

The background features several large, colorful circles in shades of teal, lime green, orange, and pink. Some circles are solid, while others are dashed outlines. A thin, light blue dashed line curves across the slide, passing behind the title.

# Đề tài: Phân loại chủ đề của bài báo tiếng Anh

Danh sách nhóm

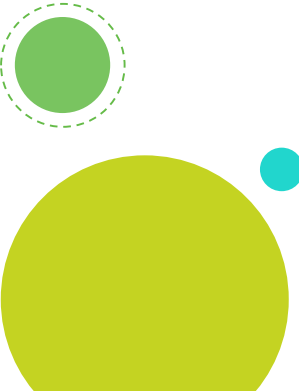

1. Đoàn Xuân Minh
2. Lê Trần Phúc Nguyên
3. Phan Quang Tấn

GVHD: Lê Đình Duy  
Phạm Nguyễn Trường An



1

Mô tả đồ án





## Input & Output:

- ⦿ Input: Đoạn văn bản của 1 bài báo tiếng Anh bất kỳ.
- ⦿ Output: Kết quả dự đoán chủ đề của bài báo.



# Dataset

- © Cách thức xây dựng: Crawl dữ liệu từ các trang báo nổi tiếng trên mạng như: Thesun, Dailymail, Telegraph.
- © Kết quả: Bộ Dataset gồm có hơn 2800 bài báo với 8 chủ đề khác nhau, mỗi của đề có khoảng 350 bài báo.

The background is white and decorated with various colorful circles and dashed lines. In the top left, there is a large orange circle with a dashed red outline, overlapping a yellow circle. Below them is a small pink circle. In the top center, a large blue number '2' is centered within a large dashed light blue circle. In the top right, there is a green circle with a white dot in the center, a small orange circle, and a yellow circle with a dashed green outline. In the bottom left, there is a green circle with a dashed green outline, a large yellow circle, and a small cyan circle. In the bottom right, there is a large cyan circle with a white dot in the center, and a cyan circle with a dashed blue outline.

2

# Preprocessing Data



## Xử lý dữ liệu:

- ◎ Remove number
- ◎ Remove punctuation
- ◎ Remove Special symbol
- ◎ Remove Stopword



# Feature Engineering

## Sử dụng TF-IDF

- © Trích xuất đặc trưng: tần suất xuất hiện của các từ vựng
- © Học các từ vựng xuất hiện nhiều trong bài báo

The background is white with various colorful circles and dashed lines. In the top left, there is a large orange circle with a dashed red outline, overlapping a yellow circle. Below them is a small pink circle. In the bottom left, there is a large green circle with a dashed green outline, overlapping a yellow circle. In the top right, there is a green circle with a white center, overlapping a yellow circle. In the bottom right, there is a large blue circle with a white center, overlapping a yellow circle. In the center, there is a large dashed blue circle containing the number 3.

3

**Model**



## Tiến hành train, test trên các Model

	LinearSVC	Multinomial NB	Logistic Regression
Accuracy (%)	87.1	78.5	85,1

# Đánh giá Model

Đánh giá Model dựa trên độ chính xác F1 - score

- ◎ Tính độ chính xác của model bằng Precision và Recall.
- ◎ Confusion Matrix

	precision	recall	f1-score	support
femail	0.81	0.65	0.72	86
films	0.89	0.89	0.89	76
health	0.77	0.77	0.77	61
lifestyle	0.84	0.82	0.83	71
sport	0.95	1.00	0.97	54
technology	0.96	0.97	0.97	78
travel	0.92	0.94	0.93	81
tvshowbiz	0.82	0.97	0.89	69
accuracy			0.87	576
macro avg	0.87	0.88	0.87	576
weighted avg	0.87	0.87	0.87	576



LinearSVC

# Đánh giá Model

	precision	recall	f1-score	support
femail	0.84	0.37	0.52	86
films	0.94	0.83	0.88	76
health	0.87	0.43	0.57	61
lifestyle	0.44	0.96	0.60	71
sport	1.00	0.98	0.99	54
technology	0.96	0.92	0.94	78
travel	0.94	0.89	0.91	81
tvshowbiz	0.80	0.96	0.87	69
accuracy			0.78	576
macro avg	0.85	0.79	0.79	576
weighted avg	0.85	0.78	0.78	576



MultinomialNB

	precision	recall	f1-score	support
femail	0.84	0.37	0.52	86
films	0.94	0.83	0.88	76
health	0.87	0.43	0.57	61
lifestyle	0.44	0.96	0.60	71
sport	1.00	0.98	0.99	54
technology	0.96	0.92	0.94	78
travel	0.94	0.89	0.91	81
tvshowbiz	0.80	0.96	0.87	69
accuracy			0.78	576
macro avg	0.85	0.79	0.79	576
weighted avg	0.85	0.78	0.78	576



Logistic Regression

The background is white with several decorative elements: a large orange circle with a dashed red outline in the top left; a yellow circle below it; a small pink circle below that; a green circle with a dashed green outline in the bottom left; a large green circle with a white dot in the top right; a yellow circle with a dashed yellow outline below it; a large blue circle with a white dot in the bottom right; and a teal circle with a dashed teal outline below that. A large, light blue dashed circle is centered on the page.

4

**Kết luận**



## Khó khăn:

- ◎ Số lượng dataset ít dẫn đến việc hạn chế về từ vựng, số lượng chủ đề.
- ◎ Thông tin cập nhật mỗi ngày, dẫn đến các từ mới xuất hiện gần đây làm kết quả dự đoán có thể sai. VD: Covid-19



## Hướng phát triển:

- ◎ Tăng thêm số lượng dataset
- ◎ Cập nhật các từ vựng mới xuất hiện
- ◎ Điều chỉnh các hyperparameter hoặc sử dụng Deep Learning để xử lý ngôn ngữ tự nhiên.