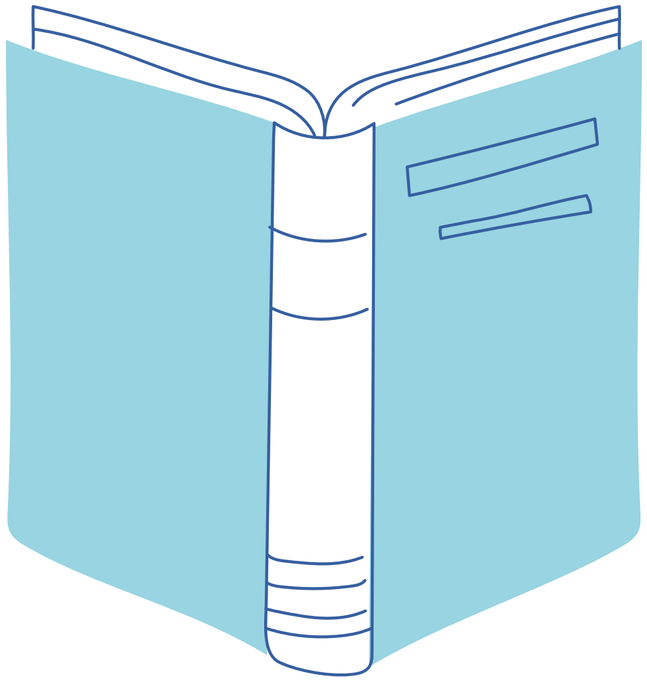
HUTECH
Đại học Công nghệ Tp.HCM

Đồ án cơ sở



Đề xuất Hashtag sách

Thành viên team:

Lâm Tấn Duy

Lê Nguyễn Anh Nhật

Lê Thái Dương

Giảng viên hướng dẫn: Lê Cung Tưởng

Giới thiệu

Trong thời đại số hóa ngày nay, việc sử dụng hashtag không còn xa lạ với mọi người, đặc biệt là trên các nền tảng mạng xã hội như Instagram, Twitter, hay Facebook.

Tuy nhiên, trong lĩnh vực sách, việc áp dụng hashtag vẫn còn nhiều hạn chế.

Với mong muốn giải quyết vấn đề này, đề tài "Ứng dụng gán nhãn hashtag sách" được đưa ra với mục tiêu tạo ra một ứng dụng hoặc nền tảng cho phép người đọc gán nhãn sách bằng hashtag một cách dễ dàng và hiệu quả.

Thu thập dữ liệu

01

Junkybooks

02

ManyBooks

03

BookBubs

Dùng thư viên BeautifulSoup của Python

Tổng quan về dữ liệu

Title

Description

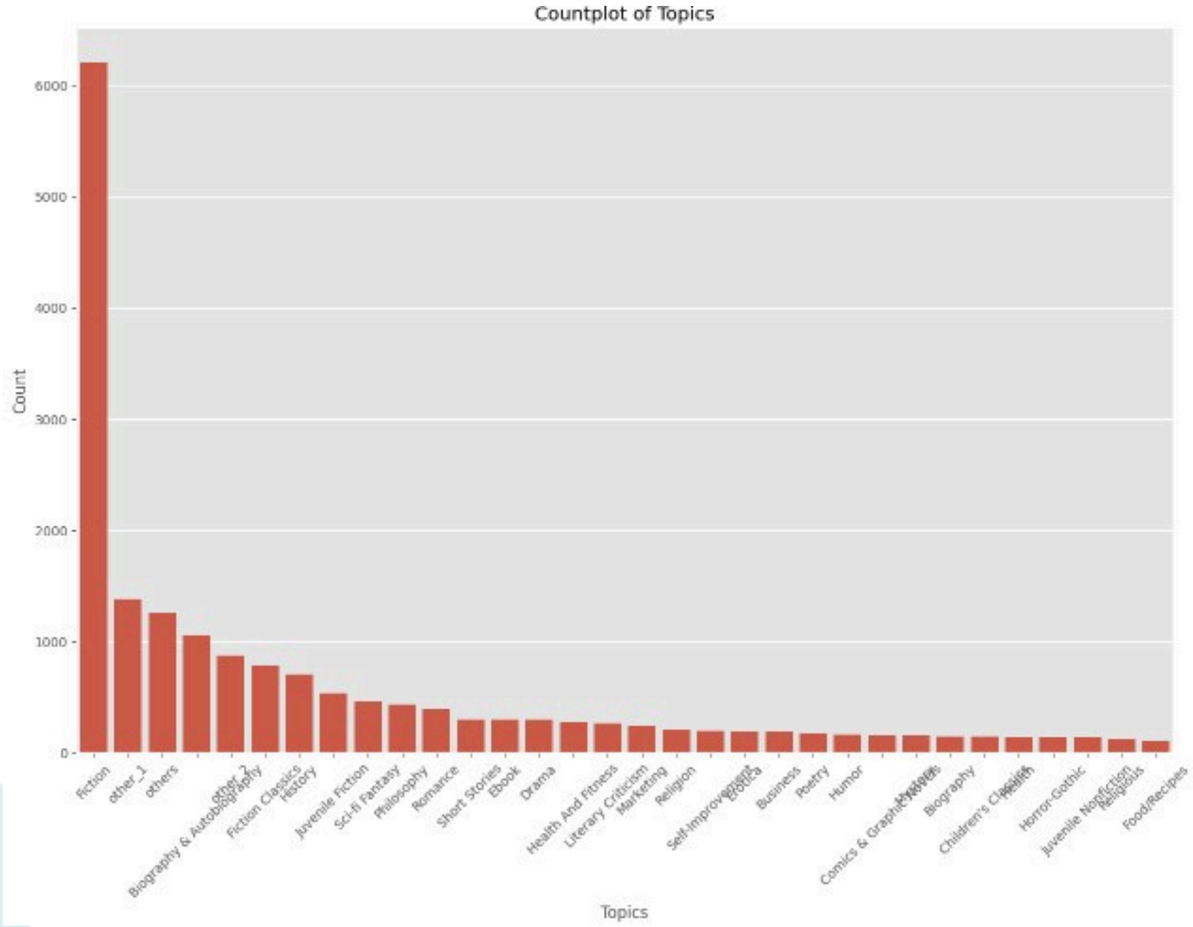
Topic

Title	Description	Topic
The Devil's Disciple	Set in Colonial America during the Revolutiona	Drama War History Fiction and Literature
Taking Chances	Spice-o-meter Rating: This fun romance is a sc	Romance
Verdi : The Story of the Little Boy who Loved the Har	This time Tapper moves his focus to Italy in his	Music Biography
Success	The contents of this volume originally appeare	Non-fiction Business
Triplanetary	ONE MAN DISCOVERED THE TRUTH—The Fall	Science Fiction Post-1930 Fiction and Literature
The Iliad	Translated by Andrew Lang.	Classic Adventure Fiction and Literature
The Four Corners of the World	This is a rather eclectic group of stories. It cor	Short Story Collection Mystery/Detective Espionage Adventure
The Jewish State	The Herzl text was originally published under t	Politics Biography History
Adrift in the Wilds	Elwood Brandon and Howard Lawrence are er	Adventure Young Readers Nautical Fiction and Literature
The Adventures of Captain Kettle	Captain Kettle is a most engaging scoundrel. Si	Short Story Collection Nautical Humor Adventure
Adventures of Huckleberry Finn	The drifting journey of Huck and his friend Jim,	Banned Books Adventure Young Readers Fiction and Literature Humo
Derelict	What was the mystery of this great ship from t	Short Story Science Fiction Post-1930
South Wind	A book totally unlike any other, inimitably hum	Fiction and Literature Travel Gay/Lesbian
Jailed for Freedom	This book deals with the intensive campaign o	Politics Women's Studies
The Facts of Reconstruction	Was the enfranchisement of the black men at	Non-fiction Politics History
Country Walks of a Naturalist with His Children	In this little book my desire has been, not so m	Nature Young Readers Fiction and Literature
The Black Tulip	This exciting novel takes place a few years afte	Adventure Fiction and Literature
Tacitus: The Histories, Volumes I and II	Translated with Introduction and Notes by W.	History
Friends in Feathers and Fur, and Other Neighbors	In this little book we have again given the initia	Nature Young Readers Instructional
Single: Miss Tennessee b/w The Cryerer	This eBook contains two short stories: Miss Te	Short Story Creative Commons Fiction and Literature Post-1930
Colonel Quaritch, V.C.	Having served~in India and Egypt, Colonel Qua	Fiction and Literature Mystery/Detective Romance
Children of the Frost	In the forests of the north -- The law of life --	Short Story Collection Fiction and Literature Short Story
The Master's Violin	A Love Story with a musical atmosphere. A pic	Fiction and Literature Music Romance
Blow The Man Down	Mayo was captain of Julius Marson's pleasure	Romance Nautical Fiction and Literature
Olivia, Mourning	Olivia wants the 80 acres in far off Michigan th	Fiction and Literature History
Chloe - Lost Girl	A missing student. A gunned-down detective. A	Mystery/Detective

Dữ liệu gồm 18000

Có 100 Topics

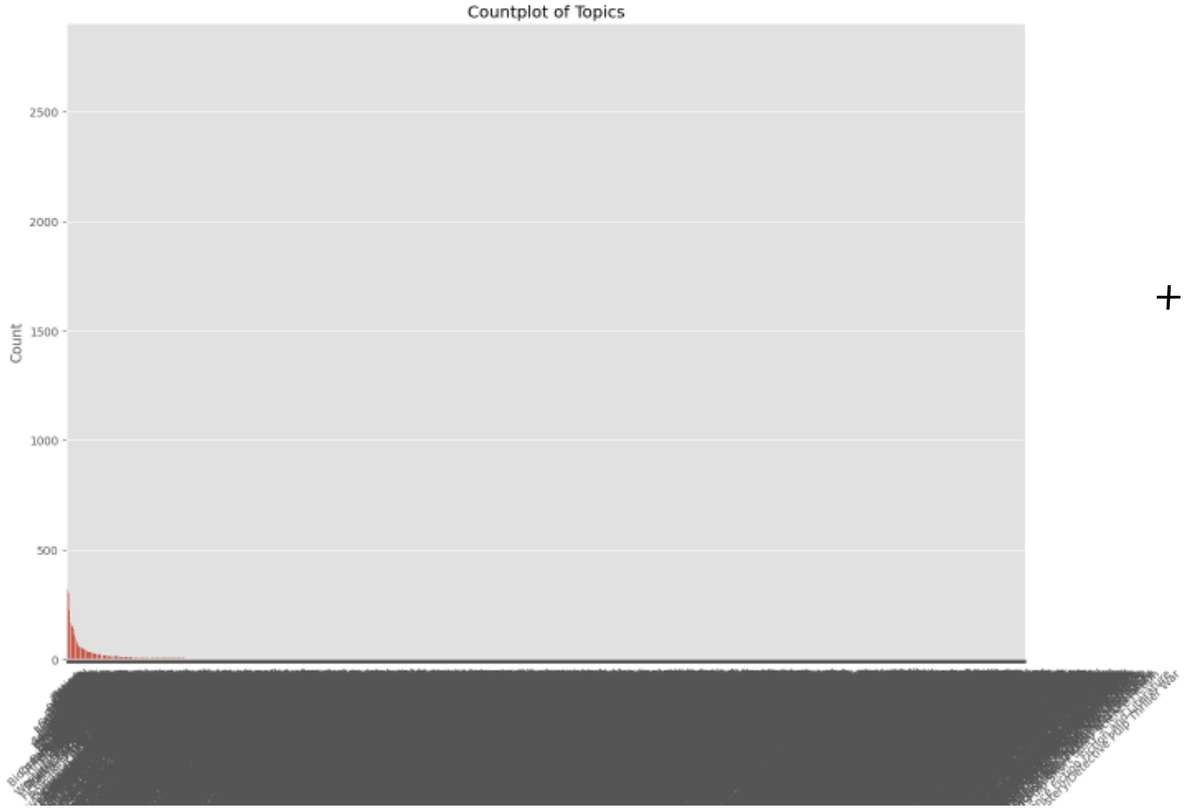
Tiền xử lý dữ liệu




```
Topic
Fiction                      6209
other_1                      1385
others                       1255
Biography & Autobiography    1049
other_2                       874
Fiction Classics              778
History                       700
Juvenile Fiction              535
Sci-fi Fantasy                459
Philosophy                    437
Romance                       389
Short Stories                 382
Ebook                         381
Drama                         295
Health And Fitness            276
Literary Criticism            261
Marketing                     238
Religion                      206
Self-Improvement              192
Erotica                       191
Business                      185
Poetry                        175
Humor                         164
Comics & Graphic Novels       157
Mystery                       154
Biography                     148
Children's Classics           144
Health                        141
Horror-Gothic                 138
Juvenile Nonfiction           134
Religious                     122
Food/Recipes                  106
Name: count, dtype: int64
```

Group các nhóm dữ liệu

Cac dữ liệu về các chủ đề rất nhiều và nhiều khó có thể đem đi dự đoán



+ Group theo cùng chủ đề

+ Group theo số lượng ít nhất được cho và other

+Group theo ngưỡng

Xử lý stopword, emoji, link,..

	Title	Description	Topic	Descr
0	The Devil's Disciple	Set in Colonial America during the Revolutiona...	Fiction	set colonial america revolutionary era play te...
1	Taking Chances	Spice-o-meter Rating: This fun romance is a so...	Romance	spiceometer rating fun romance solid maybe sca...
2	Verdi : The Story of the Little Boy who Loved ...	This time Tapper moves his focus to Italy in h...	Biography & Autobiography	time tapper moves focus italy voyage among gre...
3	Success	The contents of this volume originally appeare...	Business	contents volume originally appeared weekly art...
4	Triplanetary	ONE MAN DISCOVERED THE TRUTH—The Fall of Rome,...	Fiction	one man discovered truththe fall rome wars rac...
...	...	...	...	...
18095	Journey to the East	This book tells the tale of a man who goes on ...	other_1	book tells tale man goes wonderful amazing jou...
18096	The Monk Who Sold His Ferrari: A Fable About F...	Wisdom to Create a Life of Passion, Purpose, a...	Biography & Autobiography	wisdom create life passion purpose peace inspi...
18097	I Am that	This collection of the timeless teachings of o...	Philosophy	collection timeless teachings one greatest sag...
18098	The Berlin Phenomenology	Since the three volume edition ofHegel's Philo...	History	since three volume edition ofhegels philosophy...
18099	'I'm Telling You Stories'	This is a jubilant and rewarding collection of...	Literary Criticism	jubilant rewarding collection winterson schola...

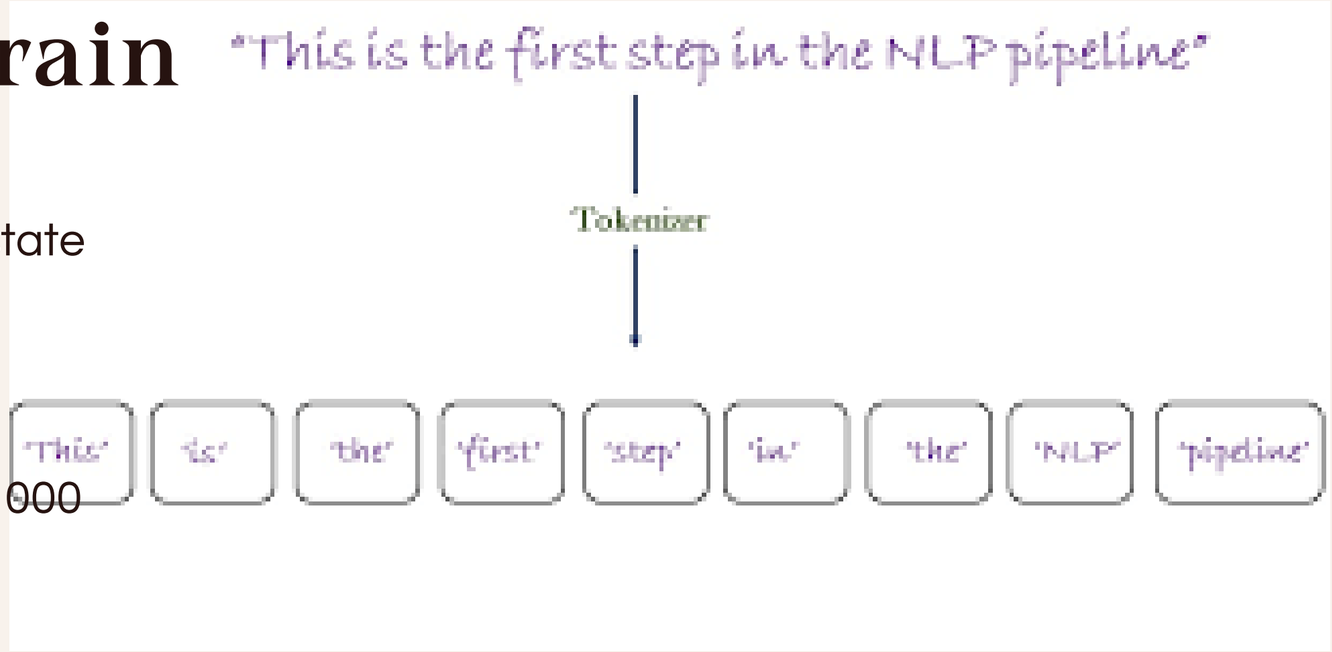
18100 rows × 4 columns

Tokenize và tập train test

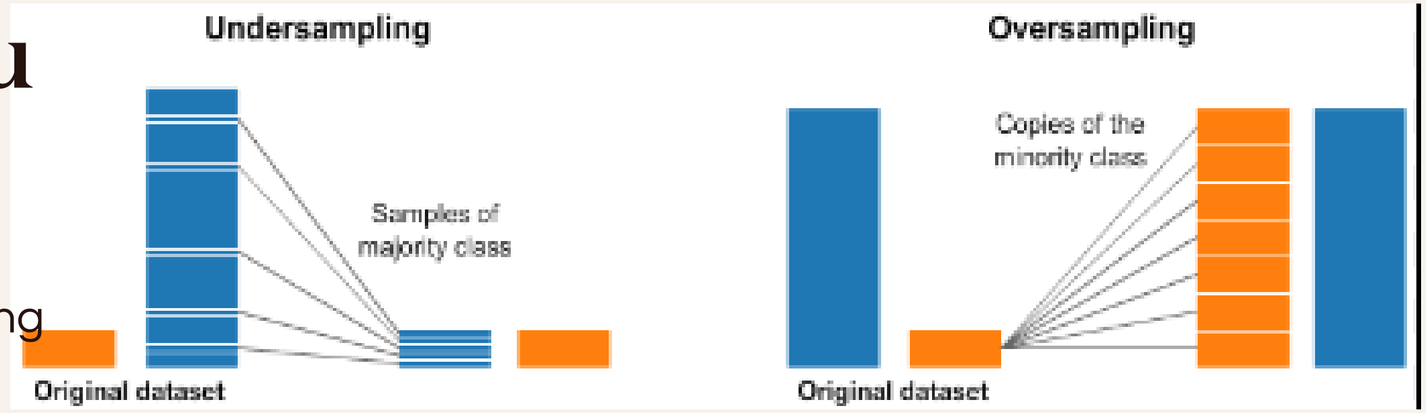
Dữ liệu được chia theo 80/20 với random state = 42.

Tách từ sau đó token như từ : “I Love You” Được tách thành “I”, “Love”, “You”. Tôi đa 1000 từ.

Chuyển hóa nó thành chuỗi số



Cân bằng dữ liệu



+ Các tập dữ liệu ít được dùng Oversampling và tập dữ liệu được cân bằng .

+ 1000 data từng loại

+ Giảm dữ liệu của các nhãn lớn nhất

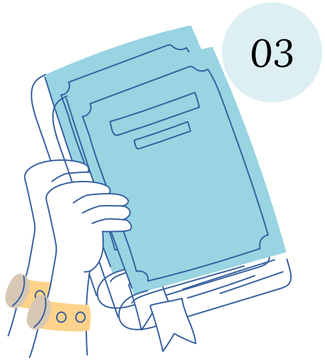
Xây dựng mô hình

Mô hình 1

Sử dụng toàn bộ nhãn cho model Bert

Mô hình 2

Train từng model cho từng nhãn khác nhau kết hợp cân bằng dữ liệu



Mô hình 3

Tập trung vào dữ liệu lớn nhất

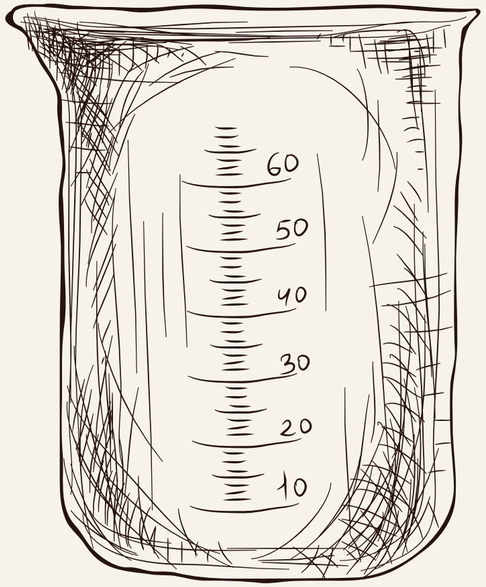
Mô hình 4

Mô hình sử dụng hàm loss khác và cân bằng dữ liệu

Mô hình 1

Sử dụng Bert với các chỉ số

Thông số	Điều chỉnh
Optimize	Adam
Learning rate	6e-6
Gradient steps	4
epochs	3
Training loop	8100
Evulation	Flat_accuracy



Mô Hình 2 3 4

Mô hình 2: Cân bằng dữ liệu về 1000 sau đó đem đi train cho từng mô hình.

Thông số	Điều chỉnh
Layer 1	Embedding
Layer 2	Flatten()
Layer 3	128, activation = relu
Layer 4	64, activation = relu
Layer 5	32, activation = relu
Layer 6	1, activation = sigmoid
Optimizer	Adam
Loss	Binary crossentropy
metrics	accuracy

25

Mô Hình 3 tập trung vào nhãn có lợi thế

Mô hình 4: Tập trung vào việc đánh giá cáo nhãn lợi thế

Dùng Pricision để làm metrics

Intuitively Understanding the Cross Entropy

$$H(P^*|P) = -\sum_i \underbrace{P^*(i)}_{\text{TRUE CLASS DISTRIBUTION}} \log \underbrace{P(i)}_{\text{PREDICTED CLASS DISTRIBUTION}}$$

Model 3

Thông số	**Điều chỉnh**
Layer 1	Embedding
Layer 2	Flatten()
Layer 3	128, activation = relu
Layer 4	Dropout(0.5)
Layer 5	64, activation = relu
Layer 6	32, activation = relu
Layer 7	1, activation = sigmoid
Optimizer	Adam
Loss	Binary crossentropy
metrics	accuracy
Early_stopping	Val_accuracy
epochs	50
Batch_size	100

Model 4

Thông số	Điều chỉnh
Layer 1	Embedding
Layer 2	Flattern
Layer 3	Relu
Layer 4	Relu
Layer 5	Relu
Layer 6	Sigmoid
Optimizer	Adam
Loss	Weights_binary_crossentropy
Metrics	Accuray, Precision
Early stopping	Val_accuracy

Kết quả

	Model 1	Model MLP 2	Model MIP 3	Mô hình 4
Accuracy	65%	44%	60%	70%

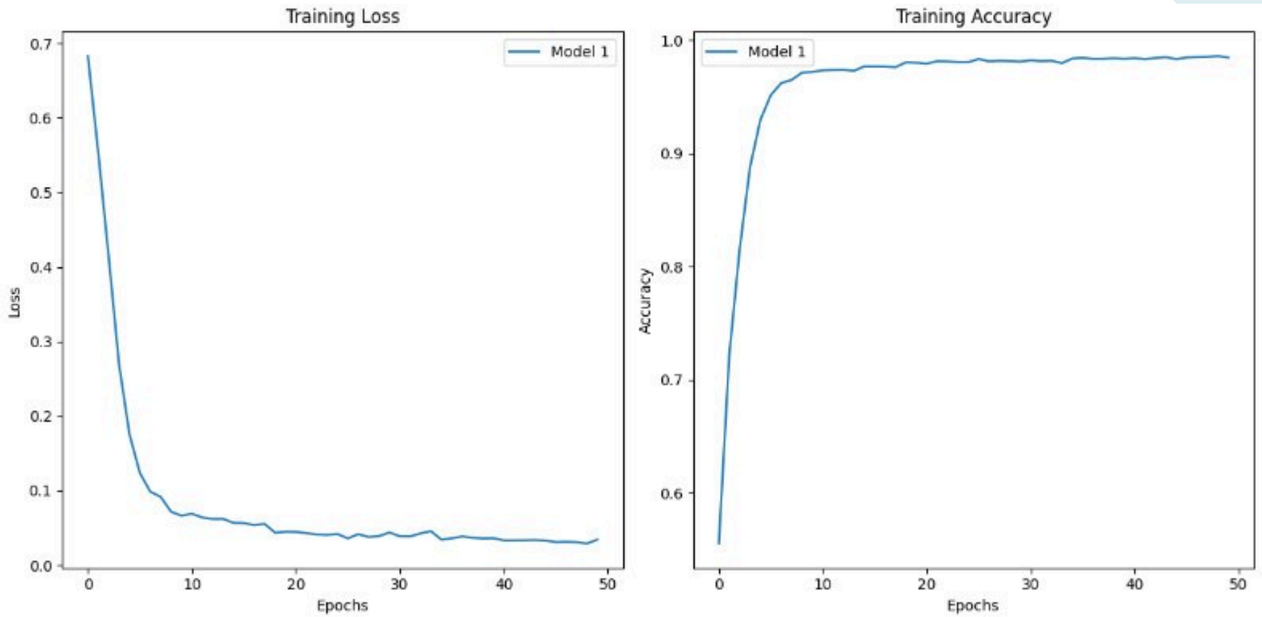
Kết Quả

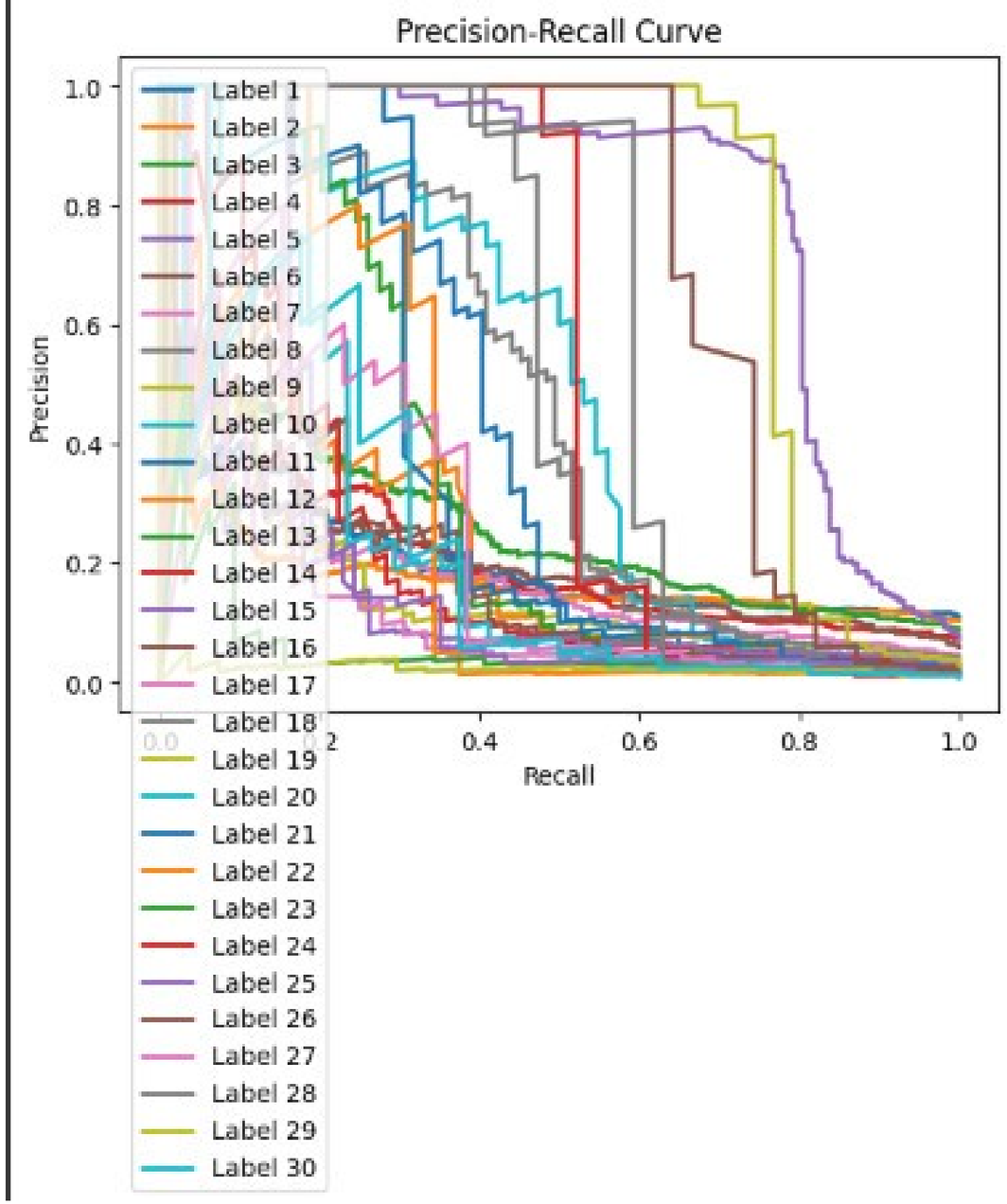
01 Độ chính xác 60 %

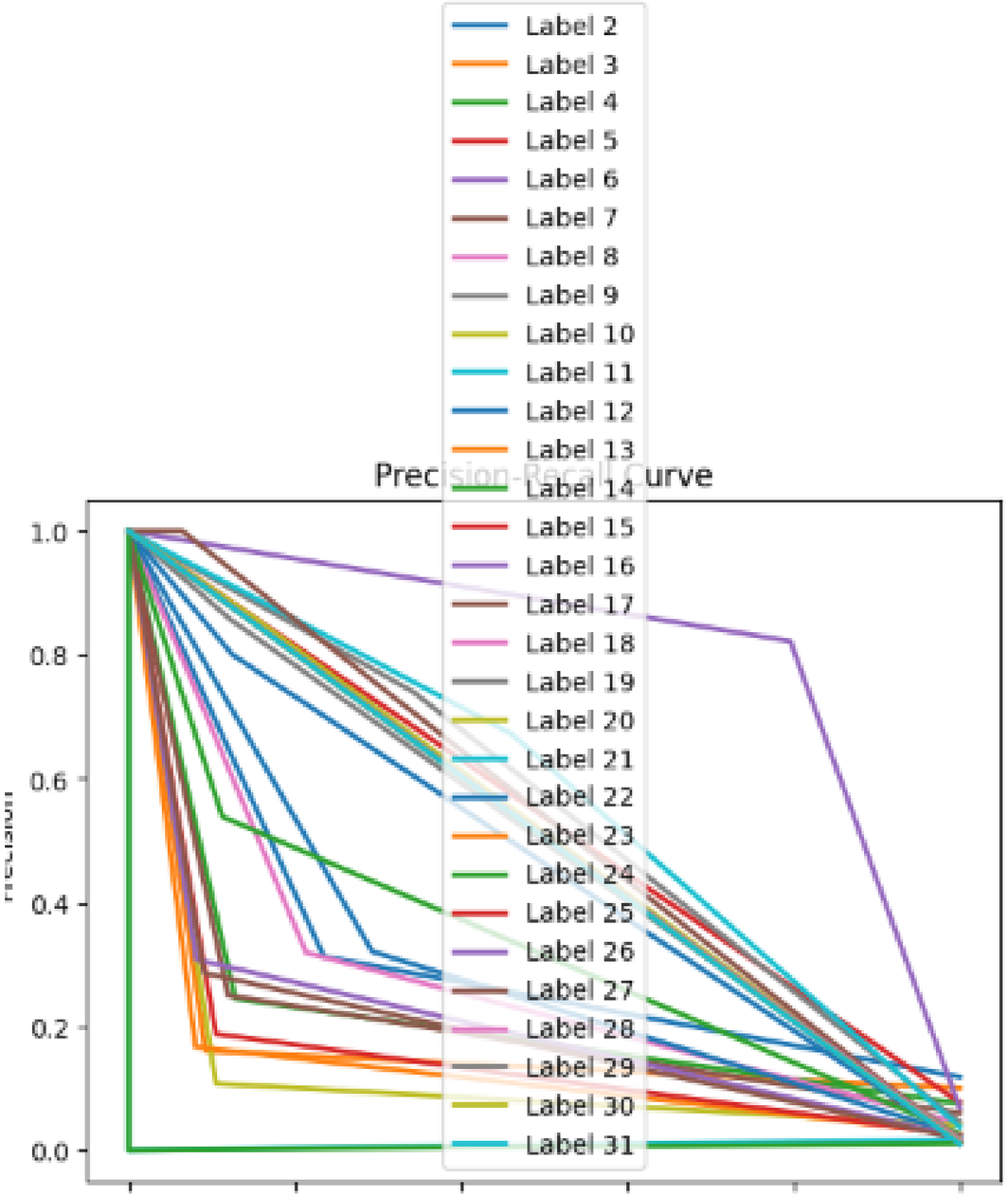
02 Độ chính xác cao cho vài Hashtag

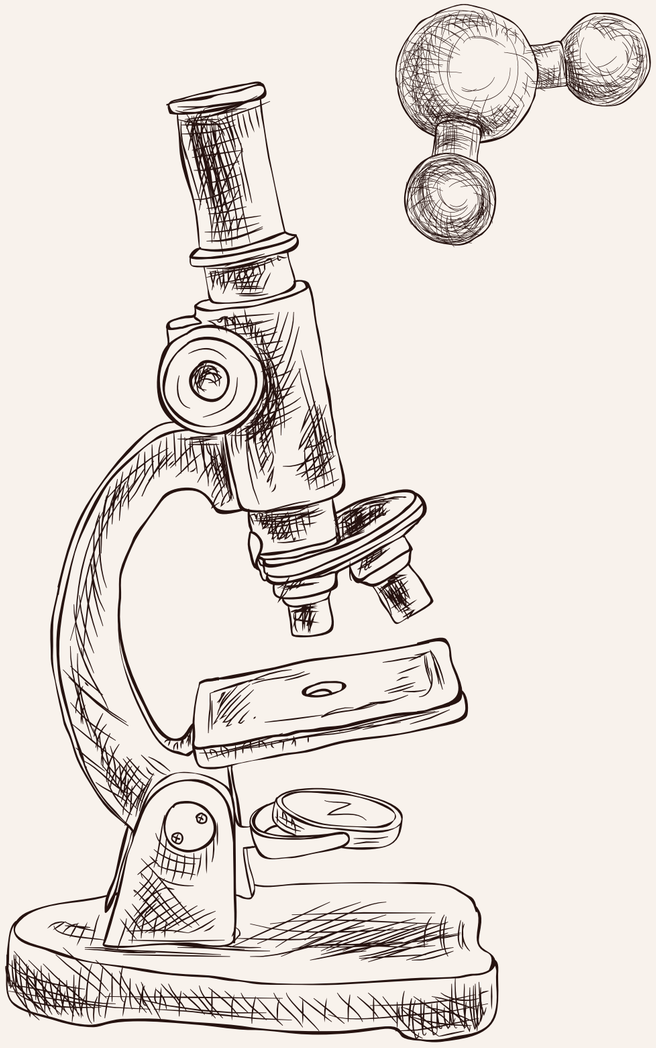
	precision	recall	f1-score	support
0	0.49	0.66	0.56	1238
1	0.17	0.48	0.25	281
2	0.06	0.26	0.10	238
3	0.11	0.55	0.18	179
4	0.12	0.52	0.20	183
5	0.46	0.82	0.59	162
6	0.14	0.54	0.23	142
7	0.12	0.52	0.20	103
8	0.29	0.49	0.37	108
9	0.12	0.58	0.20	95
10	0.24	0.54	0.33	89
11	0.08	0.53	0.14	58
12	0.13	0.55	0.22	62
13	0.10	0.48	0.16	62
14	0.12	0.53	0.20	57
15	0.05	0.43	0.10	49
16	0.12	0.64	0.21	44
17	0.06	0.30	0.10	37
18	0.30	0.59	0.40	49
19	0.78	0.79	0.78	39
20	0.08	0.59	0.14	39
21	0.13	0.69	0.21	32
22	0.12	0.48	0.20	27
23	0.04	0.16	0.06	25
24	0.35	0.50	0.41	24
25	0.13	0.56	0.21	34
26	0.43	0.71	0.54	31
27	0.21	0.31	0.25	35
28	0.80	0.74	0.77	27
29	0.03	0.15	0.04	26
30	0.12	0.57	0.20	23
micro avg	0.19	0.56	0.29	3598
macro avg	0.21	0.52	0.28	3598
weighted avg	0.29	0.56	0.36	3598
samples avg	0.25	0.56	0.32	3598

03 Độ chính xác cao cho nhãn có nhiều nhất

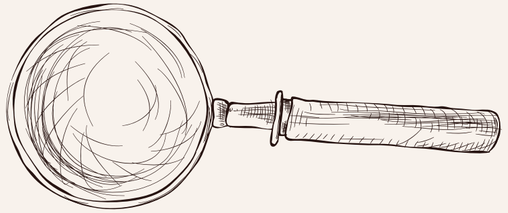


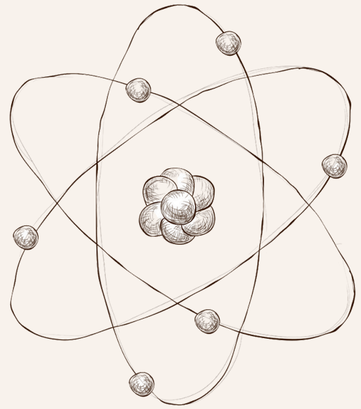

Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Label 1
Label 2
Label 3
Label 4
Label 5
Label 6
Label 7
Label 8
Label 9
Label 10
Label 11
Label 12
Label 13
Label 14
Label 15
Label 16
Label 17
Label 18
Label 19
Label 20
Label 21
Label 22
Label 23
Label 24
Label 25
Label 26
Label 27
Label 28
Label 29
Label 30

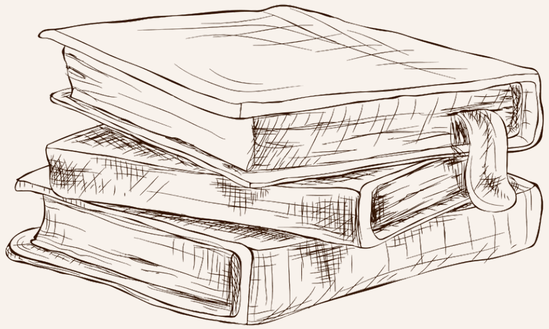

Label 2
Label 3
Label 4
Label 5
Label 6
Label 7
Label 8
Label 9
Label 10
Label 11
Label 12
Label 13
Label 14
Label 15
Label 16
Label 17
Label 18
Label 19
Label 20
Label 21
Label 22
Label 23
Label 24
Label 25
Label 26
Label 27
Label 28
Label 29
Label 30
Label 31
0.0
0.2
0.4
0.6
0.8
1.0



DEMO







Studio
Shodwe

Cảm ơn thầy và các bạn đã lắng nghe