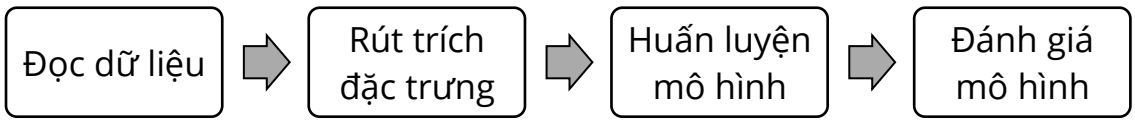
Bài thực hành 6. MÔ HÌNH PHÂN LỚP (Phần 2)

1. PHÂN LỚP VĂN BẢN

Bài toán: Nhận dạng cảm xúc từ câu phản hồi bằng văn bản của người dùng

- Phát biểu bài toán:
  - Input: Câu bình luận của người dùng.
  - Output: Nhãn cảm xúc, gồm 1 trong 3 nhãn: tích cực, tiêu cực và trung tính.
- Bộ dữ liệu: **UIT-VSFC**.
  - Công bố khoa học: K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong and N. L. Nguyen, "UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis", KSE 2018.
  - Link tải: https://drive.google.com/drive/folders/1xclbjHHK58zk2X6iqbvMPS2rcy9y9E0X?usp=drive_open.
  - Sử dụng cho các tác vụ **sentiment-based** và **topic-based**.
- Các bước thực hiện tổng quát:



  - Đọc dữ liệu:

    Dữ liệu được lưu trữ trên các file .txt, đã được phân chia sẵn thành các tập huấn luyện (train) và kiểm thử (test).

```
import pandas as pd
X_train = pd.read_csv('UIT-VSFC/train/sents.txt',  sep='\r\n',
header=None, index_col=None)
y_train = pd.read_csv('UIT-VSFC/train/sentiments.txt',  sep='\r\n',
header=None, index_col=None)
X_test = pd.read_csv('UIT-VSFC/test/sents.txt',  sep='\r\n',
header=None, index_col=None)
y_test = pd.read_csv('UIT-VSFC/test/sentiments.txt',  sep='\r\n',
header=None, index_col=None)
```

2. BÀI TẬP

- **Bài tập 1.** Đọc dữ liệu và cho biết các thông tin sau:

a) Mục tiêu/tác vụ mà bộ dữ liệu hướng tới là gì?
b) Kích thước của tập train và test?
c) Phân bố nhãn của tập train và test. Vẽ biểu đồ cột thể hiện sự phân bố nhãn trên từng tập dữ liệu.

- **Bài tập 2.** Dùng thư viện **CountVectorizer** để trích xuất đặc trưng từ dữ liệu văn bản cho tác vụ sentiment-based.
- **Bài tập 3.** Huấn luyện mô hình:
  a) Sử dụng mô hình **Logistic Regression** và **SVM** để huấn luyện.
  b) Đánh giá mô hình bằng các độ đo accuracy_score, precision, recall và macro f1-score. So sánh hiệu năng của 2 mô hình.
  c) Vẽ ma trận nhầm lẫn (confusion matrix) của 2 mô hình. Có nhận xét gì về ma trận nhầm lẫn giữa 2 mô hình vừa huấn luyện.
  Gợi ý: Code vẽ ma trận nhầm lẫn:

```
sn.set(font_scale=1.4)
sn.heatmap(cf, annot=True, annot_kws={"size": 16}, fmt='d')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

- **Bài tập 4.** Thay **CountVectorizer** bằng **TfidfVectorizer** để trích xuất đặc trưng cho dữ liệu văn bản. So sánh hiệu năng giữa 2 phương pháp trích xuất đặc trưng đối với các mô hình phân lớp?
- **Bài tập 5.** Lưu mô hình đã huấn luyện thành file.
- **Bài tập 6.** Sử dụng thêm công cụ tách từ (word segmentation) cho tiếng Việt. Việc tách từ có ảnh hưởng tới hiệu năng của mô hình hay không?
- **Bài tập 7*.** Tìm cách điều chỉnh 2 siêu tham số của phương pháp trích xuất đặc trưng TfidfVectorizer là lowercase (true và false) và ngram_ range (1 đến 3). Chọn ra bộ siêu tham số tốt nhất, và cho biết kết quả độ chính xác của mô hình trên bộ siêu tham số đó.
- **Bài tập 8*.** Liệt kê các trường hợp dự đoán sai của mô hình. Cho biết nguyên nhân nào mô hình bị dự đoán sai. Đề xuất biện pháp khắc phục.