

News classification

Nhóm 4



Project goal

Xử lý bài toán phân loại báo

Input:
Văn bản (tiêu đề, nội dung, mô tả, ... của một bài báo bất kỳ)

Output:
Nhãn của bài báo (Chính trị, kinh doanh, y tế, thể thao, khoa học)

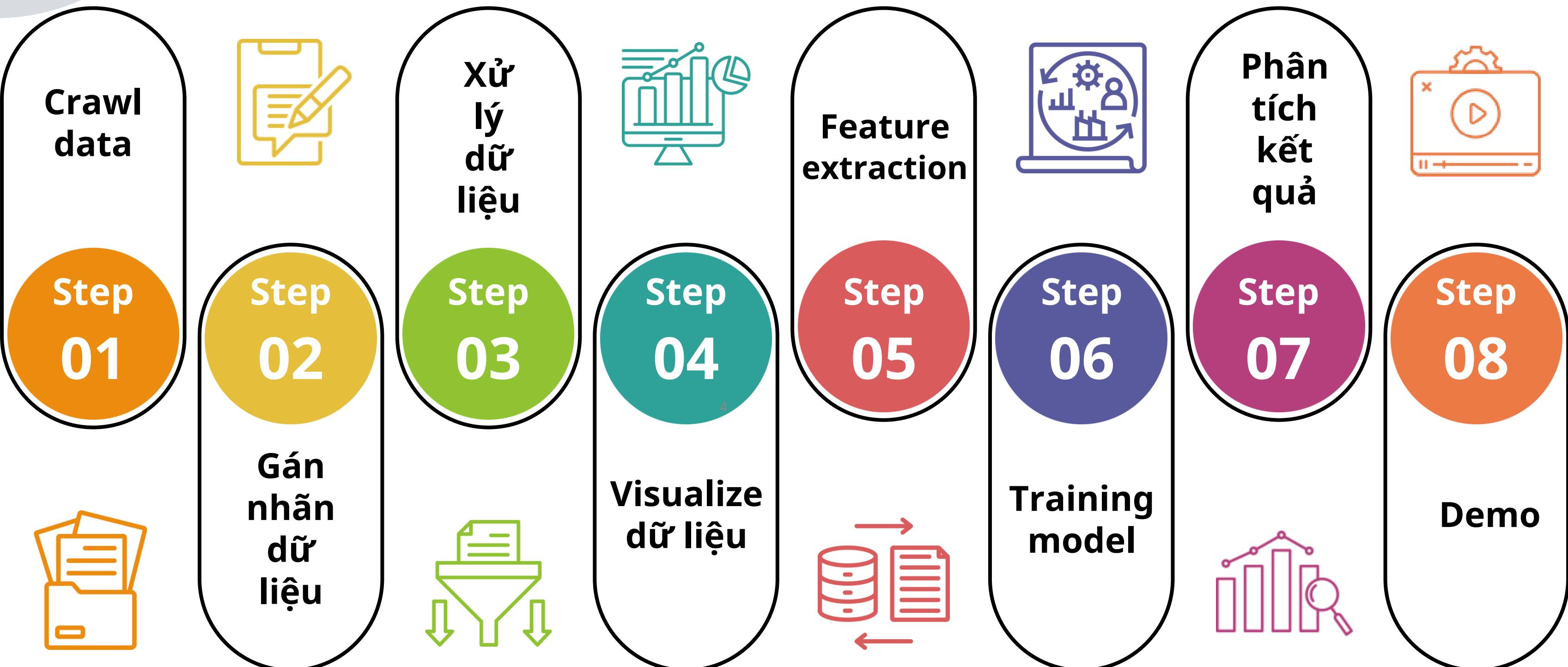
Task description

Để giải quyết bài toán đã đề ra, cần xử lý các công việc:

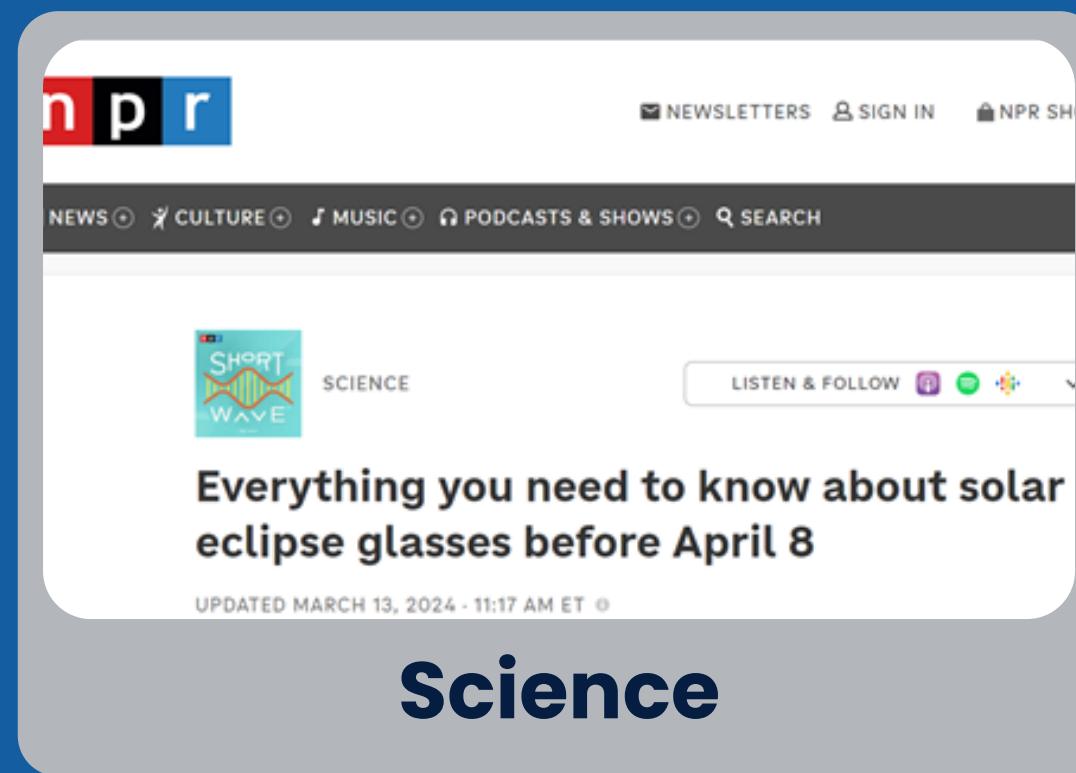
- Xây dựng một hệ thống phân loại báo chí. Dựa trên đầu vào là nội dung của bài báo, mục tiêu là có thể tự động gán nhãn cho văn bản đó.
- Dựa vào cách tiếp cận là học máy có giám sát và các mô hình học sâu, sử dụng dữ liệu báo đã được gán nhãn để chọn các thuộc tính từ đó tạo ra các mô hình phân loại có hiệu quả tốt.



Modules



Crawling data & Data annotation



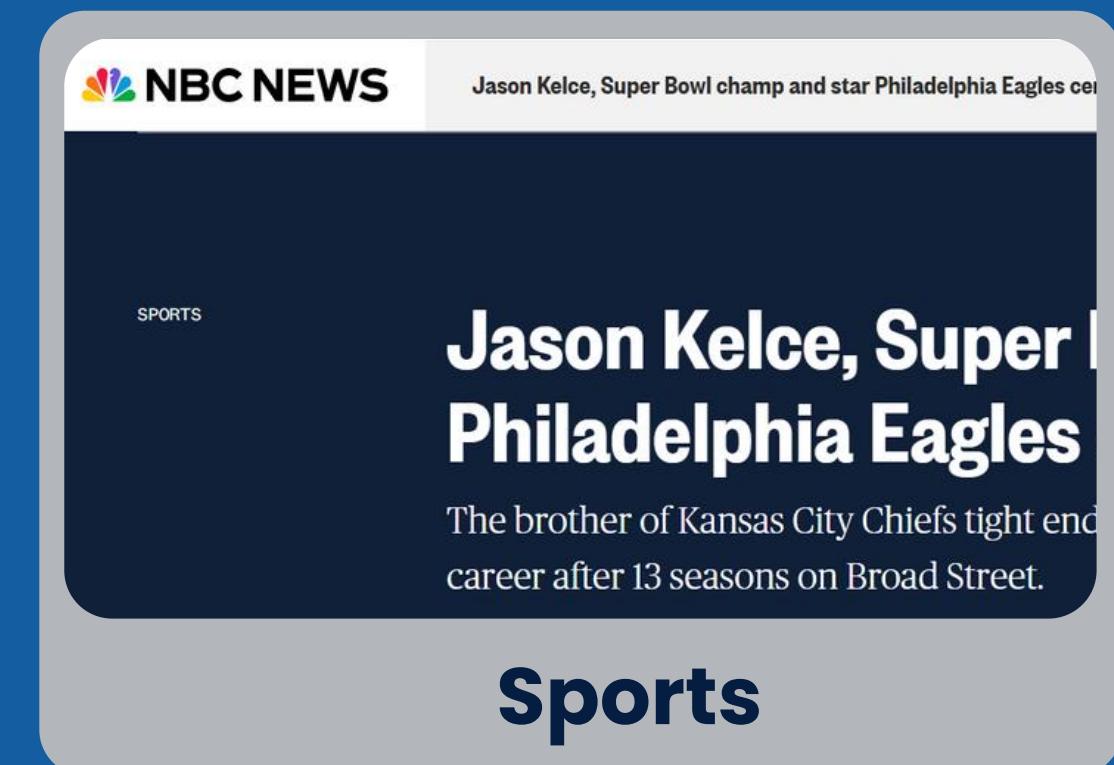
The image shows a screenshot of the NPR website's science section. At the top, there's a navigation bar with links for NEWS, CULTURE, MUSIC, PODCASTS & SHOWS, and SEARCH. Below the navigation, there's a logo for "npr SHORT WAVE SCIENCE". A main headline reads "Everything you need to know about solar eclipse glasses before April 8", with a timestamp "UPDATED MARCH 13, 2024 - 11:17 AM ET". A large blue button at the bottom of the card says "Science".



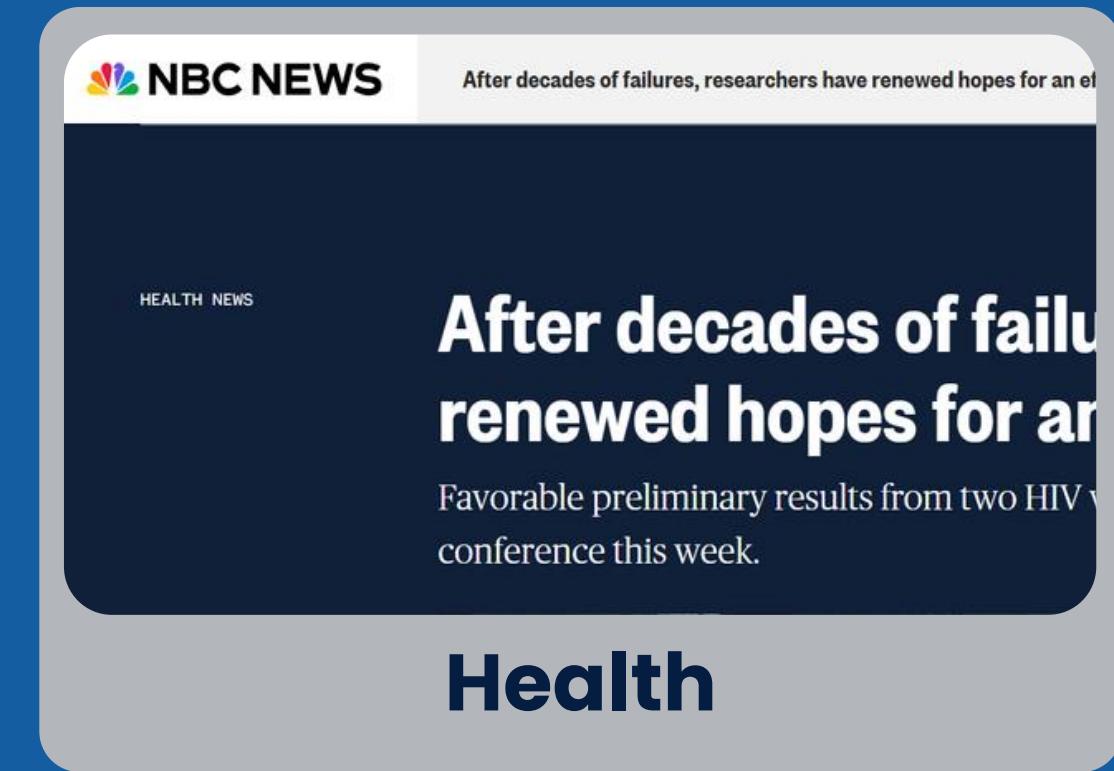
The image shows a screenshot of the NPR website's politics section. At the top, there's a navigation bar with links for NEWS, CULTURE, MUSIC, PODCASTS & SHOWS, and SEARCH. Below the navigation, there's a "POLITICS" category header. A main headline reads "Georgia on the mind of the Trump and Biden campaigns as the key state holds primary", with a timestamp "MARCH 12, 2024 - 5:00 AM ET". A large blue button at the bottom of the card says "Politics".



The image shows a screenshot of the NPR website's business section. At the top, there's a navigation bar with links for NEWS, CULTURE, MUSIC, PODCASTS & SHOWS, and SEARCH. Below the navigation, there's a "BUSINESS" category header. A main headline reads "The arts and crafts giant Joann files bankruptcy, but stores will remain open", with a timestamp "MARCH 18, 2024 - 12:10 PM ET". A large blue button at the bottom of the card says "Business".

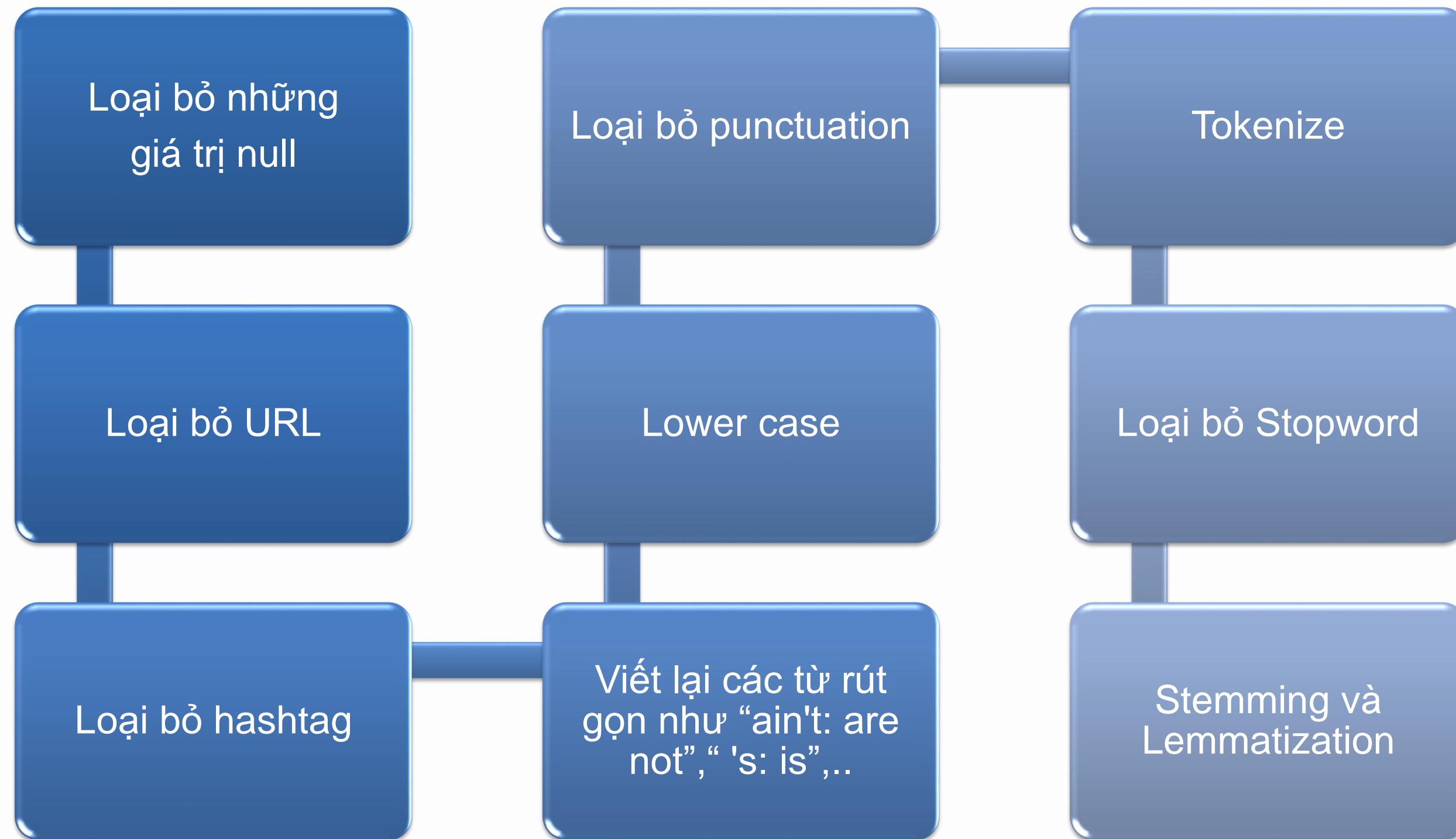


The image shows a screenshot of the NBC News website's sports section. At the top, there's a navigation bar with links for NEWSLETTERS, SIGN IN, and SEARCH. Below the navigation, there's a "SPORTS" category header. A main headline reads "Jason Kelce, Super Bowl champ and star Philadelphia Eagles center, ends career after 13 seasons on Broad Street.", with a timestamp "MARCH 12, 2024 - 5:00 AM ET". A large blue button at the bottom of the card says "Sports".



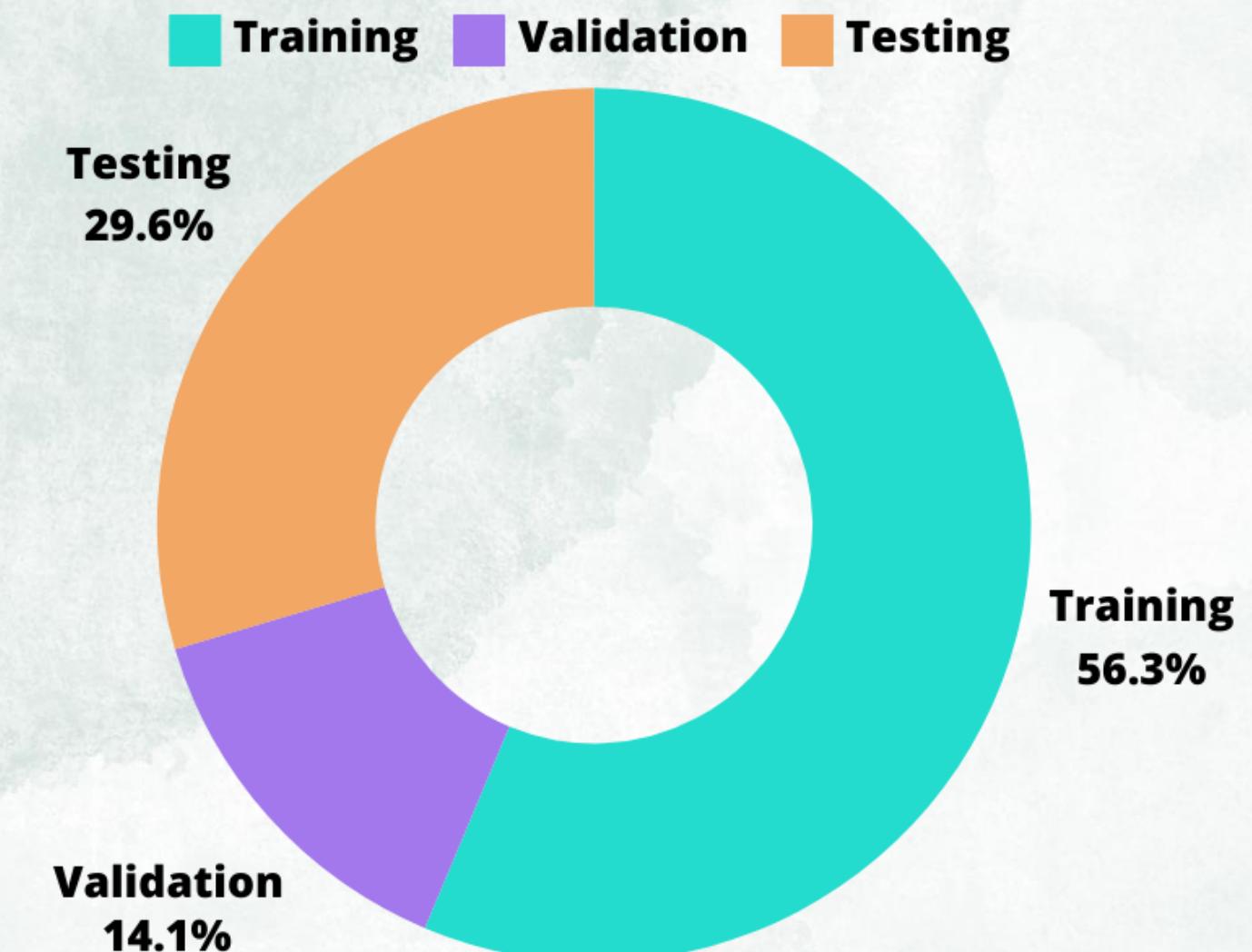
The image shows a screenshot of the NBC News website's health section. At the top, there's a navigation bar with links for NEWSLETTERS, SIGN IN, and SEARCH. Below the navigation, there's a "HEALTH NEWS" category header. A main headline reads "After decades of failure, researchers have renewed hopes for an effective HIV vaccine", with a timestamp "MARCH 18, 2024 - 12:10 PM ET". A large blue button at the bottom of the card says "Health".

Tiền xử lý dữ liệu



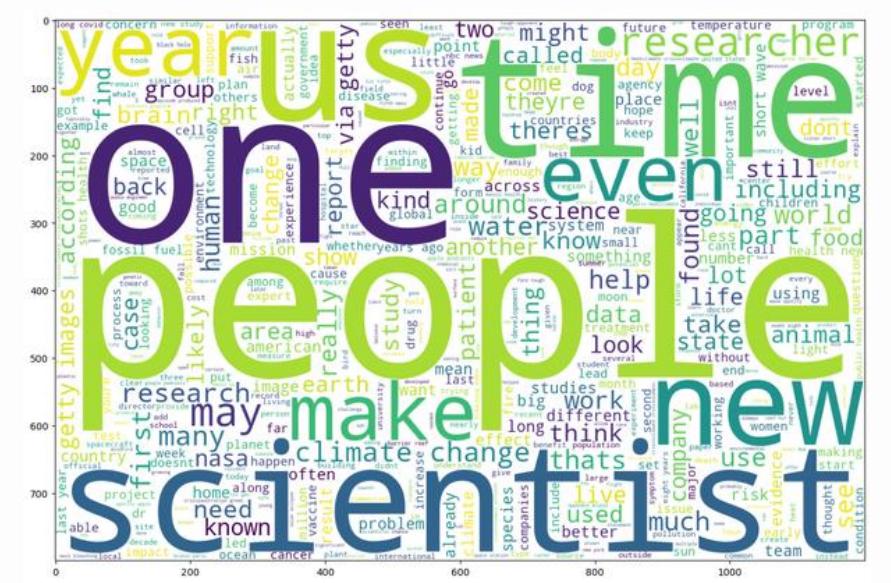
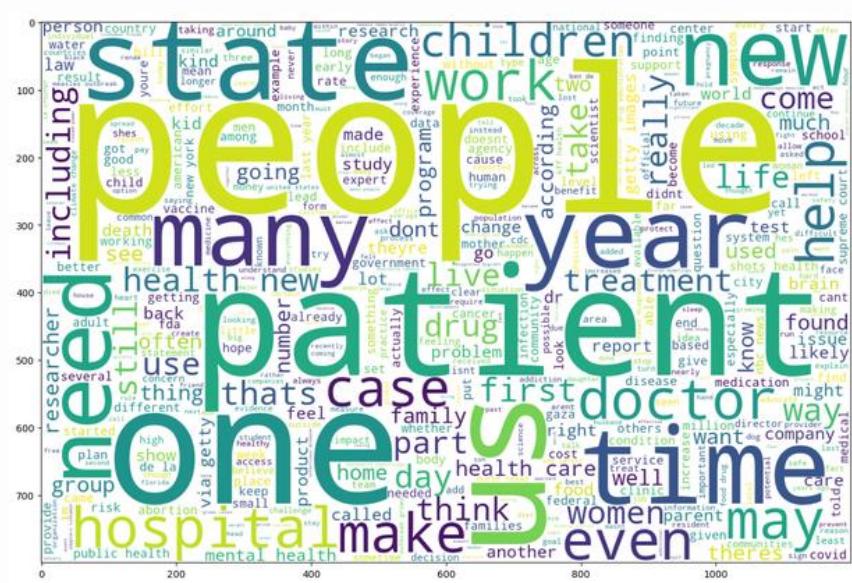
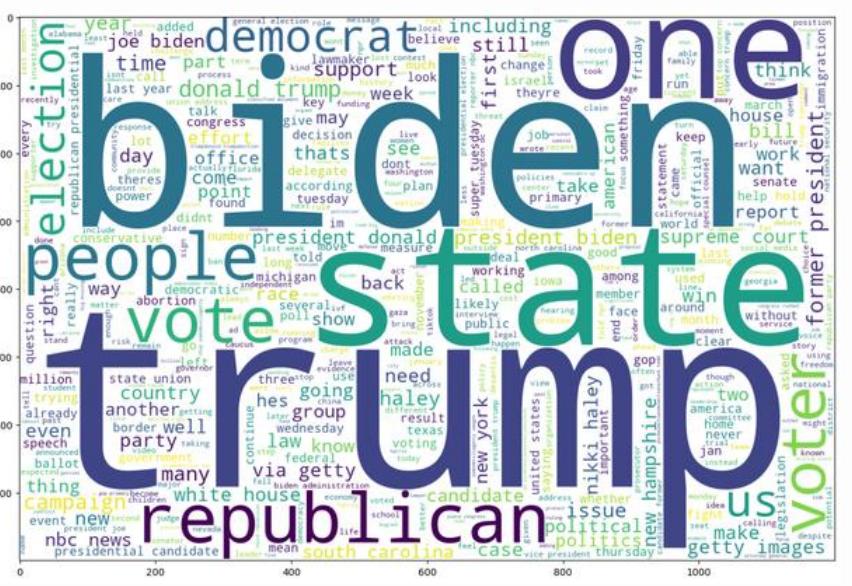
Thống kê dữ liệu

Data	Number of samples
Training	952
Validation	238
Testing	500



Visualization

Visualize những từ phổ biến của từng nhân



Machine Learning Models

SVM là model tốt nhất

Model list	Embedding method	List Params	Best accuracy
Support Vector Machine	Word2Vec	C ~ 0.4, kernel: linear gamma: scale	84,6
Random Forest	Word2vec 9	n_estimators = 46 max_depth: 5 min_samples_split: 18 min_samples_leaf: 19 max_features: sqrt	82.4
Logistic Regression	Word2vec	No params are required	83.4

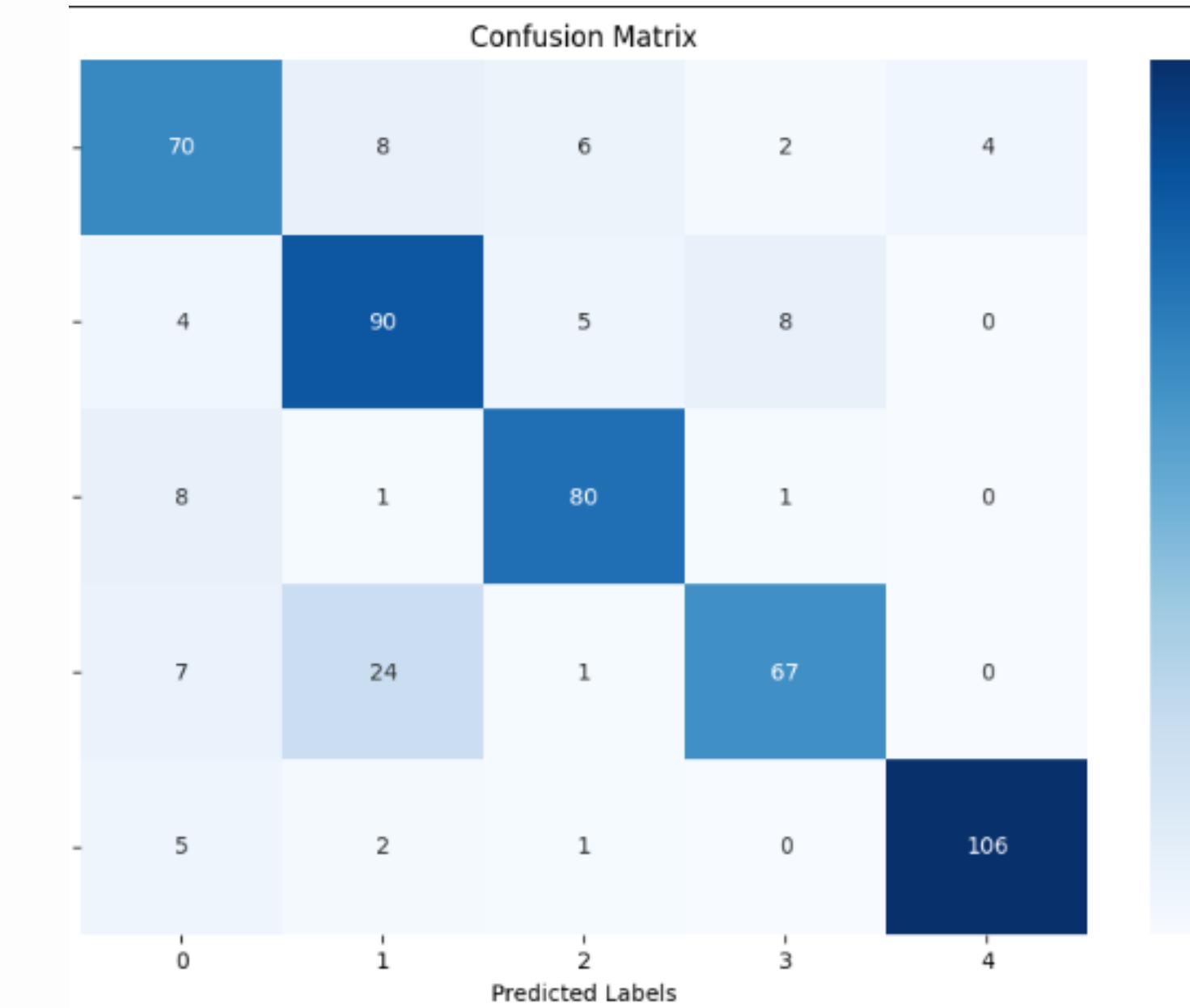
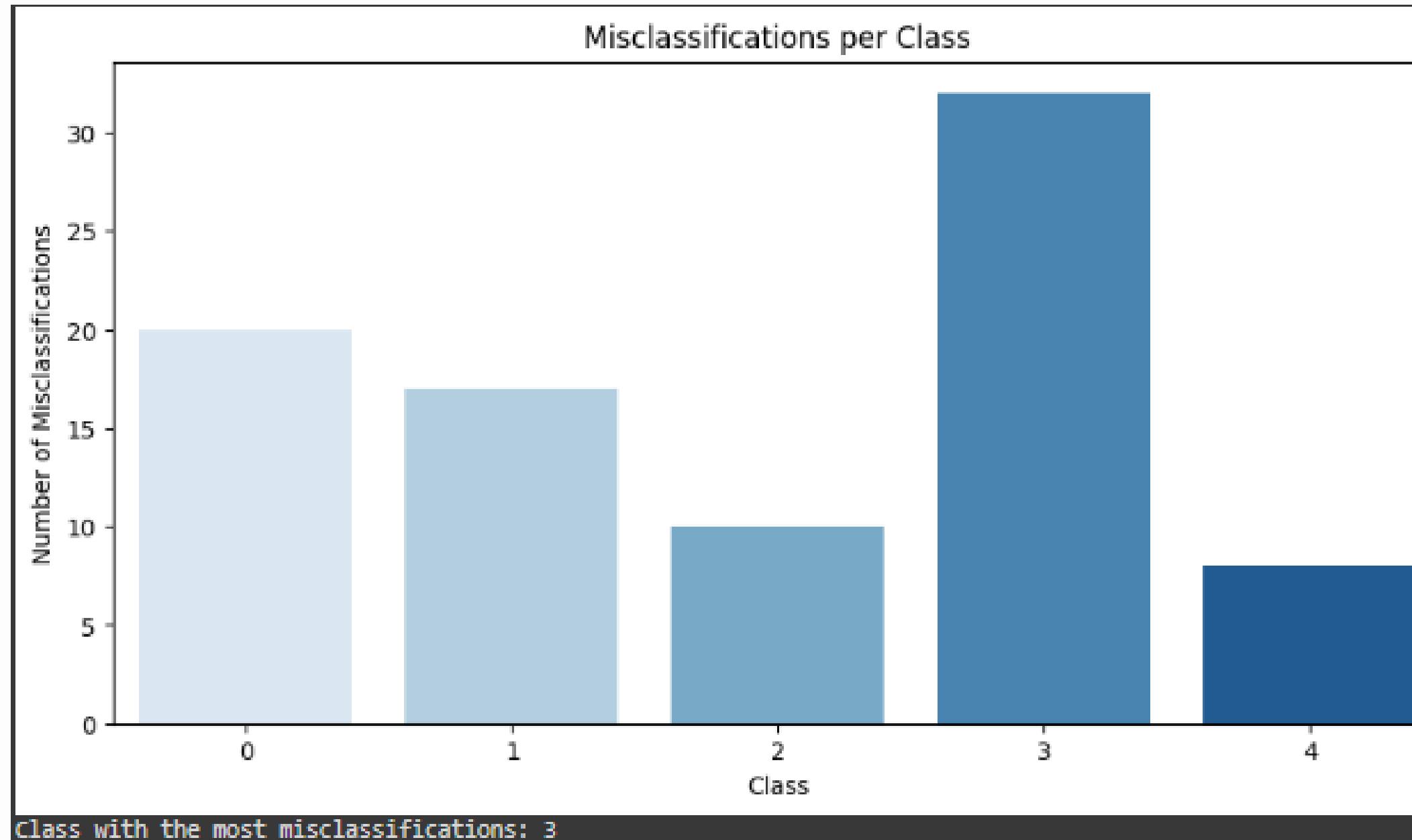
Deep Learning Models

CNN là model DL tốt nhất

Model list	Embedding method	List Params	Best accuracy
CNN	Keras Embedding Layer	embedding_dim = 300 num_filters = 100 activation= softmax batch_size = 32 9361805 params epoch = 5	80.2
LSTM	Keras Embedding Layer	embedding_dim = 300 activation='softmax' Batch_size = 32 lstm_units = 512 10667589 params epoch =	68
Bi-LSTM	Keras Embedding Layer	Batch_size = 10 gru_units = 512 embedding_dim = 300 activation= softmax 10252869 params	62

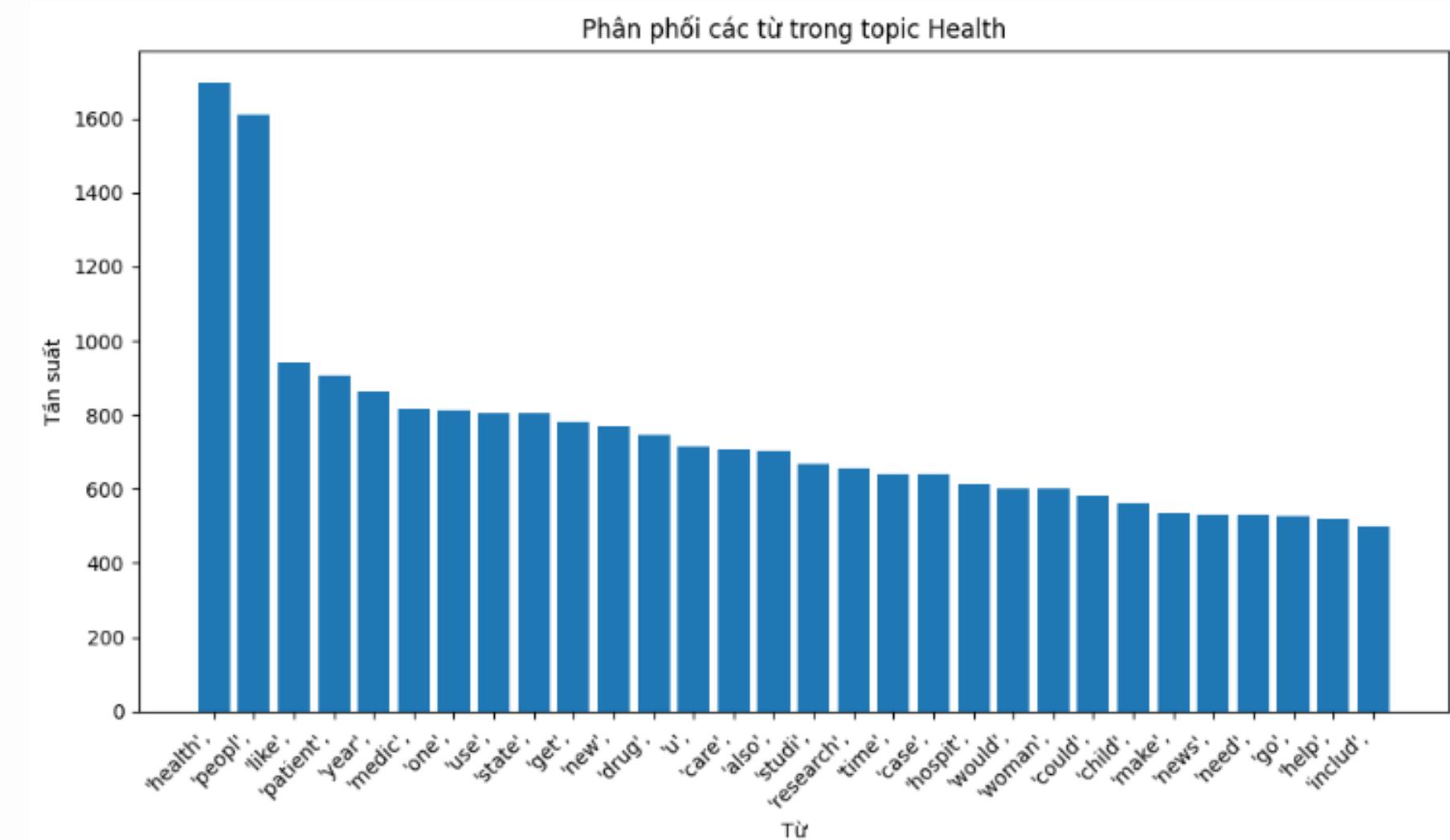
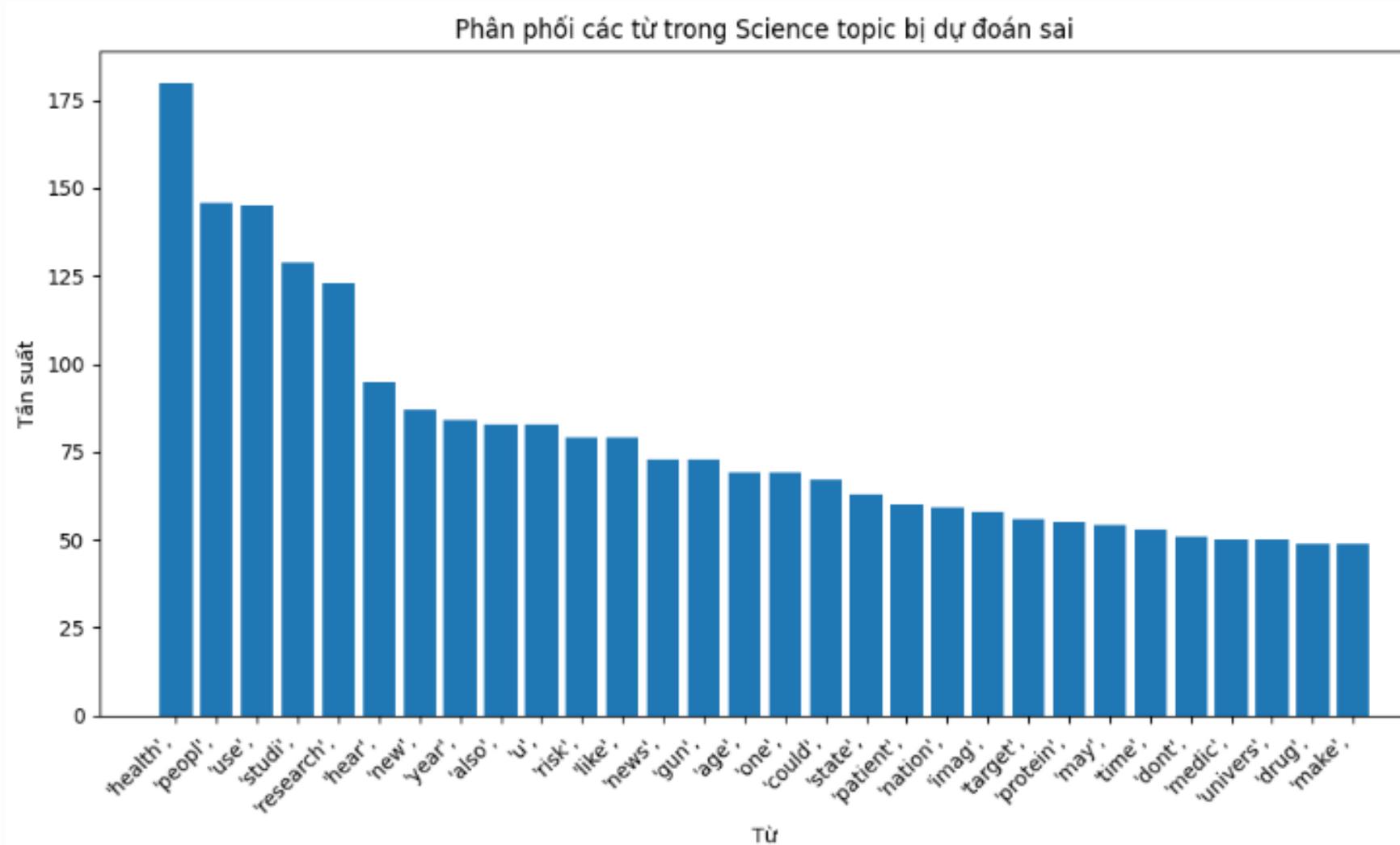
Phân tích kết quả

Hầu hết là class 3 (science) dự đoán nhầm lẫn sang class 1 (health)



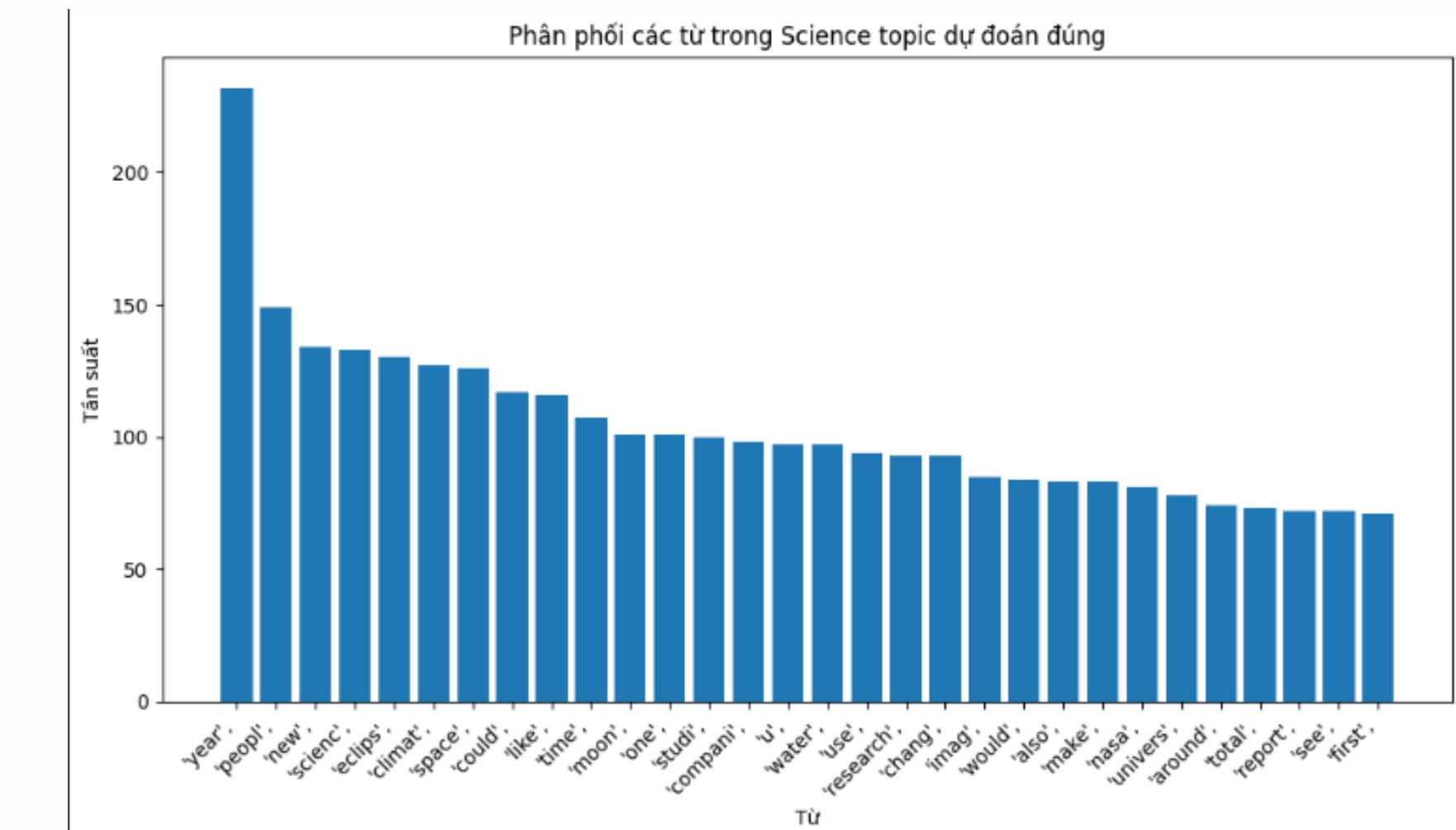
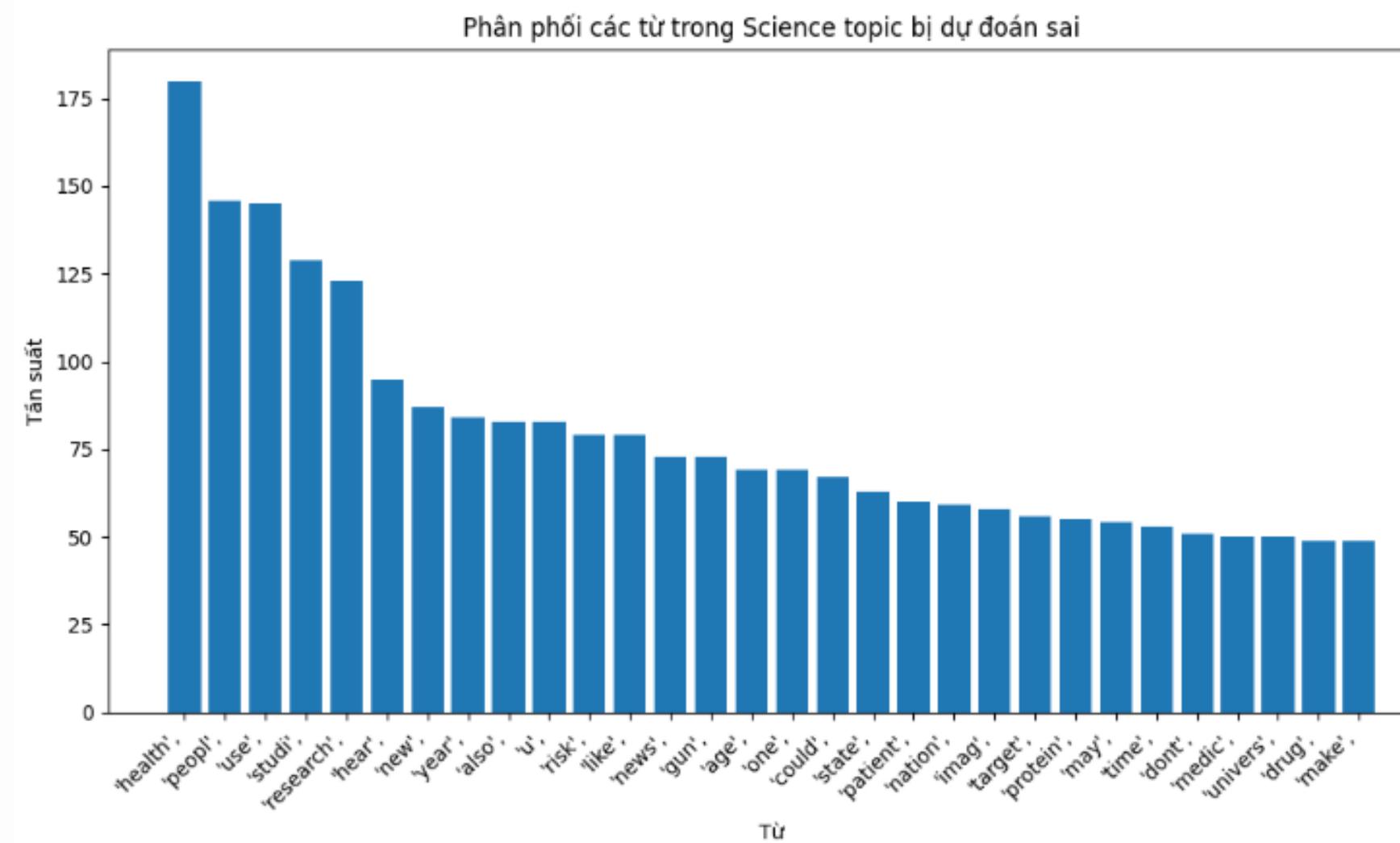
Phân tích kết quả

Ta có thể thấy phân phối của topic Science bị dự đoán sai rất giống với phân phối từ của topic Health nói chung



Phân tích kết quả

Có thể thấy phân phối các từ trong topic Science được dự đoán đúng
khá khác với phân phối các từ trong topic Science dự đoán sai



Phân tích kết quả

Có thể thấy nột số bài báo được gán 2 loại nhãn

```
[54] df3 = pd.read_csv("data.csv")
df3 = df3.dropna()
df3.shape
```

(1691, 4)

```
[55] df3.drop_duplicates(inplace=True)
df3.shape
```

(1691, 4)

```
[56] category = df3["category"]
df3.drop("category", axis=1, inplace=True)
df3.drop_duplicates(inplace=True)
df3.shape
```

(1569, 3)

npr.org/sections/science/



Justin Sullivan/Getty Images

CLIMATE

Why a town on the front line of America's energy transition isn't letting go of coal

March 28, 2024 • Kemmerer, Wyo., is on the front line of America's energy transition, with its coal plant slated to close and a nuclear plant in the works. But some think the rush to quit fossil fuels is impractical.



Kirk Siegler/NPR

HEALTH

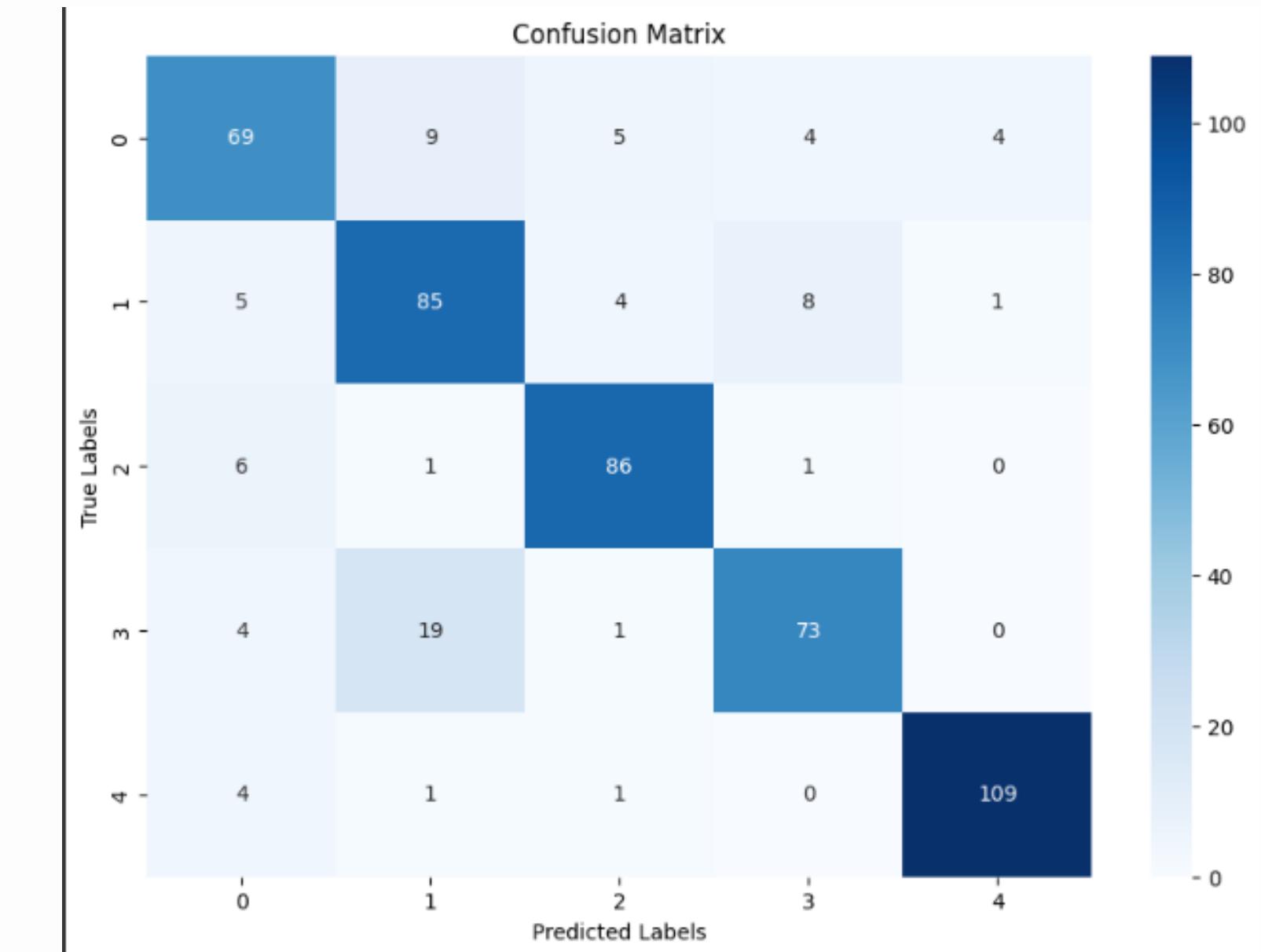
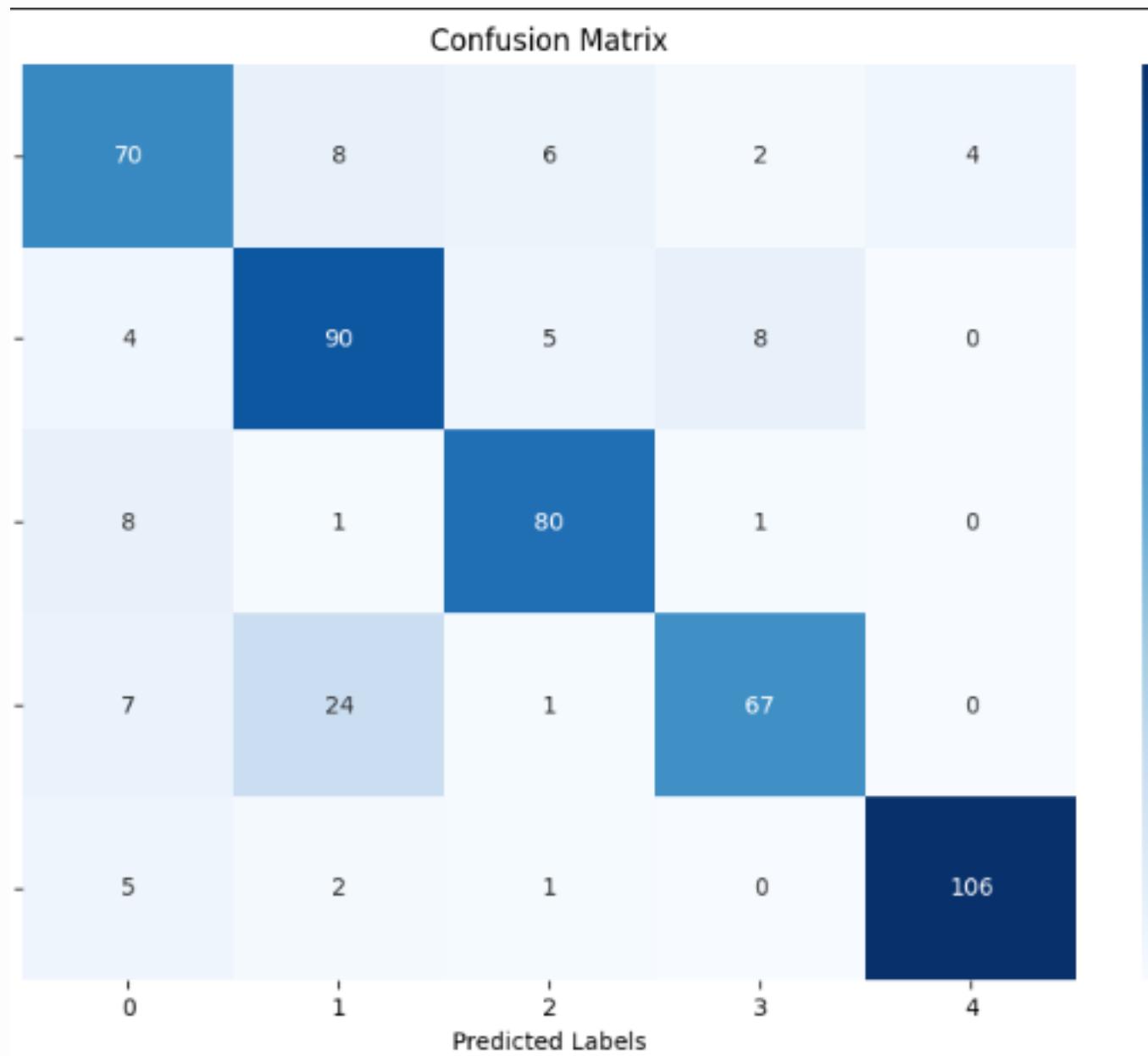
Here's what to know about dengue, as Puerto Rico declares a public health emergency

March 27, 2024 • Most people with dengue will show no signs of infection or experience only mild symptoms, but in rare cases



Phân tích kết quả

Sau khi drop những bài báo song ngữ



Demo

Dự đoán đúng

text

BNPL services have taken off among shoppers across income and credit levels for various reasons. Many are seeking cover from high credit card interest rates. Some, having burned through traditional credit options, are desperate for financial lifelines. Others are simply looking to better manage their cash flow.

The fastest uptake has been among consumers 35 and younger, who represent more than half of BNPL borrowers, LexisNexis Risk Solutions found late last year. Many are increasingly using the loans for daily essentials, not just big-ticket purchases. While some already see them as a routine tool in their wallets, others, like Whiteside, are turning away in alarm.

"I can pay on my credit cards more freely if I don't have that other consumer debt," Whiteside has since realized, referring to her existing \$10,000 card balance. After trimming her discretionary spending and sticking to home-cooked meals, she said she's been able to whittle down her BNPL debt to about \$1,200.

As BNPL usage soars, financial experts and researchers have raised alarms about risky spending on the platforms, even though they can often be used responsibly.

"I'm sure there are people who use it well, but on average, we feel it kind of replaces the credit card," said Ben Lourie, an accounting professor at the University of California, Irvine. "People are consuming extra. There's just no way around it."

Lourie and fellow researchers at UC Irvine, Stanford and Singapore Management University¹⁶

Clear Submit

Use via API · Built with Gradio

Demo

Dự đoán sai

text

While the jackpot wasn't won in the drawing on Saturday, four tickets won \$1 million each by matching the first five numbers. Those tickets were sold in Illinois, Louisiana, Michigan and Pennsylvania, according to the Powerball website.

Powerball's grand prize was last won in January, when a ticket in Michigan scored a \$842.4 million jackpot. Since then, 38 consecutive drawings have taken place without a jackpot winner, according to Powerball.

The largest Powerball jackpot – and the largest US lottery prize – ever won was \$2.04 billion by a ticket purchased in California in November 2022, according to the lottery.

Ranking second through fourth are \$1.765 billion (one ticket in California; 2023); \$1.586 billion (three tickets, 2016); and \$1.08 billion (one ticket in California, 2023). Winning the Powerball jackpot means a ticket matched all five white balls plus the red Powerball. The odds of winning any prize in a Powerball drawing are 1 in 24.9, and the odds of winning the jackpot are 1 in 292.2 million, according to the lottery.

Powerball tickets cost \$2 per play and are sold in 45 states, the District of Columbia, Puerto Rico and the US Virgin Islands. Drawings are held Mondays, Wednesdays and Saturdays at 10:59 p.m. ET in Tallahassee, Florida.

17

Clear

Submit

output

Sport

Flag

Kết luận

- Sau khi giải quyết bài toán thì model SVM là model có accuracy cao nhất với ~85%.
- Các mô hình state of the art deep learning phù hợp với bài toán như LTSM, Bi-LSTM hay GRU lại có accuracy khá thấp có thể do
- Nhãn Science là nhãn bj dự đoán sai nhiều nhất



MỞ RỘNG

- Crawl thêm dữ liệu
- Thử nghiệm các model như BERT, Attention Method, ...
- Sử dụng feature reduction như LDA, PCA có thể tăng hiệu năng tính toán
- Thử ensemble deep learning method
- Xây dựng mô hình dự đoán các nhãn có thể với những bài báo song nhãn.



Contribution



Phạm Quang Vinh

- Chọn model
- Huấn luyện model
- Tuning parameter
- Phân tích, đánh giá kết quả
- Làm slide



Vũ Minh Tiến

- Crawl data
- Gán nhãn
- Viết annotation guideline
- Demo
- Làm slide



Phạm Thị Kim Huệ

- Gán nhán
- Viết annotation guideline
- Tiền xử lý dữ liệu
- Visualize
- Làm slide



THANK YOU!

