PROPOSAL

**Nhóm 8**

|  |  |
| --- | --- |
| Họ và tên | MSSV |
| Trần Đại Nhân | 22133038 |
| Nguyễn Ngọc Thành | 22133052 |
| Nguyễn Văn Thắng | 22133054 |

**ĐỀ TÀI:**

PHÂN TÍCH & DỰ ĐOÁN CẢM XÚC REVIEW SẢN PHẨM.

## 1. Giới thiệu về bài toán và dữ liệu

**- Bài toán:**

**Bối cảnh:**

Bài toán phân tích cảm xúc (Sentiment Analysis) là một ứng dụng quan trọng trong xử lý ngôn ngữ tự nhiên (NLP).

Trong thời đại thương mại điện tử phát triển mạnh mẽ, hàng triệu đánh giá sản phẩm (review) được đăng tải mỗi ngày trên các nền tảng như Amazon, Shopee, Tiki. Những đánh giá này chứa đựng nhiều thông tin giá trị về trải nghiệm và cảm xúc của khách hàng.

Vấn đề đặt ra:

Việc đọc và phân tích thủ công số lượng lớn review là không khả thi, do đó cần một mô hình tự động có thể phân loại cảm xúc trong các đánh giá sản phẩm.

**Mục tiêu:**

Xây dựng một mô hình có khả năng phân loại các đánh giá sản phẩm thành hai nhãn: tích cực và tiêu cực.

Ngoài ra, nhóm mong muốn mở rộng theo hướng Aspect-based Sentiment Analysis (ABSA) nếu có khả năng, nhằm phân tích cảm xúc chi tiết hơn dựa trên các khía cạnh cụ thể của sản phẩm như chất lượng, giá cả, dịch vụ.

**Ứng dụng thực tế:**

* Phân tích phản hồi khách hàng.
* Gợi ý sản phẩm phù hợp.
* Cải thiện dịch vụ và chiến lược Marketing.

**- Dữ liệu:**

Bộ dữ liệu này bao gồm các đánh giá sản phẩm của khách hàng trên Amazon – một trong những nền tảng bán lẻ trực tuyến lớn nhất thế giới. Dữ liệu phản ánh trải nghiệm thực tế của khách hàng thông qua điểm đánh giá, tiêu đề review, nội dung chi tiết, cùng với nhiều thông tin metadata đi kèm. Đây là nguồn dữ liệu giá trị để phân tích mức độ hài lòng, cảm xúc và xu hướng của khách hàng.

Mô tả các cột dữ liệu chính:

* **Reviewer Name**: Tên người đánh giá.
* **Profile Link**: Đường dẫn đến hồ sơ người dùng, hỗ trợ khai thác thêm thông tin.
* **Country**: Quốc gia của người đánh giá.
* **Review Count**: Số lượng review mà người dùng đã đăng, phản ánh mức độ tương tác.
* **Review Date**: Thời gian review được đăng tải, hữu ích cho phân tích theo dòng thời gian.
* **Rating**: Điểm số thể hiện mức độ hài lòng.
* **Review Title**: Tiêu đề tóm tắt nội dung và cảm xúc chính của review.
* **Review Text**: Nội dung chi tiết phản ánh trải nghiệm của khách hàng.
* **Date of Experience**: Thời điểm khách hàng sử dụng sản phẩm/dịch vụ.

Bộ dữ liệu quy mô lớn và đa dạng này là một nguồn tài nguyên quý giá cho nhiều phân tích chuyên sâu, hỗ trợ nâng cao mức độ tương tác của khách hàng và tối ưu hóa chiến lược kinh doanh.

## 2. Kế hoạch phân tích dữ liệu

* **Input (X):** Văn bản review
* **Output (Y):** Nhãn cảm xúc (tích cực / tiêu cực).
* **Các độ đo đánh giá:** Accuracy, Precision, Recall, F1-score, AUC.
* **Các phương pháp dự kiến sử dụng:**
  + Tiền xử lý dữ liệu văn bản: làm sạch dữ liệu, loại bỏ stopwords, chuẩn hóa văn bản.
  + Trích xuất đặc trưng: Bag-of-Words, TF-IDF, Word Embeddings (Word2Vec, GloVe, hoặc Transformer-based embeddings).
  + Mô hình học máy: Logistic Regression, SVM, Random Forest.
  + Mô hình deep learning: LSTM, BiLSTM, BERT (fine-tuning).

Chia dữ liệu & Validation

* Tỷ lệ: Train 70% – Validation 15% – Test 15%
* Phương pháp kiểm định:
  + Holdout Validation: Dùng tập riêng để đánh giá.
  + Cross Validation (k-fold): Đảm bảo mô hình tổng quát, tránh overfitting.

## **Tinh chỉnh Hyperparameter**

* Mục tiêu: Tối ưu hiệu suất mô hình.
* Cách làm:
  + Dùng GridSearchCV hoặc RandomSearchCV cho Logistic Regression, SVM, RF.
  + Với Deep Learning: tinh chỉnh số layer, hidden size, dropout, learning rate.
* Tiêu chí chọn mô hình: F1-score & AUC cao nhất trên tập validation.

#### Kế hoạch thực hiện

* **Tuần 1 (29/09 – 05/10):** Làm sạch dữ liệu, khám phá dữ liệu (EDA), phân tích phân bố rating và chuẩn hóa dữ liệu đầu vào.
* **Tuần 2 (06/10 – 12/10):** Tiền xử lý văn bản (tokenization, stopwords removal, stemming/lemmatization), xây dựng đặc trưng (TF-IDF, Word2Vec).
* **Tuần 3 (13/10 – 19/10):** Huấn luyện mô hình ML (Logistic Regression, SVM, Random Forest), thử nghiệm mô hình DL cơ bản (LSTM/BiLSTM).
* **Tuần 4 (20/10 – 26/10):** Fine-tuning BERT, đánh giá – so sánh mô hình, trực quan hóa kết quả và hoàn thiện báo cáo + slide thuyết trình.

#### Phân công công việc

* **Thành viên 1:** Tiền xử lý dữ liệu, EDA, trực quan hóa kết quả.
* **Thành viên 2:** Trích xuất đặc trưng, huấn luyện mô hình học máy.
* **Thành viên 3:** Xây dựng mô hình deep learning (LSTM, BERT), đánh giá và viết báo cáo cuối cùng.