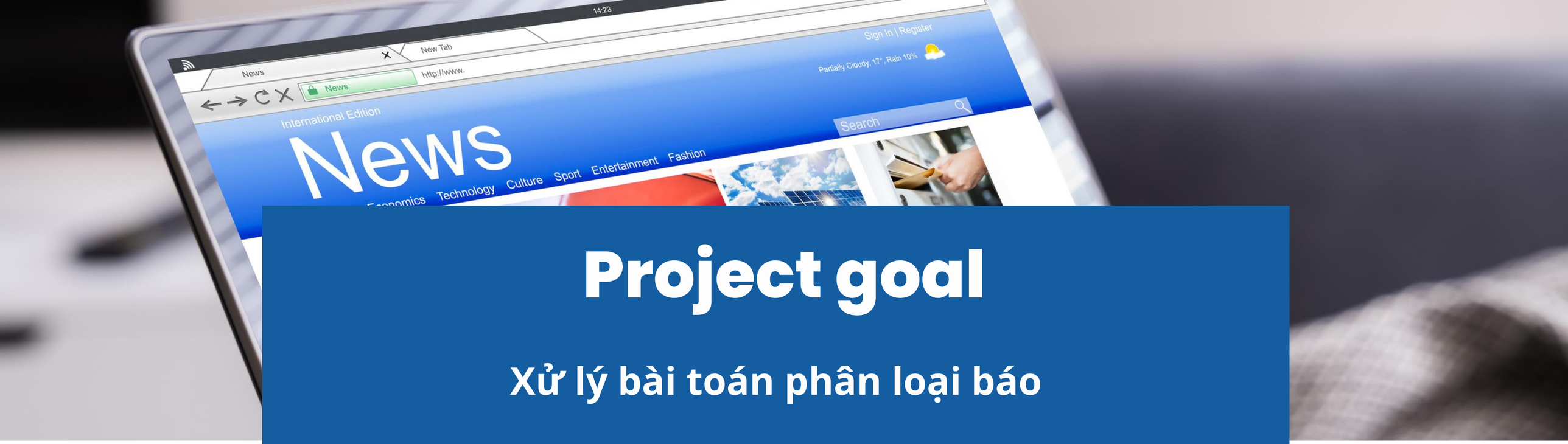
Data Mining

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

Step 01

Crawl data

Step 02

Gán nhãn dữ liệu

Step 03

Xử lý dữ liệu

Step 04

Visualize dữ liệu

Step 05

Feature extraction

Step 06

Training model

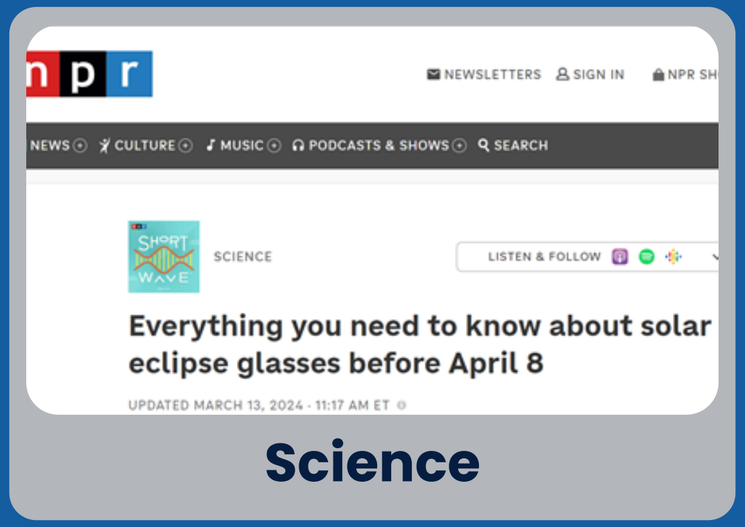
Step 07

Phân tích kết quả

Step 08

Demo

Crawling data & Data annotation

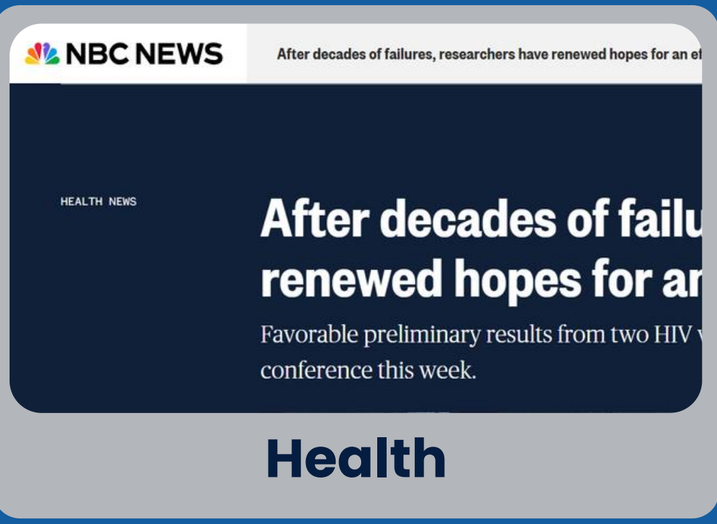


Science

Politics

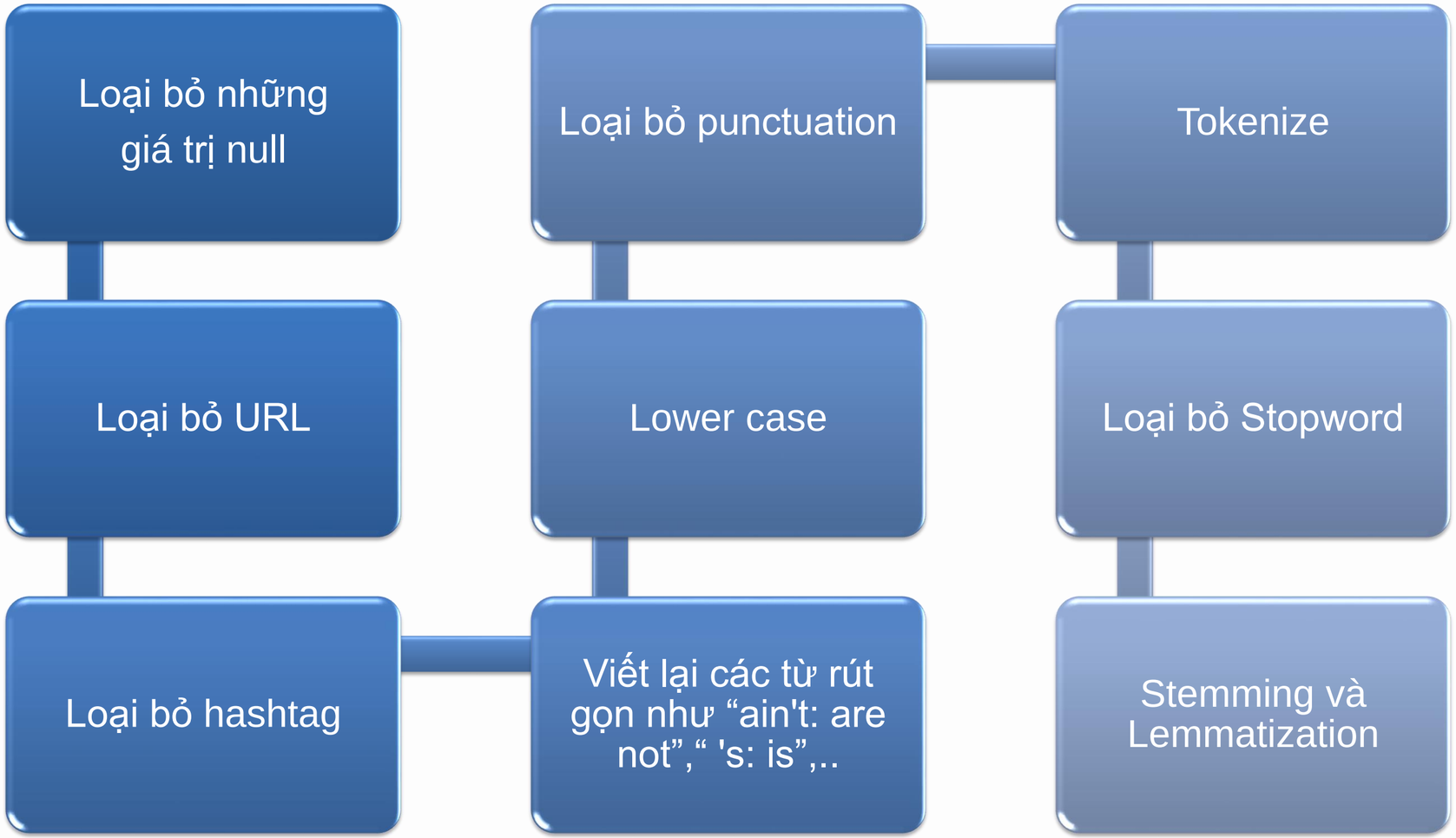
Business

Sports



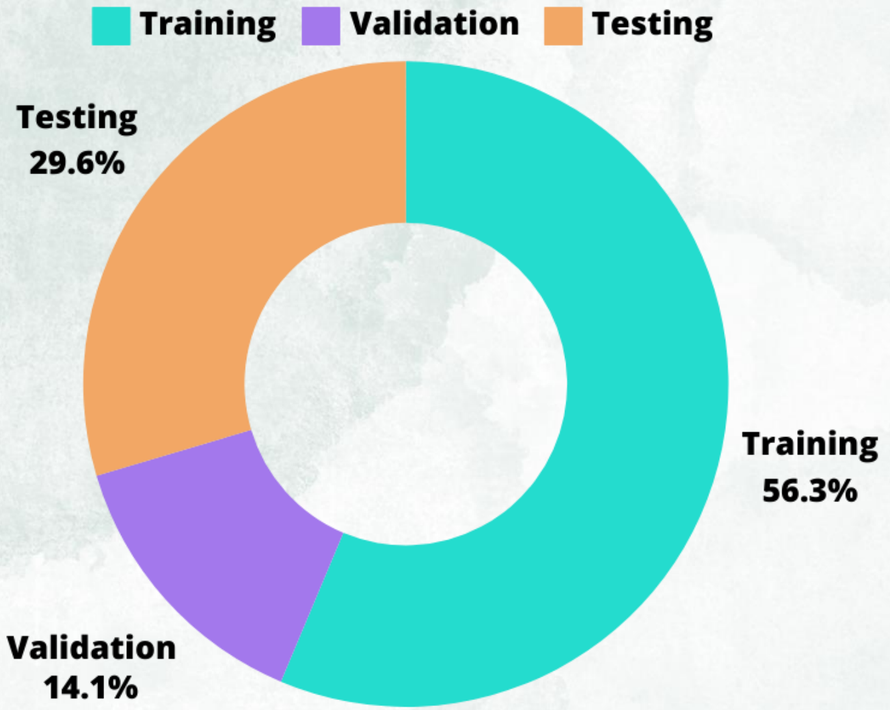
Health

Tiền xử lý dữ liệu



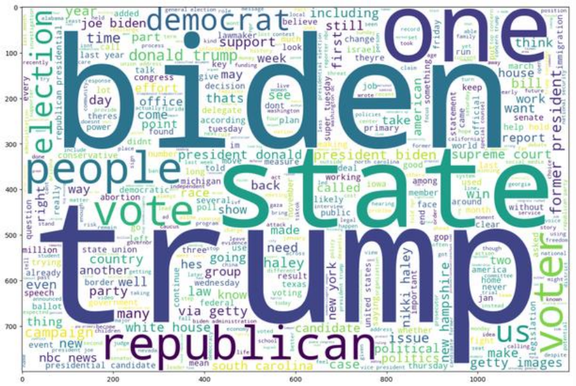
Thống kê dữ liệu

Data	Number of samples
Training	952
Validation	238
Testing	500



Visualization

Visualize những từ phổ biến của từng nhãn



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	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

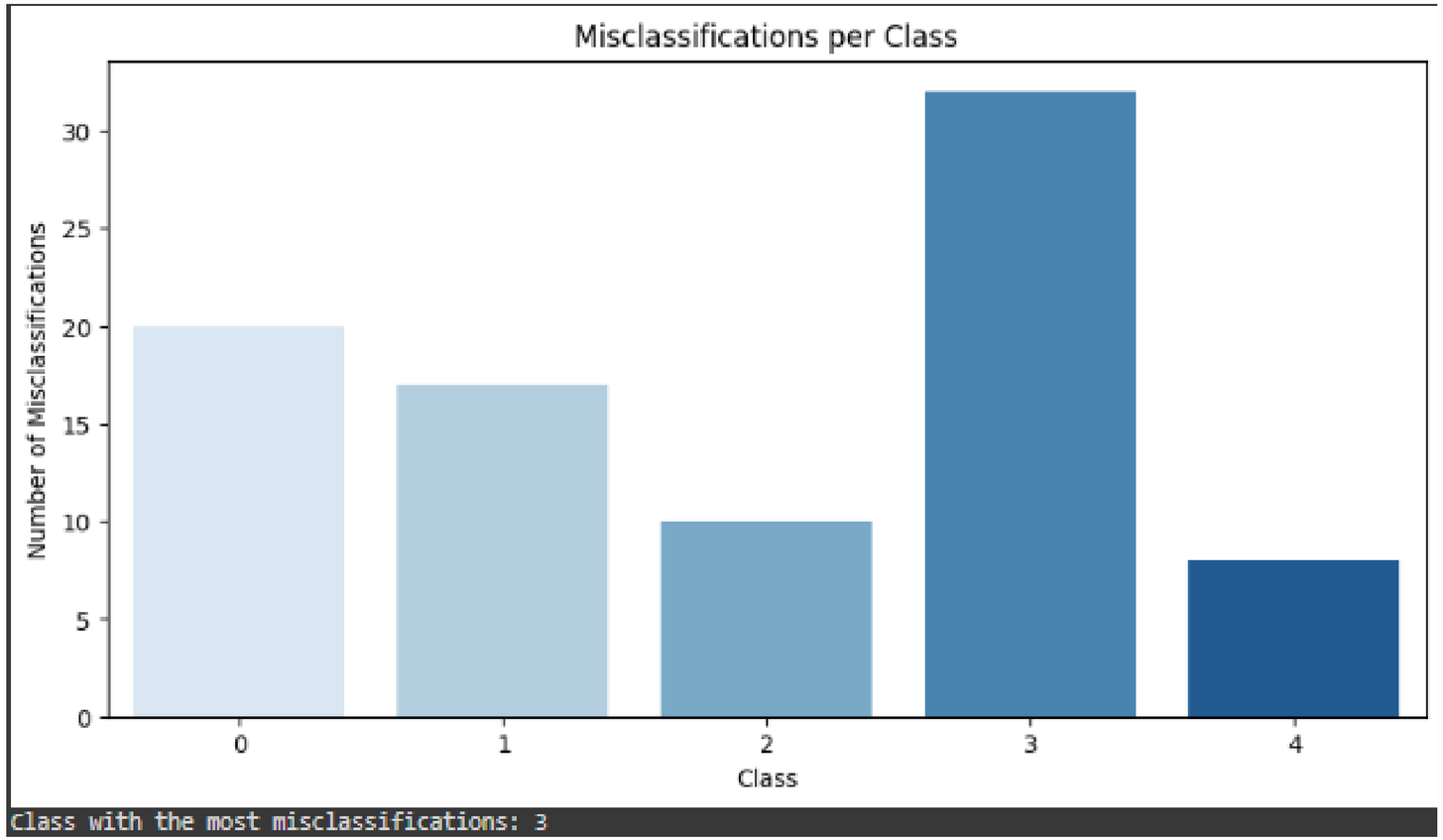
Deep Learning Models

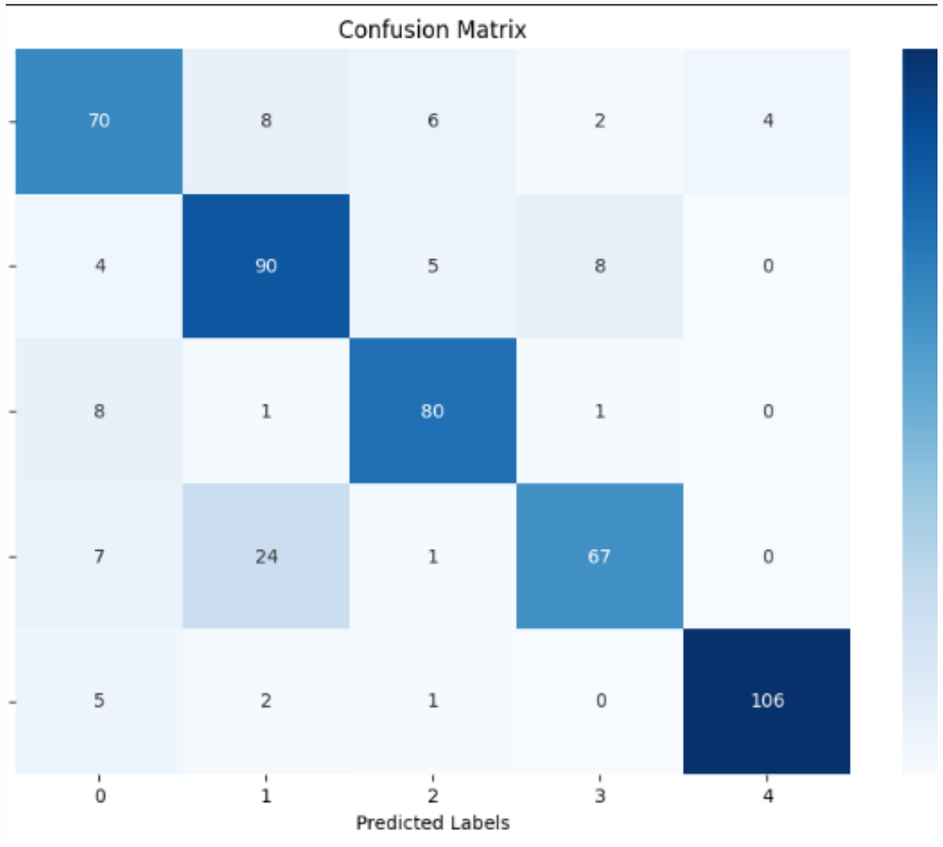
CNN là model DL tốt nhất

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 10	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

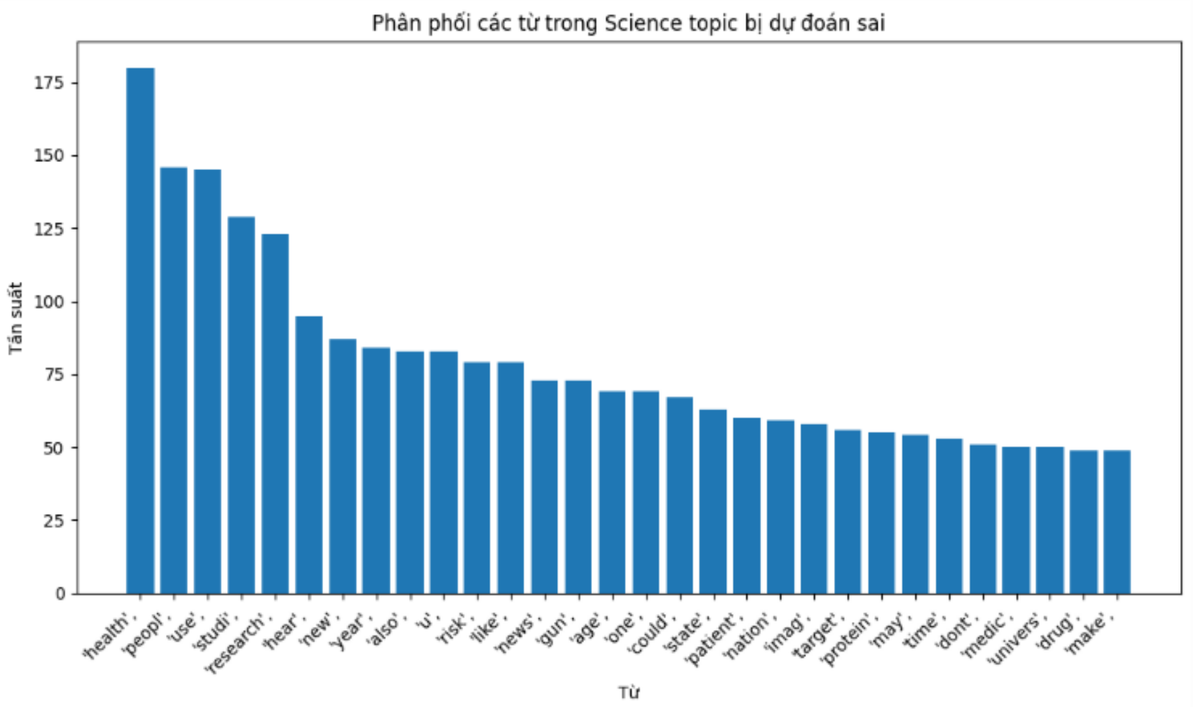
Hầu hết là class 3 (science) dự đoán nhầm lẫn sang class 1 (health)

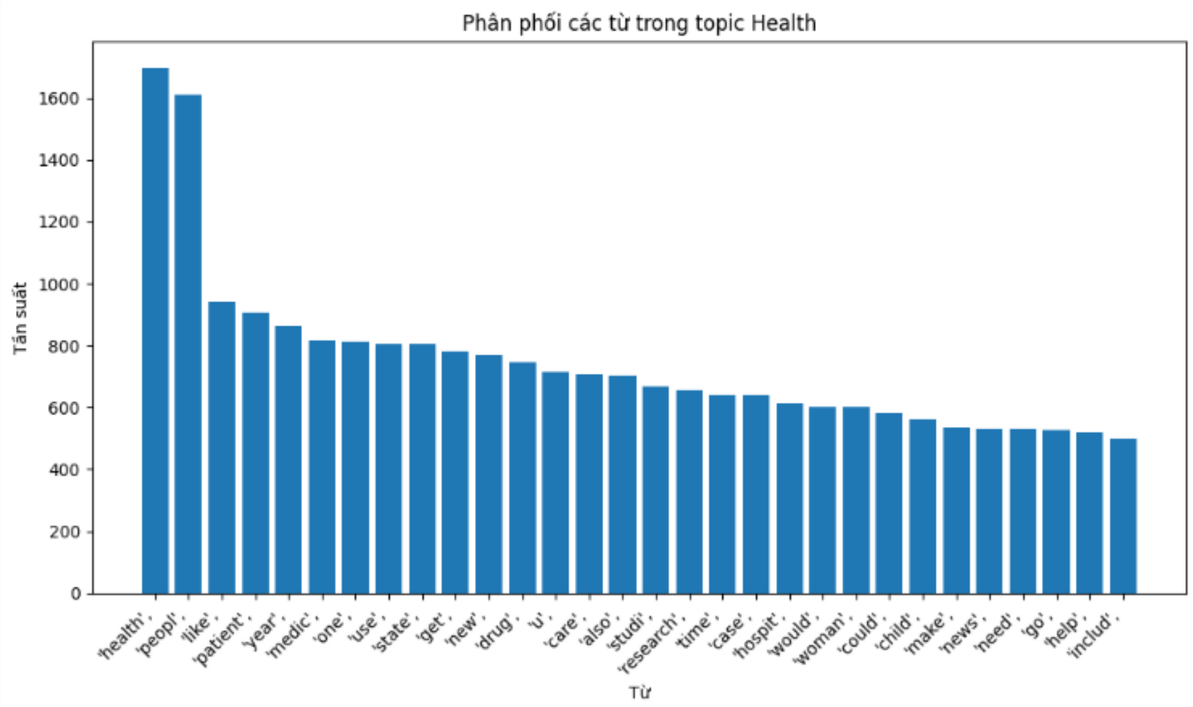




Phân tích kết quả

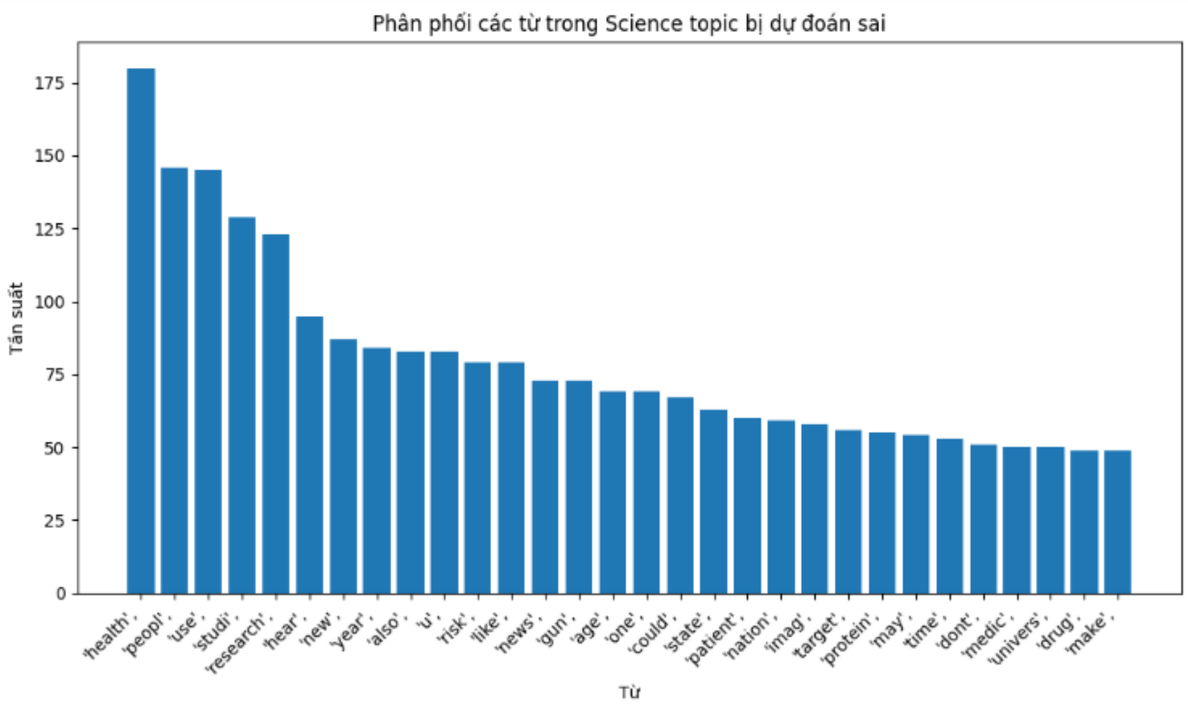
Ta có thể thấy phân phối của topic Science bị dự đoán sai rất giống với phân phối từ của topic Health nói chung

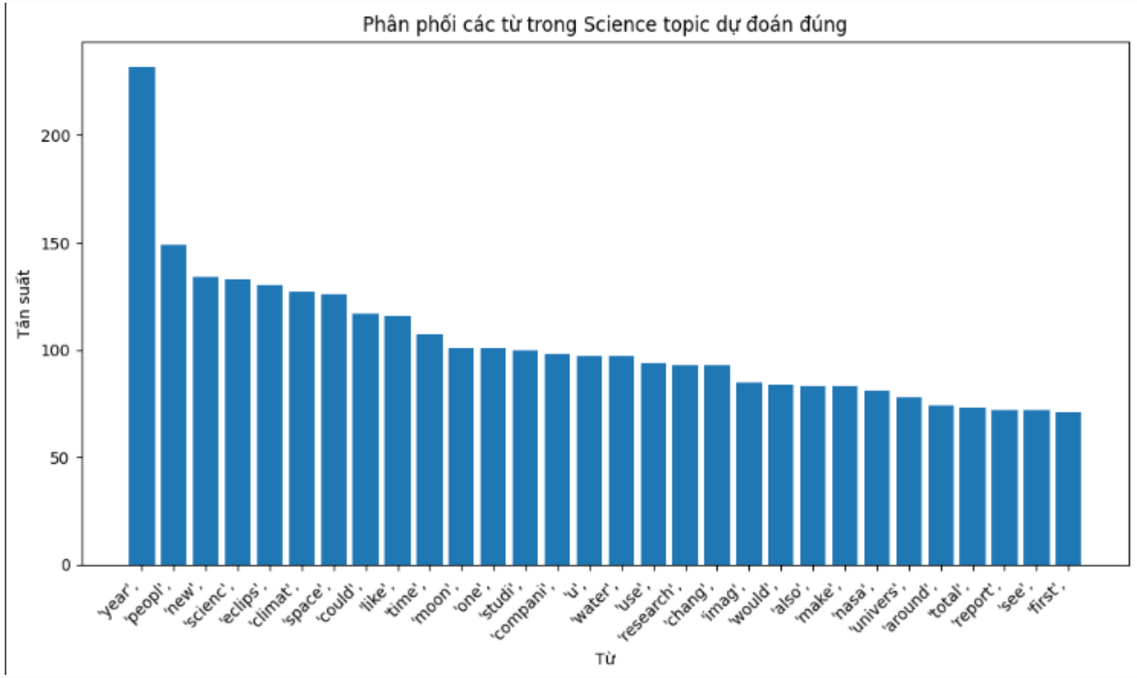




Phân tích kết quả

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

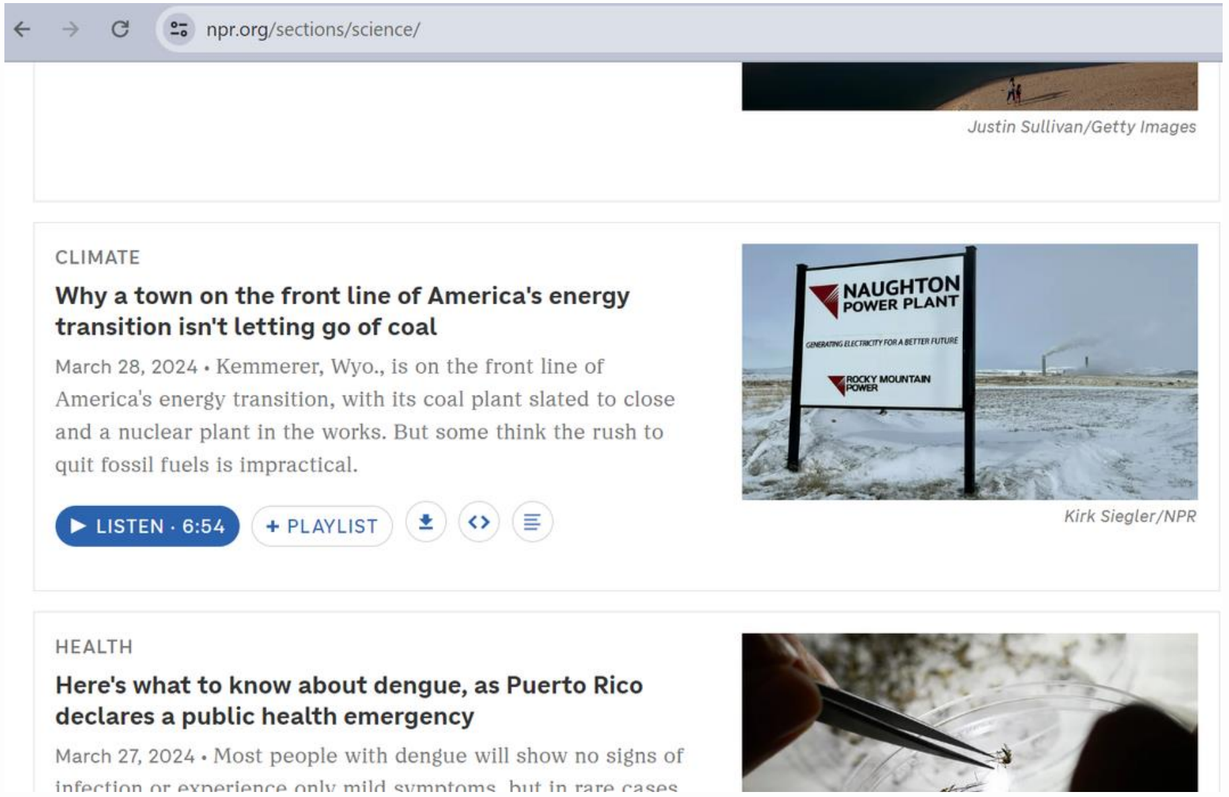




Phân tích kết quả

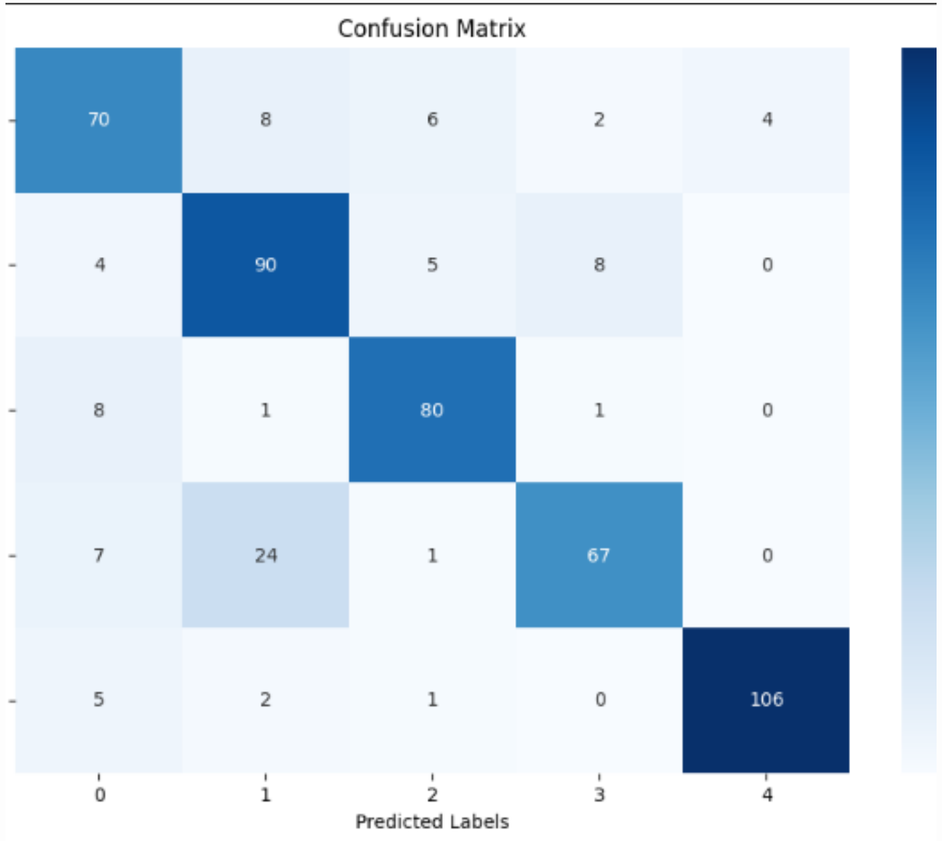
Có thể thấy một số bài báo được gán 2 loại nhãn

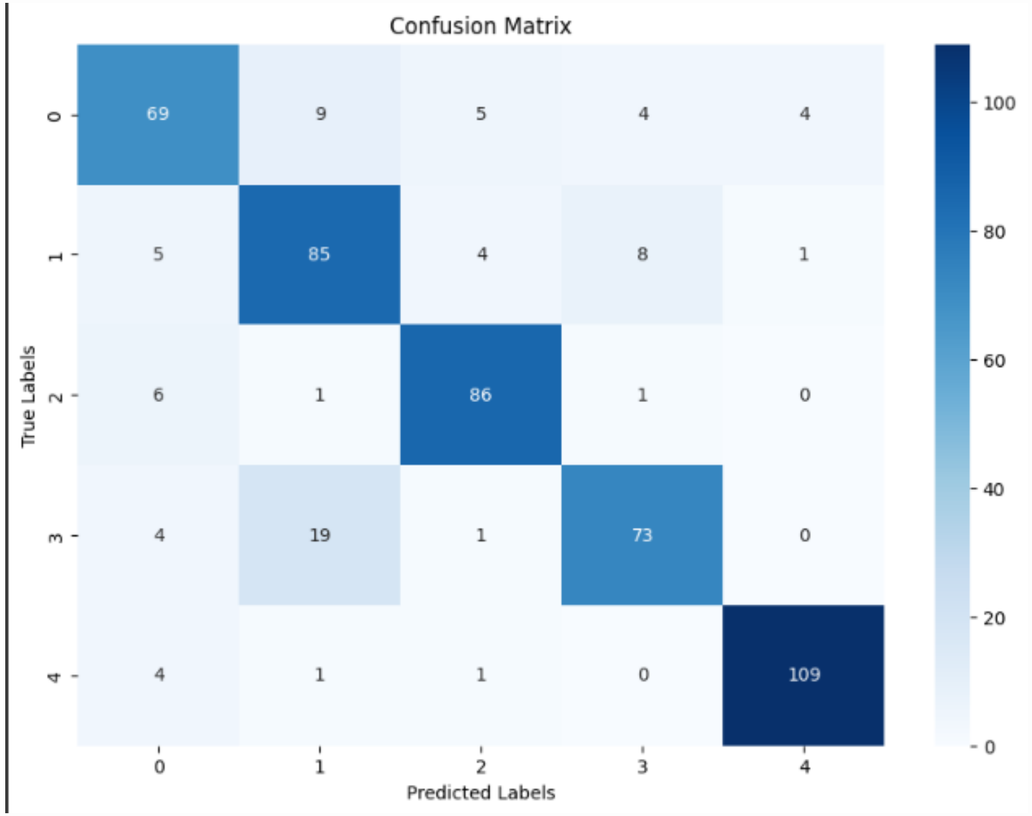
```
[54] df3 = pd.read_csv("data.csv")
     df3 = df3.dropna()
     df3.shape

     (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)
```



Phân tích kết quả

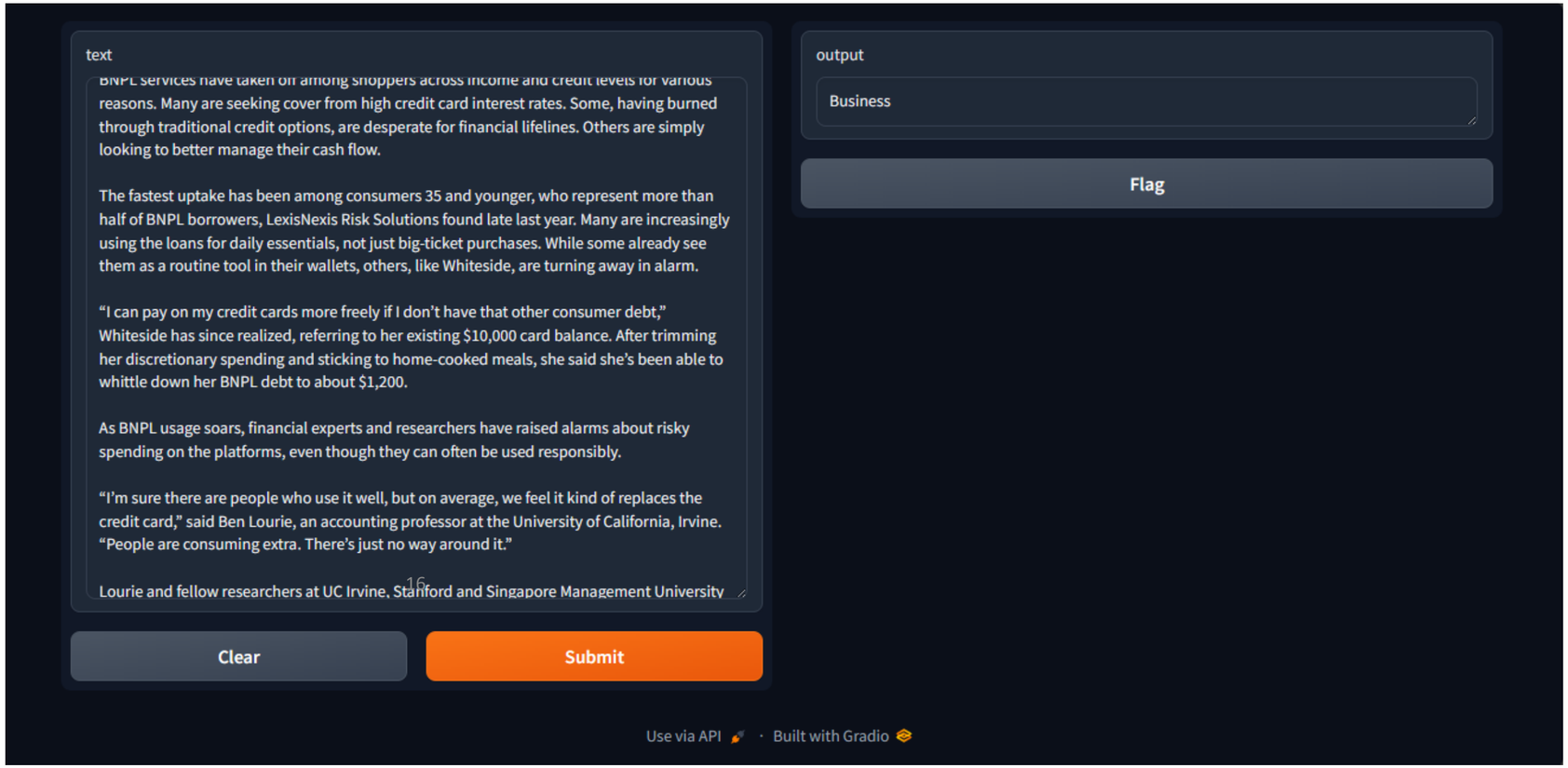
Sau khi drop những bài báo song nhãn





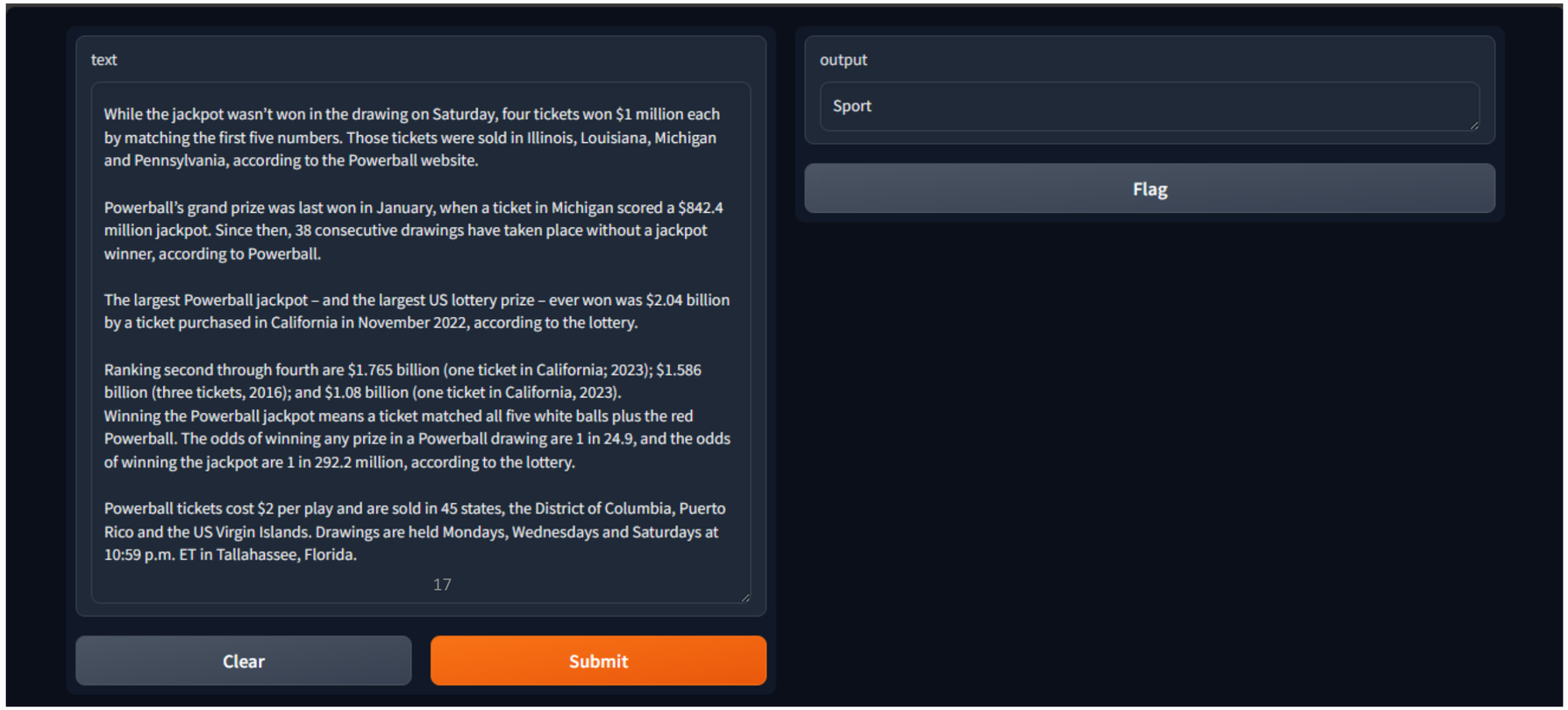
Demo

Dự đoán đúng



Demo

Dự đoán sai



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 có accuracy khá thấp.
- Nhãn Science là nhãn bị 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!


DATA MINING
knowledge
visualization
digital marketing
database
artificial
extract
online
complexity
machine learning
solution
information