Section 1:

1. Chuẩn bị dữ liệu

- Sử dụng bộ dữ liệu nhận xét (review) được gán nhãn cảm xúc tích cực, tiêu cực được lấy trên trang foody

2. Loading Dataset

- ***Text normalization (chuẩn hóa dữ liệu):*** Công đoạn loại bỏ các thành phần không cần thiết từ dữ liệu, có thể hiểu là làm sạch dữ liệu xóa đi dữ liệu rác cuối cùng nhận được đoạn văn bản chỉ có text. Ví dụ, xóa đi tag HTML, xóa link, xóa ký tự đặc biệt "\n \t &#64",...
- ***Data preprocessing (tiền xử lý dữ liệu):*** Chuyển dữ liệu/ văn bản nhận được ở giai đoạn trên thành dữ liệu đầu vào (data input) thích hợp cho đúng với mô hình (model machine learning) sử dụng phân loại. Ví dụ, các công việc cần thực hiện trước khi đưa vào thuật toán phân loại văn bản tiếng Việt như: tách từ, chuẩn hóa từ, loại bỏ stopwords, vertor hóa từ. Đây là công đoạn quan trọng trong bài toán phân loại văn bản.

3. Feature Engineering

- Ở bước này, chúng ta sẽ đưa dữ liệu dạng văn bản đã được xử lý về dạng vector thuộc tính có dạng số học. Có nhiều các khác nhau để đưa dữ liệu văn bản dạng *text* về dữ liệu dạng số mà chúng ta có thể thực hiện như:
- ***TF-IDF Vectors as features (Word level)***

```
# word level - we choose max number of words equal to 30000 except all words (100k+ words)
tfidf_vect = TfidfVectorizer(analyzer='word', max_features=30000)
tfidf_vect.fit(X_data) # learn vocabulary and idf from training set
X_data_tfidf =  tfidf_vect.transform(X_data)
# assume that we don't have test set before
X_test_tfidf =  tfidf_vect.transform(X_test)

from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=300, random_state=42)
svd.fit(X_data_tfidf)

X_data_tfidf_svd = svd.transform(X_data_tfidf)
X_test_tfidf_svd = svd.transform(X_test_tfidf)
```

4. Learn/train model (chọn model machine learning và huấn luyện):

- Naïve bayes, kNN, SVM

5. Evaluation/results (đánh giá kết quả):

- Công đoạn cuối, đánh giá kết quả nhận được, sử dụng Precision, Recall và F1-Score, accuracy

Độ chính xác (accuracy):

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total}$$

Phương pháp Precision:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

.

Phương pháp Recall

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

6. Test model, predict

7. So sanh kết quả của các mô hình đạt được và chọn mô hình.