ĐỒ ÁN TỐT NGHIỆP

Phân loại văn bản theo chủ đề ứng dụng học máy

Giảng viên hướng dẫn: TS. Ninh Khánh Duy

Sinh viên thực hiện: Nguyễn Trung Hiếu

Nội dung

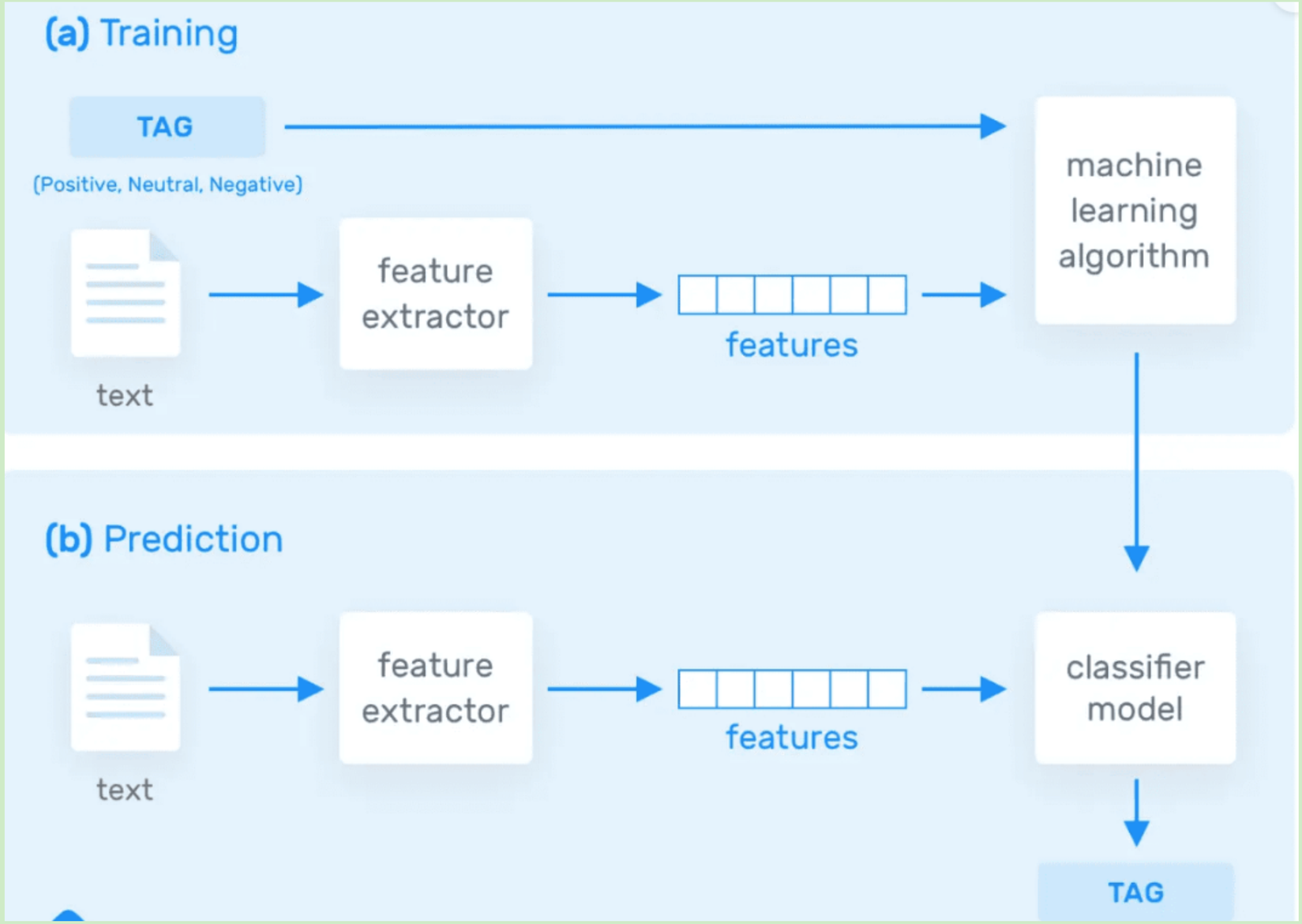
1. Tổng quan về phân loại văn bản
2. Thu thập dữ liệu và xử lý dữ liệu
3. Xây dựng mô hình
4. Kết luận

Tổng quan về phân loại văn bản

khái niệm

- Phân loại văn bản là bài toán thuộc học máy có giám sát
- Bài toán này yêu cầu có nhãn
- Mô hình sẽ học từ dữ liệu có nhãn đó, sau đó được dùng để dự đoán nhãn cho các dữ liệu mới mà mô hình chưa gặp.

Tổng quan về bài toán phân loại văn bản



Tổng quan về phân loại văn bản

Phát biểu bài toán

- Xây dựng mô hình phân loại văn bản tin tức tiếng Việt.

- Đầu vào : là nội dung của một bài báo

- Đầu ra : là chủ đề của văn bản đó

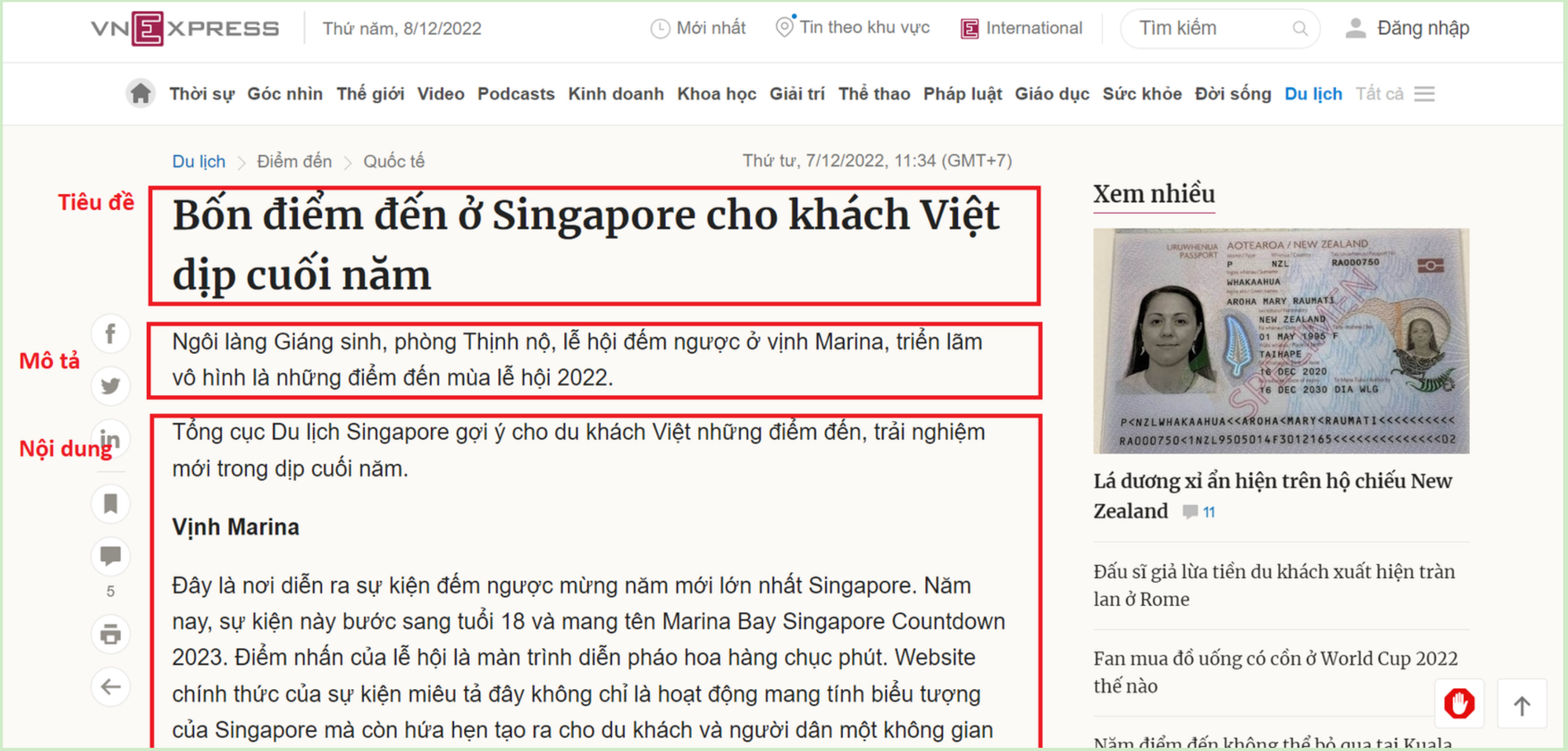
- Các chủ đề bao gồm: Chính trị xã hội, đời sống, kinh tế,sức khỏe, pháp luật...

Thu thập dữ liệu

- Nguồn dữ liệu được thu thập trên trang web VnExpress.vn
- Đây là trang báo điện tử do tập đoàn FPT thành lập và ra mắt công chúng vào năm 2001. Là một trang web được tin cậy và có nền tảng lâu đời

Thu thập dữ liệu

Với mỗi bài báo sẽ thu thập bao gồm 3 thuộc tính



Thu thập dữ liệu

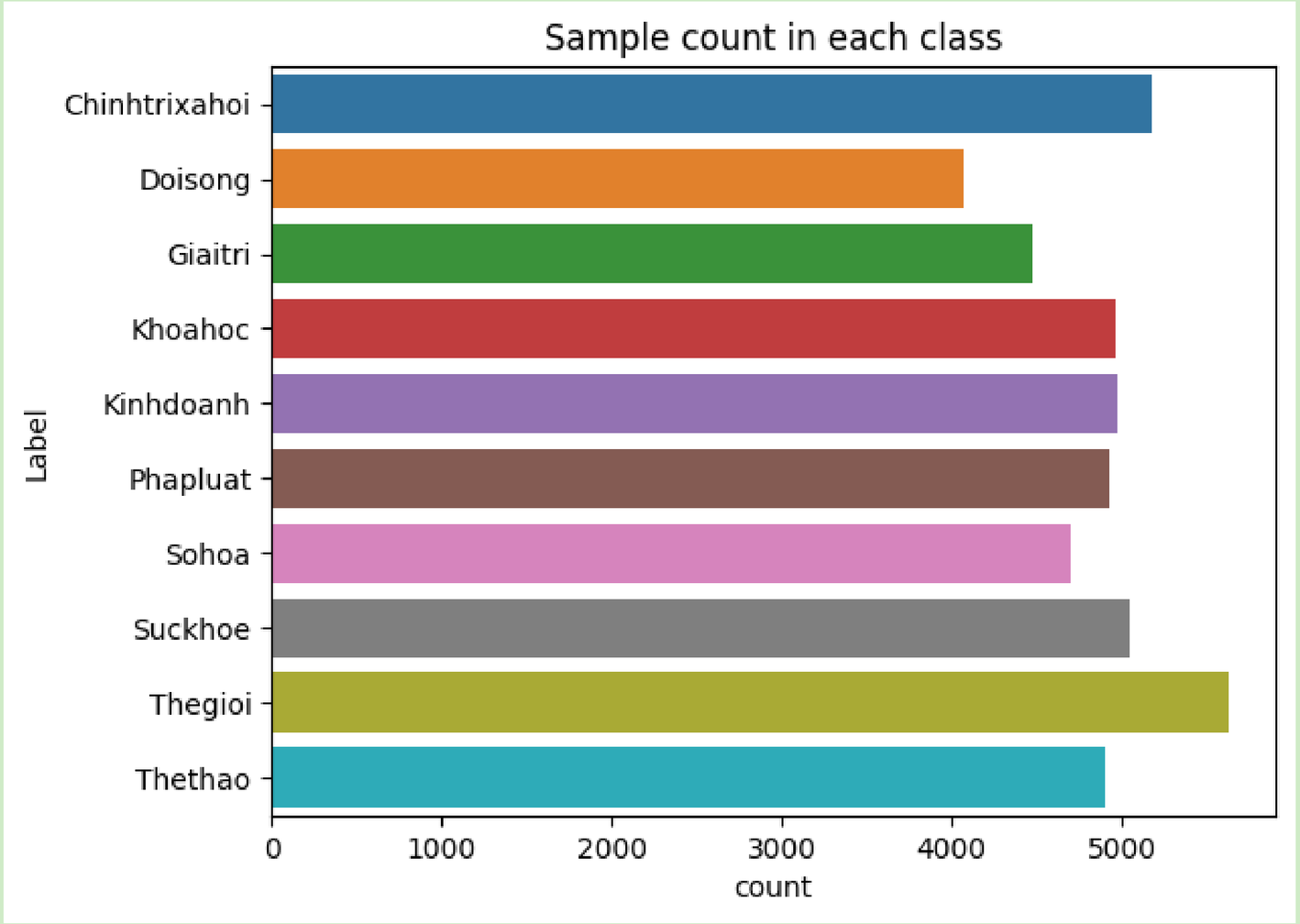
Dữ liệu thu thập bao giồm 10 chủ đề tương ứng với danh mục trong trang web VnExpress.vn.
Thu thập 5000 bài báo cho mỗi chủ đề

Nhãn dữ liệu	Chỉ mục ở trang VnExpress
Chinhtrixahoi	Thời sự
Doisong	Đời sống
Giaitri	Giải trí
Khoahoc	Khoa học
Kinhdoanh	Kinh doanh
Phapluat	Pháp luật
Sohoa	Số hóa
Suckhoe	Sức khỏe
Thegioi	Thế giới
Thethao	Thể thao

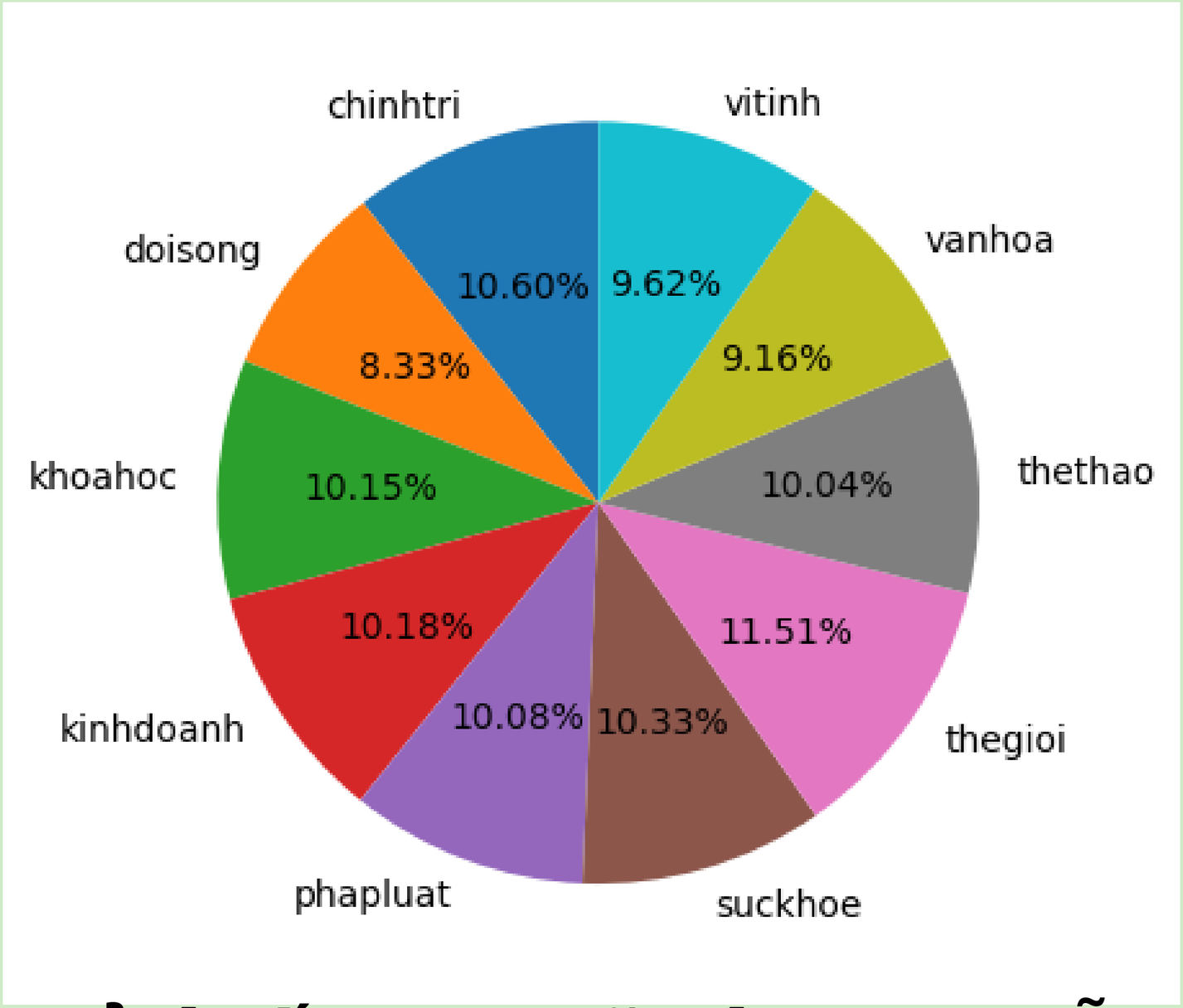
Xử lý dữ liệu

1. Chuẩn hóa kiểu gõ dấu tiếng Việt (dùng òa úy thay cho oà uý)
2. Thực hiện tách từ tiếng Việt (sử dụng thư viện tách từ như pyvi, underthesea, vncorenlp,…)
3. Đưa về văn bản lower (viết thường)
4. Xóa các ký tự đặc biệt: “.”, “,”, “;”, “)
5. Loại bỏ các stopword

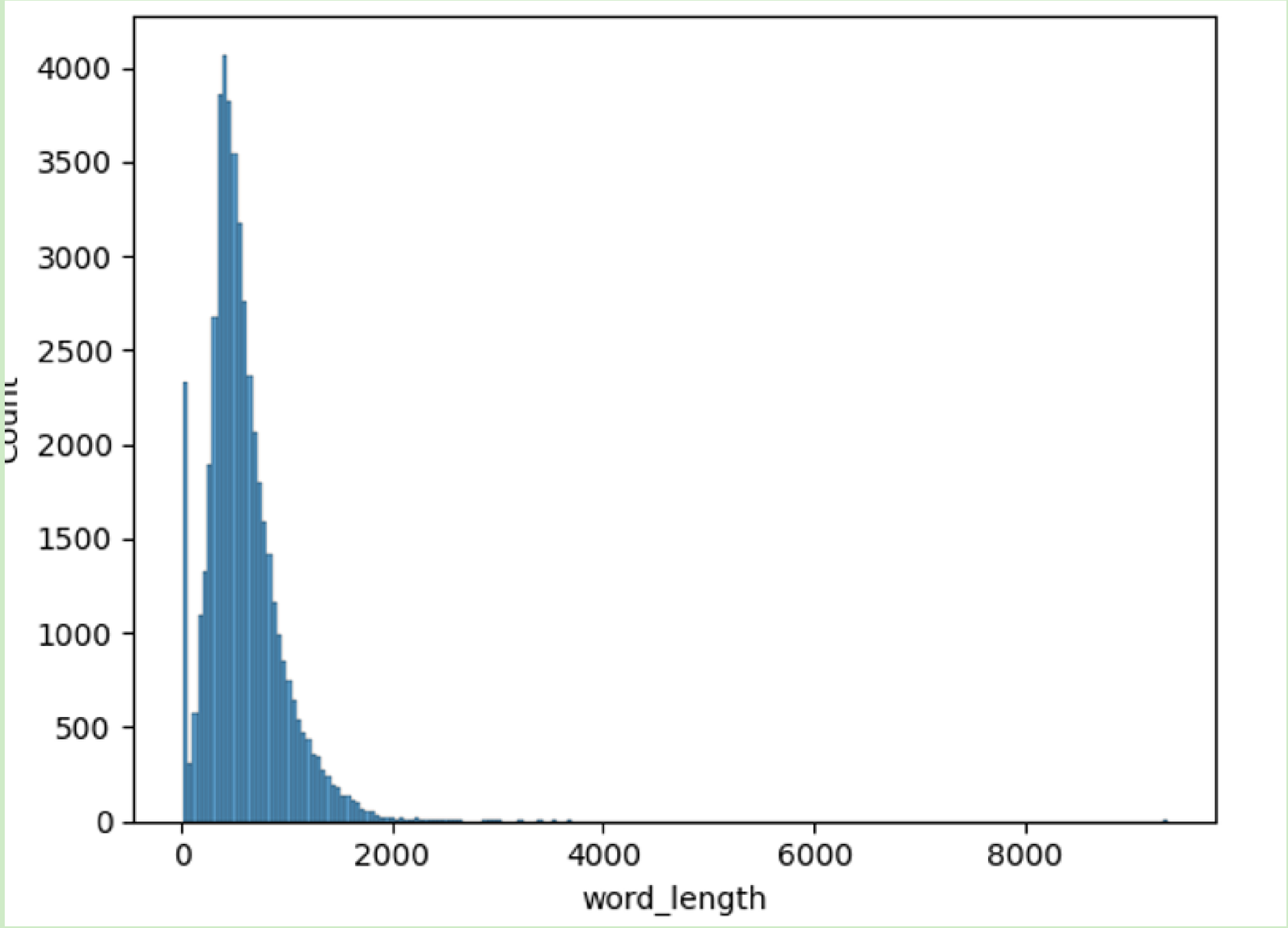
mô tả dữ liệu



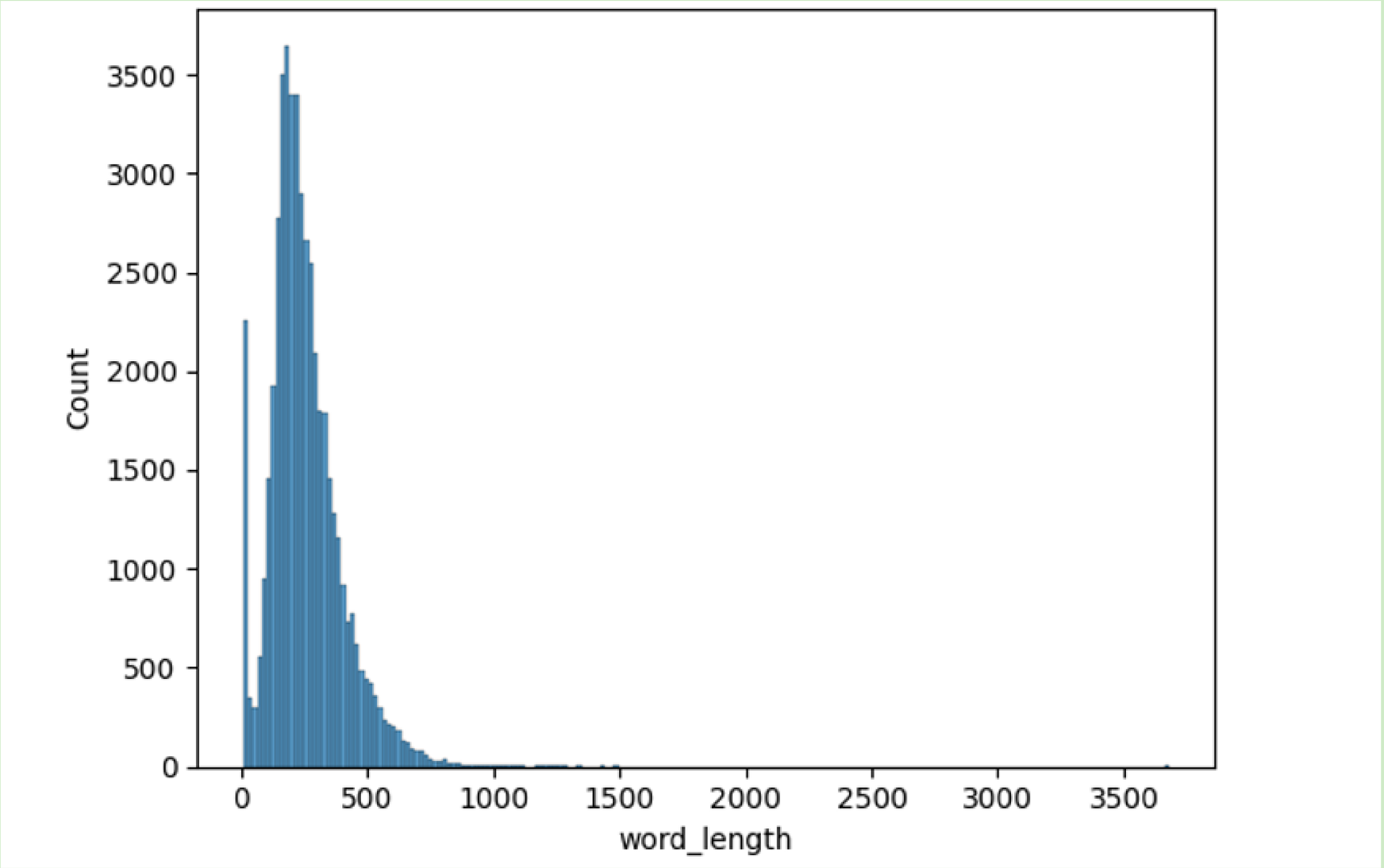
số lượng mẫu dữ liệu trong mỗi lớp



Tỉ lệ số lượng dữ liệu trong mỗi lớp

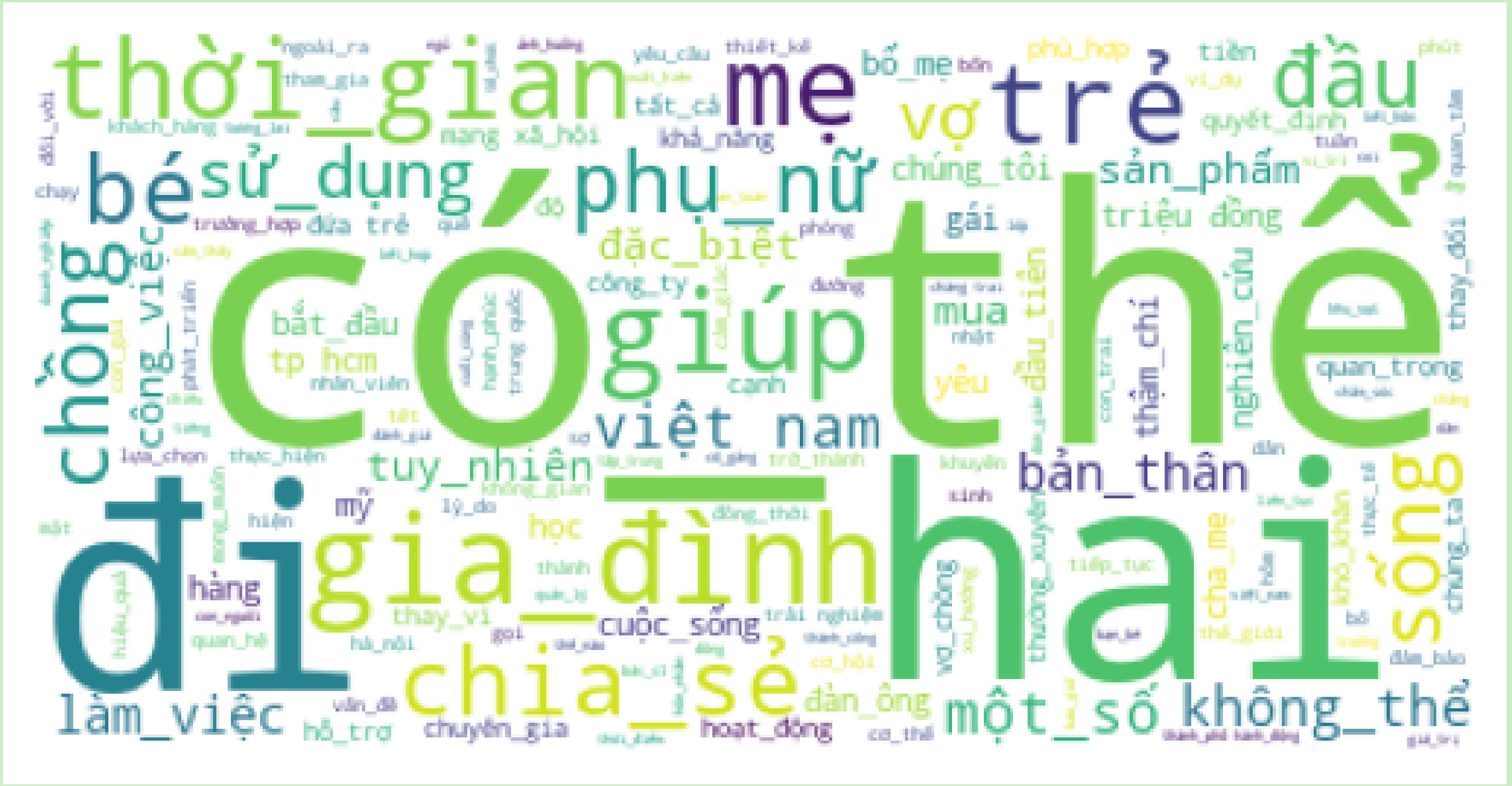


Phân bố độ dài văn bản trước khi tiền xử lý



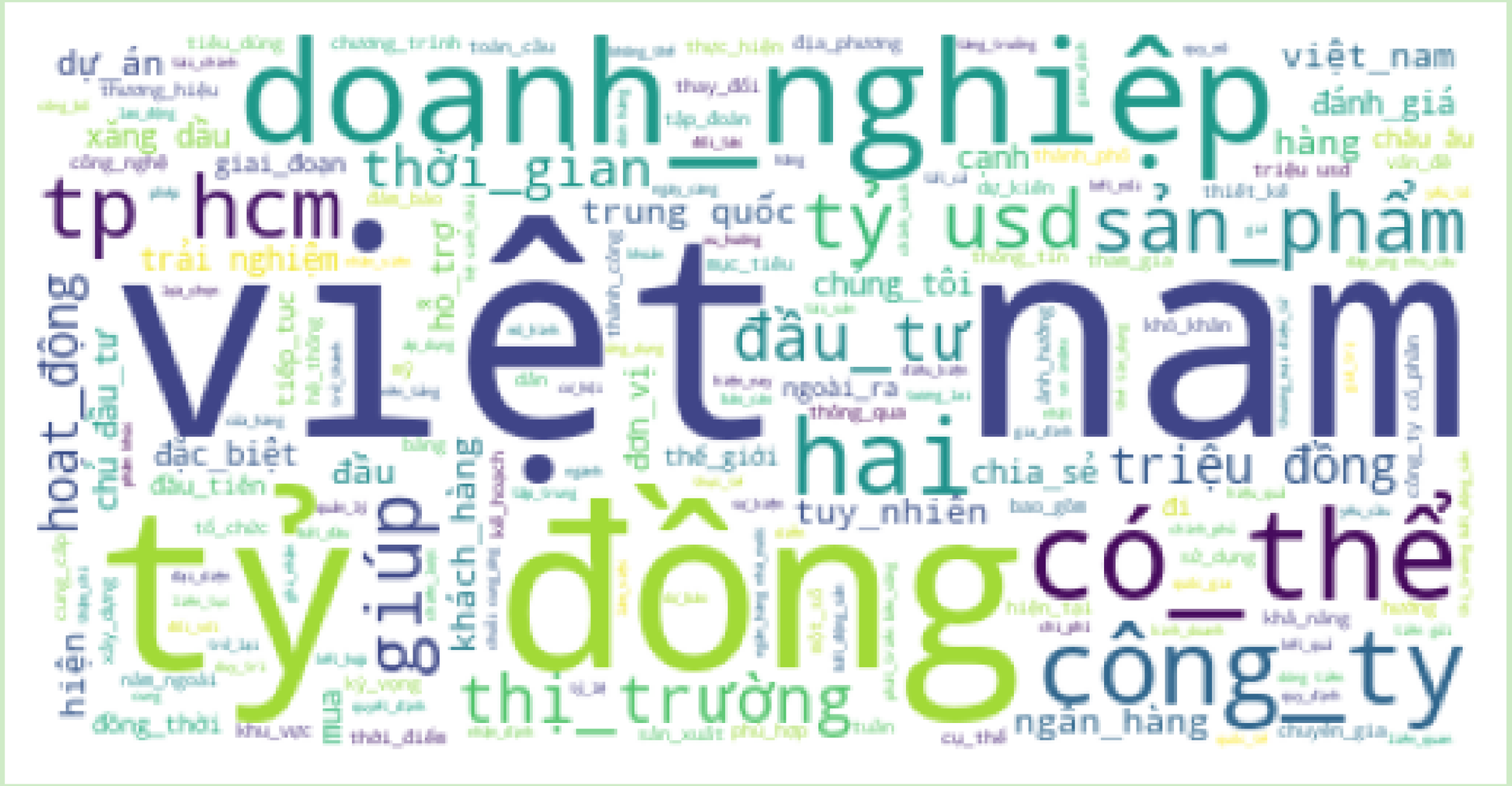
Phân bố độ dài văn bản sau khi tiền xử lý

Từ phổ biến trong lớp chính trị xã hội



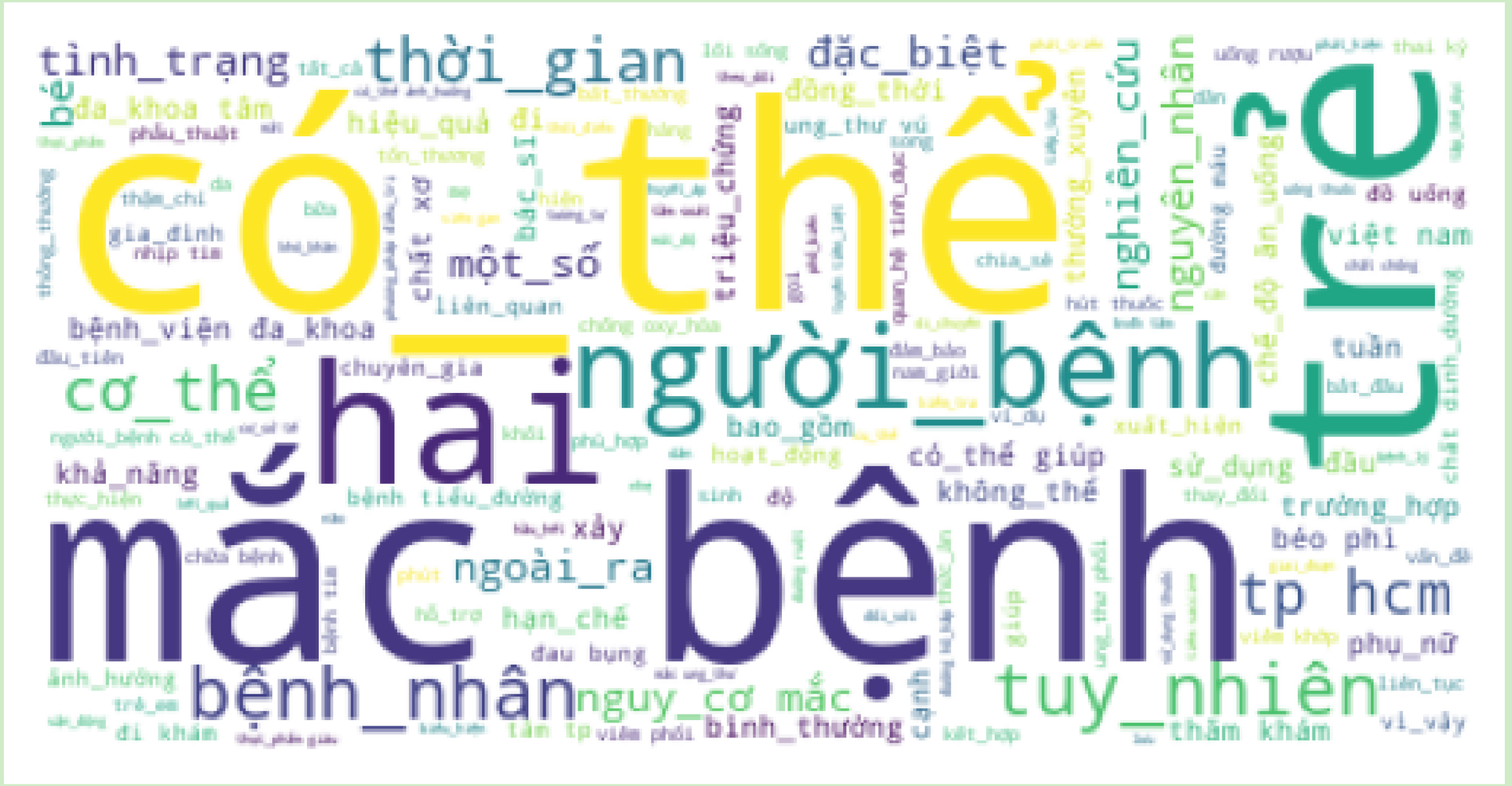
Từ phổ biến trong lớp đời sống

Từ phổ biến trong lớp khoa học

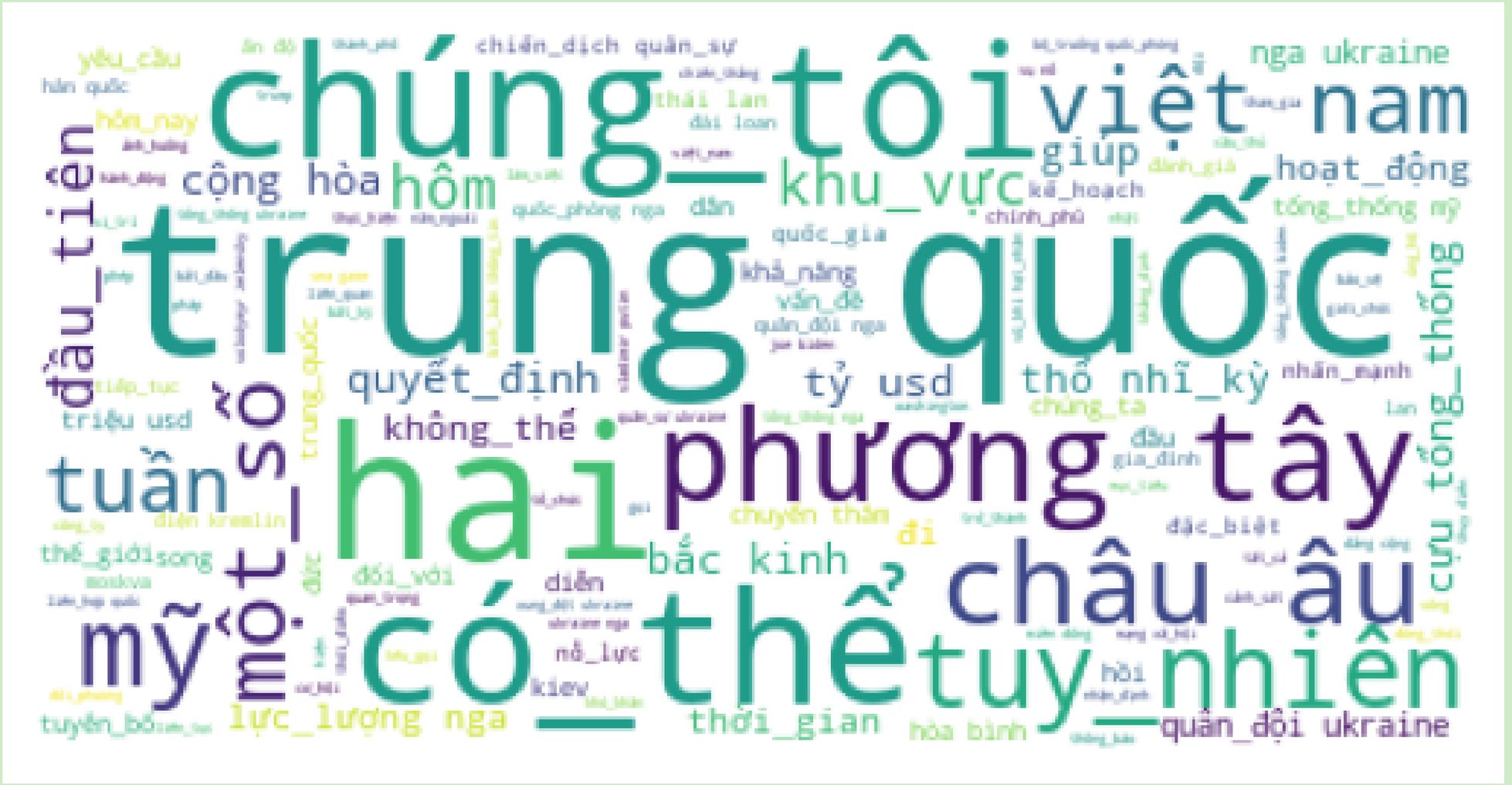


Từ phổ biến trong lớp kinh doanh

Từ phổ biến trong lớp pháp luật



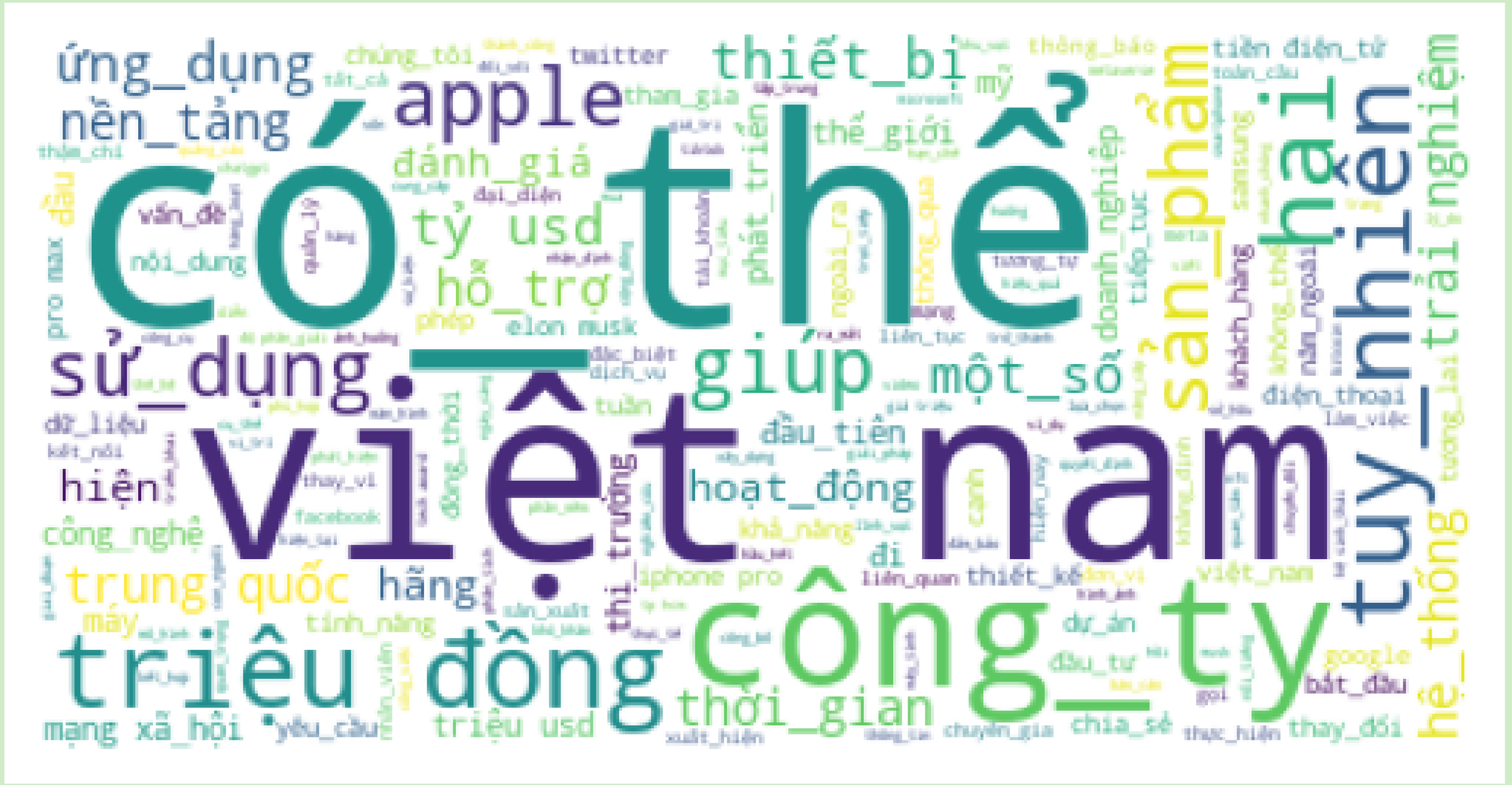
Từ phổ biến trong lớp sức khỏe



Từ phổ biến trong lớp thế giới

Từ phổ biến trong lớp thể thao

Từ phổ biến trong lớp giải trí



Từ phổ biến trong lớp số hóa

Kết quả

- TF-IDF + Uni-Gram:135.137 số từ
- TF-IDF + Bi-Gram:4.088.188 số từ
- Word2vec tự train model có 300 chiều và 137.183 từ

Kết quả

Mô hình 1:Naive bayes+ TF-IDF

	Precision(%)	Recall(%)	Accuracy(%)
TF-IDF + Uni-Gram	87.20	86.89	87.95
TF-IDF + Bi-Gram	88.7	86.29	87.04

Kết quả

Mô hình 2:SVM+ TF-IDF

	Precision(%)	Recall(%)	Accuracy(%)
TF-IDF + Uni-Gram(default parameter)	91.95	91.82	91.02
TF-IDF + Uni-Gram(Best parameter)	91.84	91.93	91.9
TF-IDF + Bi-Gram	89.42	89.03	88.87

Kết quả

Độ dài input:400 từ

LSTM_unit:128

Vocal_size :77.849

batch_size:64

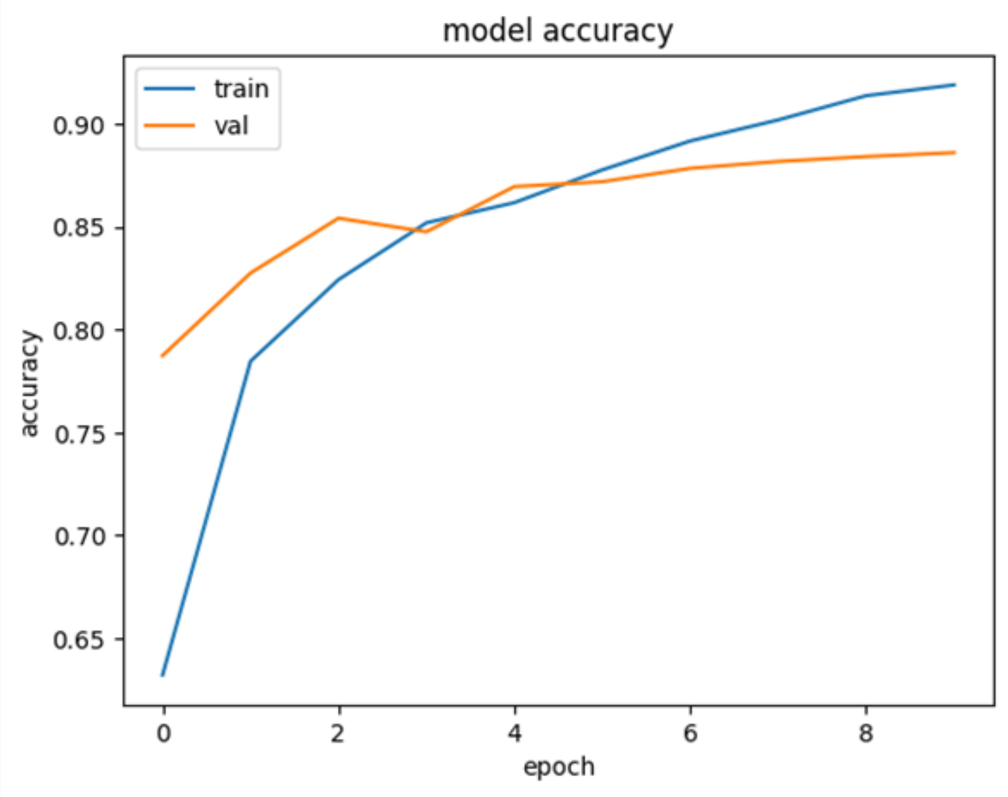
epochs:10

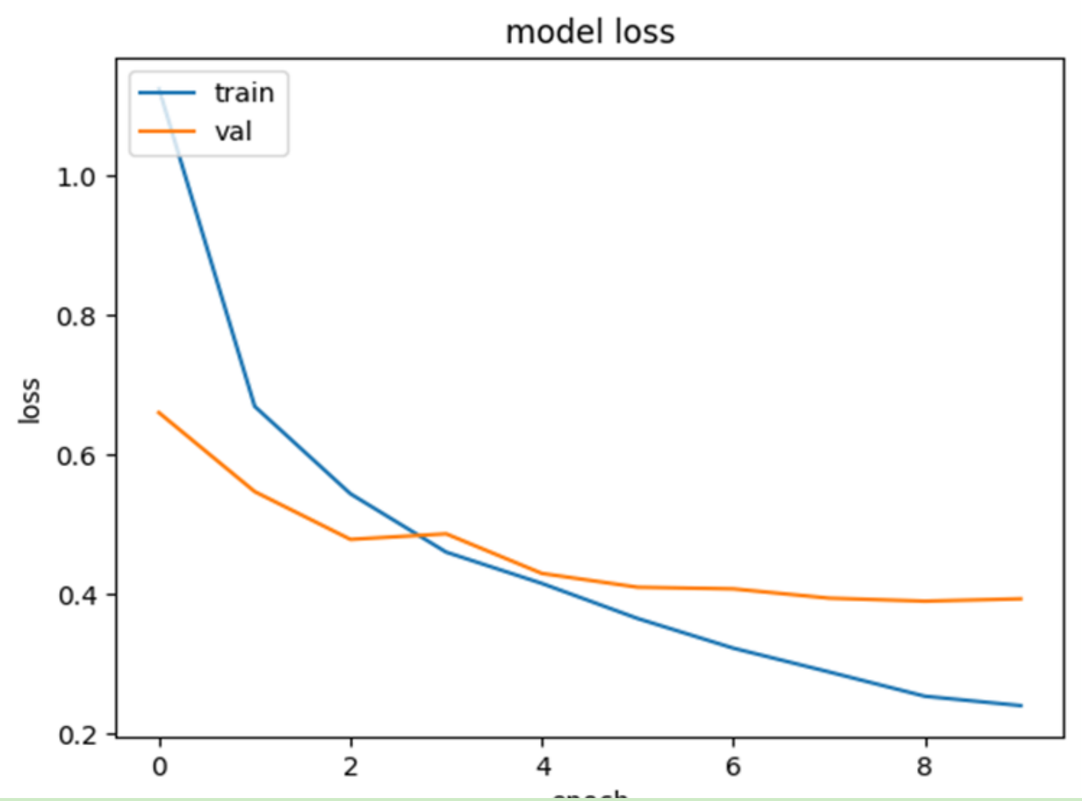
hàm chức năng:softmax

hàm mất mát :categorial_crosssentropy

Kết quả

Mô hình 3:LSTM+ Word2vec





Kết quả

	Precision	Recall	Accuracy
LSTM	87.02	87.31	87.18

Bảng so sánh kết quả

Model	Features	Precision	Recall	Accuracy
Naïve Bayes	Uni-Gram + TF-IDF	88.54%	87.84%	88.01%
	Bi-Gram + TF-IDF	89.35%	87.48%	88.00%
SVM	Uni-Gram + TF-IDF	**90.82%**	**90.96%**	**90.8%**
	Bi-Gram + TF-IDF	89.42%	89.03%	88.87%
LSTM	Word2vec	86.41%	86.32%	86.5%

So sánh kết quả

Nhận thấy mô hình SVM + TF-IDF Uni-Gram cho kết quả tốt nhất với độ chính xác 91.43%

demo