BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN

BÀI TẬP LỚN

PHÂN LOẠI ĐÁNH GIÁ CẢM XÚC PHIM

NGÀNH: CÔNG NGHỆ THÔNG TIN

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

SINH VIÊN: **NGUYỄN MINH HIẾU**

LÊ ĐỨC THẮNG

MÃ SINH VIÊN:12423049

12423032

MÃ LỚP: 124231

GV GIẢNG DẠY: **GV. NGUYỄN HOÀNG ĐIỆP**

HƯNG YÊN – 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Hưng Yên, ngày tháng năm 2025

(Ký và ghi rõ họ tên)

LỜI CAM ĐOAN

Em xin cam đoan bài tập lớn "Phân loại cảm xúc của phim" là kết quả thực hiện của nhóm em dưới sự hướng dẫn của Cô Nguyễn Hoàng Điệp.

Những phần sử dụng tài liệu tham khảo trong bài tập lớn đã được nêu rõ trong phần tài liệu tham khảo. Các kết quả trình bày trong bài tập lớn và chương trình xây dựng được hoàn toàn là kết quả do bản thân nhóm em thực hiện.

Nếu vi phạm lời cam đoan này, nhóm em xin chịu hoàn toàn trách nhiệm trước khoa và nhà trường.

Hưng Yên, ngày … tháng … năm…..

SINH VIÊN

MỤC LỤC

DANH MỤC CÁC KÝ TỰ, CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Cụm từ tiếng anh	Diễn giải
1	API	Application Programming Interface	Giao diện lập trình ứng dụng
2	AUC	Area Under the Curve	Diện tích dưới đường cong, chỉ số đánh giá khả năng phân loại của mô hình
3	BERT	Bidirectional Encoder Representations from Transformers	Mô hình học sâu dựa trên kiến trúc Transformer để hiểu ngữ cảnh ngôn ngữ
4	HTML	HyperText Markup Language	Ngôn ngữ đánh dấu siêu văn bản, thường xuất hiện dưới dạng thẻ nhiễu trong dữ liệu thô
5	IMDB	Internet Movie Database	Cơ sở dữ liệu điện ảnh trực tuyến, nguồn của bộ dữ liệu đánh giá phim được sử dụng
6	SVM	Support Vector Machine	Máy vectơ hỗ trợ - một thuật toán học máy có giám sát dùng cho phân loại
7	TF-IDF	Term Frequency - Inverse Document Frequency	Tần suất từ - Nghịch đảo tần suất văn bản, một kỹ thuật vector hóa văn bản

DANH MỤC CÁC HÌNH VẼ

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về học máy

Học máy là một lĩnh vực quan trọng trong trí tuệ nhân tạo, tập trung vào việc xây dựng các thuật toán cho phép máy tính học hỏi từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần lập trình chi tiết. Khái niệm học máy được hiểu là quá trình mà hệ thống cải thiện hiệu suất dựa trên kinh nghiệm từ dữ liệu, giúp giải quyết các vấn đề phức tạp mà lập trình truyền thống khó khăn. Các dạng học máy chính bao gồm học có giám sát, trong đó mô hình học từ dữ liệu có nhãn để dự đoán đầu ra; học không giám sát, nơi mô hình khám phá cấu trúc ẩn trong dữ liệu không nhãn; và học tăng cường, với việc mô hình tương tác với môi trường để tối ưu hóa hành động dựa trên phần thưởng. Học máy đóng vai trò cốt lõi trong trí tuệ nhân tạo, hỗ trợ các ứng dụng như nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên và dự đoán xu hướng, từ đó thúc đẩy sự phát triển của các hệ thống thông minh.

1.2. Phát biểu bài toán

Bài toán cần giải quyết là phân loại sentiment trong đánh giá phim, cụ thể là xác định xem một đánh giá là tích cực hay tiêu cực dựa trên nội dung văn bản. Đầu vào là các đánh giá phim dưới dạng văn bản thô, sau khi được tiền xử lý thành các đặc trưng số hóa, trong khi đầu ra là nhãn phân loại gồm hai lớp: positive hoặc negative. Mục tiêu của bài toán là xây dựng mô hình học máy có khả năng dự đoán chính xác sentiment của các đánh giá mới, đạt hiệu suất cao về độ chính xác và các chỉ số khác. Ý nghĩa thực tiễn của bài toán nằm ở việc hỗ trợ các nền tảng như IMDB phân tích ý kiến người dùng, cải thiện khuyến nghị phim, và giúp các nhà sản xuất phim hiểu rõ phản hồi từ khán giả để điều chỉnh chiến lược.

1.3. Mô hình học máy được lựa chọn

Các mô hình học máy được lựa chọn cho bài toán này bao gồm Logistic Regression, Linear Support Vector Machine và Decision Tree. Logistic Regression được chọn vì đây là mô hình tuyến tính đơn giản, hiệu quả cho bài toán phân loại nhị phân, đặc biệt với dữ liệu văn bản có chiều cao, và dễ diễn giải kết quả. Linear Support Vector Machine được ưu tiên nhờ khả năng tìm biên phân cách tối ưu giữa các lớp, hoạt động tốt trên dữ liệu thưa và có khả năng chống overfitting cao. Decision Tree được sử dụng vì tính trực

quan, khả năng xử lý các mối quan hệ phi tuyến tính mà không cần giả định phân bố dữ liệu, đồng thời dễ dàng hình dung quy trình quyết định.

1.3.1 Logistic Regression

Trong thống kê, logistic model (hay logit model) là một mô hình thống kê dùng để mô hình hóa log odds của một sự kiện như là một tổ hợp tuyến tính của một hay nhiều independent variables. Trong regression analysis, logistic regression (hay logit regression) ước lượng các tham số của logistic model, tức là các coefficients trong các tổ hợp tuyến tính hoặc phi tuyến.

Trong binary logistic regression, chỉ có một binary dependent variable, được mã hóa bằng indicator variable với hai giá trị 0 và 1. Các independent variables có thể là binary variables hoặc continuous variables. Xác suất tương ứng với giá trị được gán nhãn là 1 có thể thay đổi từ 0 đến 1. Hàm chuyển đổi từ log odds sang xác suất được gọi là logistic function, do đó mô hình có tên như vậy. Đơn vị đo của thang log odds được gọi là logit, viết tắt của logistic unit.

Các binary variables được sử dụng rộng rãi trong thống kê để mô hình hóa xác suất của một class hoặc event xảy ra, chẳng hạn như xác suất một đội chiến thắng hoặc một bệnh nhân khỏe mạnh. Logistic model là mô hình được sử dụng phổ biến nhất cho binary regression kể từ khoảng năm 1970.

Binary variables có thể được mở rộng thành categorical variables khi có nhiều hơn hai giá trị có thể xảy ra, ví dụ như phân loại hình ảnh là mèo, chó hay sư tử. Khi đó, binary logistic regression được tổng quát hóa thành multinomial logistic regression. Nếu các categories có thứ tự, có thể sử dụng ordinal logistic regression.

Bản thân logistic regression model chỉ mô hình hóa probability của đầu ra theo đầu vào và không trực tiếp thực hiện statistical classification. Tuy nhiên, nó có thể được sử dụng để xây dựng một classifier, ví dụ bằng cách chọn một cutoff value. Các quan sát có probability lớn hơn cutoff sẽ được phân vào một class, còn nhỏ hơn cutoff sẽ thuộc class còn lại. Đây là một cách phổ biến để xây dựng binary classifier.

Ngoài ra, còn có các analogous linear models cho binary variables sử dụng các sigmoid functions khác thay cho logistic function, nổi bật nhất là probit model. Đặc điểm xác định của logistic model là khi một independent variable tăng lên thì odds của

kết quả sẽ được nhân lên với một tỷ lệ không đổi, và mỗi independent variable có một parameter riêng. Đối với binary dependent variable, điều này dẫn đến khái niệm odds ratio.

Ở mức độ trừu tượng hơn, logistic function là natural parameter của Bernoulli distribution, và theo nghĩa này, nó là cách đơn giản nhất để chuyển một số thực thành một probability.

Các parameters của logistic regression thường được ước lượng bằng maximum likelihood estimation (MLE). Phương pháp này không có closed form solution, không giống như ordinary least squares (OLS) trong linear regression. Logistic regression đóng vai trò cơ bản đối với các binary hoặc categorical responses, tương tự như vai trò của linear regression đối với các scalar responses. Logistic regression với MLE được phát triển và phổ biến chủ yếu bởi Joseph Berkson vào năm 1944, khi ông đặt ra thuật ngữ logit.

1.3.2 Support Vector Machine – SVM

Trong machine learning, support vector machines (SVMs), còn gọi là support vector networks, là các supervised max margin models đi kèm với các learning algorithms dùng để phân tích dữ liệu cho classification và regression analysis. SVMs được phát triển tại AT&T Bell Laboratories và là một trong những mô hình được nghiên cứu nhiều nhất, dựa trên các statistical learning frameworks của VC theory do Vapnik và Chervonenkis đề xuất.

Bên cạnh việc thực hiện linear classification, SVMs còn có thể thực hiện hiệu quả non linear classification thông qua kernel trick. Phương pháp này biểu diễn dữ liệu chỉ thông qua một tập các phép so sánh pairwise similarity giữa các điểm dữ liệu ban đầu bằng cách sử dụng kernel function, hàm này biến đổi dữ liệu sang tọa độ trong một higher dimensional feature space. Nhờ đó, SVMs sử dụng kernel trick để ánh xạ một cách ngầm định các đầu vào vào không gian đặc trưng có chiều cao, nơi mà linear classification có thể được thực hiện.

Là các max margin models, SVMs có khả năng chống chịu tốt với noisy data, chẳng hạn như các ví dụ bị phân loại sai. SVMs cũng có thể được sử dụng cho các bài toán regression, trong đó mục tiêu tối ưu trở thành epsilon sensitive.

Thuật toán support vector clustering, do Hava Siegelmann và Vladimir Vapnik đề xuất, áp dụng các nguyên lý thống kê của support vectors được phát triển trong thuật toán support vector machines để phân loại unlabeled data. Các tập dữ liệu này yêu cầu các phương pháp unsupervised learning, nhằm tìm ra các natural clustering của dữ liệu thành các nhóm, và sau đó ánh xạ dữ liệu mới dựa trên các cụm này.

Sự phổ biến của SVMs có thể đến từ việc chúng thuận lợi cho theoretical analysis và có tính linh hoạt cao khi áp dụng cho nhiều loại bài toán khác nhau, bao gồm cả các bài toán structured prediction. Tuy nhiên, chưa có kết luận rõ ràng rằng SVMs có hiệu năng dự đoán tốt hơn các linear models khác như logistic regression và linear regression.

1.3.3 Decision tree

Decision tree là một cấu trúc decision support dựa trên recursive partitioning, sử dụng mô hình dạng cây để biểu diễn các decisions và các possible consequences của chúng, bao gồm chance event outcomes, resource costs và utility. Đây là một cách để biểu diễn một algorithm chỉ bao gồm các conditional control statements.

Decision trees được sử dụng rộng rãi trong operations research, đặc biệt là trong decision analysis, nhằm hỗ trợ xác định strategy có khả năng cao nhất để đạt được một goal. Bên cạnh đó, decision trees cũng là một công cụ phổ biến trong machine learning.

1.4. Ý tưởng lựa chọn mô hình

Bài toán phân loại bình luận phim theo cảm xúc (positive/negative) là một dạng bài toán phân loại văn bản nhị phân phổ biến trong xử lý ngôn ngữ tự nhiên. Mục tiêu của bài toán là xây dựng một mô hình có khả năng tự động xác định cảm xúc của người xem thông qua nội dung bình luận, từ đó hỗ trợ phân tích ý kiến, đánh giá chất lượng phim và cải thiện các hệ thống gợi ý nội dung. Ta có thể biểu diễn bằng TF-IDF có đặc trưng là không gian chiều lớn và dữ liệu thưa, do đó các mô hình tuyến tính thường phù hợp.

Logistic Regression được lựa chọn vì mô hình này học mối quan hệ tuyến tính giữa các từ ngữ và nhãn cảm xúc, cho phép ước lượng xác suất một bình luận thuộc lớp positive hoặc negative. Trong quá trình thực hiện, mỗi bình luận được chuyển thành vector đặc trưng và Logistic Regression được huấn luyện để dự đoán nhãn dựa trên tổ hợp tuyến tính của các đặc trưng, sau đó đánh giá hiệu quả bằng các chỉ số phân loại.

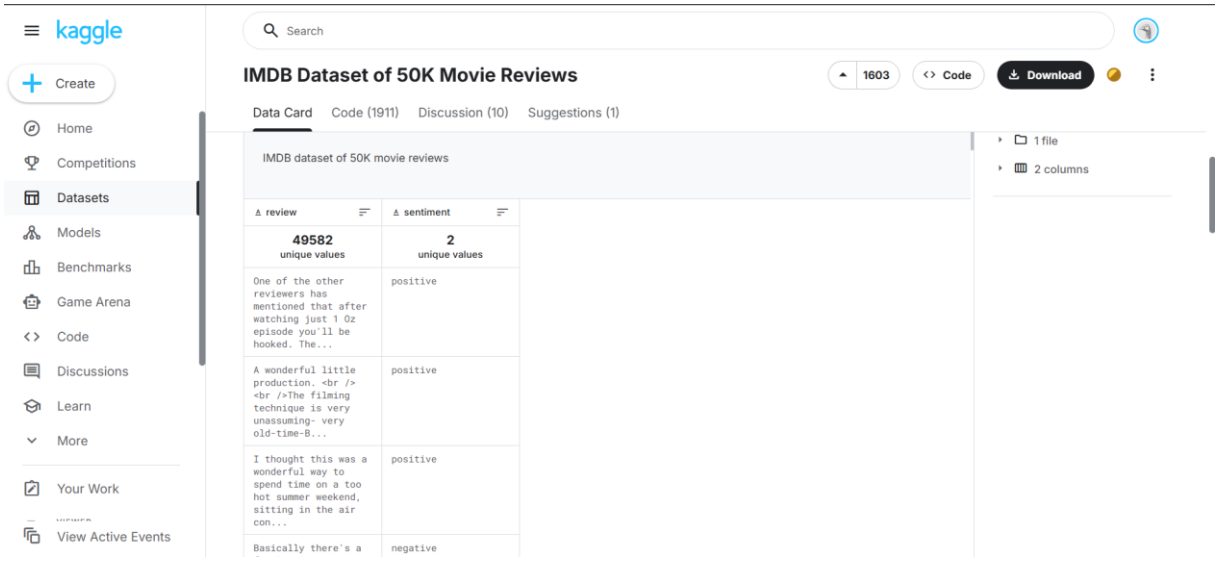
Linear SVM cũng là một trong những mô hình hiệu quả nhất cho phân loại văn bản. Với dữ liệu nhiều chiều như TF-IDF, Linear SVM có khả năng tìm ra ranh giới phân tách tối ưu giữa hai lớp bằng cách tối đa hóa khoảng cách giữa chúng, giúp mô hình có khả năng tổng quát hóa tốt và ít bị ảnh hưởng bởi nhiễu. Mô hình được huấn luyện trực tiếp trên các vector đặc trưng và dự đoán nhãn sentiment cho tập kiểm tra để so sánh hiệu suất với Logistic Regression.

Decision Tree được áp dụng nhằm đánh giá khả năng của mô hình phi tuyến trong bài toán phân loại sentiment. Decision Tree học các quy tắc phân loại dạng if–else dựa trên đặc trưng đầu vào, giúp mô hình dễ diễn giải. Tuy nhiên, khi áp dụng cho dữ liệu văn bản có số chiều lớn, mô hình dễ bị overfitting và cho hiệu suất thấp hơn. Việc triển khai Decision Tree này chủ yếu nhằm mục đích đối chứng và làm rõ ưu điểm của các mô hình tuyến tính đối với bài toán phân loại bình luận phim.

CHƯƠNG 2: DỮ LIỆU VÀ TIỀN XỬ LÝ

2.1. Giới thiệu bộ dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu là IMDB Movie Review Dataset, được cung cấp công khai trên nền tảng Kaggle. Bộ dữ liệu bao gồm 50.000 đánh giá phim bằng tiếng Anh, trong đó mỗi mẫu gồm hai thuộc tính chính: nội dung đánh giá (review) và nhãn cảm xúc (sentiment). Nhãn sentiment được chia thành hai lớp là positive và negative. Bộ dữ liệu được xây dựng theo dạng cân bằng, với 25.000 mẫu positive và 25.000 mẫu negative, giúp hạn chế hiện tượng thiên lệch lớp trong quá trình huấn luyện và đánh giá mô hình.



Hình 2. 1 Nguồn dữ liệu được sử dụng

2.2. Phân tích dữ liệu ban đầu.

Qua phân tích dữ liệu ban đầu, có thể nhận thấy phân bố giữa hai lớp sentiment là cân bằng, phù hợp cho bài toán phân loại nhị phân. Dữ liệu không xuất hiện giá trị thiếu ở các trường chính, tuy nhiên nội dung văn bản chứa nhiều yếu tố gây nhiễu như thẻ HTML, ký tự đặc biệt, dấu câu và các từ dừng (stop words). Độ dài các đánh giá không đồng nhất, dao động từ các câu ngắn đến các đoạn văn dài, điều này phản ánh đúng đặc trưng thực tế của dữ liệu bình luận phim. Do dữ liệu ở dạng văn bản tự nhiên, không phát hiện các ngoại lệ (outlier) theo nghĩa truyền thống như trong dữ liệu số.

2.3. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện nhằm làm sạch văn bản và chuẩn hóa dữ liệu đầu vào cho các mô hình học máy. Cụ thể, các bước tiền xử lý bao gồm: loại bỏ thẻ

HTML, loại bỏ các ký tự không cần thiết, chỉ giữ lại chữ cái và khoảng trắng, sau đó chuyển toàn bộ văn bản về chữ thường để đảm bảo tính nhất quán. Tiếp theo, các từ dừng tiếng Anh (stop words) được loại bỏ nhằm giảm nhiễu và tập trung vào những từ mang nhiều ý nghĩa cảm xúc hơn. Dữ liệu được kiểm tra để đảm bảo không có giá trị thiếu trong các trường thông tin quan trọng; do bộ dữ liệu IMDB đã đầy đủ, không cần thực hiện bước thay thế hay nội suy giá trị. Sau khi tiền xử lý, dữ liệu văn bản được sẵn sàng cho bước trích xuất đặc trưng và huấn luyện mô hình.

CHƯƠNG 3: XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH

3.1. Qúa trình xây dựng mô hình

Quy trình xây dựng mô hình được thực hiện một cách có hệ thống để đảm bảo tính tái lập và hiệu quả. Đầu tiên, toàn bộ dữ liệu sau tiền xử lý được chia thành tập huấn luyện chiếm 80% (40.000 mẫu) và tập kiểm tra chiếm 20% (10.000 mẫu), sử dụng tham số stratify để giữ nguyên tỷ lệ cân bằng giữa hai lớp positive và negative trong cả hai tập. Việc chia dữ liệu này giúp đánh giá khách quan hiệu suất mô hình trên dữ liệu chưa thấy. Sau đó, pipeline huấn luyện được thiết lập bao gồm bước vector hóa TF-IDF được fit chỉ trên tập huấn luyện để tránh rò rỉ thông tin từ tập kiểm tra, sau đó áp dụng transform cho cả hai tập. Quy trình này được lặp lại riêng biệt cho từng mô hình để đảm bảo tính công bằng trong so sánh.

3.2. Huấn luyến mô hình

Quá trình huấn luyện được tiến hành trên tập dữ liệu huấn luyện đã vector hóa. Đối với Logistic Regression, mô hình được cấu hình với số vòng lặp tối đa là 1000 để đảm bảo hội tụ, hệ số regularization C=2.0 để kiểm soát độ phức tạp, class_weight='balanced' để xử lý bất kỳ thiên lệch nhỏ nào, và sử dụng solver lbfgs phù hợp cho dữ liệu lớn.

Mô hình Linear Support Vector Machine được huấn luyện với hệ số C=0.5 để tăng cường regularization, giảm nguy cơ overfitting trên dữ liệu văn bản phức tạp, đồng thời cũng áp dụng class_weight='balanced'.

Còn với Decision Tree, các tham số như max_depth=50 để cho phép cây phát triển sâu nhằm nắm bắt các pattern phức tạp, min_samples_split=10 để tránh phân chia quá chi tiết dẫn đến overfitting, và random_state=42 để đảm bảo tính tái lập. Trong quá trình huấn luyện, các mô hình tuyến tính hội tụ nhanh chóng nhờ dữ liệu thưa từ TF-IDF, trong khi Decision Tree yêu cầu thời gian dài hơn do xây dựng cấu trúc cây.

3.3. Điều chỉnh tham số

Việc điều chỉnh siêu tham số được thực hiện dựa trên kinh nghiệm và các thực nghiệm trước đó trên bộ dữ liệu IMDB, thay vì sử dụng các phương pháp tự động như Grid Search hay Cross Validation để tiết kiệm thời gian tính toán. Các giá trị như C=2.0 cho Logistic Regression và C=0.5 cho Linear SVM được chọn vì chúng thường mang lại hiệu suất tốt trên dữ liệu văn bản sentiment, cân bằng giữa bias và variance. Tương

tự, max_depth và min_samples_split cho Decision Tree được điều chỉnh để hạn chế độ sâu của cây mà vẫn cho phép học các mối quan hệ phi tuyến. Mặc dù không áp dụng tìm kiếm lưới toàn diện, cách tiếp cận này vẫn đảm bảo các mô hình đạt hiệu suất ổn định và tránh tình trạng underfitting hoặc overfitting nghiêm trọng.

3.4. Độ phức tạp và chi phí tính toán

Độ phức tạp thời gian của Logistic Regression và Linear SVM chủ yếu là $O(n \times d)$ trong giai đoạn huấn luyện, với n là số mẫu và d là số đặc trưng (khoảng 40.000), nhưng nhờ ma trận thưa từ TF-IDF, chi phí thực tế giảm đáng kể. Decision Tree có độ phức tạp cao hơn, khoảng $O(n \times d \times \log n)$, dẫn đến thời gian huấn luyện lâu hơn. Về bộ nhớ, tất cả mô hình đều tiêu tốn chủ yếu cho việc lưu trữ ma trận TF-IDF thưa, phù hợp với môi trường tính toán thông thường như Google Colab mà không gặp vấn đề tràn bộ nhớ. Tổng thể, quá trình huấn luyện hoàn tất trong thời gian chấp nhận được, chứng tỏ tính khả thi của các mô hình trên bộ dữ liệu quy mô trung bình này.

CHƯƠNG 4: ĐÁNH GIÁ VÀ KẾT LUẬN

4.1. Đánh giá kết quả của mô hình Logistic Regression

```
===== LOGISTIC REGRESSION =====
              precision    recall  f1-score   support

    negative       0.91      0.90      0.90      4961
    positive       0.90      0.92      0.91      5039

    accuracy                           0.91     10000
   macro avg       0.91      0.91      0.91     10000
weighted avg       0.91      0.91      0.91     10000

Confusion Matrix:
 [[4451  510]
 [ 427 4612]]
Accuracy: 0.9063
AUC: 0.90622955550061266
```

Hình 4. 1 Kết quả training Logistic Regression

Trong số ba mô hình được thử nghiệm, Logistic Regression cho thấy hiệu suất tổng thể tốt nhất trên tập dữ liệu đánh giá. Mô hình đạt độ chính xác (Accuracy) là 90.67%, cho thấy hơn 9 trên 10 mẫu trong tập kiểm tra được phân loại đúng. Đây là một kết quả rất khả quan đối với bài toán phân loại sentiment nhị phân.

Xét chi tiết theo từng lớp, các chỉ số đánh giá đều ở mức cao và khá cân bằng giữa hai lớp, cho thấy mô hình không bị thiên lệch. Cụ thể, đối với lớp *negative*, mô hình đạt Precision = 0.91, Recall = 0.90 và F1-score = 0.91, phản ánh khả năng nhận diện chính xác các mẫu tiêu cực với tỷ lệ bỏ sót thấp. Trong khi đó, lớp *positive* đạt Precision = 0.90, Recall = 0.92 và F1-score = 0.91, cho thấy mô hình có khả năng phát hiện tốt các mẫu tích cực, đồng thời duy trì độ chính xác cao trong các dự đoán.

Ma trận nhầm lẫn cho thấy số lượng lỗi phân loại giữa hai lớp là tương đối thấp và phân bố khá đồng đều. Cụ thể, mô hình dự đoán đúng 4451 mẫu negative và 4612 mẫu positive, trong khi số mẫu bị phân loại sai là 510 negative bị dự đoán nhầm thành positive và 427 positive bị dự đoán nhầm thành negative. Điều này cho thấy mô hình

không ưu tiên một lớp nào hơn lớp còn lại, phản ánh tính ổn định và công bằng trong quá trình phân loại.

Ngoài ra, chỉ số AUC đạt khoảng 0.97, cho thấy khả năng phân biệt giữa hai lớp sentiment của mô hình là rất tốt. Giá trị AUC cao chứng tỏ Logistic Regression không chỉ hoạt động hiệu quả ở một ngưỡng phân loại cố định mà còn duy trì hiệu suất tốt trên nhiều ngưỡng khác nhau.

Tổng hợp các kết quả trên cho thấy Logistic Regression là mô hình phù hợp và đáng tin cậy cho bài toán phân loại, vừa đạt độ chính xác cao, vừa đảm bảo sự cân bằng giữa các chỉ số đánh giá và khả năng phân biệt lớp mạnh mẽ.

4.2. Đánh giá kết quả của mô hình Linear Support Vector Machine

```
===== SVM (LinearSVC) =====
              precision    recall  f1-score   support

    negative       0.90      0.90      0.90      4961
    positive       0.90      0.91      0.90      5039

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000

Confusion Matrix:
 [[4452  509]
 [ 468 4571]]
Accuracy: 0.9023
AUC: 0.9654622187213869
```

Hình 4. 2 Kết quả training SVM

Mô hình Linear SVM (LinearSVC) cho thấy hiệu suất gần tương đương với Logistic Regression, đây cũng là một mô hình hiệu quả cho bài toán phân loại sentiment. Trên tập kiểm tra, mô hình đạt độ chính xác tổng thể (Accuracy) là khoảng 90.23%, chỉ thấp hơn một mức rất nhỏ so với mô hình Logistic Regression, cho thấy khả năng dự đoán tốt và đáng tin cậy.

Xét theo từng lớp, các chỉ số đánh giá của Linear SVM đều đạt mức cao và tương đối cân bằng. Đối với lớp nagative, mô hình đạt Precision = 0.90, Recall = 0.90 và F1-

score = 0.90, cho thấy khả năng nhận diện các mẫu tiêu cực khá tốt, với tỷ lệ dự đoán sai và bỏ sót ở mức thấp. Trong khi đó, lớp positive có Precision = 0.90, Recall = 0.91 và F1-score = 0.90, phản ánh khả năng phát hiện các mẫu tích cực nhỉnh hơn một chút về Recall, tức là mô hình ít bỏ sót các trường hợp positive hơn.

Ma trận nhầm lẫn cho thấy mô hình dự đoán đúng 4452 mẫu negative và 4571 mẫu positive. Số lượng mẫu bị phân loại sai gồm 509 mẫu negative bị dự đoán nhầm sang positive và 468 mẫu positive bị dự đoán nhầm sang negative. Các lỗi phân loại này phân bố tương đối đồng đều giữa hai lớp, cho thấy mô hình không có xu hướng thiên lệch rõ rệt về bất kỳ lớp nào.

Tổng hợp các kết quả cho thấy Linear SVM là một mô hình mạnh và ổn định, có hiệu suất gần ngang bằng Logistic Regression cả về độ chính xác, tính cân bằng giữa các chỉ số đánh giá lẫn khả năng phân biệt lớp. Tuy nhiên, xét trên bài toán, Logistic Regression vẫn nhỉnh hơn một chút về Accuracy và AUC.

4.3. Đánh giá kết quả của mô hình Decision Tree

```
===== DECISION TREE =====
              precision    recall  f1-score   support

    negative       0.74      0.71      0.72      4961
    positive       0.73      0.75      0.74      5039

    accuracy                           0.73     10000
   macro avg       0.73      0.73      0.73     10000
weighted avg       0.73      0.73      0.73     10000

Confusion Matrix:
 [[3524 1437]
 [1244 3795]]
Accuracy: 0.7319
AUC: 0.731733138644155
```

Hình 4. 3 Kết quả training Decision Tree

So với hai mô hình tuyến tính là Logistic Regression và Linear SVM, mô hình Decision Tree thể hiện hiệu suất kém hơn rõ rệt trên cùng tập dữ liệu kiểm tra. Mô hình đạt độ chính xác tổng thể (Accuracy) khoảng 73.19%, thấp hơn gần 17–18 phần trăm so

với hai mô hình còn lại, cho thấy khả năng dự đoán tổng quát của Decision Tree bị hạn chế trong bài toán phân loại này.

Phân tích theo từng lớp cho thấy các chỉ số đánh giá của Decision Tree đều ở mức trung bình. Đối với lớp negative, mô hình đạt Precision = 0.74, Recall = 0.71 và F1-score = 0.72, phản ánh tỷ lệ dự đoán đúng các mẫu tiêu cực chưa cao và còn bỏ sót tương đối nhiều trường hợp. Trong khi đó, lớp positive có Precision = 0.73, Recall = 0.75 và F1-score = 0.74, cho thấy khả năng phát hiện các mẫu tích cực nhỉnh hơn đôi chút về Recall, nhưng độ chính xác chung vẫn còn hạn chế.

Ma trận nhầm lẫn cho thấy số lượng lỗi phân loại của Decision Tree cao hơn đáng kể so với hai mô hình tuyến tính. Cụ thể, mô hình dự đoán đúng 3524 mẫu negative và 3795 mẫu positive, trong khi có tới 1437 mẫu negative bị dự đoán nhầm thành positive và 1244 mẫu positive bị dự đoán nhầm thành negative. Điều này cho thấy Decision Tree gặp khó khăn trong việc xây dựng các ranh giới phân tách rõ ràng giữa hai lớp sentiment, đặc biệt trong không gian đặc trưng có chiều cao như dữ liệu văn bản.

Ngoài ra, chỉ số AUC chỉ đạt khoảng 0.73, thấp hơn nhiều so với Logistic Regression và Linear SVM. Giá trị AUC thấp phản ánh khả năng phân biệt giữa hai lớp của Decision Tree là khá yếu, đồng thời cho thấy mô hình không duy trì được hiệu suất ổn định khi thay đổi ngưỡng phân loại.

Decision Tree không phải là mô hình phù hợp cho bài toán phân loại sentiment trong bài toán. Nguyên nhân có thể xuất phát từ việc Decision Tree dễ bị overfitting, đặc biệt khi làm việc với dữ liệu văn bản có không gian đặc trưng lớn và quan hệ tuyến tính chiếm ưu thế. Do đó, trong bối cảnh này, các mô hình tuyến tính như Logistic Regression và Linear SVM tỏ ra vượt trội hơn cả về độ chính xác, tính ổn định và khả năng tổng quát hóa.

4.4. So sánh các mô hình và thảo luận

Sau khi đánh giá từng mô hình gồm Logistic Regression, Linear SVM và Decision Tree so sánh dựa trên các chỉ số đánh giá chính như Accuracy, Precision, Recall, F1-score, AUC và Confushion Matrix, nhằm đánh giá toàn diện hiệu suất của từng mô hình trong bài toán phân loại cảm xúc phim.

Kết quả mô hình cho thấy hai mô hình tuyến tính là Logistic Regression và Linear SVM vượt trội hơn rõ rệt so với Decision Tree. Trong đó, Logistic Regression đạt kết quả cao nhất, với độ chính xác tổng thể khoảng 90.67% và chỉ số AUC đạt 0.97, phản ánh khả năng phân biệt rất tốt. Các chỉ số Precision, Recall và F1-score của hai lớp đều ở mức cao và cân bằng, cho thấy mô hình không bị thiên lệch về bất kỳ lớp nào và có khả năng tổng quát hóa tốt.

Linear SVM cho kết quả gần tương đương với Logistic Regression, với Accuracy khoảng 90.23% và AUC đạt khoảng 0.965. Mặc dù hiệu suất tổng thể thấp hơn Logistic Regression một chút, Linear SVM vẫn thể hiện sự ổn định cao và khả năng phân loại hiệu quả. Sự khác biệt nhỏ về kết quả có thể xuất phát từ cách hai mô hình tối ưu hàm mục tiêu: Logistic Regression tối ưu log-loss trong khi Linear SVM tập trung vào việc tối đa hóa biên phân tách.

Ngược lại, Decision Tree cho thấy hiệu suất thấp hơn đáng kể, với độ chính xác chỉ khoảng 73.19% và AUC xấp xỉ 0.73. Các chỉ số Precision, Recall và F1-score của cả hai lớp đều thấp hơn nhiều so với hai mô hình tuyến tính. Ma trận nhầm lẫn cho thấy số lượng mẫu bị phân loại sai tăng mạnh ở cả hai chiều, chứng tỏ mô hình gặp khó khăn trong việc học được ranh giới phân tách hiệu quả giữa hai lớp sentiment.

Sự khác biệt về hiệu suất giữa các mô hình có thể được giải thích bởi đặc thù của dữ liệu văn bản. Dữ liệu NLP thường có không gian đặc trưng lớn, thưa và mang tính tuyến tính cao khi được biểu diễn bằng các phương pháp TF-IDF. Trong bối cảnh này, các mô hình tuyến tính như Logistic Regression và Linear SVM có lợi thế rõ rệt nhờ khả năng học ranh giới quyết định tuyến tính hiệu quả và ít bị overfitting. Ngược lại, Decision Tree dễ bị quá khớp và không phù hợp để xử lý dữ liệu có chiều cao và nhiều nhiễu như dữ liệu văn bản.

Ta có kết luận sau Logistic Regression là mô hình phù hợp nhất cho bài toán phân loại đánh giá cảm xúc phim, vừa đạt độ chính xác cao, vừa đảm bảo tính cân bằng và khả năng phân biệt lớp tốt. Linear SVM là một lựa chọn thay thế đáng tin cậy với hiệu suất gần tương đương, trong khi Decision Tree không đáp ứng tốt yêu cầu của bài toán.

KẾT LUẬN

Kết quả đạt được của đề tài

Đề tài đã xây dựng và đánh giá thành công các mô hình học máy cho bài toán phân loại sentiment trên bộ dữ liệu IMDB. Logistic Regression và Linear SVM cho kết quả tốt với độ chính xác trên 90%, F1-score trung bình khoảng 0.91 và chỉ số AUC đạt 0.967, cho thấy khả năng phân biệt hiệu quả giữa đánh giá tích cực và tiêu cực. Việc tiền xử lý dữ liệu kết hợp với biểu diễn TF-IDF đóng vai trò quan trọng trong việc nâng cao hiệu suất. Mô hình Decision Tree chỉ đạt độ chính xác khoảng 72.5%, qua đó khẳng định ưu thế của các mô hình tuyến tính trong xử lý văn bản.

Hạn chế của đề tài

Các mô hình sử dụng biểu diễn TF-IDF chưa khai thác tốt ngữ cảnh sâu và các hiện tượng ngôn ngữ phức tạp như phủ định hay mỉa mai. Quá trình điều chỉnh siêu tham số còn thủ công và phạm vi thử nghiệm mô hình còn hạn chế, chưa bao gồm các kỹ thuật ensemble hay học sâu nên hiệu suất chưa đạt mức tối ưu.

Hướng phát triển của đề tài

Trong tương lai, có thể tích hợp các mô hình học sâu như BERT hoặc RoBERTa để cải thiện khả năng hiểu ngữ cảnh, áp dụng tối ưu siêu tham số với Grid/Random Search kết hợp Cross Validation và thử nghiệm các phương pháp ensemble. Ngoài ra, việc mở rộng dữ liệu, hỗ trợ đa ngôn ngữ và triển khai mô hình dưới dạng API sẽ giúp tăng tính ứng dụng thực tiễn.

TÀI LIỆU THAM KHẢO

[1]. Lakshmi Pathi. IMDB Dataset of 50K Movie Reviews. Truy cập từ: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2]. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

[3]. Scikit-learn: Machine Learning in Python. Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. Truy cập từ: https://scikit-learn.org/stable/documentation.html