TRƯỜNG ĐẠI HỌC THỦY LỢI



Báo CÁO BÀI TẬP LỚN MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Đề tài: Phân tích cảm xúc bình luận

*NGƯỜI HƯỚNG DẪN: PGS.TS.Nguyễn Quang Hoan*

|  |  |
| --- | --- |
| Họ và tên | MSV |
| Lương Tuấn Anh | 2251262571 |

Hà Nội, Năm 2025

MỤC LỤC

[DANH MỤC THUẬT NGỮ, KÝ HIỆU 3](#_Toc201760096)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 3](#_Toc201760097)

[1.1. Tóm tắt đề tài 3](#_Toc201760098)

[1.2. Lý do chọn đề tài 4](#_Toc201760099)

[1.3. Mục tiêu và phạm vi nghiên cứu 4](#_Toc201760100)

[1.4. Phương pháp nghiên cứu và ý nghĩa khoa học, thực tiễn 4](#_Toc201760101)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 5](#_Toc201760102)

[2.1. Tổng quan về xử lý ngôn ngữ tự nhiên (NLP) 5](#_Toc201760103)

[2.1.1. Khái niệm NLP 5](#_Toc201760104)

[2.1.2. Ứng dụng của NLP trong thực tế 5](#_Toc201760105)

[2.2. Phân tích cảm xúc 6](#_Toc201760106)

[2.3. Tiền xử lý dữ liệu trong NLP 6](#_Toc201760107)

[2.4. Biến đổi đặc trưng bằng TF-IDF 7](#_Toc201760108)

[2.5. Thuật toán Logistic Regression 7](#_Toc201760109)

[2.6. Các phương pháp đánh giá kết quả 8](#_Toc201760110)

[CHƯƠNG 3: BỘ DỮ LIỆU & TIỀN XỬ LÝ & TRỰC QUAN HÓA 8](#_Toc201760111)

[3.1. Mô tả bộ dữ liệu IMDB 8](#_Toc201760112)

[3.2. Quá trình tiền xử lý dữ liệu 9](#_Toc201760113)

[3.2.1. Luồng xử lý dữ liệu 9](#_Toc201760114)

[3.2.2. Tiền xử lý 9](#_Toc201760115)

[3.3. Trực quan hóa 12](#_Toc201760116)

[CHƯƠNG 4: XÂY DỰNG, HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH 13](#_Toc201760117)

[4.1. Kiến trúc mô hình Logistic Regression 13](#_Toc201760118)

[4.2. Quá trình huấn luyện 13](#_Toc201760119)

[4.3. Kết quả thực nghiệm 14](#_Toc201760120)

[4.4. Phân tích kết quả 15](#_Toc201760121)

[CHƯƠNG 5: DEMO 16](#_Toc201760122)

[KẾT LUẬN 17](#_Toc201760123)

[Bibliography 18](#_Toc201760124)

# DANH MỤC THUẬT NGỮ, KÝ HIỆU

|  |  |  |
| --- | --- | --- |
| **STT** | **Ký hiệu viết tắt** | **Chữ viết đầy đủ** |
| 1 | NLP | Natural Language Processing |
| 2 | TF-IDF | Term Frequency-Inverse Document Frequency |
| 3 | LR | Logistic Regression |
| 4 | NLTK | Natural Language Toolkit |
| 5 | IMDB | Internet Movie Database |
| 6 | AI | Artificial Intelligence |
| 7 | HTML | HyperText Markup Language |
| 8 | OMW | Open Multilingual Wordnet |
| 9 | LSTM | Long Short Term Memory |
| 10 | BERT | Bidirectional Encoder Representations from Transformers |

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

## Tóm tắt đề tài

Đề tài nghiên cứu về phân tích cảm xúc trên tập dữ liệu IMDB bằng phương pháp học máy truyền thống, trong đó sử dụng quy trình xử lý ngôn ngữ tự nhiên để chuẩn hóa dữ liệu review, trích xuất các đặc trưng bằng TF-IDF và áp dụng với mô hình Logistic Regression để phân loại Review thành tích cực hoặc tiêu cực. Kết quả nghiên cứu giúp đánh giá hiệu quả của việc kết hợp NLP cùng với Logistic Regression trong bài toán phân tích cảm xúc, hướng đến việc ứng dụng thực tế trong việc tự động hóa xử lý phản hồi của khách hàng trên các website phim ảnh.

## Lý do chọn đề tài

Việc đánh giá cảm xúc trong review phim đang ngày càng trở nên cần thiết trong thời đại số hiện nay, khi lượng phản hồi của người xem trên internet ngày một tăng. Bộ dữ liệu IMDB được sử dụng rộng rãi trong nghiên cứu NLP nhờ tính đa dạng về ngữ nghĩa và độ lớn phù hợp. Do đó, lựa chọn đề tài ‘Phân tích cảm xúc bình luận’ nhằm tạo ra một công cụ tự động xác định nhanh trạng thái tích cực hay tiêu cực của các reviewer, góp phần nâng cao trải nghiệm của người dùng và hiệu quả trong việc quản lý các nội dung trên nền tảng trực tuyến.

## Mục tiêu và phạm vi nghiên cứu

Mục tiêu chính của nghiên cứu lần này là để xây dựng lên một hệ thống có khả năng tự động nhận diện cảm xúc trong review phim trên tập dữ liệu IMDB. Cụ thể, nghiên cứu hướng tới việc đánh giá hiệu quả của quy trình xử lý NLP kết hợp trích xuất đặc trưng của TF-IDF cùng với thuật toán Logistic Regression trong việc phân tích cảm xúc của các bình luận.

Nghiên cứu được giới hạn trong việc sử dụng tập dữ liệu review phim IMDB, gồm khoảng 50,000 review được gán nhãn ‘positive’ và ‘negative’. Phạm vi bao gồm toàn bộ quy trình tiền xử lý dữ liệu, trích xuất đặc trưng, huấn luyện, đánh giá độ hiệu quả của mô hình và thử nghiệm trên các review mới.

## Phương pháp nghiên cứu và ý nghĩa khoa học, thực tiễn

Nghiên cứu được sử dụng kết hợp phương pháp thực nghiệm trên dữ liệu thực tế cùng với kỹ thuật xử lý ngôn ngữ tự nhiên. Cụ thể, quy trình bao gồm: làm sạch các review thô bằng cách chuẩn hóa, tokenization, loại bỏ stopwords, lemmatization, trích xuất các đặc trưng bằng TF-IDF rồi sử dụng thuật toán Logistic Regression để huấn luyện và đánh giá kết quả trên tập test bằng các thước đo như accuracy, precision, recall, F1-score.

Kết quả nghiên cứu góp phần khẳng định tính hiệu quả của việc kết hợp NLP với Logistic Regression trong bài toán phân tích cảm xúc bình luận. Về mặt khoa học, đề tài này giúp tổng kết kiến thức về xử lý ngôn ngữ tự nhiên trên tập dữ liệu review quy mô lớn. Về mặt thực tiễn, nghiên cứu tạo ra nền tảng giúp tự động hóa việc đánh giá review, phục vụ nhu cầu trong ngành truyền thông, phim ảnh và thương mại điện tử

# CƠ SỞ LÝ THUYẾT

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

### Khái niệm NLP

Xử lý ngôn ngữ tự nhiên (NLP) là một công nghệ máy học, cung cấp cho máy tính khả năng diễn giải, tương tác và hiểu được ngôn ngữ của con người. 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]

NLP là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc nghiên cứu và phân tích ngôn ngữ tự nhiên, với mục tiêu hiểu và xử lý thông tin trong các ngôn ngữ mà con người sử dụng hàng ngày. NLP sử dụng một số phương pháp và công nghệ để cho phép máy tính hiểu, tạo ra và tương tác với ngôn ngữ tự nhiên một cách tự động. [2]

### Ứng dụng của NLP trong thực tế

NLP đã có những ứng dụng rất đa dạng trong cuộc sống hàng ngày của chúng ta. Dưới đây là một số ứng dụng của NLP:

* **Tự động hóa dịch thuật**: Công nghệ NLP giúp tự động dịch thuật giữa các ngôn ngữ khác nhau, giúp người dùng truy cập thông tin từ trang web, tài liệu hoặc tin nhắn trong ngôn ngữ mà họ không thạo
* **Tự động trả lời email**: NLP có thể được sử dụng để phân tích và trả lời các email theo cách tự động, giúp tiết kiệm thời gian và năng lực làm việc.
* **Phân loại tin tức**: NLP có thể giúp tổ chức và phân loại tin tức từ các nguồn thông tin khác nhau, giúp người dùng có thể dễ dàng theo dõi và tiếp cận thông tin theo quan tâm cá nhân
* **Phân tích cảm xúc xã hội**: NLP có thể được sử dụng để phân tích cảm xúc từ các bài đăng trên mạng xã hội, giúp các công ty và tổ chức hiểu được ý kiến và phản hồi của khách hàng.

## Phân tích cảm xúc

Phân tích cảm xúc là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên nhằm xác định trạng thái cảm xúc hoặc thái độ được diễn đạt trong đoạn văn bản, giúp phân loại ý kiến thành tích cực, tiêu cực hoặc trung lập.

Trong nghiên cứu lần này, bài toán được giới hạn trong việc phân tích cảm xúc nhị phân, xác định rõ review phim thuộc nhóm tích cực (positive) hay tiêu cực (negative). Phân tích cảm xúc có thể thực hiện ở nhiều mức độ khác nhau: ở mức **Từ** để xác định ý nghĩa cảm xúc của từng từ ngữ riêng lẻ, ở mức **Câu** để đánh giá cảm xúc chung của cả câu, và ở mức **Đoạn** để xác định cảm xúc tổng thể của cả một đoạn văn. Việc lựa chọn mức phân tích phù hợp sé giúp cải thiện độ chính xác và ý nghĩa của kết quả dự đoán.

## Tiền xử lý dữ liệu trong NLP

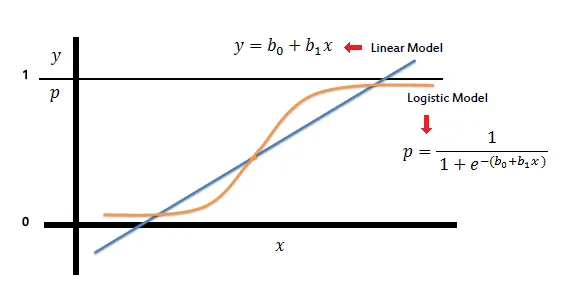
Tiền xử lý dữ liệu trong NLP là một bước quan trọng giúp chuẩn hóa đoạn văn bản thô thành dạng phù hợp để trích xuất đặc trưng và huấn luyện mô hình. Trước hết, toàn bộ đoạn review được chuẩn hóa về chữ thường (**Lowercasing**) nhằm tạo sự nhất quán trong bộ dữ liệu, giúp mô hình không bị nhầm lẫn giữa các từ giống nhau nhưng khác về kiểu chữ. Tiếp đó là loại bỏ các thẻ HTML, dấu câu, các ký tự đặc biệt để giảm thiệu độ nhiễu trong review, chỉ giữ lại các phần có ý nghĩa. Sau đó, đoạn review được tạo thành các token (từ) bằng phương pháp **Tokenization**, giúp biểu diễn đoạn văn dưới dạng dãy từ phục vụ cho việc trích xuất đặc trưng. Các từ dừng (**Stopwords**) ví dụ như: ‘and’, ‘is’, ‘the’, … sẽ được lọc ra nhằm giữ lại những thành phần có đóng góp tích cực trong việc xác định trạng thái của cảm xúc. Cuối cùng, kỹ thuật **Lemmatization** được sử dụng để chuẩn hóa các từ về dạng gốc, giúp giảm thiểu số lượng từ trong từ điển, tăng hiệu quả tổng quát hóa của mô hình trong bài toán phân tích cảm xúc.

## Biến đổi đặc trưng bằng TF-IDF

Sau khi đã hoàn thành bước tiền xử lý ở trên, đoạn review sẽ cần được biểu diễn dưới dạng số để phục vụ cho các thuật toán học máy. Phương pháp TF-IDF được sử dụng nhằm xác định mức độ quan trọng của mỗi từ trong toàn bộ tập dữ liệu. Cách tính toán TF-IDF kết hợp tần suất xuất hiện của từ đó trong tài liệu (TF) cùng mức độ hiếm của từ trong toàn bộ tập tài liệu (IDF). Kỹ thuật này giúp nhấn mạnh các từ có ý nghĩa đặc trưng trong đoạn review, đồng thời hạn chế tác động của các từ phổ biến nhưng lại ít ý nghĩa, từ đó sẽ giúp cải thiện đáng kể hiệu quả trong bài toán phân tích cảm xúc

## Thuật toán Logistic Regression

Logistic Regression là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm. Hồi quy Logistics làm việc dựa trên nguyên tắc của hàm sigmoid – một hàm phi tuyến tự chuyển đầu vào của nó thành một xác suất thuộc về một trong hai lớp nhị phân.



Logistic Regression sử dụng hàm phi tuyến để xác định xác suất của hai lớp 0 và 1

Hàm sigmoid:

Hàm mất mát Cross-Entropy:

Trong đó:

* n: số lượng mẫu dữ liệu trong tập huấn luyện.
* : giá trị thực tế của đầu ra thứ i.
* : xác suất dự đoán thuộc lớp 1 của mô hình cho đầu vào thứ i.

Hàm Cross-Entropy đo lường khoảng cách giữa hai phân phối xác suất và . Khi mô hình dự đoán chính xác, tức là nếu thì càng gần 1 và nếu thì càng gần 0, sau đó hàm mất mát sẽ tiến gần về 0.

Trong quá trình huấn luyện, chúng ta sẽ tìm cách cập nhật bộ trọng số ‘w’ sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự báo tốt nhất. [3]

## Các phương pháp đánh giá kết quả

Để đánh giá hiệu quả của mô hình trong bài toán phân tích cảm xúc bình luận, có nhiều thước đo được sử dụng. Accuracy là tỷ lệ phần trăm số dự đoán đúng trên toàn bộ tập dữ liệu. Precision đo tỷ lệ dự đoán đúng trong số tất cả các kết quả được dự báo là tích cực (positive). Recall xác định tỷ lệ phát hiện đúng trong tổng số kết quả thực tế tích cực. F1-score là thước đo kết hợp giữa Precision và Recall, giúp đánh giá toàn diện hiệu suất mô hình. Ngoài ra, Confusion Matrix được sử dụng để tổng kết số lượng dự đoán đúng, sai trên cả hai nhãn, giúp đánh giá chi tiết mức độ nhầm lẫn của mô hình trên từng lớp.

# BỘ DỮ LIỆU & TIỀN XỬ LÝ & TRỰC QUAN HÓA

## Mô tả bộ dữ liệu IMDB

Bộ dữ liệu IMDB được sử dụng trong nghiên cứu có nguồn gốc từ kho dữ liệu review phim trên Internet Movie Database, một trong những website đánh giá phim lớn nhất thế giới. Bộ dữ liệu này được công bố công khai nhằm mục đích nghiên cứu trong lĩnh vực xử lý ngôn ngữ tự nhiên, đặc biệt là với bài toán phân tích cảm xúc. Bộ dữ liệu bao gồm tổng cộng 50.000 đoạn review phim được chia đều thành hai nhãn: 25.000 review tích cực (positive) và 25.000 review tiêu cực (negative). Mỗi review là dạng đoạn văn bản thô, có độ dài khác nhau, có chứa các thẻ HTML, dấu câu, ký tự đặc biệt, đòi hỏi quy trình tiền xử lý phù hợp để có thể trích xuất được thông tin cần thiết phục vụ cho quá trình huấn luyện mô hình.

## Quá trình tiền xử lý dữ liệu

### Luồng xử lý dữ liệu

Preprocessing

INPUT

Tokenization, Lemmatization, TF-IDF

Split train/test

Negative - Positive

Logistic Regression

Accuracy, F1-score, Confusion Matrix

### Tiền xử lý

A screen shot of a computer

AI-generated content may be incorrect.

Đầu tiên tải về các tài nguyên của NLTK (thư viện xử lý ngôn ngữ tự nhiên)

* Stopwords: là tập các từ dừng trong tiếng anh (VD: a, an, the, is, and, …) những từ thường không có nhiều ý nghĩa trong việc phân tích
* Wordnet: Là cơ sở dữ liệu về từ vựng tiếng Anh (từ, nghĩa, từ gần nghĩa)
* Omw-1.4: OMW – bộ dữ liệu bổ sung cho Wordnet

A screen shot of a computer screen

AI-generated content may be incorrect.

Ta sẽ tiến hành lấy danh sách toàn bộ stopwords tiếng Anh trong NLTK, sau đó sẽ biến chúng thành một tập hợp để có thể tra cứu nhanh hơn khi lọc.

Tạo một Lemmatizer để chuẩn hóa từ (VD: running -> run, better -> good, cats -> cat), điều này sẽ giúp cho đoạn text trở nên đồng nhất và dễ dàng phân tích hơn khi tạo features

A computer code with text

AI-generated content may be incorrect.

Hàm clean\_text trên đây sẽ tiền hành các công việc sau: Biến toàn bộ text thành chữ thường để tránh khác biệt kiểu chữ, xóa toàn bộ thẻ HTML giúp đoạn text sạch hơn và xóa toàn bộ các ký tự đặc biệt bằng cách thay thế bằng dấu cách. Sau đó sử dụng Tokenization để tách các từ và lọc ra các stopwords đã được chuẩn bị từ trước để giữ lại các phần có nghĩa rồi sử dụng Lemmatization để chuyển từ về dạng gốc. Cuối cùng ghép các token còn lại thành một đoạn text mới được làm sạch.

VD:

Input: "This is an <br> example of cats running quickly."

Output: "example cat run quickly"

A computer code on a dark background

AI-generated content may be incorrect.

Ta tiến hành đọc dữ liệu, dữ liệu IMDB-Dataset.csv sẽ có hai cột, một cột review và một cột sentiment. Sau khi qua bước tiền xử lý ở trên thì đoạn text mới được làm sạch sẽ được viết ở một cột mới là clean\_review.

A computer screen shot of a program code

AI-generated content may be incorrect.

Tiền hành chia tập train/test với tỉ lệ 80/20 với 80% train và 20% test. Khởi tạo công cụ TfidfVectorizer và chỉ lấy 5000 từ phổ biến nhất trong toàn bộ corpus, việc này sẽ giúp tiết kiệm bộ nhớ và tránh hiện tượng overfitting. Sau đó áp dụng TF-IDF lên tập train qua hàm fit\_transform, biến toàn bộ X\_train thành một ma trận số (mỗi review sẽ biến thành một vector dài 5000) và đây sẽ chính là đầu vào để train model, nhãn ở đây sẽ là y với 2 giá trị 1 và 0 biểu diễn cho positive và negative.

## Trực quan hóa

A blue and pink squares

AI-generated content may be incorrect.

Bộ dữ liệu gồm 50.000 review, chia đều cho 2 nhãn Positive và Negative mỗi nhãn sẽ có 25.000 review

A graph of a number of words

AI-generated content may be incorrect.

Dưới đây là top 20 từ có tổng TF-IDF cao nhất trong toàn bộ dataset

# XÂY DỰNG, HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

## Kiến trúc mô hình Logistic Regression

Logistic Regression là một mô hình hồi quy tuyến tính được sử dụng để thực hiện bài toán phân loại nhị phân. Mô hình tạo ra một tổng có trọng số:

Sau đó, z sẽ được đưa qua hàm Sigmoid:

Hàm Sigmoid sẽ giúp kết quả được chuẩn hóa về khoảng xác suất [0, 1]

Hàm loss được sử dụng trong quá trình huấn luyện là Binary Cross – Entropy:

## Quá trình huấn luyện

Mô hình được huấn luyện bằng thuật toán tối ưu hóa Gradient được tích hợp trong scikit-learn. Việc lựa chọn tham số:

* max-iter = 200: Số vòng lặp tối đa, giúp thuật toán có đủ thời gian để hội tụ
* max-features = 5000: cho biết đầu vào của mỗi review sẽ là một vector với độ dài là 5000

Biến toàn bộ review thành một vector số bằng TfidfVectorizer, áp dụng Logistics Regression trên ma trận đặc trưng đó bằng để học ra các hệ số . Khi loss đủ thấp hoặc đạt đủ số vòng lặp đã thiết lập từ trước, quá trình huấn luyện sẽ kết thúc.

## Kết quả thực nghiệm

A screenshot of a graph

AI-generated content may be incorrect.

Với độ chính xác tổng quát accuracy = 0.89, cho thấy rằng mô hình dự đoán đúng 89% trên tập test, kết quả tốt trong bài toán phân tích cảm xúc trên bộ dữ liệu IMDB.

Với nhãn Negative

* Precision = 0.90: Khi dự đoán nhãn ‘Negative’. Có 90% khả năng dự đoán đúng
* Recall = 0.87: Trong số review thật sử là ‘Negative’, có tổng số 87% được dự đoán đúng
* F1-score = 0.89: Đánh giá tổng hòa cả Precision và Recall

Với nhãn Positive

* Precision = 0.88: Khi dự đoán ‘Positive’, xác suất đúng là 89%
* Recall = 0.90: Model nhận diện được 90% review ‘Positive’
* F1-score = 0.89: Đạt hiệu suất cao trong việc nhận dạng review tích cực

Macro avg và weighted avg = 0.89: điều này chứng tỏ rằng model không bị lệch quá nhiều về một phía, phân bố kết quả khá ổn định trên cả hai nhãn.

A blue squares with white text

AI-generated content may be incorrect.

* Có 4337 số review là ‘Negative’ được dự đoán đúng
* Có 624 số review là ‘Negative’ nhưng bị dự đoán nhầm sang ‘Positive’
* Có 483 số review là ‘Positive’ nhưng bị dự đoán thành ‘Negative’
* Có 4556 số review ‘Positive’ được dự đoán đúng

## Phân tích kết quả

Ưu điểm:

* Đơn giản, dễ sử dụng, tốc độ huấn luyện nhanh
* Cho kết quả tương đối chính xác trên bài toán nhị phân

Nhược điểm

* Khi dữ liệu có nhiều đặc trưng phi tuyến, hiệu quả sẽ bị hạn chế
* Cần phải xử lý kỹ phần trích xuất đặc trưng (như TF-IDF) để có thể đạt hiệu quả cao hơn

Với kết quả thu được là độ chính xác 89% phù hợp với kỳ vọng ban đầu, mô hình hoàn toàn có đủ khả năng để làm baseline trong bài toán sentiment analysis.

Để cải tiến cho mô hình có thể đạt được hiệu quả cao hơn, có thể kết hợp sử dụng với Word Embedding hoặc áp dụng các mô hình Deep Learning để đạt hiệu suất cao hơn.

# DEMO

Dưới đây là phần thử nghiệm thực tế của hệ thống phân tích cảm xúc trên bộ dữ liệu IMDB. Sau khi mô hình LR được huấn luyện thành công, hệ thống được tích hợp thành một quy trình thử nghiệm giúp người dùng có thể nhập đoạn review bằng tiếng Anh và nhận kết quả dự đoán ngay lập tức.

Khi thử nghiệm, review do người dùng nhập sẽ được truyền qua toàn bộ quy trình xử lý: chuẩn hóa ký tự, loại bỏ thẻ HTML, dấu câu, ký tự đặc biệt, tách từ, loại bỏ từ dừng và chuẩn hóa về dạng gốc. Đoạn review sạch sau đó sẽ được vector hóa bằng TF-IDF tạo thành một ma trận đặc trưng phù hợp với đầu vào của mô hình. Tiếp đó, mô hình LR sẽ thực hiện dự đoán xác suất thuộc về nhãn ‘Positive’ hoặc ‘Negative’. Kết quả sẽ hiển thị trực tiếp trên màn hình khi người dùng nhấn nút ‘Predict Sentiment’ sau khi đã gõ xong review.

A screenshot of a video game

AI-generated content may be incorrect.

# KẾT LUẬN

Đề tài nghiên cứu về phân tích cảm xúc trên bộ dữ liệu IMDB bằng NLP đã trình bày toàn bộ quy trình thực hiện, từ việc khảo sát vấn đề, tìm hiểu cơ sở lý thuyết, lựa chọn công cụ, cho đến thực nghiệm và đánh giá kết quả. Trên cơ sở đó, nghiên cứu đã sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên như chuẩn hóa dữ liệu, Tokenization, loại bỏ stopwords và lemmatization để làm sạch dữ liệu review. Sau đó đặc trưng được trích xuất bằng phương pháp TF-IDF, giúp biểu diễn review dưới dạng số, phục vụ cho bài toán học máy. Mô hình Logistic Regression được lựa chọn để thực hiện nhiệm vụ phân tích cảm xúc, tận dụng ưu điểm về tính hiệu quả, tốc độ huấn luyện cùng khả năng diễn giải kết quả rõ ràng.

Kết quả thực nghiệm trên bộ dữ liệu IMDB cho thấy hệ thống đạt độ chính xác cao trong việc xác định trạng thái cảm xúc tích cực hay tiêu cực trong review phim. Các thước đo hiệu quả như accuracy, precision, recall, f1-score đều đạt mức cao, khẳng định tính khả thi của phương pháp nghiên cứu. Phần Demo thực tế giúp minh họa rõ ràng khả năng ứng dụng của hệ thống trong việc tự động phân tích phản hồi của người xem, phục vụ nhu cầu thực tế trong ngành công nghiệp điện ảnh, truyền thông và nhiều lĩnh vực khác.

Bên cạnh kết quả tích cực đạt được, nghiên cứu vẫn còn những hạn chế nhất định, như việc chỉ giới hạn trong bài toán nhị phân hay phụ thuộc phần lớn vào chất lượng của quy trình tiền xử lý dữ liệu. Trong tương lai có thể mở rộng ra bài toán đa lớp, tích hợp thêm các mô hình học sâu như LSTM, Transformer hoặc BERT để nâng cao độ chính xác, cũng như thử nghiệm trên nhiều bộ dữ liệu da dạng khác.

# Bibliography

|  |  |
| --- | --- |
| [1] | "Xử lý ngôn ngữ tự nhiên (NLP) là gì?," AWS, [Online]. Available: https://aws.amazon.com/vi/what-is/nlp/. |
| [2] | Đ. M. Đức, "NLP là gì? Toàn bộ kiến thức về Natural Language Processing từ A-Z," 2023. |
| [3] | "Bài 6: Logistic Regression (Hồi quy Logistic)," *Trí tuệ nhân tạo.* |
| [4] | V. H. Tiệp, "Bài 10: Logistic Regression," *Machine Learning cơ bản,* 2017. |