**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**Báo cáo tổng kết**

**ĐỀ TÀI: Phân tích đánh giá sản phẩm trên nền tảng thương mại điện tử bằng Machine Learning.**

**Chủ nhiệm đề tài : 3121560024 – Nguyễn Quang Dương**

**Các thành viên tham gia :**

**3121560024 – Nguyễn Quang Dương**

**3123410151 – Nguyễn Dương Khang**

**3124720066 – Nguyễn Gia Thành**

**Môn : Phương pháp luận nghiên cứu khoa học.**

**Giảng viên hướng dẫn** : **Đỗ Như Tài.**

**TPHCM, Ngày 18 tháng 5 năm 2025.**

# Mục lục

[Mục lục 1](#_Toc198667065)

[Mục lục hình 4](#_Toc198667066)

[Mục lục bảng 5](#_Toc198667067)

[Tóm Tắt 6](#_Toc198667068)

[Chương 1 : Tổng quan vấn đề. 8](#_Toc198667069)

[1.1. Lý do chọn đề tài. 8](#_Toc198667070)

[1.2. Vấn đề nghiên cứu. 8](#_Toc198667071)

[1.3. Mục tiêu nghiên cứu. 9](#_Toc198667072)

[1.4. Câu hỏi nghiên cứu. 9](#_Toc198667073)

[1.5. Phạm vi nghiên cứu. 9](#_Toc198667074)

[Chương 2 : Lược khảo tài liệu 11](#_Toc198667075)

[2.1. Nghiên cứu liên quan. 11](#_Toc198667076)

[2.2. Cơ sở lý thuyết. 13](#_Toc198667077)

[2.2.1. Tổng quan về phân tích cảm xúc. 13](#_Toc198667078)

[2.2.2. Các phương pháp được dùng trong phân tích cảm xúc. 14](#_Toc198667079)

[a. Mô hình học máy truyền thống (Machine Learning). 14](#_Toc198667080)

[b. Mô hình học sâu (Deep Learning). 14](#_Toc198667081)

[c. Bảng so sánh. 15](#_Toc198667082)

[2.2.3 Mô hình LSTM (Long Short-Term Memory). 16](#_Toc198667083)

[a. Cấu trúc của LSTM : 16](#_Toc198667084)

[b. Cở sở toán học 18](#_Toc198667085)

[2.2.4. Mô hình Naive Bayes. 19](#_Toc198667086)

[a. Cấu trúc của mô hình Naive Bayes. 19](#_Toc198667087)

[b. Cơ sở toán học. 20](#_Toc198667088)

[2.2.5. Mô hình Logistic Regression. 21](#_Toc198667089)

[a. Cấu trúc cơ bản. 22](#_Toc198667090)

[b. Cơ sở toán học. 22](#_Toc198667091)

[2.2.6. Thuật toán Random Forest. 23](#_Toc198667092)

[a. Cấu trúc của Random Forest. 24](#_Toc198667093)

[2.2.7. Thông số đánh giá mô hình. 25](#_Toc198667094)

[2.3. Phân tích các nghiên cứu trước và định hướng tính kế thừa, phát triển. 26](#_Toc198667095)

[2.3.1. Điểm mạnh của cá nghiên cứu trước. 26](#_Toc198667096)

[a. Mô hình LSTM. 26](#_Toc198667097)

[b. Mô hình Naive Bayes. 27](#_Toc198667098)

[c. Mô hình Logistic Regression. 27](#_Toc198667099)

[d. Thuật toán Random Forest. 28](#_Toc198667100)

[2.3.2. Hạn chế của nghiên cứu. 29](#_Toc198667101)

[a. Mô hình LSTM. 29](#_Toc198667102)

[b. Mô hình Naive Bayes. 30](#_Toc198667103)

[c. Mô hính Logistic Regression. 30](#_Toc198667104)

[d. Thuật toán Random Forest. 31](#_Toc198667105)

[2.3.3. Kế thừa và phát triển. 32](#_Toc198667106)

[Chương 3 : Phương pháp nghiên cứu 33](#_Toc198667107)

[3.1. Thiết kế nghiên cứu. 33](#_Toc198667108)

[3.2. Đối tượng và mẫu nghiên cứu. 33](#_Toc198667109)

[3.3. Phương pháp thu thập và tiền xử lý dữ liệu. 34](#_Toc198667110)

[3.3.1. Thu thập dữ liệu. 34](#_Toc198667111)

[3.3.2. Tiền xử lý. 34](#_Toc198667112)

[a. Gán nhãn cảm xúc : 34](#_Toc198667113)

[b. Làm sạch văn bản : 34](#_Toc198667114)

[c. Khám phá dữ liệu (EDA) : 35](#_Toc198667115)

[d. Xử lý mất cân bằng dữ liệu: 38](#_Toc198667116)

[e. Biến đổi văn bản thành vector: 38](#_Toc198667117)

[3.3.3. Các công cụ hỗ trợ. 38](#_Toc198667118)

[a. Phần mềm hỗ trợ. 38](#_Toc198667119)

[b. Thư viện hỗ trợ. 38](#_Toc198667120)

[Chương 4 : Thực nghiệm và thảo luận. 40](#_Toc198667121)

[4.1. Thiết lập thực nghiệm. 40](#_Toc198667122)

[4.2. Kết quả mô hình. 40](#_Toc198667123)

[4.2.1. Đánh giá mô hình. 40](#_Toc198667124)

[4.2.2. Biểu đồ mô hình. 43](#_Toc198667125)

[4.3. So sánh với các nghiên cứu trước. 47](#_Toc198667126)

[4.4. Ý nghĩa thực tiễn. 49](#_Toc198667127)

[4.5. Hạn chế và kế hoạch phát triển cho nghiên cứu tiếp theo. 50](#_Toc198667128)

[Chương 5 : Kết luận và hướng phát triển. 51](#_Toc198667129)

[5.1. Kết luận. 51](#_Toc198667130)

[5.2. Câu hỏi nghiên cứu. 52](#_Toc198667131)

[5.3. Hạn chế của nghiên cứu hiện tại. 54](#_Toc198667132)

[5.4. Hướng phát triển. 55](#_Toc198667133)

[Tài liệu tham khảo 56](#_Toc198667134)

# Mục lục hình

[Hình 1. Cấu trúc của LSTM 17](#_Toc198667197)

[Hình 2. Xử lý chuỗi của LSTM 17](#_Toc198667198)

[Hình 3. Công thức tỏng quát cho forward pass của LSTM 18](#_Toc198667199)

[Hình 4. Cấu trúc của mô hình Naive Bayes ( Multinomial Naive Bayes ). 19](#_Toc198667200)

[Hình 5. Mô hình hồi quy Logistic. 22](#_Toc198667201)

[Hình 6. Hàm dư đoán xác suất. 23](#_Toc198667202)

[Hình 7. Hàm mất mát. 23](#_Toc198667203)

[Hình 8. Cấu trúc Random Forest. 24](#_Toc198667204)

[Hình 9. Công thức dự đoán (bài toán phân loại) của Random Forest. 25](#_Toc198667205)

[Hình 10.Công thức dự đoán (bài toán hồi quy) của Random Forest. 25](#_Toc198667206)

[Hình 11. Biểu đồ phân bố số lượng đánh giá theo nhãn 35](#_Toc198667207)

[Hình 12. Ma trận tương quan giữa số sao, độ dài văn bản và số lượng từ. 35](#_Toc198667208)

[Hình 13. Ma trận phân tán. 37](#_Toc198667209)

[Hình 14. So sánh độ chính xác của 4 mô hình. 43](#_Toc198667210)

[Hình 15. Confusion Matrix của 4 mô hình. 44](#_Toc198667211)

# Mục lục bảng

[Bảng 1. So sánh ưu nhược điểm các mô hình học sâu. 15](#_Toc198667212)

[Bảng 2. So sánh mô hình học máy và học sâu trong phân tích cảm xúc. 16](#_Toc198667213)

[Bảng 3. So sánh kết quả của 4 mô hình. 41](#_Toc198667214)

[Bảng 4. Đánh giấ hiệu suất tổng quát của 4 mô hình. 41](#_Toc198667215)

[Bảng 5. Tổng kết 4 mô hình. 43](#_Toc198667216)

# Tóm Tắt

Bài nghiên cứu về đề tài “Phân tích đánh giá sản phẩm trên nền tảng thương mại điện tử bằng Machine Learning” nhằm mục tiêu khảo sát và so sánh hiệu quả của các mô hình học máy và học sâu trong bài toán phân loại cảm xúc văn bản.

Mở đầu, nhóm cung cấp một số bài báo nghiên cứu liên quan đến phương pháp xử lý và tiến hành, rút ra ý chính từ những bài báo đó, đặt ra câu hỏi nghiên cứu. Sau đó nghiên cứu trình bày khái niệm phân tích cảm xúc và các phương pháp phổ biến trong việc xử lý và phân tích văn bản dạng chuỗi. Tiếp theo, bài nghiên cứu giới thiệu tổng quan về các mô hình học máy và học sâu – với học sâu được xem là một bước tiến vượt bậc dựa trên nền tảng của học máy, giúp cải thiện hiệu suất chất lượng ouput.

Từ đó, nhóm lựa chọn ba mô hình học máy tiêu biểu gồm: Random Forest, Naive Bayes, Logistic Regression, cùng với một mô hình học sâu cơ bản là LSTM để tiến hành thực nghiệm. Mỗi mô hình được trình bày với phần mô tả về cấu trúc, nguyên lý hoạt động và cơ sở toán học (có hình ảnh minh họa).

Sau phần lý thuyết, nghiên cứu tiến hành thu thập dữ liệu từ tập Amazon Fine Food Reviews, thực hiện các bước tiền xử lý dữ liệu như làm sạch văn bản, gán nhãn cảm xúc, vector hóa dữ liệu đầu vào bằng TF-IDF hoặc tokenizer, lemmatization.

Sau khi xử lý xong dữ liệu, nhóm bắt đầu quá trình huấn luyện và đánh giá mô hình, từ đó rút ra kết luận, lập bảng về hiệu suất của từng mô hình trong việc phân loại đánh giá sản phẩm thành ba nhóm cảm xúc: tích cực, trung tính và tiêu cực.

Và cuối cùng bài nghiên cứu sẽ nêu ra cho bạn biết hạn chế mà nghiên cứu còn vướng phải và hướng khắc phục, phát triển trong tương lai. Đồng thời đưa ra câu trả lời cho các câu hỏi nghiên cứu trước đó.

# Chương 1 : Tổng quan vấn đề.

## 1.1. Lý do chọn đề tài.

Trong kỷ nguyên số, người tiêu dùng luôn tiến tới các phương thức tiện lợi, đơn giản nhưng hiệu quả cao và phương thức mua hàng trực tuyến là phổ biến nhất, chính vì lẻ đó xu hướng tham khảo các đánh giá trực tuyến trước khi quyết định mua hàng. Những đánh giá này phản ánh mức độ cảm nhận thực tế của người dùng, đồng thời cũng là nguồn dữ liệu quan trọng giúp cải thiện sản phẩm và dịch vụ chăm sóc khách hàng. Do đó, việc phân tích đánh giá sản phẩm bằng kỹ thuật xử lý ngôn ngữ tự nhiên kết hợp với nhiều mô hình học máy, học sâu là cần thiết.

## 1.2. Vấn đề nghiên cứu.

Trong bối cảnh thương mại điện tử phát triển, 1 sản phẩm có hàng ngàn đánh giá từ người dùng, viẹc hiểu được cảm xúc, mức độ hài lòng hay phàn nàn cảu khách hàng từ đoạn văn tự do (text) là một thách thức to lớn.Đa số đánh giá là phi cấu trúc, chứa nhiều từ ngữ chưa được chuẩn hóa, các ký tự đặc biệt (biểu tượng cảm xúc, các dấu như @,#,!,...).

Từ đó nảy sinh vấn đề : Người mua khó khăn trong việc lựa chọn sản phẩm dựa vào việc phải đọc qua hàng ngàn đánh giá, người bán không thể biết được có bao nhiêu đánh giá hài lòng với sản phẩm đã mua để có thể cải thiện dịch vụ 🡪 Cần công cụ tự động phân tích xử lý dữ liệu lớn, phân loại cảm xúc của văn bản, chia nhận xét về sản phẩm thành 3 lớp (positive, negative, neural).

## 1.3. Mục tiêu nghiên cứu.

Xây dựng mô hình xử lý, phân tích, đánh giá cảm xúc từ các bình luận sản phẩm trên nền tảng thương mại điện tử dựa vào tệp dữ liệu Amazon Fine Food Reviews (Kaggle).

## 1.4. Câu hỏi nghiên cứu.

Các đặc trưng ngôn ngữ nào thường xuất hiện trong đánh giá tích cực/tiêu cực/trung tính ?

Mô hình học máy nào có thể cho ra hiệu quả tốt nhất trong việc phân loại đánh giá sản phẩm ?

Có thể thay thế mô hình học máy tốt nhát sau nghiên cứu bằng mô hình học sâu để giải quyết vấn đề hay không và có đạt được hiệu quả như mô hình học máy ?

## 1.5. Phạm vi nghiên cứu.

Đối tượng nghiên cứu : Các đánh giá sản phẩm được người dùng đăng tải trên nền tảng thương mại điện tử Amazon được tổng hợp thành file Amazon Food Reviews trên Kaggle.

Phạm vi nghiên cứu :

Không gian : Dữ liệu từ tập tin csv trên Kaggle.

Thời gian : Các đánh giá trong năm 2012.

Mô hình nghiên cứu :

Sử dụng mô hình học máy (Logitic Regression, Naive Bayes, Random Forest)

Tiền xử lý dữ liệu, mã hóa nhãn thành số và one-hot, cân bằng dữ liệu(oversampling).

Đưa ra các thông số kết luận accuracy tổng thể và f1-score, recall, precision của từng loại nhãn (positive, neural, negative)

Sử dụng mô hình học sâu (LSTM)

Dán nhãn thông qua text (positive, negative, neural)

Dán nhãn cho từng đánh giá thông qua score

Dữ liệu ban đầu

Lọc dữ liệu (năm 2012)

# Chương 2 : Lược khảo tài liệu

## 2.1. Nghiên cứu liên quan.

Trong những năm gần đây, do ngành thương mại điện tử nổi dậy khá mạnh mẽ, một số trang thương mại nổi tiếng như Amazon, Shopee,...

Và từ đó các vấn đề nghiên cứu xoay quanh chủ đề này dần nổ ra, một vấn đề khá được quan tâm và chủ trọng là việc phân tích cảm xúc (sentiment analysis) trên nền tảng thương mại điện tử. Nhiều công trình được tạo ra với mục đích tổng hợp ý kiến khách hàng về các mặt hàng xuất hiện trên sàn nhắm hỗ trợ cho các doanh nghiệp cải thiện chất lượng sản phẩm và chiến lược marketing để tối ưu lượng hàng hóa tiêu thụ.

Một số bài báo khá tiêu biểu :

“ *Text based Sentiment Analysis using LSTM* ” của *Dr. G. S. N. Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi và Mounika Balusonti.*[1]

Kết luận chính : LSTM có độ chính xác vượt trội hơn các mô hình truyền thống như Naive Bayes, SVM.

*"On the Role of Text Preprocessing in Neural Network Architectures"* của *Jose Camacho-Collados và Mohammad Taher Pilehvar.*[2]

Kết luận chính :

Bài báo tập trung đánh giá vai trò của các kỹ thuật **tiền xử lý văn bản** trong hiệu quả của các mô hình học sâu như CNN, LSTM, và các mô hình embedding hiện đại (như fastText). Tác giả đặt ra câu hỏi liệu các bước tiền xử lý truyền thống như :

* **Tokenization** (tách từ).
* **Lemmatization/Stemming** (chuẩn hóa từ).
* **Lowercasing** (chuyển về chữ thường).
* **Loại bỏ stopwords, punctuation.**

có còn cần thiết trong bối cảnh các mô hình hiện đại có khả năng học biểu diễn ngữ nghĩa mạnh mẽ.

*“Sentiment Analysis using Logistic Regression”* của *George B. Aliman và cộng sự*.[3]

Kết luận chính :

**Logistic Regression là một mô hình hiệu quả và đơn giản** cho bài toán phân tích cảm xúc văn bản, đặc biệt trong các tình huống mà tập dữ liệu không quá lớn hoặc không có yêu cầu cao về độ phức tạp mô hình.

Trong thí nghiệm trên dữ liệu cảm xúc (Twitter), Logistic Regression đã đạt được độ chính xác tương đối cao, dễ huấn luyện, và tối ưu tốt khi kết hợp với các kỹ thuật xử lý ngôn ngữ như Bag-of-Words hoặc TF-IDF.

Mặc dù không mạnh bằng các mô hình deep learning hiện đại như LSTM hay BERT, Logistic Regression vẫn là một baseline tốt và đáng tin cậy, đặc biệt với khối lượng dữ liệu hạn chế và yêu cầu tính diễn giải cao.

*“Multinomial Naive Bayes Classification Model for Sentiment Analysis”* của *Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon và Anees Ahmed.*[4]

Kết luận chính :

**Thuật toán Multinomial Naive Bayes (MNB)** hoạt động hiệu quả và đáng tin cậy trong nhiệm vụ phân loại cảm xúc văn bản. Với dữ liệu đầu vào được xử lý bằng kỹ thuật Bag-of-Words, MNB đạt độ chính xác cao trong việc phân loại các câu đánh giá thành các nhóm cảm xúc tích cực, tiêu cực và trung lập. Bài báo khẳng định rằng MNB là một phương pháp đơn giản nhưng mạnh mẽ, phù hợp cho các hệ thống phân tích cảm xúc trong môi trường có tài nguyên tính toán hạn chế hoặc cần triển khai nhanh.

*"Sentiment Analysis for Film Reviews Based on Random Forest"* của *tạp chí Dean Francis Press.*[5]

Kết luận chính :

**Random Forest** chứng minh khả năng phân loại mạnh mẽ nhờ đặc điểm kết hợp nhiều cây quyết định và giảm thiểu overfitting, giúp cải thiện độ ổn định và độ tin cậy của dự đoán.

So với các thuật toán khác như Naive Bayes hoặc Logistic Regression, Random Forest có độ chính xác cao hơn và khả năng tổng quát hóa tốt hơn trong bài toán phân tích cảm xúc văn bản.

Kết quả của nghiên cứu cho thấy Random Forest là lựa chọn khả thi và hiệu quả cho các ứng dụng phân tích cảm xúc trong lĩnh vực phim ảnh hoặc các nội dung đánh giá từ người dùng.

## 2.2. Cơ sở lý thuyết.

### 2.2.1. Tổng quan về phân tích cảm xúc.

Phân tích cảm xúc là một ứng dụng khá phổ biến của xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định và phân loại cảm xúc, thái độ người dùng thông qua văn bản (tích cực, tiêu cực, trung tính). Úng dụng thực tiễn trong các tác vụ như phân tích, thống kê phản hồi về sản phẩm, đánh giá chất lượng sản phẩm, và trên mạng xã hội.

### 2.2.2. Các phương pháp được dùng trong phân tích cảm xúc.

#### a. Mô hình học máy truyền thống (Machine Learning).

Sử dụng các mô hình truyền thống như Naive Bayes (biến thể Multional Naive Bayes), Random Forest, Logistic Regression,... kết hợp sử dụng TF-IDF để vecto hóa dữ liệu.

Ưu điểm của phương pháp này là phân tích hiệu quả hơn phương pháp lexicon-based. Bên cạnh đó, các mô hình học máy truyền thống này thường cần trích suất đặc trưng thủ công, và không thể hiểu quá sâu được ngữ cảnh của câu.

#### b. Mô hình học sâu (Deep Learning).

Để cải thiện và khắc phục các hạn chế của mô hình học máy truyền thống, người ta sử dụng các mô hình học sâu để có thể khai thác ngữ cảnh và tự động học các đăc trưng dữ liệu tốt hơn. Một số mô hình học sâu khá phổ biến như : Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Transfomer,... Ưu điẻm của mô hình học sâu : tự động học các đặc trưng tốt (Không cần trích xuất đặc trưng thủ công), tự học hỏi các đặc trưng phức tạo, phi tuyến tính từ dữ liệu, đọ chính xác cao khi xử lý với lượng dữ liệu lớn, khả năng nắm bắt mối quan hệ ngữ cảnh giữa các từ trong câu. Với ưu điểm xử lý và đưa ra độ chính xác cao từ dữ liệu cũng đồng nghĩa với việc chúng sẽ tiêu tốn khá nhiều tài nguyên, nếu dữ liệu quá nhỏ, nhóm mô hình này dễ bị overfitting, cần nhiều kỹ năng và kiến thức chuyên sâu để tối ưu và xử lý dữ liệu.

|  |  |  |
| --- | --- | --- |
| **Mô hình** | **Ưu điểm chính** | **Nhược điểm** |
| **CNN** | Tốt trong trích xuất đặc trưng cục bộ (n-grams), nhanh | Không xử lý tốt ngữ cảnh dài |
| **RNN / LSTM / GRU** | Ghi nhớ chuỗi, xử lý tuần tự tốt | Chậm, khó song song, dễ mất ngữ cảnh xa |
| **Transformer** | Xử lý song song, học ngữ cảnh xa tốt | Cần nhiều tài nguyên |
| **BERT** | Biểu diễn ngữ nghĩa hai chiều mạnh mẽ | Cần fine-tuning, khá nặng |
| **GPT** | Tạo văn bản tự nhiên, tốt cho nhiệm vụ sinh văn bản | Thiên về sinh hơn là phân loại |

Bảng 1. So sánh ưu nhược điểm các mô hình học sâu.

#### c. Bảng so sánh.

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Học máy truyền thống (Machine Learning)** | **Học sâu (Deep Learning)** |
| **Cách hoạt động** | Dựa vào đặc trưng trích xuất thủ công từ dữ liệu (feature engineering) | Tự động học đặc trưng từ dữ liệu thô |
| **Trích xuất đặc trưng** | Phải thiết kế thủ công (TF-IDF, Bag-of-Words, v.v.) | Mô hình tự học đặc trưng thông qua mạng nơ-ron |
| **Khả năng xử lý ngữ cảnh** | Kém – khó nắm được ngữ cảnh câu dài hoặc cấu trúc ngôn ngữ phức tạp | Tốt – đặc biệt với LSTM, Transformer, BERT |
| **Yêu cầu dữ liệu** | Có thể hoạt động tốt với tập dữ liệu nhỏ | Cần dữ liệu lớn để đạt hiệu quả cao |
| **Tài nguyên tính toán** | Ít tài nguyên, huấn luyện nhanh | Cần GPU/TPU, tốn bộ nhớ và thời gian huấn luyện |
| **Hiệu suất** | Tốt với bài toán đơn giản, dữ liệu nhỏ | Tốt hơn trong bài toán phức tạp, dữ liệu lớn |
| **Giải thích mô hình (Interpretability)** | Dễ hiểu, có thể giải thích quyết định mô hình | Khó giải thích (mô hình “hộp đen”) |
| **Dễ triển khai** | Dễ triển khai, ít phức tạp | Cần kiến thức chuyên sâu, cấu hình phức tạp |
| **Ví dụ mô hình** | Naive Bayes, Logistic Regression, SVM, Random Forest | CNN, RNN, LSTM, GRU, Transformer, BERT, GPT |

Bảng 2. So sánh mô hình học máy và học sâu trong phân tích cảm xúc.

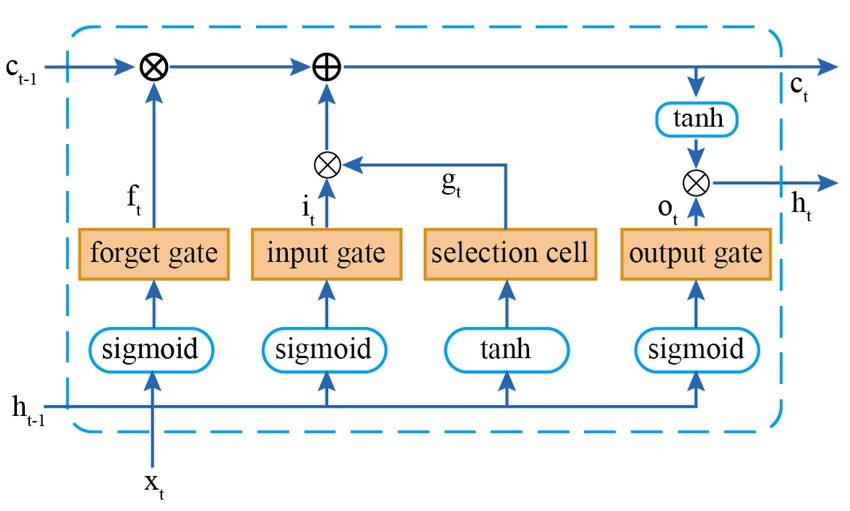
### 2.2.3 Mô hình LSTM (Long Short-Term Memory).

**LSTM** là một biến thể của mô hình **RNN (Recurrent Neural Network)** được tạo ra để khắc phục hiện tượng *vanishing gradient* [[1]](#footnote-1) khi xử lý chuỗi dài như văn bản, âm thanh,...

#### Cấu trúc của LSTM :

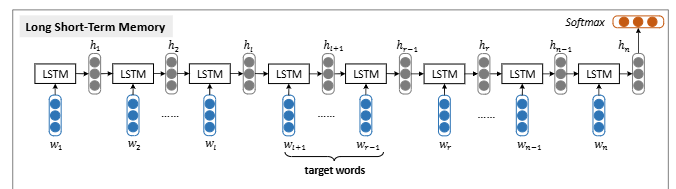
LSTM bao gồm ba cổng chính :

* **Cổng quên (Forget Gate)**: quyết định thông tin nào cần loại bỏ khỏi trạng thái trước.
* **Cổng đầu vào (Input Gate)**: quyết định thông tin mới nào sẽ được thêm vào.
* **Cổng đầu ra (Output Gate)**: quyết định giá trị đầu ra dựa trên trạng thái hiện tại.



Hình . Cấu trúc của LSTM

Quy trình xử lý dữ liệu chuỗi của LSTM :



Hình . Xử lý chuỗi của LSTM

Ưu điểm của LSTM trong phân tích cảm xúc :

* Ghi nhớ thông tin lâu hơn RNN thông thường → hiểu được ngữ cảnh dài trong câu.
* Phù hợp với dữ liệu chuỗi như văn bản, đặc biệt khi các từ có phụ thuộc ngữ nghĩa xa nhau.
* Hiệu quả cao hơn so với các mô hình học máy truyền thống.

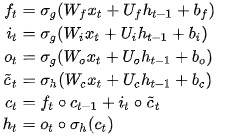
Hạn chế :

* Tốc độ huấn luyện chậm hơn mô hình truyền thống.
* Cần nhiều dữ liệu hơn để hoạt động tốt.
* Khó tối ưu khi không có kinh nghiệm.

#### Cở sở toán học

Long Short-Term Memory (LSTM) là một kiến trúc mạng nơ-ron hồi tiếp (RNN) được thiết kế nhằm khắc phục vấn đề mất mát thông tin dài hạn (vanishing gradient) trong các mạng RNN truyền thống. Một điểm nổi bật của LSTM là sự xuất hiện của các cổng (gates) để kiểm soát dòng thông tin đi qua mạng, trong đó forget gate đóng vai trò quan trọng trong việc quyết định phần thông tin nào trong trạng thái cell cần được giữ lại hay loại bỏ tại mỗi bước thời gian.

Công thức tổng quát cho quá trình tính toán bước tiến (forward pass) của một đơn vị LSTM có forget gate được viết gọn như sau:



Hình . Công thức tỏng quát cho forward pass của LSTM

* σ là hàm sigmoid, tạo ra các giá trị trong khoảng [0,1] dùng để điều khiển luồng thông tin qua các cổng.
* tanh là hàm hyperbolic tangent, giúp chuẩn hóa dữ liệu đầu vào.
* ⊙ là phép nhân Hadamard (phép nhân từng phần tử).
* W∗,U∗, b∗​ là các ma trận trọng số và vector bias được học trong quá trình huấn luyện.

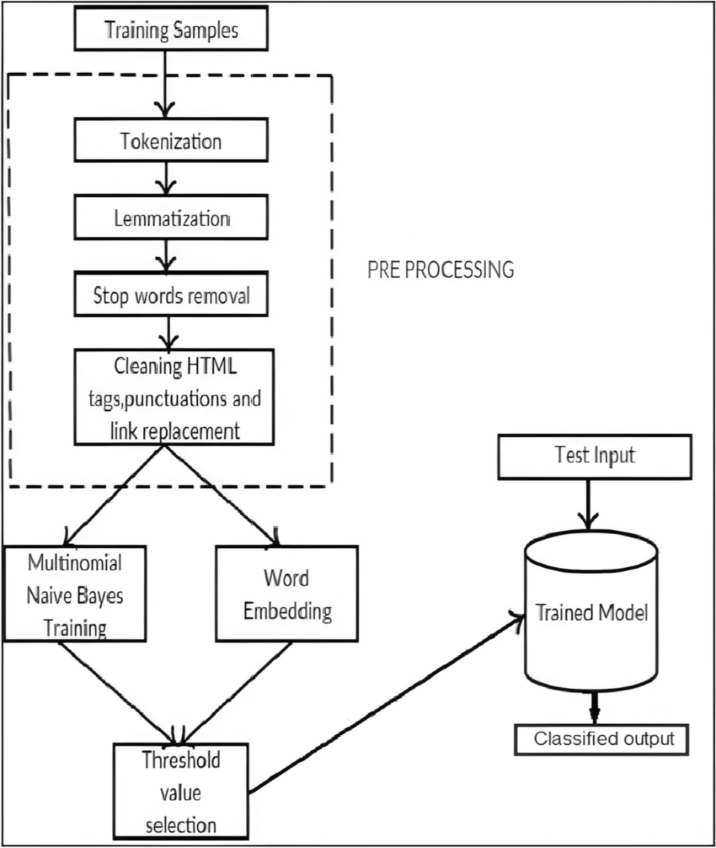
Khởi đầu trạng thái cell và trạng thái ẩn được đặt là c0 = 0, h0 = 0.

Mô hình này cho phép mạng học được các mối quan hệ dài hạn trong chuỗi dữ liệu bằng cách điều chỉnh lưu giữ hoặc quên thông tin qua forget gate.

### 2.2.4. Mô hình Naive Bayes.

Naive Bayes là một mô hình học máy/ mô hình thống kê dựa trên định lý Bayes, thường được áp dụng trong các bài toán phân loại văn bản do tính đơn giản, hiệu quả và tốc độ huấn luyện nhanh. Trong đó, Multinomial Naive Bayes (MNB**)** đặc biệt phù hợp với dữ liệu rời rạc như văn bản, nơi các đặc trưng được biểu diễn dưới dạng tần suất xuất hiện từ ngữ (Bag-of-Words).

#### Cấu trúc của mô hình Naive Bayes.



Hình 4. Cấu trúc của mô hình Naive Bayes ( Multinomial Naive Bayes ).

#### b. Cơ sở toán học.

**Định lý Bayes :**

Cho một mẫu dữ liệu có vector đặc trưng x = (x1, x2, ..., xn) ta cần dự đoán nhãn lớp Ck ∈ {C1, C2, ..., Ck}.

Công thức định lý Bayes : P(Ck | x) = (P(x | Ck) \* P(Ck)) / P(x).

Trong đó :

* P(Ck | x) : xác suất mẫu thuộc lớp Ck khi biết đặc trưng x.
* P(x | Ck) : xác suất xuất hiện đặc trưng x khi biết lớp Ck.
* P(Ck) : xác suất tiên nghiệm của lớp Ck.
* P(x) : xác suất xuất hiện đặc trưng x.

**Giả thiết độc lập có điều kiện :**

Mô hình Naive Bayes giả sử các đặc trưng xi độc lập có điều kiện với nhau khi biết lớp Ck, tức là : P(x | Ck) = Π (i=1 đến n) P(xi | Ck).

**Hàm quyết định phân loại :**

Khi phân loại một mẫu mới, ta chọn lớp có xác suất hậu nghiệm lớn nhất : Ĉ = argmax\_Ck P(Ck | x) = argmax\_Ck P(Ck) \* Π (i=1 đến n) P(xi | Ck)

Do P(x) không phụ thuộc lớp nên có thể bỏ qua khi tìm lớp tối ưu.

**Ứng dụng với dữ liệu dạng văn bản (Multinomial Naive Bayes) :**

Với dữ liệu văn bản, đặc trưng thường là số lần xuất hiện của từng từ trong từ điển V.

Giả sử fi là tần suất của từ wi trong văn bản.

Xác suất của văn bản d với tập từ V được tính :

P(d | Ck) = Π (i=1 đến |V|) P(wi | Ck)fi.

**Xử lý từ chưa xuất hiện :**

Để tránh xác suất bằng 0 khi từ không xuất hiện trong lớp học, sử dụng Laplace smoothing : P(wi | Ck) = (Nwi,Ck + α) / (NCk + α|V|).

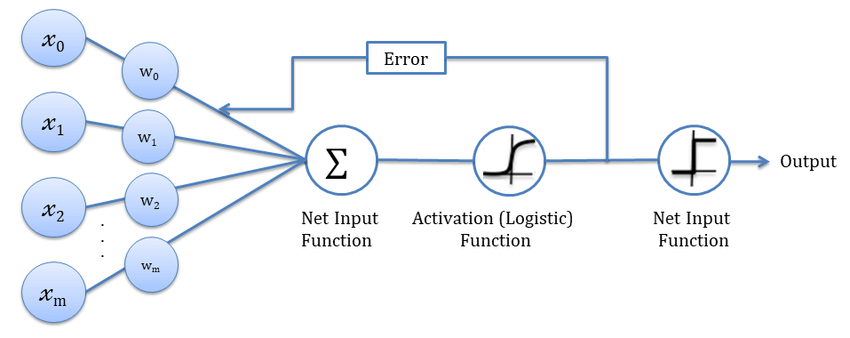
Trong đó:

* Nwi,Ck là số lần từ wi xuất hiện trong các tài liệu lớp Ck.
* NCk là tổng số từ trong tất cả tài liệu lớp Ck.
* α là tham số smoothing (thường chọn α=1).

### 2.2.5. Mô hình Logistic Regression.

**Logistic Regression** là một mô hình hồi quy được sử dụng để dự đoán xác suất của một biến nhị phân (binary) hoặc đa nhị phân (multinomial) dựa trên một hoặc nhiều biến đầu vào (biến đặc trưng). Mô hình này dùng hàm logistic (hay còn gọi là hàm sigmoid) để chuyển đổi tổ hợp tuyến tính của các biến đầu vào thành một giá trị nằm trong khoảng từ 0 đến 1, đại diện cho xác suất thuộc về một lớp cụ thể.

#### Cấu trúc cơ bản.



Hình 5. Mô hình hồi quy Logistic.

**Đầu vào (input) :** Vecto đặc trưng x=(x1,x2,...,xn) của một mẫu dữ liệu.

**Hệ số trọng số và bias :** Mỗi đặc trung xi tương ứng với một trọng số wi và có thêm hệ số bias (intercept) b.

**Tổ hợp tuyến tính (Linear combination) :** Mô hình tính tổ hợp tuyến tính các đặc trưng : z= w1x1+w2x2+…+wnxn+b=wTx+b.

**Hàm sigmoid (logistic function) :** Kết quả z được đưa qua hàm sigmoid để chuyển thành xác suất trong khoảng [0,1] : σ (z) =1/(1+e-x)

**Dự đoán :** σ (z) >= 0.5 🡪 phân vào lớp 1 (positive), σ (z) < 0.5 🡪 phân vào lớp 0 (negative).

#### Cơ sở toán học.

**Hàm dư đoán xác suất :**

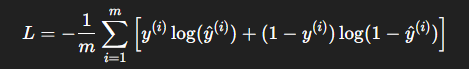


Hình 6. Hàm dư đoán xác suất.

Trong đó:

* x=(x1,x2,...,xn) là vecto đặc trưng đầu vào.
* w=(w1,w2,…,wn) là vecto trọng số (hệ số hồi quy).
* b là hệ số chệch.
* σ (z) =1/(1+e-x) là hàm sigmoid (logistic).

**Hàm mất mát (**thường dùng **hàm log-loss (binary cross-entropy)):**



Hình 7. Hàm mất mát.

Trong đó :

* y^(i) = P(y=1|x(i)).

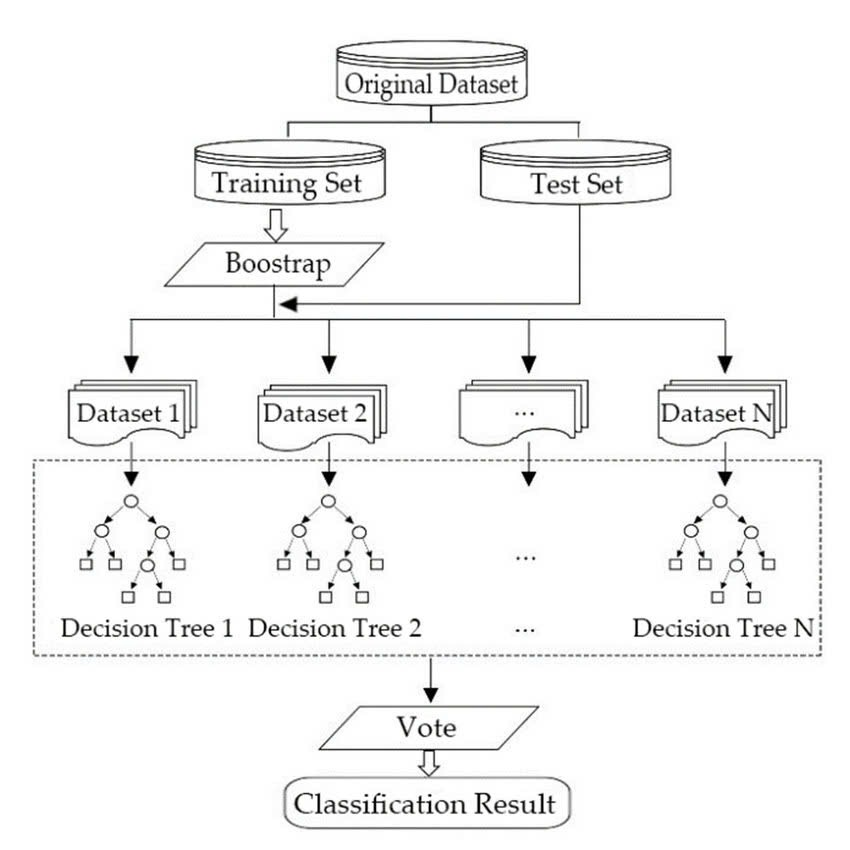
**Tối ưu hóa:** dùng các thuật toán như Gradient Descent để tìm w,b sao cho L nhỏ nhất.

### 2.2.6. Thuật toán Random Forest.

**Random Forest** là một thuật toán học máy thuộc nhóm ensemble learning, được sử dụng phổ biến trong các bài toán phân loại và hồi quy. Thuật toán này xây dựng nhiều cây quyết định (decision trees) độc lập trên các mẫu dữ liệu con (bootstrap samples) và tập con các đặc trưng, sau đó kết hợp kết quả của các cây con để đưa ra dự đoán cuối cùng bằng cách bỏ phiếu đa số (đối với phân loại) hoặc trung bình (đối với hồi quy).

Điểm mạnh của Random Forest là khả năng giảm thiểu overfitting so với việc sử dụng một cây quyết định đơn lẻ, đồng thời vẫn giữ được độ chính xác và khả năng tổng quát hóa tốt trên dữ liệu mới.

#### Cấu trúc của Random Forest.



Hình . Cấu trúc Random Forest.

1. Cơ sở toán học.

Giả sử ta có tập dữ liệu huấn luyện D={(xi,yi)}i=1N với xi∈Rd là vector đặc trưng và yi​ là nhãn (cho phân loại) hoặc giá trị liên tục (cho hồi quy).

**Bootstrap sampling :**  
Với mỗi cây t, tạo tập con huấn luyện ​ Di bằng cách lấy mẫu ngẫu nhiên có hoàn lại từ D.

**Xây dựng cây quyết định :**  
Với mỗi nút của cây ttt, thuật toán:

* Chọn ngẫu nhiên một tập con đặc trưng Ft⊂{1,2,...,d} có kích thước nhỏ hơn d.
* Tìm điều kiện phân chia trên Ft sao cho tối ưu hàm mục tiêu (ví dụ, giảm độ hỗn loạn thông qua chỉ số Gini, entropy, hoặc sai số bình phương).

**Dự đoán của cây t :**  
Gọi hàm dự đoán của cây t là ht(x).

**Dự đoán của Random Forest :**

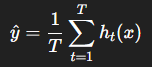
* Với bài toán phân loại, dự đoán là nhãn được đa số cây dự đoán:



Hình 9. Công thức dự đoán (bài toán phân loại) của Random Forest.

Với T là số cây trong rừng

* Với bài toán hồi quy, dự đoán là trung bình dự đoán của các cây.



Hình 10.Công thức dự đoán (bài toán hồi quy) của Random Forest.

### 2.2.7. Thông số đánh giá mô hình.

* Precision: Độ chính xác của các dự đoán dương tính.

Precisionclass​=TPclass ​/ ( FPclass + ​TPclass​ )​.

* Recall: Khả năng tìm đúng các trường hợp dương tính.

Recallclass​=TPclass ​/ ( FNclass + ​TPclass​ )​

* F1-score: Trung bình điều hòa của Precision và Recall.

F1-scoreclass=( 2 \* Precisionclass \* Recallclass ) / ( Precisionclass +Recallclass )

* Accuracy: Tỷ lệ dự đoán đúng trên toàn bộ dữ liệu.

Accuracy=Số mẫu dự đoán đúng / Tổng số mẫu

## 2.3. Phân tích các nghiên cứu trước và định hướng tính kế thừa, phát triển.

### 2.3.1. Điểm mạnh của cá nghiên cứu trước.

#### a. Mô hình LSTM.

Nghiên cứu của Dr. G. S. N. Murthy và các cộng sự (2020) đã úng dụng mô hình LSTM (Long Short-Term Memory) để phân tích cảm xúc từ văn bản, đặc biệt trên các đánh giá sản phẩm từ Ãmazon và nhận xét phim từ IMDb. Một số điểm mạnh và nổi bật của bài nghiên cứu này bao gồm :

* **Ứng dụng LSTM trong phân tích cảm xúc :** Nghiên cứu đã sử dụng LSTM, một loại mạng nơ-ron hồi tiếp, để xử lý và phân tích chuỗi dữ liệu văn bản dài, giúp mô hình nắm bắt được ngữ cảnh và mối quan hệ giữa các từ trong câu.
* **Tiền xử lý dữ liệu hiệu quả :** Quá trình tiền xử lý bao gồm việc chuyển đổi văn bản thành chữ thường, loại bỏ dấu câu và mã hóa từ vựng thành các chỉ số số học, giúp chuẩn hóa dữ liệu đầu vào và cải thiện hiệu suất mô hình.
* **Sử dụng embedding từ :** Nghiên cứu đã áp dụng kỹ thuật embedding từ để chuyển đổi từ ngữ thành các vector số học, giúp mô hình hiểu được ngữ nghĩa và mối quan hệ giữa các từ.
* **Đánh giá mô hình trên tập dữ liệu chuẩn :** Mô hình được đánh giá trên các tập dữ liệu chuẩn như IMDb và Amazon, cho thấy khả năng phân loại cảm xúc hiệu quả với độ chính xác cao.
* **Ứng dụng thực tiễn** **:** Nghiên cứu nhấn mạnh tầm quan trọng của phân tích cảm xúc trong việc hiểu và khai thác thông tin từ dữ liệu văn bản, hỗ trợ các quyết định kinh doanh và nghiên cứu thị trường.

#### b. Mô hình Naive Bayes.

Điểm mạnh của bài nghiên cứu "Multinomial Naive Bayes Classification Model for Sentiment Analysis" của các tác giả Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon và Anees Ahmed :

* **Hiệu suất cao :** Mô hình Multinomial Naive Bayes (MNB) đạt độ chính xác khoảng 90% trong việc phân loại cảm xúc từ các bài đánh giá phim, cho thấy hiệu quả vượt trội của phương pháp này trong xử lý văn bản .
* **Giải pháp cải tiến mô hình :** Nghiên cứu đề xuất các giải pháp heuristic đơn giản để khắc phục một số vấn đề của MNB, đặc biệt là khi dữ liệu văn bản không hoàn toàn tuân theo giả định phân phối đa thức, giúp cải thiện độ chính xác tổng thể của mô hình .
* **Phân tích từ vựng chi tiết :** Bài báo cung cấp phân tích về các từ ngữ tích cực và tiêu cực có ảnh hưởng lớn đến kết quả phân loại, giúp hiểu rõ hơn về cách mô hình đưa ra quyết định .
* **Khả năng mở rộng :** MNB là mô hình đơn giản, dễ triển khai và có khả năng mở rộng tốt, phù hợp với các ứng dụng thực tế yêu cầu xử lý nhanh chóng và hiệu quả.

#### c. Mô hình Logistic Regression.

Bài nghiên cứu "Sentiment Analysis using Logistic Regression" của các tác giả George B. Aliman, Tanya Faye S. Nivera, Jensine Charmille A. Olazo, Daisy Jane P. Ramos, Chris Danielle B. Sanchez, Timothy M. Amado, Nilo M. Arago, Romeo L. Jorda Jr., Glenn C. Virrey, và Ira C. Valenzuela sau khi đọc qua và tổng hợp ta thấy nó có 3 điểm mạnh :

* **Khả năng xử lý ngôn ngữ đa dạng :** Bài nghiên cứu tập trung vào phân tích cảm xúc trong ba ngôn ngữ: tiếng Anh, tiếng Philippines và Taglish (sự kết hợp giữa tiếng Anh và tiếng Philippines). Việc này cho thấy mô hình Logistic Regression có khả năng ứng dụng trong môi trường ngôn ngữ đa dạng.
* **Phương pháp tiếp cận đơn giản và hiệu quả :** Logistic Regression là một mô hình học máy đơn giản, dễ triển khai và hiệu quả trong các bài toán phân loại nhị phân như phân tích cảm xúc. Việc sử dụng mô hình này giúp giảm thiểu độ phức tạp trong quá trình huấn luyện và triển khai.
* **Kết quả đáng tin cậy :** Mặc dù chi tiết về độ chính xác không được nêu rõ trong phần tóm tắt, nhưng việc lựa chọn Logistic Regression cho thấy mô hình đạt được kết quả khả quan trong việc phân loại cảm xúc từ các đánh giá văn bản.

#### d. Thuật toán Random Forest.

Bài báo nghiên cứu "Sentiment Analysis for Film Reviews Based on Random Forest" của tạp chí Dean Francis Press có những điểm mạnh sau :

* **Hiệu quả cao :** Mô hình Random Forest thường cho độ chính xác tốt, như bài báo đề cập đạt khoảng 86% trên tập dữ liệu đánh giá phim, cho thấy khả năng phân loại cảm xúc hiệu quả.
* **Giảm overfitting :** Nhờ kỹ thuật ensemble và bootstrap sampling, mô hình tránh được tình trạng overfitting so với các cây quyết định đơn lẻ.
* **Khả năng xử lý dữ liệu đa chiều :** Random Forest có thể làm việc tốt với dữ liệu có nhiều đặc trưng mà không cần quá nhiều bước tiền xử lý phức tạp.
* **Độ ổn định :** Mô hình không nhạy cảm quá mức với dữ liệu nhiễu hoặc ngoại lệ.
* **Giải thích được mức độ quan trọng đặc trưng :** Cho phép phân tích xem từ/câu nào ảnh hưởng nhiều đến quyết định phân loại cảm xúc.

### 2.3.2. Hạn chế của nghiên cứu.

#### a. Mô hình LSTM.

Mặc dù nghiên cứu của Dr. G. S. N. Murthy và cộng sự (2020) đã ứng dụng thành công mô hình LSTM trong phân tích cảm xúc văn bản, nhưng vẫn còn một số hạn chế như sau:

* **Phân loại cảm xúc thường chỉ gồm 2 lớp (positive/negative) :** Nghiên cứu tập trung chủ yếu vào phân loại nhị phân, chưa khai thác tốt lớp trung lập (neutral), trong khi đánh giá trung lập rất phổ biến trong các hệ thống thương mại điện tử.
* **Vấn đề mất cân bằng dữ liệu chưa được xử lý rõ ràng :** Các tập dữ liệu thực tế thường bị mất cân bằng về số lượng các lớp cảm xúc, điều này ảnh hưởng đến hiệu quả dự đoán các lớp ít phổ biến như neutral hoặc negative.
* **Tiền xử lý dữ liệu còn khá cơ bản :** Việc tiền xử lý dữ liệu như loại bỏ dấu câu, chuẩn hóa chữ thường chưa bao gồm các bước nâng cao như xử lý từ viết tắt, biểu tượng cảm xúc, hoặc lemmatization, làm hạn chế khả năng hiểu sâu ngữ cảnh của mô hình.
* **Chưa đề cập rõ đến việc phân tích lỗi và hạn chế thực tiễn :** Nghiên cứu chưa tập trung nhiều vào việc phân tích nguyên nhân lỗi của mô hình hay đề xuất cải tiến cụ thể cho các trường hợp khó phân loại.

#### b. Mô hình Naive Bayes.

Điểm yếu của bài nghiên cứu "Multinomial Naive Bayes Classification Model for Sentiment Analysis" của các tác giả Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon và Anees Ahmed :

* **Giả định độc lập đặc trưng :** MNB giả định rằng các đặc trưng (từ ngữ) trong văn bản là độc lập với nhau, điều này không luôn đúng trong thực tế và có thể ảnh hưởng đến độ chính xác của mô hình.
* **Xử lý dữ liệu mất cân bằng :** Mô hình có thể gặp khó khăn khi xử lý các tập dữ liệu mất cân bằng, nơi mà một số lớp cảm xúc có số lượng mẫu ít hơn đáng kể so với các lớp khác, dẫn đến khả năng phân loại kém cho các lớp thiểu số.
* **Thiếu so sánh với các mô hình hiện đại :** Nghiên cứu không thực hiện so sánh hiệu suất của MNB với các mô hình học sâu hiện đại như LSTM hoặc BERT, vốn đã chứng minh hiệu quả vượt trội trong các nhiệm vụ phân tích cảm xúc.
* **Không đề cập đến xử lý ngôn ngữ đa dạng :** Bài báo tập trung vào dữ liệu đánh giá phim và không thảo luận về khả năng áp dụng mô hình cho các ngôn ngữ hoặc lĩnh vực khác, điều này hạn chế tính tổng quát của nghiên cứu.

#### Mô hính Logistic Regression.

Bên cạnh đó, bài nghiên cứu *"Sentiment Analysis using Logistic Regression"* của các tác giả George B. Aliman, Tanya Faye S. Nivera, Jensine Charmille A. Olazo, Daisy Jane P. Ramos, Chris Danielle B. Sanchez, Timothy M. Amado, Nilo M. Arago, Romeo L. Jorda Jr., Glenn C. Virrey, và Ira C. Valenzuela vẫn để lại nhiều hạn chế tương ứng :

* **Hạn chế trong việc xử lý ngữ cảnh phức tạp :** Logistic Regression chủ yếu dựa vào các đặc trưng tuyến tính và có thể không hiệu quả trong việc nắm bắt các mối quan hệ phi tuyến hoặc ngữ cảnh phức tạp trong văn bản, đặc biệt là trong các ngôn ngữ có cấu trúc linh hoạt như Taglish.
* **Thiếu so sánh với các mô hình tiên tiến hơn :** Bài nghiên cứu không đề cập đến việc so sánh hiệu suất của Logistic Regression với các mô hình học sâu hiện đại như LSTM, BERT hoặc các mô hình dựa trên transformer, vốn đã chứng minh hiệu quả vượt trội trong các nhiệm vụ xử lý ngôn ngữ tự nhiên.
* **Không đề cập đến kỹ thuật xử lý dữ liệu mất cân bằng :** Trong các bài toán phân tích cảm xúc, dữ liệu thường bị mất cân bằng giữa các lớp (ví dụ: nhiều đánh giá tích cực hơn tiêu cực). Bài nghiên cứu không nêu rõ liệu có áp dụng các kỹ thuật như oversampling, undersampling hoặc sử dụng trọng số lớp để xử lý vấn đề này hay không.

#### d. Thuật toán Random Forest.

Bài báo nghiên cứu "Sentiment Analysis for Film Reviews Based on Random Forest" của tạp chí Dean Francis Press vẫn còn nhiều hạn chế :

* **Không rõ ràng về mô hình bên trong (black-box) :** Mặc dù tốt hơn nhiều mô hình khác, Random Forest vẫn khó để giải thích chi tiết quyết định của từng cây đối với từng mẫu.
* **Yêu cầu tính toán lớn :** Khi có nhiều cây và dữ liệu lớn, thời gian huấn luyện và dự đoán có thể tăng cao.
* **Có thể bị giảm hiệu quả khi dữ liệu bị lệch :** Nếu dữ liệu mất cân bằng nghiêm trọng giữa các lớp cảm xúc, Random Forest có thể cần thêm các kỹ thuật xử lý dữ liệu mất cân bằng.
* **Không tận dụng được chuỗi thời gian hoặc cấu trúc ngữ cảnh :** Với dữ liệu dạng văn bản dài hoặc có mối quan hệ theo chuỗi, Random Forest không tận dụng tốt các đặc điểm này so với các mô hình mạng nơ-ron như LSTM.

### 2.3.3. Kế thừa và phát triển.

Nghiên cứu của tôi kế thừa các ưu điểm của nghiên cứu trên đồng thời phát triển thêm các điểm mới như sau:

* **Sử dụng mô hình LSTM cơ bản :** Tiếp tục sử dụng LSTM nhờ ưu điểm về khả năng xử lý chuỗi và ghi nhớ ngữ cảnh, đồng thời giữ mô hình đơn giản giúp tiết kiệm tài nguyên tính toán.
* **Sử dụng lại mô hình Naive Bayes, Logistic Regression, Random Forest :** Tổng kết, so sánh kết quả để đưa ra mô hình tốt nhất trong 3 mô hình rồi đem so sánh với LSTM.
* **Phân loại 3 lớp cảm xúc (positive, neutral, negative) :** Mở rộng so với phân loại nhị phân, nhằm phản ánh chính xác hơn thực tế các đánh giá sản phẩm trên nền tảng thương mại điện tử.
* **Tiền xử lý dữ liệu kỹ lưỡng hơn :** Bổ sung các bước tiền xử lý nâng cao như loại bỏ từ dừng (stop words), chuẩn hóa từ viết tắt, xử lý biểu tượng cảm xúc, và lemmatization để cải thiện chất lượng dữ liệu đầu vào.
* **Đánh giá chi tiết theo từng lớp :** Sử dụng các chỉ số Precision, Recall, F1-score cho từng lớp cảm xúc riêng biệt, đặc biệt tập trung vào lớp neutral vốn thường bị bỏ qua.
* **Phân tích và thảo luận kết quả kỹ lưỡng :** Thực hiện phân tích lỗi và so sánh kết quả với nghiên cứu trước, từ đó đề xuất hướng phát triển mô hình trong tương lai.

# Chương 3 : Phương pháp nghiên cứu

## 3.1. Thiết kế nghiên cứu.

Nghiên cứu được thực hiện vầ thiết kế trên phương pháp định lượng, sử dụng các mô hình và thuật toán của học máy (Machine Learning) để phân loại cảm xúc mà người dùng để lại thông qua score (thang 5) và văn bản đánh giá. Việc phân tích được thực hiện trên dữ liệu văn bản thực tế được thu thập trên nền tảng trang thương mại điện tử (Amazon). Nghiên cứu sử dụng các kỹ thuật tièn xử lý văn bản, khám phá dữ liệu (EDA), vecto hóa, huấn luyện mô hình học sâu (LSTM), và đánh giá hiệu suất mô hình (accuracy, F1-score, recall, precision).

## 3.2. Đối tượng và mẫu nghiên cứu.

Đối tượng nghiên cứu : Các đánh giá dạng văn bản do người dùng để lại sau khi sử dụng sản phẩm trên nền tảng thương mại điện tử.

Mẫu nghiên cứu :

* Dữ liệu được trích suất từ tập Amazon Fine Food Reviews công khai trên Kaggle[6] với hơn 500,000 đánh giá thực tế.
* Lọc các đánh giá trong năm 2012 ta thu được 198,659 đánh giá văn bản. Các đánh giá này thuộc về khoảng 5,000 sản phẩm ở nhiều danh mục sản phẩm khác nhau như thực phẩm, điện tử, thời trang,...
* Mỗi đánh giá bao gồm thông tin : nội dung đánh giá (Text), số sao (score), ngày đăng (Time), và ID sản phẩm (ProductId).

## 3.3. Phương pháp thu thập và tiền xử lý dữ liệu.

### 3.3.1. Thu thập dữ liệu.

Dữ liệu được thu thập từ tập [*Amazon Fine Food Reviews*](https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews), một tập dữ liệu công khai trên nền tảng Kaggle, bao gồm hơn 500.000 đánh giá thực tế của người dùng đối với các sản phẩm thực phẩm. Để đảm bảo tính nhất quán và giới hạn phạm vi nghiên cứu, nhóm tiến hành lọc các đánh giá được để lại trong năm 2012, từ đó thu được 198,659 đánh giá văn bản. Mỗi mục dữ liệu bao gồm các thông tin: nội dung đánh giá (Text), số sao (Score), ngày đăng (Time), và ID sản phẩm (ProductId). Lọc lấy dữ liệu trong năm 2012 vì đây là khoảng thời gian gần hiện tại nhất của tập tin.

### 3.3.2. Tiền xử lý.

Các bước được thục hiện như sau :

#### Gán nhãn cảm xúc :

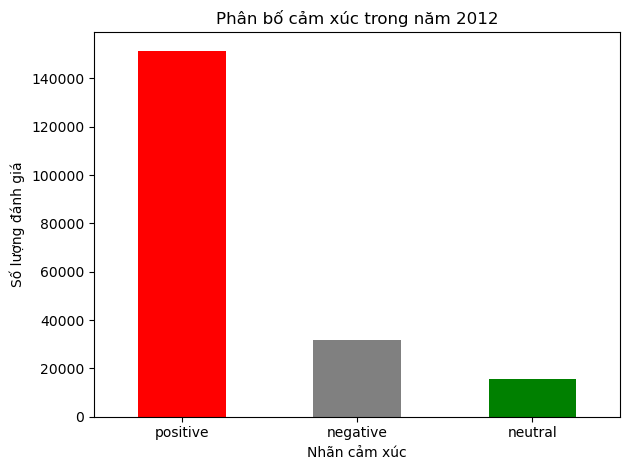
Dựa vào số sao (Score) từ 1 đến 5, dữ liệu được gán thành ba nhãn cảm xúc:

* Từ 1 đến 2 sao → **negative.**
* 3 sao → **neutral.**
* Từ 4 đến 5 sao → **positive.**

#### Làm sạch văn bản :

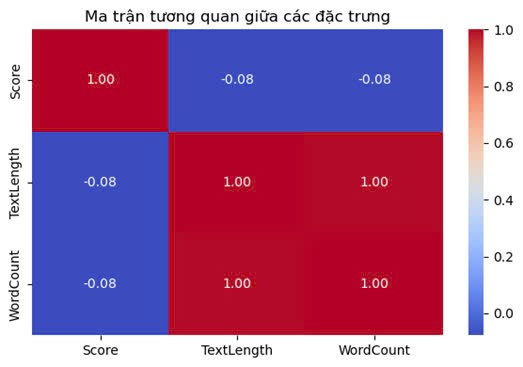
Loại bỏ các ký tự đặc biệt, dấu câu, chữ số và chuyển tất cả về chữ thường. Loại bỏ stopwords (từ dừng) để giảm nhiễu. Áp dụng kỹ thuật lemmatization để đưa từ về dạng gốc.

#### Khám phá dữ liệu (EDA) :



Hình 11. Biểu đồ phân bố số lượng đánh giá theo nhãn

Thông qua biểu đồ, ta thấy biểu đồ mất cân bằng vì số đánh giá có nhãn (positive) chiếm (~76%) chiếm hơn ¾ tổng dữ liệu của tập tin.



Hình 12. Ma trận tương quan giữa số sao, độ dài văn bản và số lượng từ.

**Mối quan hệ giữa TextLength và WordCount :**

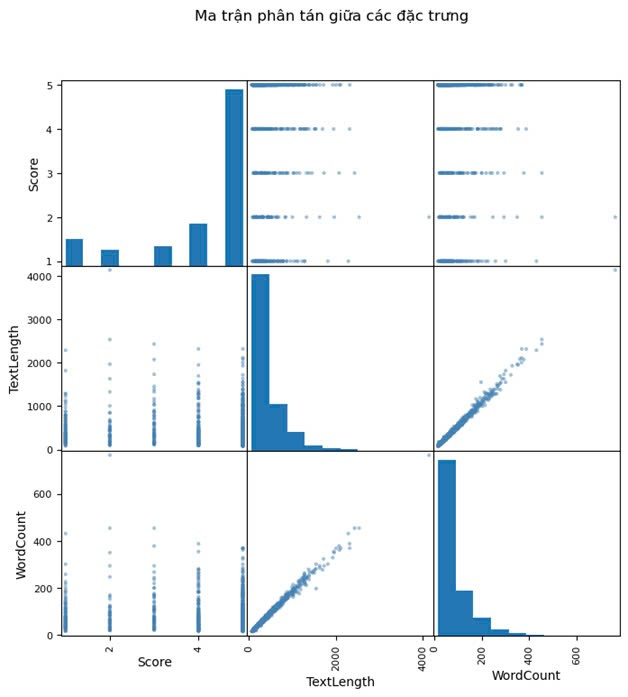
* Giá trị tương quan là **1.00**.
* Màu sắc đỏ đậm nhất.
* **Nhận xét:** Điều này cho thấy có một mối tương quan tuyến tính dương cực kỳ mạnh mẽ và hoàn hảo giữa TextLength (độ dài văn bản) và WordCount (số lượng từ). Khi độ dài văn bản tăng lên, số lượng từ cũng tăng lên theo một tỷ lệ gần như cố định. Đây là điều hoàn toàn hợp lý và phù hợp với quan sát từ ma trận phân tán trước đó.

**Mối quan hệ giữa Score và TextLength :**

* Giá trị tương quan là **-0.08**.
* Màu sắc xanh lam rất nhạt, gần với màu trắng.
* **Nhận xét:** Giá trị tương quan này rất gần với 0 và mang dấu âm nhỏ. Điều này cho thấy có một mối tương quan tuyến tính **rất** yếu và hơi tiêu cực giữa Score và TextLength. Nói cách khác, độ dài văn bản có rất ít hoặc hầu như không có mối liên hệ tuyến tính rõ ràng với Score.

**Mối quan hệ giữa Score và WordCount :**

* Giá trị tương quan là **-0.08**.
* Màu sắc xanh lam rất nhạt, gần với màu trắng.
* **Nhận xét:** Tương tự như mối quan hệ giữa Score và TextLength, giá trị tương quan này cũng rất gần với 0 và hơi âm. Điều này chỉ ra một mối tương quan tuyến tính rất yếu và hơi tiêu cực giữa Score và WordCount. Số lượng từ cũng có rất ít mối liên hệ tuyến tính rõ ràng với Score.



Hình 13. Ma trận phân tán.

Nhận xét :

* **"Score"** có vẻ là một biến rời rạc/phân loại, còn **"TextLength"** và **"WordCount"** là các biến liên tục và có mối tương quan rất cao với nhau.
* Phân phối của **"TextLength"** và **"WordCount"** bị lệch nặng về phía giá trị nhỏ.
* Không có mối quan hệ tuyến tính đơn giản và mạnh mẽ giữa **"Score"** và hai đặc trưng còn lại. Score có thể xuất hiện với nhiều độ dài và số lượng từ khác nhau, mặc dù có thể có một xu hướng nhỏ là các văn bản dài hơn có thể có sự đa dạng hơn về Score hoặc các Score cao hơn có thể xuất hiện trên phạm vi độ dài lớn hơn. Mối tương quan mạnh nhất trong biểu đồ này là giữa **WordCount** và **TextLength**.

#### Xử lý mất cân bằng dữ liệu:

Áp dụng kỹ thuật **oversampling** đối với nhãn neutral và negative nhằm cân bằng số lượng giữa các lớp, giúp mô hình học tốt hơn và giảm thiên lệch.

#### Biến đổi văn bản thành vector:

Áp dụng kỹ thuật Tokenization và Padding để chuẩn hóa độ dài câu. Sử dụng embedding layer trong mô hình LSTM để ánh xạ văn bản sang không gian vector.

### 3.3.3. Các công cụ hỗ trợ.

#### a. Phần mềm hỗ trợ.

Quá trình nghiên cứu và phân tích dữ liệu được thực hiện trên môi trường **Jupyter Notebook**, sử dụng ngôn ngữ lập trình **Python** với các phiên bản phổ biến như 3.10.8, 3.12.9 hoặc trực tiếp trên **Python**. Việc sử dụng Jupyter giúp thuận tiện trong việc trực quan hóa kết quả và kiểm thử các mô hình.

#### b. Thư viện hỗ trợ.

Để thực hiện các bước tiền xử lý, trực quan hóa dữ liệu, xây dựng và đánh giá mô hình, bài nghiên cứu sử dụng các thư viện Python sau :

* **Pandas, NumPy** : Hỗ trợ thao tác và xử lý dữ liệu dạng bảng và mảng.
* **Matplotlib, Seaborn** : Dùng cho trực quan hóa dữ liệu và biểu đồ trong giai đoạn EDA.
* **Scikit-learn** : Cung cấp công cụ tiền xử lý, chia tập dữ liệu, đánh giá mô hình, và cài đặt các thuật toán học máy truyền thống như Naive Bayes, Logistic Regression, Random Forest.
* **TensorFlow, Keras** : Được sử dụng để xây dựng và huấn luyện mô hình học sâu (LSTM).
* **Imbalanced-learn** : Hỗ trợ các kỹ thuật xử lý mất cân bằng dữ liệu, đặc biệt là phương pháp oversampling.
* **Nltk** : Dùng để loại bỏ các stopword (and, the, a, an,...), và để dùng lemmatization để biến các từ về dạng gốc (running🡪run).

# Chương 4 : Thực nghiệm và thảo luận.

## 4.1. Thiết lập thực nghiệm.

**Tập dữ liệu :** Sử dụng dữ liệu từ bộ Amazon Fine Food Reviews, chỉ lọc các đánh giá trong năm 2012 với tổng cộng 198,659 đánh giá.

**Tiền xử lý :** Gồm gán nhãn cho từng đánh giá (3 lớp : positve, negative, neural), làm sạch văn bản, tokenization, padding, dùng NLP (thư viện nltk) để xóa các stopword và dùng công cụ lemmatization.

**Mục tiêu** **:** Phân loại cảm xúc thông qua văn bản thành 1 trong 3 loại nhãn và chọn ra mô hình xử lý dữ liệu tốt nhất (LSTM, Random Forest, Logistic Regression, Naive Bayes).

**Chia dữ liệu tập huấn và kiểm tra mô hình :** Tập huấn (80%) và kiểm tra (20%).

**Xử lý mất cân bằng :** Oversampling cho 2 nhãn có giá trị thấp (negative và neural) 🡪 Cân bằng lại dữ liệu (sao chép nhiều mẫu hơn của 2 nhãn để train).

## 4.2. Kết quả mô hình.

### 4.2.1. Đánh giá mô hình.

Để có cơ sở so sánh với mô hình LSTM, nhóm đã thử nghiêm ba mô hình học máy truyền thống Naive Bayes, Logistic Regression, và Random Forest. Dữ liệu được chia giống với LSTM để có thể đem lại giá trị thông số khách quan nhất (80% huấn luyện và 20% kiểm tra mô hình).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Nhãn** | **Precision** | **Recall** | **F1-score** |
| **Naive Bayes** | Positive | 0.96 | 0.76 | 0.85 |
|  | Neutral | 0.24 | 0.60 | 0.34 |
|  | Negative | 0.58 | 0.71 | 0.64 |
| **Logistic Regression** | Positive | 0.96 | 0.81 | 0.88 |
|  | Neutral | 0.30 | 0.65 | 0.41 |
|  | Negative | 0.66 | 0.76 | 0.70 |
| **Random Forest** | Positive | 0.89 | 0.99 | 0.94 |
|  | Neutral | 0.95 | 0.41 | 0.57 |
|  | Negative | 0.88 | 0.67 | 0.76 |
| **LSTM** | Positive | 0.95 | 0.91 | 0.93 |
|  | Neutral | 0.47 | 0.59 | 0.52 |
|  | Negative | 0.74 | 0.78 | 0.76 |

Bảng 3. So sánh kết quả của 4 mô hình.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Accuracy (%)** | **Precision (Macro)** | **Recall (Macro)** | **F1-score (Macro)** |
| **Naive Bayes** | 73.86 | 0.59 | 0.69 | 0.61 |
| **Logistic Regression** | 79.14 | 0.64 | 0.74 | 0.66 |
| **Random Forest** | 89.18 | 0.90 | 0.69 | 0.76 |
| **LSTM** | 86.74 | 0.72 | 0.76 | 0.74 |

Bảng 4. Đánh giấ hiệu suất tổng quát của 4 mô hình.

**Nhận xét :**

**Positive :**

* Tất cả mô hình đều đạt precision và recall cao với nhãn *positive*.
* Naive Bayes và Logistic Regression có precision rất cao (0.96), nhưng recall thấp hơn so với Random Forest (0.99) và LSTM (0.91).
* LSTM và Random Forest là hai mô hình có F1-score cao nhất (0.93 và 0.94), chứng tỏ rất phù hợp với việc nhận diện cảm xúc tích cực.

**Neutral :**

* Đây là nhãn yếu nhất của cả 4 mô hình.
* Naive Bayes và Logistic Regression có precision thấp (0.24 và 0.30) nhưng recall khá cao (0.60 và 0.65), tức là dự đoán đúng nhiều nhưng dễ nhầm lẫn với nhãn khác.
* Random Forest có precision rất cao (0.95) nhưng recall thấp (0.41), cho thấy nó chỉ chắc chắn khi rất tự tin.
* LSTM có F1-score cao nhất (0.52), tuy chưa xuất sắc nhưng vẫn vượt trội hơn các mô hình còn lại cho nhãn *neutral*.

**Negative :**

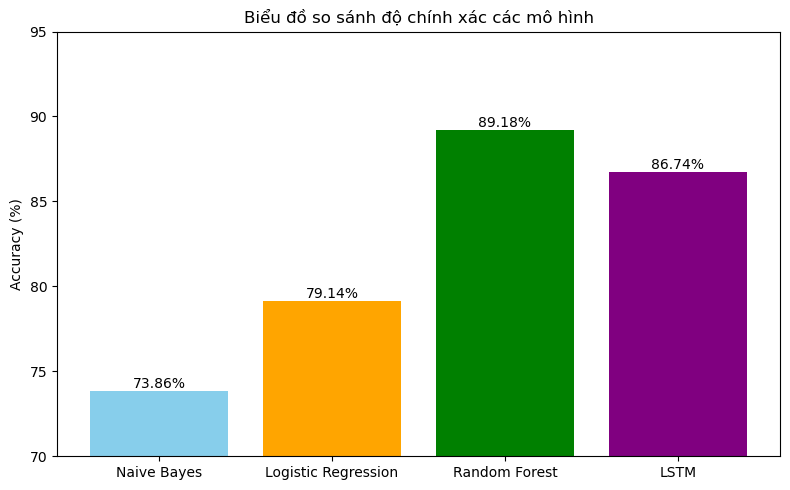
* Random Forest và LSTM dẫn đầu với F1-score 0.76.
* Logistic Regression theo sau với 0.70.
* Naive Bayes yếu nhất (F1-score 0.64), cho thấy khó phân biệt các cảm xúc tiêu cực.

**Tổng kết mô hình :**

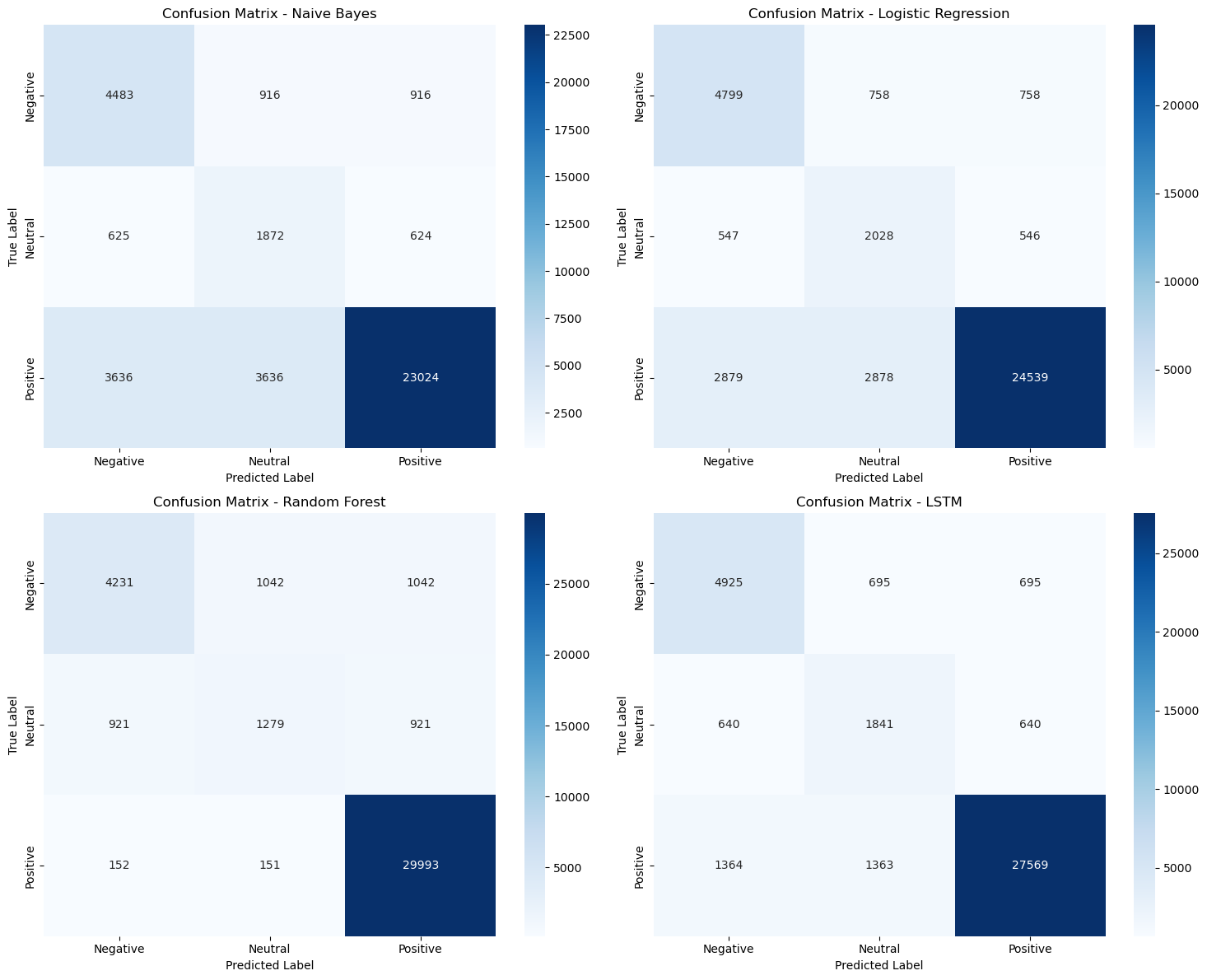
|  |  |
| --- | --- |
| **Mô hình** | **Nhận xét tổng quát** |
| **Naive Bayes** | Nhanh, đơn giản, hiệu quả ở nhãn *positive*, nhưng yếu ở *neutral* và *negative*. |
| **Logistic Regression** | Cân bằng, hiệu suất khá ở cả 3 nhãn, nhưng xử lý *neutral* vẫn chưa tốt. |
| **Random Forest** | Mạnh mẽ ở *positive*, chính xác cao, nhưng overfitting nhẹ và recall *neutral* kém. |
| **LSTM** | Toàn diện nhất, cân bằng precision/recall, vượt trội ở cả 3 nhãn. Tuy nhiên tốn tài nguyên và thời gian huấn luyện. |

Bảng 5. Tổng kết 4 mô hình.

### 4.2.2. Biểu đồ mô hình.



Hình 14. So sánh độ chính xác của 4 mô hình.



Hình 15. Confusion Matrix của 4 mô hình.

**Mô hình Naive Bayes :**

**Ưu điểm :** Có khả năng nhận diện một số lượng đáng kể các mẫu Negative (4483) và Positive (23024) một cách chính xác.

**Hạn chế :**

* Mức độ chính xác cho lớp Neutral khá thấp (1872).
* Số lượng lớn các mẫu Positive bị phân loại sai thành Negative (3636) và Neutral (3636).
* Số lượng các mẫu Negative và Neutral bị phân loại sai cũng khá cao.

**Nhận xét tổng quan:** Naive Bayes có vẻ là mô hình yếu nhất trong số bốn mô hình này cho nhiệm vụ phân loại ba lớp, đặc biệt là trong việc phân biệt giữa các lớp, dẫn đến nhiều lỗi phân loại sai.

**Mô hình Logistic Regression :**

**Ưu điểm :**

* Hiệu suất tốt hơn Naive Bayes cho các lớp Negative (4799) và Neutral (2028).
* Số lượng lỗi phân loại sai từ lớp Positive sang Negative (2879) và Neutral (2879) đã giảm so với Naive Bayes.

**Hạn chế :** Mặc dù tốt hơn Naive Bayes, mô hình này vẫn còn khá nhiều lỗi phân loại sai, đặc biệt là với lớp Positive bị nhầm lẫn với Negative và Neutral.

**Nhận xét tổng quan:** Logistic Regression cho thấy sự cải thiện đáng kể so với Naive Bayes, đạt được độ chính xác tốt hơn trên các lớp.

**Thuật toán Random Forest :**

**Ưu điểm nổi bật :** Cực kỳ xuất sắc trong việc phân loại lớp Positive (29993 mẫu đúng). Số lượng mẫu Positive bị phân loại sai thành Negative (152) và Neutral (151) là rất ít. Điều này cho thấy mô hình này cực kỳ giỏi trong việc nhận diện các trường hợp tích cực.

**Hạn chế :**

* Tuy nhiên, hiệu suất của nó đối với các lớp Negative (4231) và Neutral (1279) lại kém hơn so với Logistic Regression và LSTM (sẽ phân tích sau).
* Số lượng mẫu Neutral bị phân loại sai thành Negative (921) và Positive (921) khá cao.

**Nhận xét tổng quan :** Random Forest có một sự đánh đổi rõ ràng: độ chính xác cực cao cho lớp Positive, nhưng lại không hiệu quả bằng các mô hình khác cho lớp Negative và Neutral. Điều này có thể hữu ích trong các ứng dụng mà việc xác định chính xác lớp Positive là ưu tiên hàng đầu.

**Mô hình LSTM :**

**Ưu điểm :**

* Hiệu suất tổng thể tốt nhất và cân bằngnhất trong số bốn mô hình.
* Đạt được số lượng mẫu Negative được phân loại đúng cao nhất (4925).
* Đạt được số lượng mẫu Positive được phân loại đúng rất cao (27569), chỉ kém một chút so với Random Forest nhưng lại có ít lỗi phân loại sai hơn từ Positive sang Negative/Neutral so với Naive Bayes và Logistic Regression.
* Hiệu suất cho lớp Neutral (1841) cũng khá tốt, mặc dù vẫn còn một số lỗi.

**Hạn chế :** Vẫn còn một số mẫu Neutral và các lớp khác bị phân loại sai, nhưng nhìn chung là ít hơn.

**Nhận xét tổng quan :** Mô hình LSTM cho thấy hiệu suất mạnh mẽ và ổn định trên cả ba lớp, với khả năng phân loại đúng cao cho cả Negative và Positive, đồng thời giảm thiểu đáng kể các lỗi phân loại sai.

**So sánh tổng thể giữa 4 mô hình:**

* LSTM dường như là mô hình có hiệu suất tốt nhất và cân bằng nhất cho nhiệm vụ phân loại ba lớp này, thể hiện khả năng mạnh mẽ trên tất cả các lớp.
* Random Forest nổi bật với khả năng nhận diện lớp Positive cực kỳ chính xác, nhưng lại yếu hơn ở các lớp còn lại.
* Logistic Regression thể hiện hiệu suất tốt hơn Naive Bayes và có thể là lựa chọn chấp nhận được nếu không yêu cầu độ chính xác quá cao.
* Naive Bayes cho thấy hiệu suất thấp nhất và cần được cải thiện đáng kể nếu được sử dụng.

**Tóm lại :** Thông qua thực nghiêm từ nguồn dữ liệu công khai thực tế Amazon Food Reviews và được chọn lọc theo tiêu chí trong năm 2012, ta thu được kết quả và lựa chọn cuối cùng là mô hình LSTM.

## 4.3. So sánh với các nghiên cứu trước.

Kết quả thực nghiệm từ nghiên cứu này cho thấy mô hình **LSTM** đạt hiệu suất cao nhất trong việc phân loại cảm xúc thành ba lớp (tích cực, trung lập, tiêu cực) với độ chính xác 86.74%, đồng thời giữ được sự cân bằng giữa các nhãn qua các chỉ số precision, recall và F1-score. So sánh với các nghiên cứu trước, có thể nhận thấy một số điểm tương đồng và khác biệt đáng chú ý.

Trong nghiên cứu *“Text based Sentiment Analysis using LSTM”* của *Dr. G. S. N. Murthy et al*, LSTM cũng được chứng minh có độ chính xác vượt trội hơn so với các mô hình truyền thống như Naive Bayes và SVM. Nghiên cứu của nhóm tác giả này tập trung vào phân loại nhị phân (positive/negative), còn nghiên cứu hiện tại được xử lý mất cân bằng dữ liệu, còn mở rộng lên phân loại ba lớp, cho thấy tính khả thi của LSTM trong việc xử lý cảm xúc phức tạp hơn.

Ngoài ra, nghiên cứu *“On the Role of Text Preprocessing in Neural Network Architectures”* của *Jose Camacho-Collados và Mohammad Taher Pilehvar*đặt ra nghi vấn về sự cần thiết của các bước tiền xử lý truyền thống như lemmatization, stopword removal trong thời đại của các mô hình học sâu hiện đại. Trong nghiên cứu này, việc giữ lại một số bước tiền xử lý cơ bản (như chuyển về chữ thường, loại bỏ ký tự đặc biệt và tách từ) vẫn giúp cải thiện hiệu quả mô hình, cho thấy tiền xử lý vẫn đóng vai trò quan trọng tùy theo tính chất của dữ liệu và mô hình sử dụng.

Đối với mô hình tuyến tính, nghiên cứu *“Sentiment Analysis using Logistic Regression”* của *George B. Aliman et al*cho thấy Logistic Regression là một lựa chọn đơn giản nhưng hiệu quả trong các bài toán cảm xúc, đặc biệt khi dữ liệu không quá lớn. Trong nghiên cứu này, Logistic Regression đạt độ chính xác 79.14%, thấp hơn đáng kể so với LSTM, và cũng gặp khó khăn trong việc nhận diện cảm xúc trung lập – một điểm yếu được quan sát rõ trong báo cáo F1-score. Điều này cho thấy Logistic Regression vẫn là một baseline tốt, nhưng chưa đủ mạnh để xử lý các đặc điểm ngữ nghĩa sâu như mô hình học sâu.

Tương tự, nghiên cứu *“Multinomial Naive Bayes Classification Model for Sentiment Analysis”* của *Muhammad Abbas et al* nhấn mạnh sự đơn giản và hiệu quả của thuật toán MNB. Tuy nhiên, trong thực nghiệm hiện tại, Naive Bayes chỉ đạt 73.86% độ chính xác – mức thấp nhất trong số các mô hình được so sánh. Mặc dù đã được cân băng dữ liệu, và cơ bản dễ triển khai, chạy nhanh, MNB có xu hướng hoạt động kém hơn trong các bài toán phân loại ba lớp, đặc biệt khi dữ liệu chứa nhiều biểu hiện cảm xúc trung tính và cấu trúc ngôn ngữ phức tạp.

Cuối cùng, nghiên cứu *“Sentiment Analysis for Film Reviews Based on Random Forest”* của *Dean Francis Press* cho thấy Random Forest có độ chính xác cao nhờ khả năng tổng quát hóa tốt. Trong nghiên cứu này, Random Forest cũng đạt hiệu suất tương đối cao (89.18% accuracy), tuy nhiên recall của lớp neutral vẫn thấp, cho thấy vấn đề mất cân bằng cảm xúc không được xử lý tốt như trong mô hình LSTM (dù đã được xử lý mất cân bằng ở bước tiền xử lý).

Tổng kết, nghiên cứu này khẳng định rằng LSTM, với khả năng học các đặc trưng ngữ nghĩa theo chuỗi, vượt trội hơn rõ rệt các mô hình truyền thống trong bài toán phân tích cảm xúc đa lớp. Kết quả phù hợp và mở rộng các nghiên cứu trước đây, đóng góp thêm góc nhìn về vai trò của tiền xử lý, cấu trúc mô hình, và hiệu năng trên dữ liệu đánh giá sản phẩm thực tế.

## 4.4. Ý nghĩa thực tiễn.

Nghiên cứu này mang nhiều ý nghĩa thực tiễn cuộc sống trong bối cảnh mà ngành thương mại điện tử đang dần phát triển và thay thế các mô hình thương mại truyền thống và số lượng đánh giá về sản phẩm ngày một gia tăng :

* **Hỗ trợ người tiêu dùng :** Hệ thống phân tích cảm xúc giúp người dùng nhanh chóng nắm bắt xu hướng đánh giá chung nhất của một sản phẩm mà không phải bỏ ra quá nhiểu thời gian để tổng hợp, từ đó giúp tiết kiệm thòi gian và đưa ra quyết định nhanh chóng và hợp lý hơn.
* **Hỗ trợ doanh nghiệp và các nhà bán lẻ :** Tìm ra mô hình có tiềm năng phát triển nhất trong việc phân tích và tổng hợp ý kiến khách hàng giúp các doanh nghiệp tùy chỉnh kế hoạch marketing kịp thời, từ đó nâng cao chất lượng dịch vụ và trải nghiệm khách hàng.
* **Khả năng mở rộng và tích hợp với chatbot :** Mô hình LSTM có thể được triển khai trên các nền tảng thương mại điện tử lớn (Shopee, Lazada, Amazon,...) hoặc các hệ thống chăm sóc khách hàng.

## 4.5. Hạn chế và kế hoạch phát triển cho nghiên cứu tiếp theo.

**Hạn chế :**

* **Việc mất cân bằng dữ liệu :** Mặc dù đã sử dụng kỹ thuật Oversampling, nhưng tỷ lệ lớp neural vẫn thấp, ảnh hưởng đến hiệu quả phân loại và làm giảm thông số recall dẫn đến ảnh hưởng đến chỉ số F1-score của lớp này.
* **Mô hình LSTM cơ bản :** Mô hình sử dụng chưa được khai thác các kỹ thuật nâng caonhuw attetion, pretrained embedding (GloVe, BERT), hoặc các kiến trúc phức tạp hơn (BiLSTM, Transformer), do đó hiệu năng vẫn có thể cải thiện.
* **Cần nguồn tài nguyên GPU lớn :** Gây tiêu tốn khá nhiều tài nguyên mỗi lần chạy mô hình LSTM 🡪 gây tiêu tốn khá nhiều kinh phí.
* **Thời gian huấn luyện và suy luận :** Chỉ làm việc với tập tin như Amazon Fine Food Reviews (đã lọc trong năm 2012) mà đã tiêu tốn khá nhiều thời gian (~6 giờ) ảnh hưởng khá năng ứng dụng vào trong thực tế trên quy mô vừa và lớn.

Đề xuất nghiên cứu tiếp theo :

* **Mở rộng ngôn ngữ và miền dữ liệu :** Triển khai mô hình trên nhiề dữ liệu ngon ngữ khác nhau (như Tiếng Việt) và trên các ngành hàng khác như thời trang, công nghệ, dịch vụ ăn uống, giải trí,... để đánh giá tính khái quát.
* **Ứng dụng Embedding hiện đại :** Kết hợp LSTM với vecto từ học sẵn như Word2Vec, GloVe hoặc BERT để tăng khả năng học học ngôn ngữ giúp hiểu rõ ngữ cảnh hơn.
* **Thử nghiệm trên mô hình mới :** So sánh với các mô hình hiện đại hơn như BiLSTM, GRU, Transformer để cải thiện các thông số và hiệu quả phân loại.

# Chương 5 : Kết luận và hướng phát triển.

## 5.1. Kết luận.

Trong nghiên cứu này, nhóm đã thực hiện phân tích cảm xúc các đánh giá sản phẩm từ tập dữ liệu **Amazon Fine Food Reviews**. Mục tiêu chính là xây dựng một hệ thống phân loại cảm xúc văn bản thành ba nhóm: **tích cực**, **trung lập**, và **tiêu cực**. Dựa trên phương pháp nghiên cứu định lượng, nhóm đã tiến hành:

* Tiền xử lý dữ liệu văn bản (loại bỏ stopwords, chuyển chữ thường, token hóa, v.v.).
* Gán nhãn cảm xúc dựa trên điểm đánh giá (rating).
* Phân tích dữ liệu khám phá (EDA) nhằm hiểu rõ đặc điểm phân bố cảm xúc và độ dài văn bản.
* Biến đổi văn bản sang dạng số với tokenizer và embedding.
* Xử lý mất cân bằng dữ liệu bằng kỹ thuật oversampling.
* Triển khai và so sánh hiệu quả của bốn mô hình học máy và học sâu :
  + Multinomial Naive Bayes (**73,86%** )
  + Logistic Regression (**79,14%**)
  + Random Forest (**89,18%**)
  + LSTM (**86,74%**)

Kết quả thực nghiệm cho thấy mô hình LSTM đạt hiệu quả phân loại cao nhất, khẳng định sức mạnh của mạng nơ-ron hồi tiếp (RNN) trong việc nắm bắt ngữ cảnh và xử lý chuỗi văn bản. Đồng thời, các mô hình truyền thống như Logistic Regression và Random Forest vẫn giữ vai trò quan trọng với độ chính xác cạnh tranh và thời gian huấn luyện nhanh chóng.

## 5.2. Câu hỏi nghiên cứu.

**Các đặc trưng ngôn ngữ nào thường xuất hiện trong đánh giá tích cực, tiêu cực và trung tính ?**

Dựa trên kết quả phân tích từ tập dữ liệu và các kỹ thuật như WordCloud, n-grams, và tần suất từ :

**Đánh giá tích cực** :

* Thường chứa các từ mang cảm xúc rõ ràng và mang tính khen ngợi như: *"delicious"*, *"excellent"*, *"amazing"*, *"love"*, *"perfect"*, *"best"*, *"great"*, *"tasty"*.
* Câu thường có cấu trúc đơn giản, tích cực, nhiều tính từ miêu tả tích cực.
* Thường xuất hiện dấu chấm than (!), biểu hiện sự hài lòng cao.

**Đánh giá tiêu cực** :

* Thường chứa từ mang nghĩa chê bai như: *"bad"*, *"terrible"*, *"disappointed"*, *"worst"*, *"awful"*, *"waste"*, *"hate"*.
* Câu thường mang cấu trúc phủ định hoặc chứa cảm xúc mạnh.
* Có xu hướng sử dụng từ cảm thán tiêu cực và dấu chấm than (!).

**Đánh giá trung tính** :

* Thường chứa các từ mô tả sản phẩm một cách khách quan, ít cảm xúc như: *"arrived"*, *"package"*, *"received"*, *"price"*, *"quantity"*, *"shipping"*.
* Câu có xu hướng đưa thông tin hơn là thể hiện cảm xúc.
* Dễ gây nhầm lẫn với đánh giá tích cực/tiêu cực nhẹ nếu chỉ nhìn bề ngoài.

**Mô hình học máy nào cho hiệu quả tốt nhất trong việc phân loại đánh giá sản phẩm ?**

Trong khuôn khổ nghiên cứu, ba mô hình học máy truyền thống được thử nghiệm: Multinomial Naive Bayes, Logistic Regression, và Random Forest. Kết quả như sau :

* **Random Forest** đạt hiệu quả tốt nhất trong số các mô hình học máy với độ chính xác **89,18%** :
* Mạnh trong xử lý đặc trưng không tuyến tính.
* Giảm thiểu overfitting thông qua trung bình hóa nhiều cây quyết định.
* Phù hợp với các bài toán phân loại nhiều lớp và dữ liệu có nhiễu nhẹ.

**Có thể thay thế mô hình học máy tốt nhất sau khi nghiên cứu bằng mô hình học sâu để giải quyết vấn đề hay không và hiệu quả ra sao?**

Có thể thay thế, và việc sử dụng mô hình học sâu như **LSTM** là một lựa chọn hợp lý trong bối cảnh phù hợp. Trong nghiên cứu này, **Random Forest** là mô hình học máy đạt độ chính xác cao nhất (Accuracy 89,18%), cho thấy khả năng tổng quát hóa tốt và hiệu quả trong phân loại đánh giá sản phẩm.

Tuy nhiên, mô hình **LSTM** – dù có độ chính xác tổng thể thấp hơn (86,74%) – lại vượt trội hơn về Recall, đặc biệt ở lớp cảm xúc tiêu cực. Điều này có ý nghĩa thực tiễn quan trọng trong các hệ thống phân tích đánh giá, nơi việc không bỏ sót các phản hồi tiêu cực từ người dùng là ưu tiên hàng đầu (ví dụ trong quản lý dịch vụ khách hàng, đánh giá chất lượng sản phẩm).

Bên cạnh đó, **LSTM** có khả năng học ngữ cảnh tuần tự và mối quan hệ ngữ nghĩa phức tạp trong câu, điều mà các mô hình học máy truyền thống khó nắm bắt được khi chỉ dựa vào đặc trưng thủ công như TF-IDF hoặc Bag-of-Words.

Do đó, mô hình học sâu có thể thay thế mô hình học máy trong nhiều tình huống, đặc biệt khi:

* Có tập dữ liệu lớn và đa dạng.
* Yêu cầu hiểu ngữ nghĩa sâu hoặc xử lý ngôn ngữ tự nhiên phức tạp.
* Mục tiêu là tối ưu hóa Recall hoặc hiệu suất trên các lớp nhạy cảm.

Tuy nhiên, cần lưu ý rằng **LSTM** tốn thời gian huấn luyện nhiều hơn, yêu cầu tài nguyên tính toán cao hơn, và cần tối ưu hóa tốt hơn để đạt hiệu quả vượt trội so với các mô hình truyền thống.

## 5.3. Hạn chế của nghiên cứu hiện tại.

* **Thời gian huấn luyện lâu**: Mô hình LSTM có độ trễ cao trong cả huấn luyện và suy luận, gây trở ngại khi áp dụng vào các hệ thống thời gian thực hoặc trên thiết bị có tài nguyên hạn chế.
* **Dữ liệu mất cân bằng**: Dù đã áp dụng kỹ thuật oversampling, dữ liệu cảm xúc vẫn chưa thật sự cân bằng giữa ba lớp, có thể ảnh hưởng đến độ chính xác tổng thể.
* **Chưa sử dụng embedding nâng cao**: Nghiên cứu sử dụng embedding cơ bản. Các mô hình embedding hiện đại như GloVe, Word2Vec hoặc BERT chưa được áp dụng.
* **Không kiểm tra hiệu suất mô hình trên dữ liệu thời gian thực**: Việc đánh giá chỉ thực hiện trên tập dữ liệu tĩnh, chưa kiểm tra khả năng tổng quát hóa hoặc áp dụng trong tình huống thực tế.

## 5.4. Hướng phát triển.

Trong tương lai, nghiên cứu có thể mở rộng theo các hướng sau :

* Tối ưu hóa mô hình LSTM về thời gian huấn luyện và inference bằng cách:
  + Giảm chiều dài chuỗi đầu vào (truncation).
  + Áp dụng early stopping và tuning siêu tham số.
  + Chuyển sang GPU hoặc dùng TensorFlow Lite khi triển khai thực tế.
* Thử nghiệm mô hình ngôn ngữ hiện đại như BERT, RoBERTa hoặc DistilBERT để cải thiện khả năng hiểu ngữ cảnh sâu hơn.
* Xây dựng ứng dụng thực tế như hệ thống đánh giá tự động hoặc dashboard phân tích cảm xúc cho doanh nghiệp thương mại điện tử.
* Mở rộng sang đa ngôn ngữ để ứng dụng vào đánh giá sản phẩm trên các nền tảng quốc tế.
* Kết hợp đặc trưng phi văn bản như thời gian đăng đánh giá, độ dài đánh giá, hoặc sản phẩm cụ thể nhằm cải thiện độ chính xác của mô hình.

# Tài liệu tham khảo

[1] Dr. G. S. N. Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, Mounika Belusonti, and Aditya Institute of Technology and Management, “Text based Sentiment Analysis using LSTM,” *Int. J. Eng. Res.*, vol. V9, no. 05, p. IJERTV9IS050290, May 2020, doi: 10.17577/IJERTV9IS050290.

[2] J. Camacho-Collados and M. T. Pilehvar, “On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis,” Aug. 23, 2018, *arXiv*: arXiv:1707.01780. doi: 10.48550/arXiv.1707.01780.

[3] “(PDF) Sentimental Analysis using Logistic Regression.” Accessed: May 02, 2025. [Online]. Available: https://www.researchgate.net/publication/353514662\_Sentimental\_Analysis\_using\_Logistic\_Regression?utm\_source=chatgpt.com

[4] M. Abbas, A. Kamran, Memon, A. A. Jamali, Saleemullah Memon, and Anees Ahmed, “Multinomial Naive Bayes Classification Model for Sentiment Analysis,” 2019, *Unpublished*. doi: 10.13140/RG.2.2.30021.40169.

[5] D. Zheng, “Sentiment Analysis for Film Reviews Based on Random Forest,” *Sci. Technol. Eng. Chem. Environ. Prot.*, vol. 1, no. 7, Jun. 2024, doi: 10.61173/5t8epb44.

[6] “SNAP: Web data: Amazon Fine Foods reviews.” Accessed: May 18, 2025. [Online]. Available: https://snap.stanford.edu/data/web-FineFoods.html

[7] J. Brownlee, A. Tam, and Z. M. Chng, “Deep Learning with Python, Second Edition”.

[8] J. Brownlee, “Machine Learning Mastery With Python”.

[9] T. Nguyen, “Topic Modelling and Sentiment Analysis of Customer Reviews for B2C E-commerce Platforms in Vietnam: A Comparative Study of Lazada, Shopee, Tiki, and Sendo,” *Sr. Indep. Study Theses*, Jan. 2023, [Online]. Available: https://openworks.wooster.edu/independentstudy/10623

[10] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends® Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, Jul. 2008, doi: 10.1561/1500000011.

[11] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc., 2009.

1. Vanishing gradient là hiện tượng gradient trở nên rất nhỏ khi lan truyền ngược trong mạng nơ-ron, khiến mô hình không học được các quan hệ xa trong chuỗi, đặc biệt nghiêm trọng với RNN. LSTM ra đời để khắc phục vấn đề này nhờ cơ chế bộ nhớ và các cổng điều khiển thông tin. [↑](#footnote-ref-1)