1. Khởi tạo dữ liệu

- Đọc các comment vào biến X

- Gắn nhãn tương ứng (Tốt, Xấu) vào biến y

2. Tiền xử lí dữ liệu

- One-hot y

- Loại các stop word, các kí tự đặc biệt, dấu câu, xuống dòng,…. => một comment “sạch”

- Tokenizer dữ liệu (Do các comment đều là các từ nên ko thể fit ngay vào model được) ta sẽ mã hóa bằng số

VD: Fox go to the fox town => 1 2 3 4 1 5

- để cho tất cả các dữ liệu có cùng độ dài, ta cho qua “pad\_sequences”

VD: pad\_sequences([[1, 2, 3], [3, 4, 5, 6], [7, 8]])

=> array([[0, 1, 2, 3],

[3, 4, 5, 6],

[0, 0, 7, 8]], dtype=int32)

- tách X và y ra thành X\_train, y\_train, X\_test, y\_test (test size: 0.2)

3. Build model

Layer 1: Embedding(input: vocalbulary\_size, output: 16)

Layer 2: GRU(8 units, dropout=0.2, recurrent\_dropout=0.2)

Layer 3: Dense(2) tương ứng với 2 class “Tốt” và “Xấu”

Complier(Adam, loss=categorical\_crossentropy)

4. Train model với batch\_size= 30

Kết quả: accuraccy: 92% sau 7 epochs