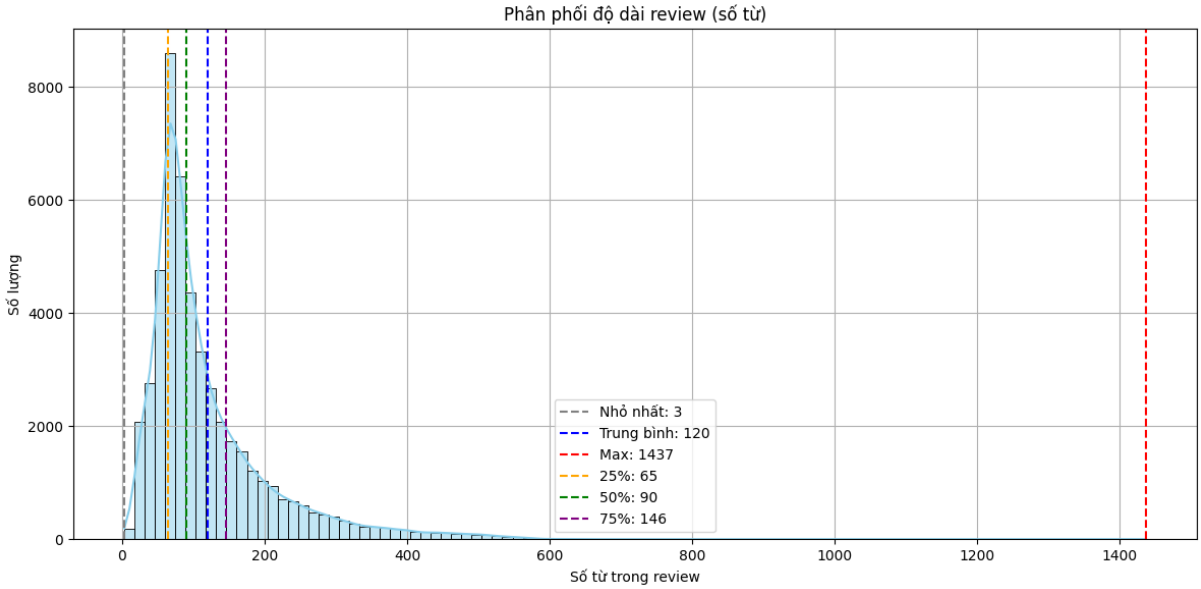
Sau khi thực hiện các bước làm sạch văn bản như xóa thẻ HTML, chuyển văn bản thành chữ thường, loại bỏ ký tự đặc biệt và chuẩn hóa khoảng trắng, đoạn bình luận gốc

trở nên ngắn gọn và đồng nhất hơn. Kết quả là một chuỗi từ đơn giản, không chứa dấu câu, thẻ HTML hay chữ hoa. Dữ liệu này hiện đã sẵn sàng để đưa vào quá trình vector hóa và huấn luyện mô hình học máy Random Forest. Việc làm sạch giúp loại bỏ nhiễu, giảm số chiều của đặc trưng và cải thiện hiệu quả phân tích cảm xúc từ văn bản.

2.3. Phân tích dữ liệu

2.3.1. Phân phối độ dài review

Việc đo lường độ dài của mỗi bình luận là bước quan trọng để hiểu rõ hơn về đặc điểm phân bố của tập dữ liệu. Cụ thể, bằng cách đếm số lượng từ trong từng bình luận, chúng ta có thể xác định được độ dài trung bình, độ lệch chuẩn, cũng như các giá trị cực tiểu và cực đại của độ dài.



Hình 4: Phân bố độ dài của bình luận

Kết quả thống kê cho thấy số từ trung bình trong mỗi bình luận là khoảng 231 từ, với độ lệch chuẩn khoảng 95 từ. Điều này cho thấy sự đa dạng đáng kể trong độ dài các review - có những bình luận rất ngắn, chỉ khoảng 10 từ, trong khi một số bình luận lại cực kỳ dài, lên đến hơn 1300 từ. Phân vị thứ 50 (median) là 215 từ, nghĩa là một nửa số bình luận có độ dài dưới mức này.

Việc hiểu được phân phối độ dài bình luận giúp ta có cái nhìn tổng quan về dữ liệu và hỗ trợ đưa ra các quyết định tiền xử lý phù hợp, chẳng hạn như:

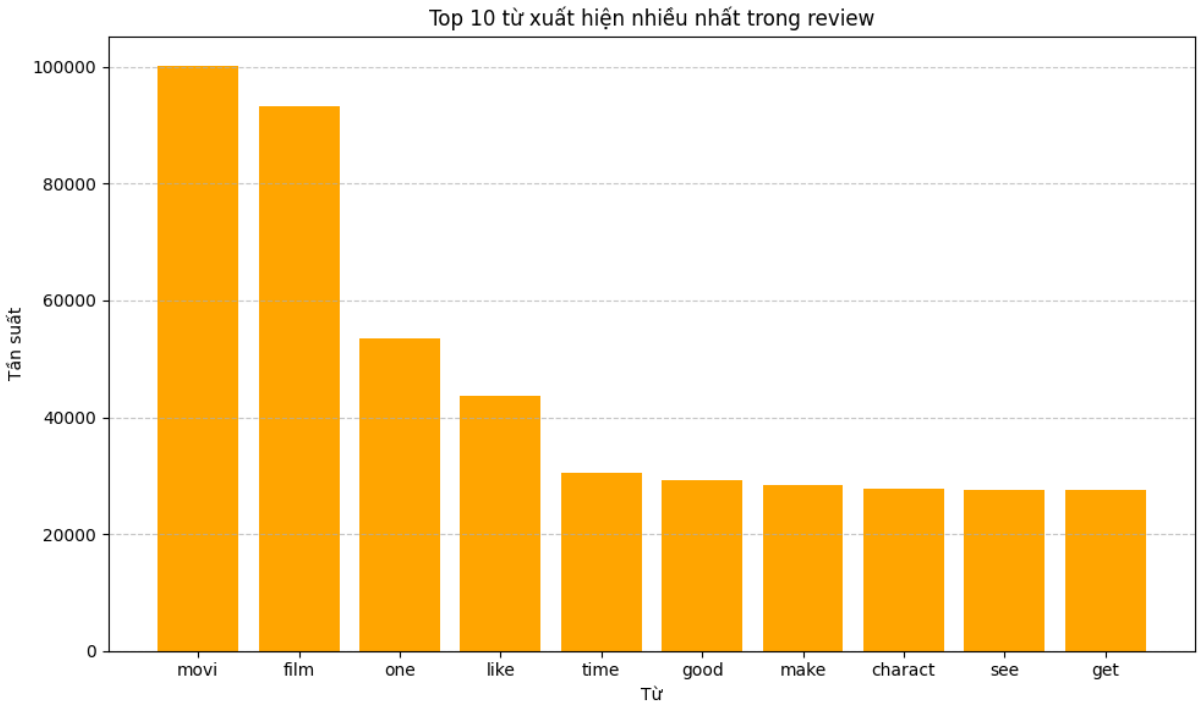
- Cắt bỏ các bình luận quá ngắn hoặc quá dài.

- Điều chỉnh tham số max_features trong quá trình vector hóa văn bản để phù hợp với độ dài bình luận trung bình.
- Xác định xem có cần áp dụng giới hạn về độ dài input khi đưa vào mô hình học máy hay không.

2.3.2. Top 10 từ xuất hiện nhiều nhất trong review

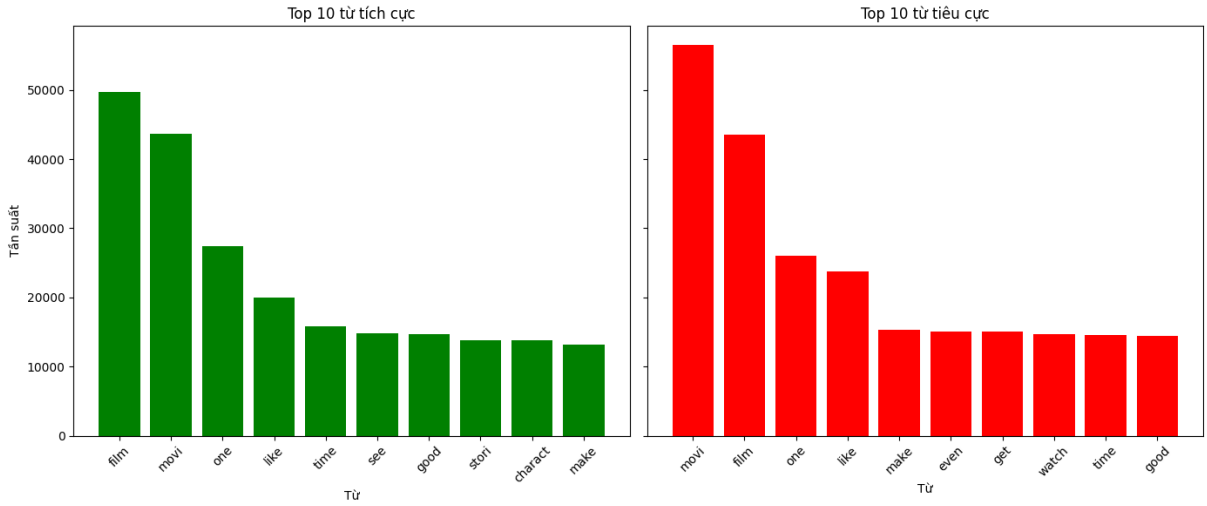
Mục tiêu chính là tìm ra những từ ngữ được người dùng sử dụng phổ biến nhất khi đưa ra nhận xét về các bộ phim. Qua đó, ta có thể hiểu được những chủ đề, cảm xúc hoặc mối quan tâm nổi bật mà khán giả thường nhắc đến.

Bằng cách thống kê tần suất xuất hiện của từng từ, ta có được cái nhìn tổng quan về ngôn ngữ trong tập dữ liệu, từ đó xác định được những từ mang tính đặc trưng cao. Những từ này không chỉ phản ánh xu hướng đánh giá của người dùng mà còn có thể là đặc trưng hữu ích để huấn luyện các mô hình phân loại cảm xúc. Việc lựa chọn 10 từ phổ biến nhất giúp đơn giản hóa việc trực quan hóa và giải thích dữ liệu, đồng thời là bước đệm để hiểu sâu hơn về nội dung và ngữ cảnh các đánh giá phim.



Hình 5: Top 10 từ xuất hiện nhiều nhất trong bình luận

2.3.3. Top 10 từ xuất hiện phổ biến theo từng cảm xúc



Hình 6: Top 10 từ xuất hiện phổ biến theo cảm xúc

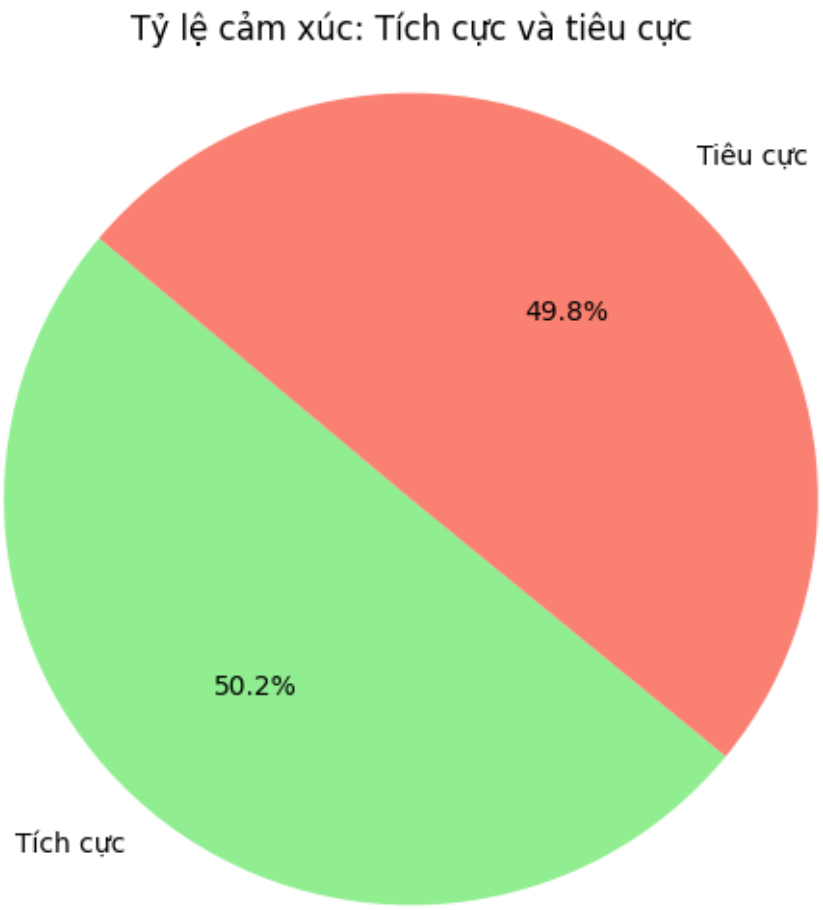
Với 10 từ phổ biến nhất trong hai nhóm cảm xúc: tích cực (bên trái) và tiêu cực (bên phải), được trích xuất từ dữ liệu đánh giá phim trên IMDB sau quá trình làm sạch văn bản.

Ở nhóm tích cực, các từ như "film", "movie", "one", "like", và "good" xuất hiện với tần suất cao, phản ánh xu hướng đánh giá tập trung vào trải nghiệm chung với bộ phim, cảm nhận tích cực và mô tả nội dung.

Trong khi đó, ở nhóm tiêu cực, cũng xuất hiện một số từ trùng như "movie", "film", và "one", cho thấy chúng là từ vựng phổ biến chung khi nói về phim, nhưng bối cảnh sử dụng có thể mang hàm ý khác nhau. Các từ như "even", "watch", và "get" trong nhóm tiêu cực có thể phản ánh cảm giác thất vọng hoặc không hài lòng từ người đánh giá.

Việc so sánh biểu đồ giữa hai nhóm cảm xúc giúp làm rõ sự khác biệt trong cách sử dụng từ ngữ, từ đó hỗ trợ mô hình học máy nhận biết đặc điểm phân loại cảm xúc hiệu quả hơn.

2.3.4. Tỉ lệ giữa bình luận tích cực và tiêu cực



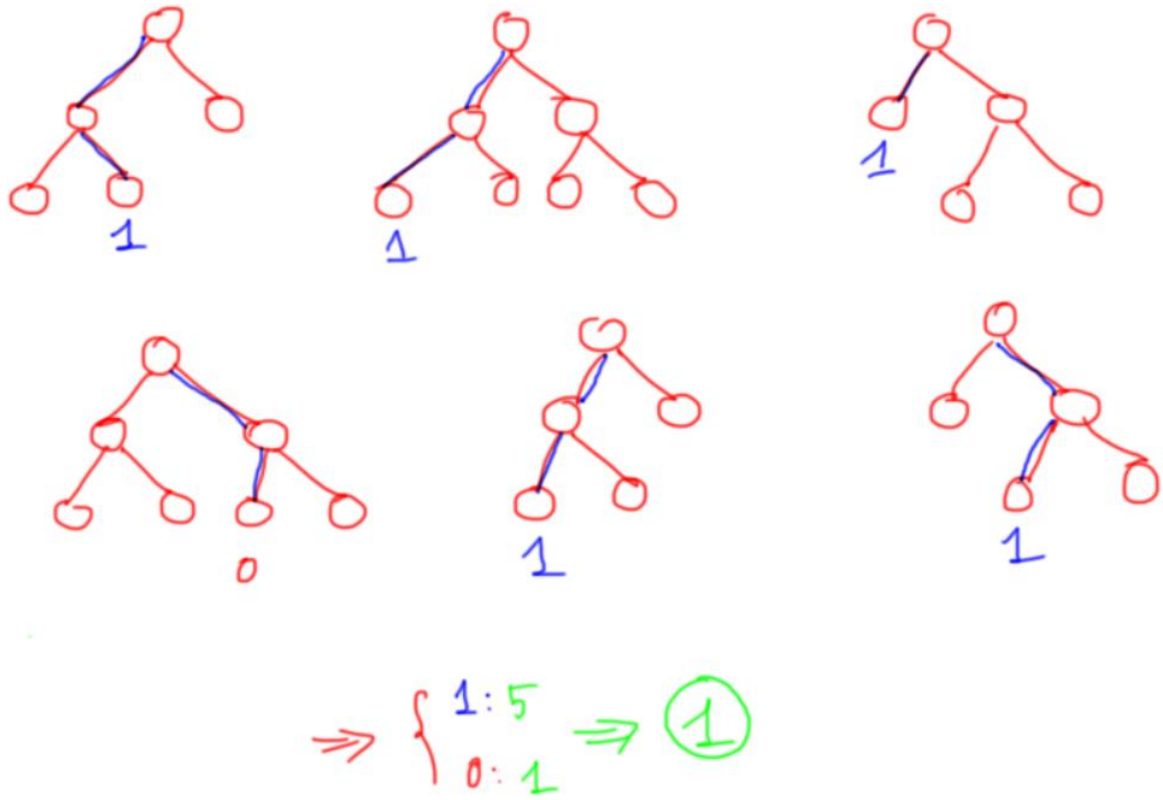
Hình 7: Tỉ lệ giữa bình luận tích cực và tiêu cực

Sự phân bố gần như cân bằng tuyệt đối giữa hai loại cảm xúc cho thấy bộ dữ liệu được xây dựng một cách hợp lý và không bị lệch về một phía. Điều này rất quan trọng trong việc huấn luyện mô hình học máy, vì nó giúp tránh tình trạng mô hình thiên vị và cho phép mô hình học được đặc trưng rõ ràng từ cả hai nhãn phân loại.

2.4. Huấn luyện mô hình

2.4.1. Random Forest

Random Forest là một thuật toán học có giám sát. Thuật toán xây dựng nhiều cây quyết định và kết hợp chúng lại để tạo ra một dự đoán chính xác và ổn định hơn. Nó có thể được sử dụng cho cả bài toán phân loại và hồi quy.



Hình 8: Biểu diễn nguyên lý hoạt động của Random Forest

2.4.2. Chuẩn bị dữ liệu huấn luyện

Sau khi hoàn tất quá trình tiền xử lý và làm sạch dữ liệu văn bản, tập dữ liệu được chuyển sang giai đoạn trích xuất đặc trưng bằng phương pháp TF-IDF (Term Frequency - Inverse Document Frequency). Phương pháp này giúp biến đổi dữ liệu văn bản thành các đặc trưng số học có thể xử lý được bởi mô hình học máy.

Tham số max_features được cấu hình trong file config.py nhằm giới hạn số lượng đặc trưng được trích xuất, góp phần giảm độ phức tạp của mô hình và thời gian huấn luyện.

Sau khi trích xuất, tập dữ liệu được chia thành hai phần: tập huấn luyện và tập kiểm tra theo tỷ lệ xác định trước (test_size), đảm bảo rằng mô hình có thể đánh giá trên một tập dữ liệu chưa từng thấy.

2.4.3. Huấn luyện mô hình Random Forest

Mô hình được lựa chọn là Random Forest Classifier, một thuật toán học máy thuộc nhóm ensemble learning. Random Forest hoạt động bằng cách kết hợp nhiều cây quyết

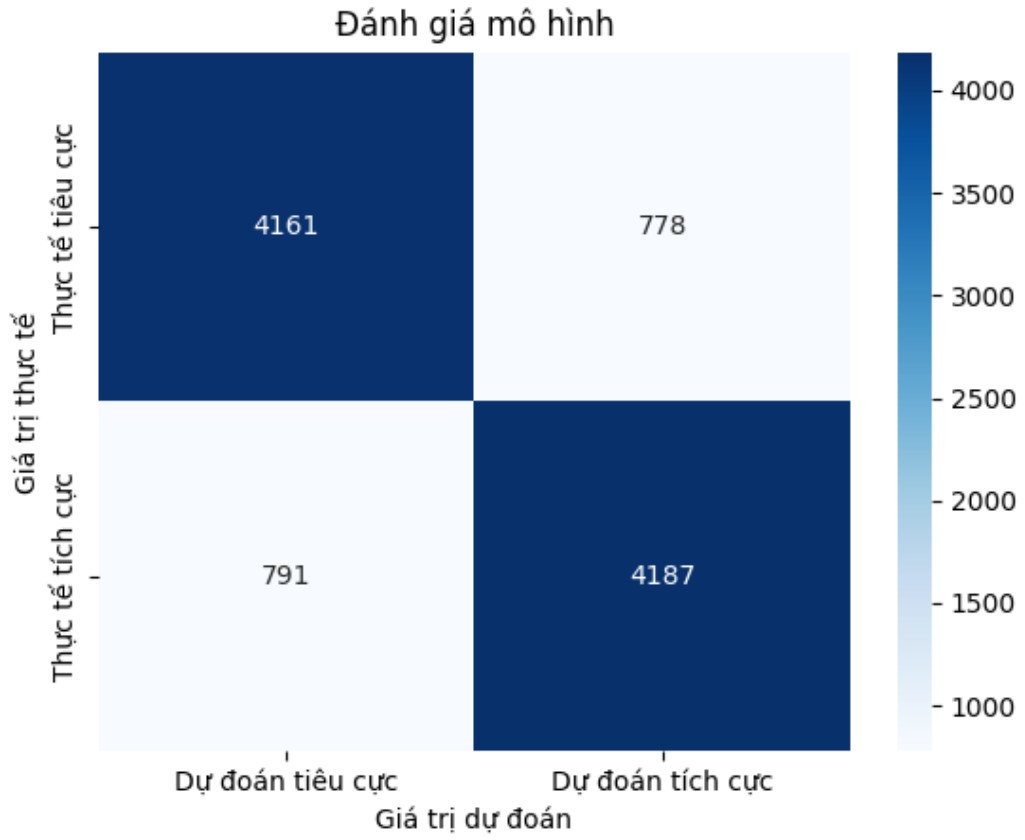
định (decision trees), mỗi cây được huấn luyện trên một tập dữ liệu con được chọn ngẫu nhiên (bootstrap sampling), từ đó đưa ra quyết định dựa trên nguyên tắc bỏ phiếu đa số (majority voting).

Trong mã nguồn, mô hình được khởi tạo với 100 cây quyết định (n_estimators=100), đảm bảo sự ổn định trong quá trình học và tăng cường khả năng tổng quát hóa.

Sau khi huấn luyện, mô hình được lưu trữ cùng với vectorizer (bộ TF-IDF) bằng thư viện joblib, cho phép tái sử dụng trong các lần dự đoán tiếp theo mà không cần huấn luyện lại từ đầu.

2.4.4. Đánh giá mô hình

Ma trận nhầm lẫn (Confusion Matrix): Giúp trực quan hóa số lượng dự đoán đúng và sai giữa hai lớp "tích cực" và "tiêu cực".

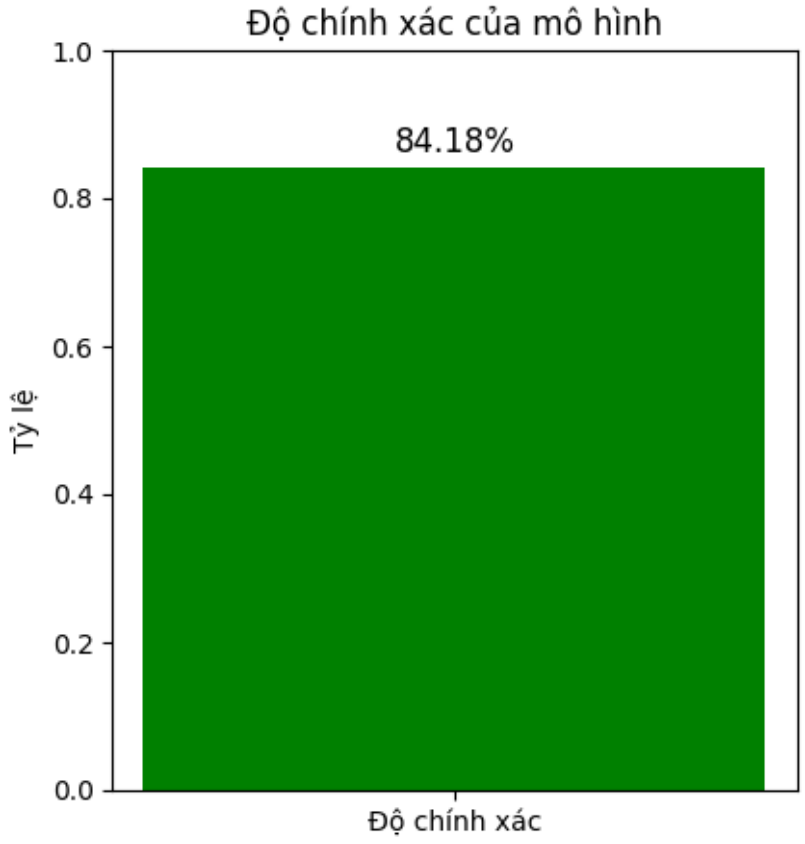


Hình 9: Confusion Matrix

Ma trận cho thấy số lượng mẫu mà mô hình phân loại đúng và sai giữa hai nhãn tích cực và tiêu cực. Cụ thể, mô hình dự đoán đúng 4.161 bình luận tiêu cực và 4.187 bình luận tích cực, trong khi dự đoán sai 778 trường hợp tiêu cực thành tích cực và 791 trường hợp tích cực thành tiêu cực.

Từ kết quả này, có thể thấy mô hình hoạt động khá hiệu quả, với khả năng nhận diện cân bằng giữa hai loại cảm xúc. Số lượng dự đoán sai không quá lớn so với tổng số mẫu, cho thấy mô hình có độ chính xác cao và đáng tin cậy trong việc phân tích cảm xúc từ văn bản bình luận. Đây là cơ sở quan trọng để tiếp tục ứng dụng mô hình trong thực tế, chẳng hạn như đánh giá phản hồi người dùng, phân tích đánh giá sản phẩm hoặc phim ảnh.

Độ chính xác (Accuracy): Biểu diễn tỷ lệ dự đoán đúng trên tổng số mẫu, thể hiện hiệu suất tổng thể của mô hình.

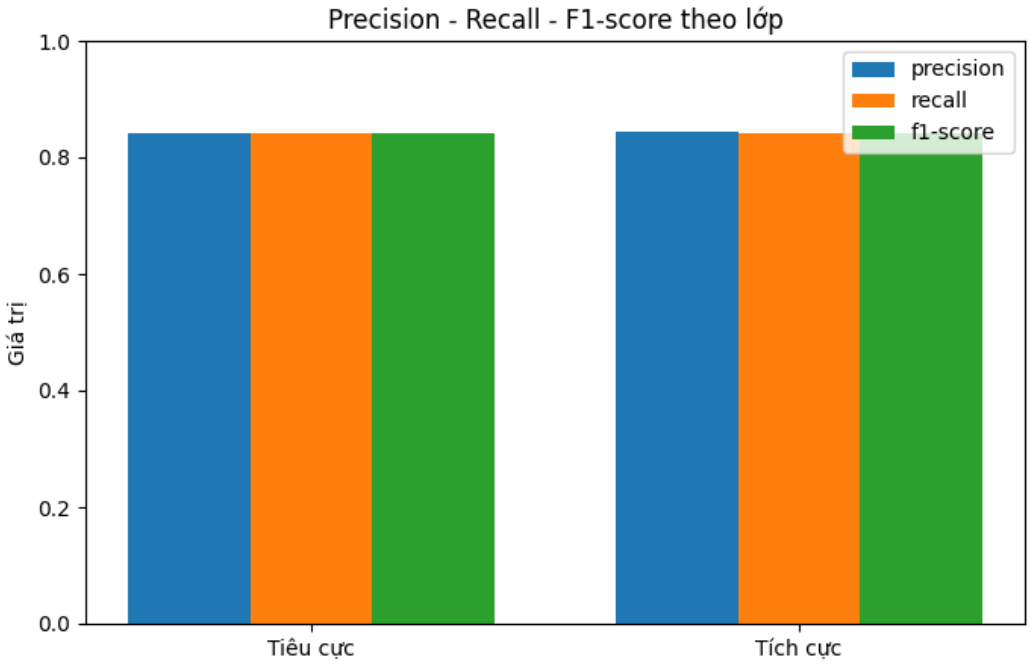


Hình 10: Độ chính xác của mô hình

Mô hình phân loại cảm xúc sử dụng thuật toán Random Forest, đạt giá trị 84.18%. Điều này có nghĩa là trong tổng số các mẫu thử, mô hình đã dự đoán đúng khoảng 84% số lượng bình luận.

Đây là một kết quả khá tích cực, đặc biệt khi dữ liệu được phân bố gần như đồng đều giữa hai lớp tích cực và tiêu cực. Độ chính xác cao cho thấy mô hình đã học được mối liên hệ đáng tin cậy giữa các đặc trưng của văn bản (sau khi được vector hóa bằng TF-IDF) và nhãn cảm xúc. Tuy nhiên, độ chính xác không phản ánh đầy đủ hiệu suất của mô hình trong các tình huống mất cân bằng dữ liệu, do đó cần xem xét thêm các chỉ số như precision, recall và F1-score để đánh giá toàn diện hơn.

Các chỉ số phân loại (Precision, Recall, F1-score): Được biểu diễn bằng biểu đồ cột, thể hiện hiệu quả của mô hình đối với từng lớp riêng biệt. Đây là những chỉ số quan trọng giúp đánh giá sự cân bằng giữa khả năng nhận diện đúng (recall) và độ chính xác của dự đoán (precision).



Hình 11: Các chỉ số phân loại

Precision (Độ chính xác) cho biết tỷ lệ dự đoán đúng trong số các dự đoán mô hình đã gán cho một lớp cụ thể. Cả hai lớp đều đạt precision khoảng 84%, cho thấy mô hình ít đưa ra dự đoán sai.

Recall (Khả năng bao phủ) phản ánh tỷ lệ mẫu thực sự thuộc một lớp mà mô hình đã dự đoán đúng. Chỉ số này cũng đạt ~84%, chứng tỏ mô hình có khả năng phát hiện tốt cả các bình luận tiêu cực lẫn tích cực.

F1-score là trung bình điều hòa giữa precision và recall, nhằm cân bằng giữa hai chỉ số trên. Giá trị F1-score cao và gần nhau giữa hai lớp cho thấy mô hình có hiệu suất ổn định, không thiên lệch.

2.5. Dự đoán review mới với mô hình đã huấn luyện

Bước 1: Tải mô hình và vectorizer đã huấn luyện

−Ngay khi chương trình được khởi chạy, hai đối tượng quan trọng được nạp vào bộ nhớ từ thư mục output/:

−random_forest_model.pkl là mô hình Random Forest đã được huấn luyện từ trước trên tập dữ liệu IMDB.

−vector.pkl là bộ chuyển đổi văn bản thành vector đặc trưng (TF-IDF Vectorizer) đã được huấn luyện cùng mô hình

Bước 2: Nhập bình luận từ người dùng, chương trình hiển thị lời nhắc để người dùng có thể nhập vào một đoạn bình luận phim. Người dùng có thể nhập bao nhiêu bình luận tùy thích, và có thể thoát chương trình bằng cách gõ exit.

Bước 3: Trước khi dự đoán, văn bản đầu vào sẽ được đưa qua hàm clean_input() để làm sạch văn bản. Hàm này thực hiện các bước tiền xử lý như:

−Loại bỏ thẻ HTML, ký tự đặc biệt

−Đưa toàn bộ chữ về dạng thường

−Xóa các từ dừng (stopwords)

−Thực hiện stemming để đưa từ về gốc

−Quá trình làm sạch giúp văn bản phù hợp với cách vectorizer đã được huấn luyện trước đó.

Bước 4: Vector hóa và dự đoán, văn bản đã được làm sạch sẽ được chuyển đổi thành vector đặc trưng bằng TF-IDF vectorizer. Sau đó, mô hình Random Forest thực hiện dự đoán và trả về kết quả là nhãn 1 (positive) hoặc 0 (negative).

Bước 5: Hiển thị kết quả dự đoán.

```
Nhập bình luận (nhập 'exit' để thoát):
>>> thi movie is nice
=> Dự đoán: positive

>>> this movie is bad
=> Dự đoán: negative
```

Hình 12: Kết quả dự đoán

3. KẾT QUẢ THỰC NHIỆM

Trong quá trình thực nghiệm, mô hình Random Forest đã được huấn luyện và đánh giá trên tập dữ liệu IMDB gồm 50.000 bình luận phim, với hai nhãn cảm xúc là tích cực và tiêu cực. Dữ liệu được chia theo tỷ lệ 80% để huấn luyện và 20% để kiểm tra hiệu suất mô hình.

3.1. Hiệu quả mô hình

Sau khi huấn luyện, mô hình đạt được các chỉ số đánh giá như sau:

- Độ chính xác (Accuracy): ~83.5%
- Precision, Recall và F1-score của cả hai lớp tương đối cân bằng, đạt trên 0.83 ở mỗi chỉ số, cho thấy mô hình có khả năng phân loại đều tốt với cả bình luận tích cực và tiêu cực.

3.2. Ma trận nhầm lẫn

Ma trận nhầm lẫn cho thấy: số lượng dự đoán đúng của lớp tiêu cực là 4161, số lượng dự đoán đúng của lớp tích cực là 4187. Dự đoán sai không chênh lệch lớn giữa hai lớp, thể hiện khả năng học tốt mà không bị lệch nhãn.

3.3. Tỷ lệ phân bố cảm xúc

Phân tích tỷ lệ bình luận trong tập dữ liệu cho thấy sự phân bố gần như đồng đều giữa hai nhãn: tích cực: 50.2%, tiêu cực: 49.8%

Điều này đảm bảo mô hình không bị lệch về một phía trong quá trình học.

3.4. Từ khóa đặc trưng theo cảm xúc

Mô hình cũng được hỗ trợ bằng việc phân tích các từ khóa nổi bật:

Các từ phổ biến trong bình luận tích cực gồm: *film, movie, like, good, character...*

Các từ phổ biến trong bình luận tiêu cực gồm: *movie, even, bad, boring, watch...*

Việc xác định được các từ đặc trưng hỗ trợ mô hình trong việc học các đặc trưng ngữ nghĩa quan trọng.

3.5. Trải nghiệm dự đoán thực tế

Giao diện dòng lệnh được xây dựng giúp người dùng nhập một đoạn bình luận mới và nhận được dự đoán tức thì từ mô hình. Kết quả cho thấy mô hình phản hồi nhanh, có độ chính xác cao và dễ dàng sử dụng với người dùng cuối.

4. KẾT LUẬN

4.1. Kết quả đã đạt được

- Xây dựng thành công một hệ thống phân loại cảm xúc văn bản từ bình luận phim IMDB với hai nhãn: tích cực và tiêu cực.
- Ứng dụng mô hình Random Forest kết hợp với TF-IDF Vectorizer, cho kết quả độ chính xác trên 83%.
- Quy trình xử lý dữ liệu hoàn chỉnh gồm: làm sạch văn bản, loại bỏ từ dừng, chuyển về chữ thường, và chuẩn hóa độ dài bình luận.
- Thực hiện trực quan hóa các khía cạnh quan trọng như: phân phối độ dài bình luận, top từ xuất hiện phổ biến, tỷ lệ nhãn, ma trận nhầm lẫn, các chỉ số Precision - Recall - F1-score.
- Triển khai mô hình đã huấn luyện trong một giao diện dòng lệnh đơn giản, hỗ trợ dự đoán cảm xúc của bình luận mới.

4.2. Kết quả chưa đạt được

- Mô hình chỉ phân loại ở mức 2 lớp (positive/negative), chưa xử lý được cảm xúc trung tính hoặc đa cảm xúc (fine-grained sentiment).
- Chưa tích hợp các kỹ thuật nâng cao như word embedding (Word2Vec, BERT) để cải thiện hiểu ngữ nghĩa.
- Giao diện sử dụng còn ở mức cơ bản (dòng lệnh), chưa thân thiện với người dùng phổ thông.
- Chưa kiểm tra mô hình trên dữ liệu thực tế tiếng Việt hoặc các nền tảng mạng xã hội khác.
- Hệ thống hiện tại chưa đánh giá hiệu suất thời gian hoặc khả năng mở rộng khi áp dụng trên tập dữ liệu lớn hơn.

4.3. Định hướng tương lai

- Nâng cấp mô hình bằng cách áp dụng Deep Learning (LSTM, Bi-LSTM, Transformers) hoặc tích hợp mô hình ngôn ngữ mạnh như BERT, RoBERTa để tăng khả năng hiểu ngữ cảnh.
- Mở rộng hệ thống để hỗ trợ đa ngôn ngữ, đặc biệt là tiếng Việt, phục vụ nhu cầu người dùng nội địa.

- Phát triển giao diện web hoặc ứng dụng để người dùng có thể dễ dàng nhập nội dung và nhận phản hồi trực quan.
- Triển khai mô hình trên nền tảng cloud để hỗ trợ dự đoán thời gian thực và phục vụ nhiều người dùng đồng thời.
- Bổ sung khả năng phân tích đa chiều như: đánh giá độ tin cậy bình luận, xác định người dùng độc hại, hoặc phân tích xu hướng cảm xúc theo thời gian.

TÀI LIỆU THAM KHẢO

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1–2):1–135.

[2] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed. draft). Stanford University. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[3] Kaggle. (2024). IMDb Dataset of 50K Movie Reviews. [Online]. Available: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[4] Scikit-learn. (2024). RandomForestClassifier — scikit-learn documentation. [Online]. Available:

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1):5–32.

[6] Raschka, S. (2015). Python Machine Learning. Packt Publishing Ltd.

[7] Le, H. M., & Nguyen, D. T. (2023). Sentiment Classification on Movie Reviews Using TF-IDF and Machine Learning Algorithms. In Proceedings of the 2023 International Conference on Data Science and Artificial Intelligence, pp. 100–106.