# Phân tích đánh giá sản phẩm trên nền tảng thương mại điện tử bằng Machine Learning

**Nguyễn Gia Thành, Nguyễn Dương Khang, Nguyễn Quang Dương**

## Giới thiệu

### Định nghĩa vấn đề:

- **Input:** Dữ liệu dạng văn bản (các đánh giá sản phẩm).
- **Output:** Đưa ra nhận định về đánh giá đó chia làm 3 loại (positive, neutral, negative).

### Thách thức:

- Đa dạng ngôn ngữ & giọng điệu
- Dữ liệu không cân bằng
- Nhiễu & lỗi chính tả
- Khó nhận diện ý định thầm lặng
- Quá khớp (Overfitting)

### Mục tiêu:

- Xây dựng và tìm kiếm mô hình có hiệu suất cao trong xử lý, phân tích, phân loại các đánh giá trên nền tảng thương mại điện tử từ tệp dư liệu Amazon Fine Food Reviews trên Kaggle.

## Dữ liệu

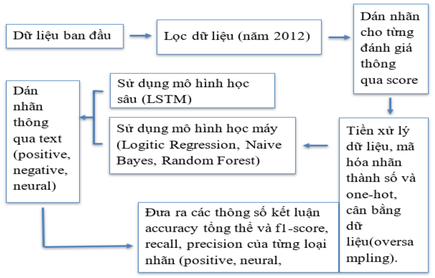
- Số lượng: 198,659 đánh giá (trên tổng số hơn 500,000 đánh giá).
- Nguồn tập dữ liệu: Kaggle

https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

## PHƯƠNG PHÁP NGHIÊN CỨU

### 1.Tiền xử lý dữ liệu

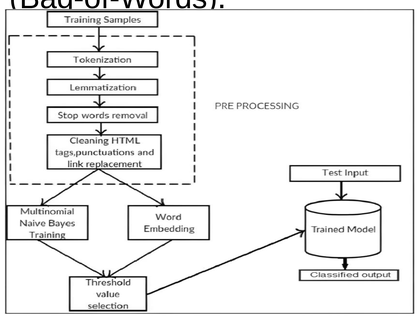
- Dán nhãn theo 'Score': 1,2 sao → negative; 3 sao → neural; 4,5 sao → positive.
- Giữ nguyên văn bản gốc (Raw text)
- Chuyển chữ hoa → thường (Lowercasing)
- Loại bỏ stopwords ("i", "have", "of", "and",…)
- Lemmatization (rút gọn từ về gốc running -> run)
- Xử lý mất cân bằng (Oversampling): tư động tăng số mẫu của negative và neural.




## 2.Mô hình.

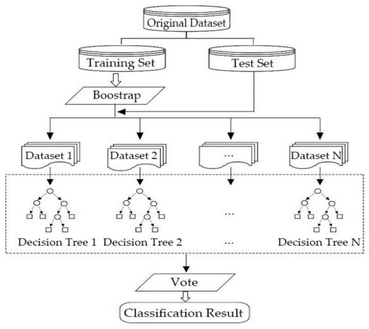
### 2.1. Multinomial Naïve Bayes.

Naive Bayes là một mô hình học máy/ mô hình thống kê dựa trên định lý Bayes. 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).



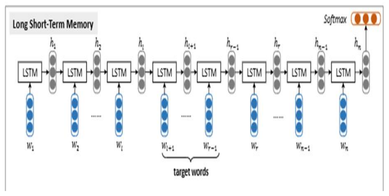

### 2.2. Random Forest.

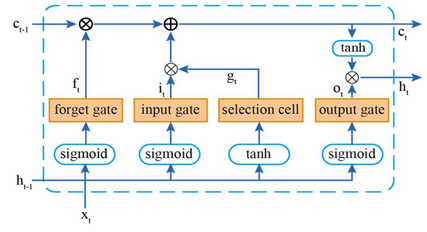
**Random Forest** là một thuật toán học máy thuộc nhóm ensemble learning, được sử dụng phổ biến trong các bài toán phân loại và hồi quy. Thuật toán này xây dựng nhiều cây quyết định (decision trees) độc lập trên các mẫu dữ liệu con (bootstrap samples) và tập con các đặc trưng, sau đó kết hợp kết quả 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).




### 2.3. LSTM.

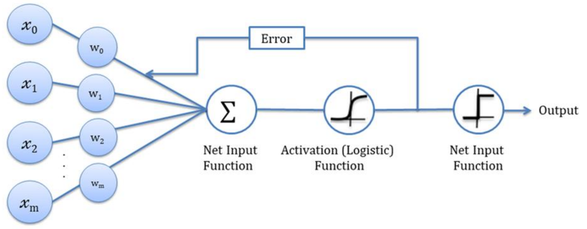
Long Short-Term Memory (LSTM) là một kiến trúc mạng nơ-ron hồi tiếp (RNN) được thiết kế nhằm khắc phục vấn đề mất mát thông tin dài hạn (vanishing gradient) trong các mạng RNN truyền thống.




### 2.4. 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.

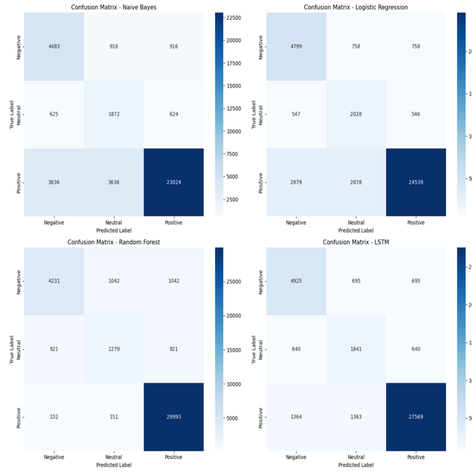



## KẾT QUẢ ĐẠT ĐƯỢC

Kết quả huấn luyện các mô hình khác nhau trên tập dữ liệu Amazon Food Reviews.

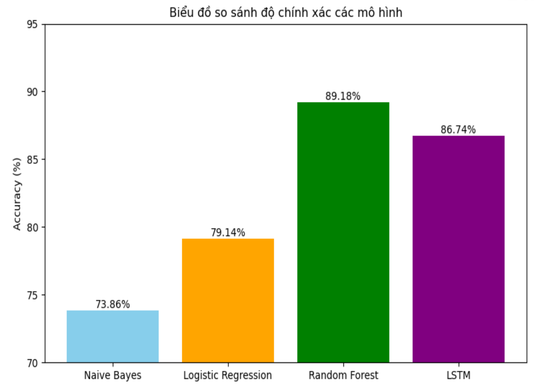
| Mô hình | Nhãn | Precision | Recall | F1-score |
|---|---|---|---|---|
| Naive Bayes | Positive | 0.96 | 0.76 | 0.85 |
| | Neutral | 0.24 | 0.60 | 0.34 |
| | Negative | 0.58 | 0.71 | 0.64 |
| Logistic Regression | Positive | 0.96 | 0.81 | 0.88 |
| | Neutral | 0.30 | 0.65 | 0.41 |
| | Negative | 0.66 | 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 |


| 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 |


- **"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.
- Mối tương quan mạnh nhất trong biểu đồ này là giữa **WordCount** và **TextLength**.

### Ma trận phân tán




**Độ đo đánh giá:** độ đo (metric) của bài là độ chính xác (accuracy)