

# KHAI THÁC THÔNG TIN

## ĐỀ TÀI 1

**Nghiên cứu về việc áp dụng các mô hình học sâu/ transformer  
trong việc truy hồi thông tin tiếng việt  
giúp tìm kiếm các văn bản tương đồng về mặt nội dung**

GVHD: TS. Phạm Thế Anh Phú

Phan Thị Anh - 2541861001

Trần Sỹ Huy- 2541861009

Phan Hoàng Kha - 2541861010

Võ Hà Nam - 2541861017

Quách Đình Nhân – 2541861020

# NỘI DUNG

1. Giới thiệu Doc2Vec
2. Giới thiệu PhoBERT
3. Tiền Xử lý văn bản
4. Vector hoá dữ liệu sử dụng mô hình PhoBERT
5. Tính toán tương đồng và chọn top- rank bằng cosine
6. Đánh giá thực nghiệm và hướng phát triển

# 1 - GIỚI THIỆU DOC2VEC

## a. Giới thiệu:

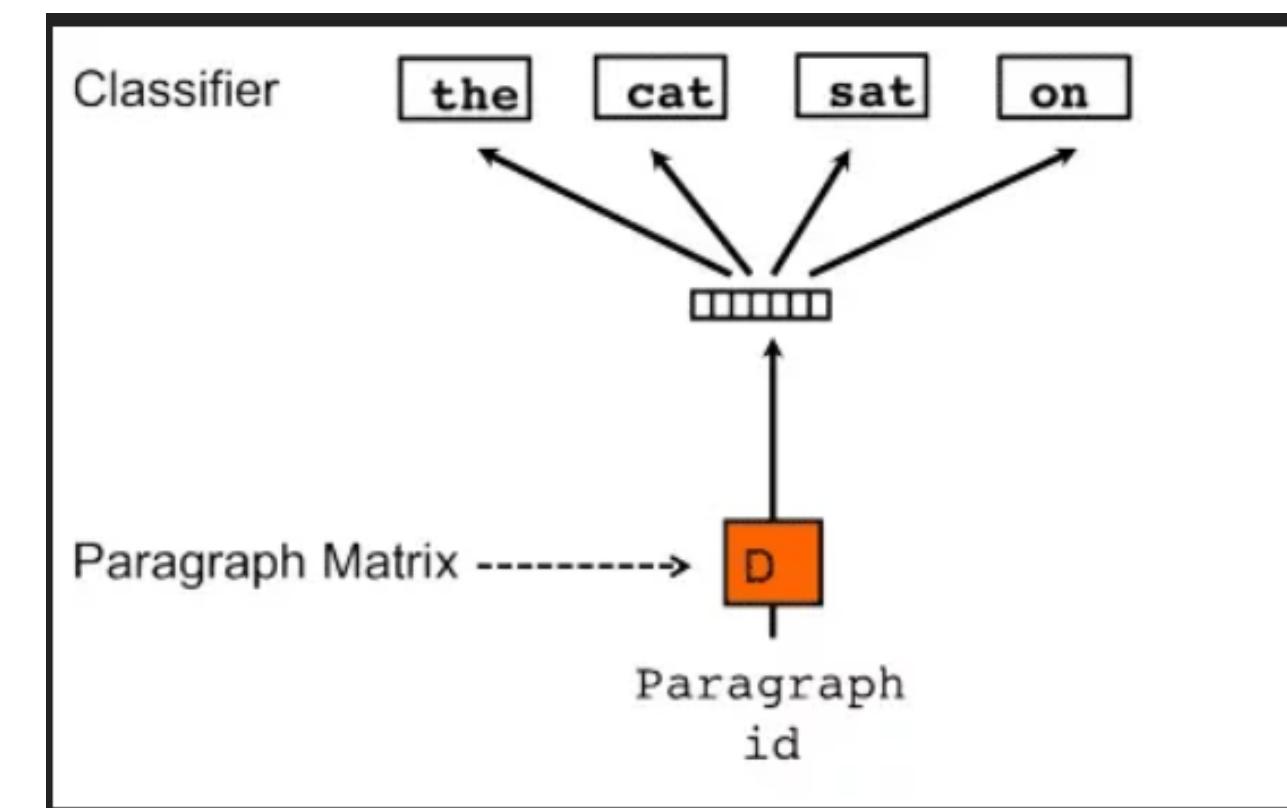
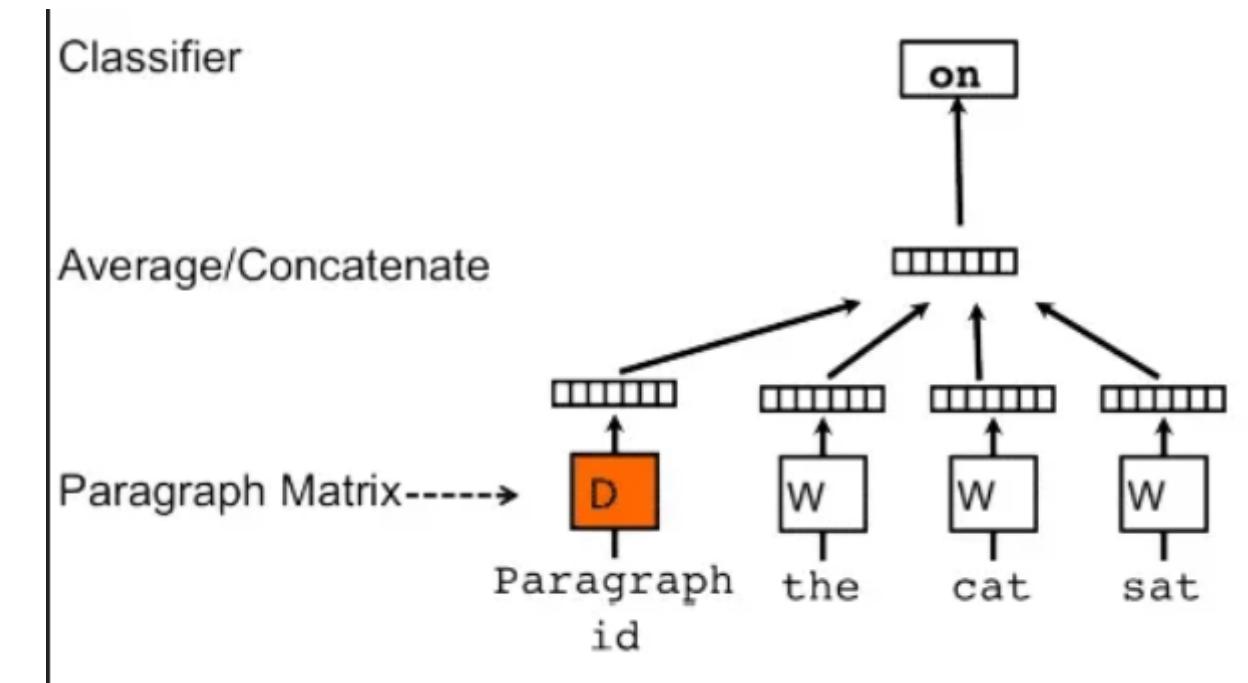
- Năm 2014, ông Tomas Mikolov (một trong những tác giả của Word2Vec) và ông Lê Quốc Vinh đã giới thiệu phương pháp Doc2Vec.
- Doc2vec (Document-to-Vector):
  - Đây là phương pháp giúp chuyển đổi dữ liệu dạng text theo cấp tài liệu/văn bản sang vector giúp máy tính có hiểu hiểu được.
  - Đây là một phần mở rộng của mô hình Word2Vec.

# 1 - GIỚI THIỆU DOC2VEC

## b. Phương pháp:

Có hai phương pháp của Doc2vec chính là:

- Distributed Memory (PV-DM) – Paragraph Vector – Distributed Memory.
- Distributed Bag of Words (PV-DBOW) – Paragraph Vector – Distributed Bag of Words.



# 1 - GIỚI THIỆU DOC2VEC

## c. Đánh giá:

- Ưu điểm:
  - Biểu diễn ngữ nghĩa tốt hơn mô hình truyền thống.
  - Khả năng hiểu theo ngữ cảnh.
- Nhược điểm:
  - Không áp dụng cho tài liệu quá ngắn.
  - Dữ liệu không được xử lý tốt khi tài liệu gấp nhiều lối chính tả, ...

# 1 - GIỚI THIỆU DOC2VEC

## d. So sánh Doc2Vec và Word2Vec:

Nội dung	Word2Vec	Doc2Vec
Mục đích	<ul style="list-style-type: none"><li>Chuyển đổi từ thành Vecto</li></ul>	<ul style="list-style-type: none"><li>Chuyển đổi văn bản/tài liệu thành Vecto</li></ul>
Độ dài văn bản xử lý	<ul style="list-style-type: none"><li>Thích hợp cho cụm từ, câu ngắn</li></ul>	<ul style="list-style-type: none"><li>Tốt cho đoạn văn, tài liệu dài</li></ul>
Ưu điểm	<ul style="list-style-type: none"><li>Nhanh, đơn giản</li><li>Hiệu quả cho bài toán ngữ nghĩa từ</li></ul>	<ul style="list-style-type: none"><li>Gán vector cho cả văn bản</li></ul>
Nhược điểm	<ul style="list-style-type: none"><li>Không biểu diễn được ngữ nghĩa toàn văn bản</li></ul>	<ul style="list-style-type: none"><li>Phức tạp hơn</li><li>Cần nhiều dữ liệu để học tốt</li></ul>

## 2 - GIỚI THIỆU PhoBERT

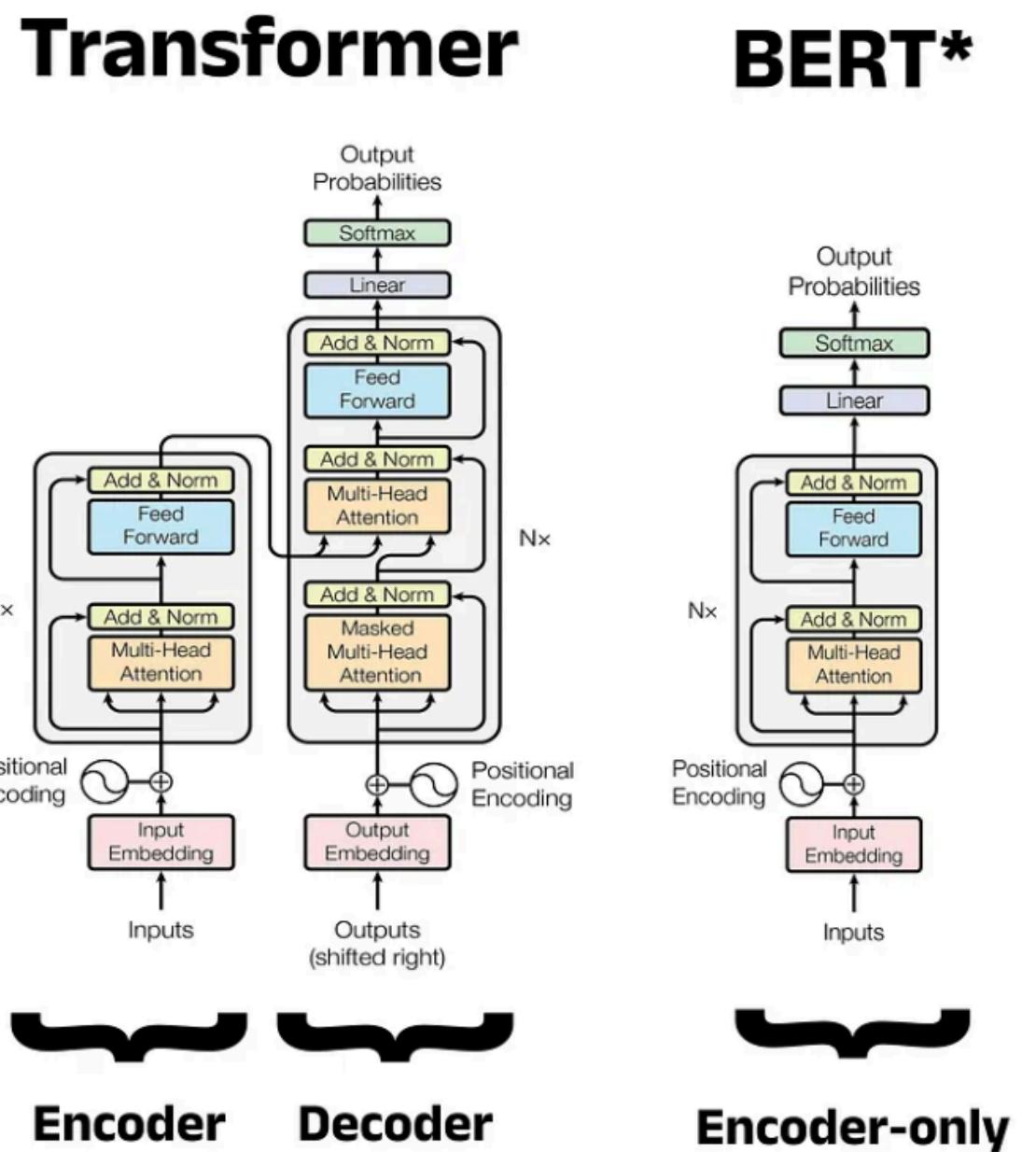
**BERT là gì?**

**BERT – Bidirectional Encoder Representation from Transformer**

do Google AI phát triển dựa trên kiến trúc transformers.

BERT được giới thiệu vào tháng 10/2018 qua bài báo "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" do Jacob Devlin và các cộng sự ở Google AI language.

# 2 - GIỚI THIỆU PhoBERT



## Bốn đặc tính chính của BERT

- **Encoder-only architecture**
- **Pre-training approach:** Tiền huấn luyện dựa trên khối dữ liệu lớn (Ví dụ: Wikipedia) trước khi áp dụng cho bài toán thực tế
- **Model fine-tuning:** sau khi Pre-training, mô hình có thể được fine-tune trên các tác vụ cụ thể như phân loại cảm xúc, nhận dạng thực thể, trả lời câu hỏi... chỉ với một vài bước huấn luyện thêm và một lượng dữ liệu nhỏ hơn.
- **Use of bidirectional context:** BERT xử lý đồng thời cả hai hướng—giúp nắm bắt tốt hơn ngữ cảnh tổng thể và ý nghĩa của từ trong câu.

## 2 - GIỚI THIỆU PhoBERT

Mô hình **PhoBERT** được hai tác giả Dat Quoc Nguyen và Anh Tuan Nguyen giới thiệu vào năm 2020 với hai version PhoBERT<sub>base</sub> và PhoBERT<sub>large</sub>, là mô hình large-scale monolingual language đầu tiên cho tiếng việt.

### HẠN CHẾ CỦA BERT:

- Ngữ cú tiếng việt Wikipedia là dữ liệu duy nhất để huấn luyện cho monolingual language models.
- Tất cả các mô hình monolingual và multilingual dựa trên BERT đều không nhận ra được sự khác biệt giữa âm tiết và từ tiếng Việt.

## 2 - GIỚI THIỆU PhoBERT

### **Giải quyết hạn chế của BERT:**

**1-** Sử dụng 20GB tài liệu pre-training data set. Bộ dữ liệu này là tổng hợp của hai bộ ngữ cú (corpora):

- a. Ngữ cú tiếng Việt từ Wikipedia ~1GB
- b. Bộ dữ liệu ~19GB ngữ cú tin tức tiếng Việt.

**2-** Sử dụng RDRsegmenter (Nguyen et al., 2018) từ VnCoreNLP (Vu et al., 2018) để phân tích từ và câu trong bộ dữ liệu pre-training.

# 2 - GIỚI THIỆU PhoBERT

Đánh giá mô hình PhoBERT đạt SOTA với 4 tác vụ NLP đó là:

- POS tagging
- dependency parsing
- Named-entity recognition (NER)
- Natural language inference (NLI)

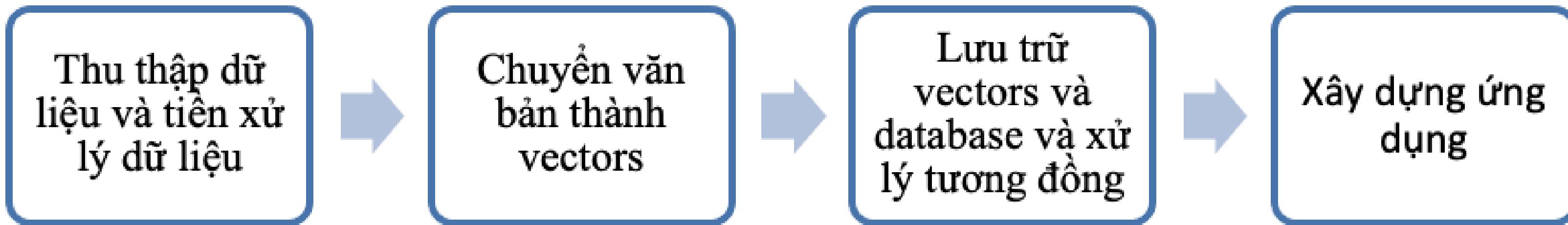
POS tagging (word-level)		Dependency parsing (word-level)	
Model	Acc.	Model	LAS / UAS
RDRPOSTagger (Nguyen et al., 2014a) [♣]	95.1	–	–
BiLSTM-CNN-CRF (Ma and Hovy, 2016) [♣]	95.4	VnCoreNLP-DEP (Vu et al., 2018) [★]	71.38 / 77.35
VnCoreNLP-POS (Nguyen et al., 2017) [♣]	95.9	jPTDP-v2 [★]	73.12 / 79.63
jPTDP-v2 (Nguyen and Verspoor, 2018) [★]	95.7	jointWPD [★]	73.90 / 80.12
jointWPD (Nguyen, 2019) [★]	96.0	Biaffine (Dozat and Manning, 2017) [★]	74.99 / 81.19
XLM-R <sub>base</sub> (our result)	96.2	Biaffine w/ XLM-R <sub>base</sub> (our result)	76.46 / 83.10
XLM-R <sub>large</sub> (our result)	96.3	Biaffine w/ XLM-R <sub>large</sub> (our result)	75.87 / 82.70
PhoBERT <sub>base</sub>	96.7	Biaffine w/ PhoBERT <sub>base</sub>	78.77 / 85.22
PhoBERT <sub>large</sub>	96.8	Biaffine w/ PhoBERT <sub>large</sub>	77.85 / 84.32

NER (word-level)		NLI (syllable- or word-level)	
Model	F <sub>1</sub>	Model	Acc.
BiLSTM-CNN-CRF [♦]	88.3	–	–
VnCoreNLP-NER (Vu et al., 2018) [♦]	88.6	BiLSTM-max (Conneau et al., 2018)	66.4
VNER (Nguyen et al., 2019b)	89.6	mBiLSTM (Artetxe and Schwenk, 2019)	72.0
BiLSTM-CNN-CRF + ETNLP [♠]	91.1	multilingual BERT (Devlin et al., 2019) [■]	69.5
VnCoreNLP-NER + ETNLP [♠]	91.3	XLM <sub>MLM+TLM</sub> (Conneau and Lample, 2019)	76.6
XLM-R <sub>base</sub> (our result)	92.0	XLM-R <sub>base</sub> (Conneau et al., 2020)	75.4
XLM-R <sub>large</sub> (our result)	92.8	XLM-R <sub>large</sub> (Conneau et al., 2020)	79.7
PhoBERT <sub>base</sub>	93.6	PhoBERT <sub>base</sub>	78.5
PhoBERT <sub>large</sub>	94.7	PhoBERT <sub>large</sub>	80.0

Nguồn: [1] Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2019), North American Chapter of the Association for Computational Linguistics.

# KẾ HOẠCH THỰC HIỆN



# **3 - TIỀN XỬ LÝ DỮ LIỆU**

# 3 - TIỀN XỬ LÝ DỮ LIỆU

Nguồn: VnExpress, Tuổi Trẻ, Vietnamnet

Mục đích: Làm dữ liệu đầu vào cho truy hồi văn bản

Dữ liệu phải đủ sạch và đúng định dạng

The screenshot shows a web browser displaying a Kaggle dataset page. The URL in the address bar is [kaggle.com/datasets/sarahhimiko/vietnamese-online-news-csv-dataset](https://kaggle.com/datasets/sarahhimiko/vietnamese-online-news-csv-dataset). The page title is "Vietnamese Online News .csv dataset". On the left, there is a sidebar with navigation links: Create, Home, Competitions, **Datasets** (which is selected), Models, Benchmarks, Code, Discussions, Learn, and More. The main content area contains a brief description: "A dataset that consists of 150K+ news". Below this is a "Data Card" section with tabs for Data Card, Code (9), Discussion (0), and Suggestions (0). The "About Dataset" section includes a note about the dataset being converted from .json to .csv, a link to the original .json file, and a summary of the dataset's purpose for practicing Natural Language Processing in Vietnamese. To the right, there are sections for Usability (10.00), License (Database: Open Database, Cont...), Expected update frequency (Annually), and Tags (Earth and Nature, Text, Exploratory Data Analysis, NLP, Text Pre-Processing, Vietnamese). At the bottom, there is a "Data Explorer" section showing a file named "Fixed\_news\_dataset.csv" (434.68 MB) with a "View Active Events" button.

# 3 - TIỀN XỬ LÝ DỮ LIỆU

## Underthesea

- Loại bỏ khoảng trắng thừa
- Tách từ tiếng Việt (word tokenization)

```
> <pre>from underthesea import word_tokenize</pre>
df["content"] = df["content"].astype(str).str.strip()
df["content"] = df["content"].apply(lambda x: word_tokenize(x, format="text"))

[6]<pre>df[["id", "content"]].head()</pre>
[7]<table border="1">
<thead>
<tr><th>id</th><th>content</th></tr>
</thead>
<tbody>
<tr>0 218270 Chiều 31/7 , Công_an tỉnh Thừa_Thiên - Huế đã ...
<tr>1 218269 Gần đây , Thứ_trưởng Bộ Phát triển Kỹ Thuật số...
<tr>2 218268 Kết_quả thi tốt_nghiệp THPT năm 2022 cho thấy ...
<tr>3 218267 Thống_đốc Kentucky_Any Beshear hôm 31/7 cho_h...
<tr>4 218266 Vụ tai_nạn giao_thông liên_hoàn trên phố đi bộ...
```

# 3 - TIỀN XỬ LÝ DỮ LIỆU

## Lưu dữ liệu dưới dạng CSV

- 📝 Làm sạch dữ liệu (Pandas)
  - Bỏ dòng thiếu, trùng
  - Chuyển chữ thường

### 1. Tải dữ liệu (Load Data)

Đọc dữ liệu tin tức từ file CSV gốc. Dữ liệu bao gồm các cột: id, title, content, author, topic, source, url, v.v. Nguồn: [https:/](https://)

```
# Import thư viện pandas để xử lý dữ liệu
import pandas as pd

# Đường dẫn file dữ liệu gốc và file đầu ra
BASE_DATA_DIR = "../data/raw_data/Original_news_dataset.csv"
BASE_OUTPUT_DATA_DIR = "../data/processed_data/Original_news_dataset.csv"

# Đọc dữ liệu từ file CSV
df = pd.read_csv(BASE_DATA_DIR, dtype=str, nrows=2000)
df.head()
```

# 3 - TIỀN XỬ LÝ DỮ LIỆU

- Văn bản đã chuẩn hóa và tách từ
- Dữ liệu lưu thành file sạch, thống nhất
- Sẵn sàng cho bước tiếp theo:  
Chuyển đổi văn bản thành vector  
(PhoBERT)
- Liên kết dữ liệu – mô hình học sâu

Unnamed: 0	id	author	content	picture_count	processed	source		
0	0	218270	NaN	Chiều 31/7, Công an tỉnh Thừa Thiên - Huế đã c...	3	0	docbao.vn	Tên cướp tiệm vàng tại Huế
1	1	218269	(Nguồn: Sina)	Gần đây, Thứ trưởng Bộ Phát triển Kỹ thuật số...	1	0	vtc.vn	BỎ qua mạng 5G, Nga tiến th
2	2	218268	Hồ Sỹ Anh	Kết quả thi tốt nghiệp THPT năm 2022 cho thấy ...	3	0	thanhnien.vn	Địa phương nào đứng đầu
3	3	218267	Ngọc Ánh	Thống đốc Kentucky Andy Beshear hôm 31/7 cho h...	1	0	vnexpress	Người chết trong mưa lũ
4	4	218266	HÀI YẾN - MINH LÝ	Vụ tai nạn giao thông liên hoàn trên phố đi bộ...	12	0	soha	Hải Phòng: Hình ảnh xe "điề
...	...	...	...	...	...	...	...	
993	1993	215812	NaN	Mới đây, Khánh Loan đã phủ nhận tin đồn sắp cư...	6	1	laodong	Ca sĩ Khánh Loan vượt bi
994	1994	215811	Anh Vũ	Quỳnh Nga kể từ sau loạt vai diễn phản diện là...	0	1	eva.vn	Không phải váy hở hang
995	1995	215810	NaN	Bệnh đậu mùa khi có thể lây sang người khi có ...	1	1	baoquocte	Bệnh đậu mùa khi và những
997	1997	215808	BONGDAPLUS	Đội tuyển Aerobic trẻ Bình Dương tham dự giải ...	1	1	bongdaplus	Bình Dương xếp nhì cụm
999	1999	215805	Mi Vân	Tối 30/7, chung kết cuộc thi Hoa hậu Hoàn vũ T...	16	1	dantri	"Người đẹp đến từ bãi rác

## 4. Vector hoá dữ liệu sử dụng phoBERT

- Các kỹ thuật chính & các bước thực hiện
  - Tokenization (Tách từ)
  - Text → Vector
- Những điểm khác biệt:
  - So với các kỹ thuật chuyển text → vector khác

## 4. Vector hoá dữ liệu sử dụng phoBERT

- Tokenization là gì?
  - Chuyển văn bản → Thành phần nhỏ hơn:
    - **Ký tự:**
    - **Từ:** Cái được học
    - **Subword:** Tách ra tiền tố, hậu tố, từ gốc
  - Ví dụ qua slide kết bên

# 4. Vector hoá dữ liệu sử dụng phoBERT

## Types Of Tokenization

"Machine",  
"learning",  
"is", "fun", ":"

**Word-Based**

"ma",  
"chine,",  
"learn", "ing"

**Subword-Based**

"M", "a", "c",  
"h", "i", "n",  
"e", "l", "e",  
"a", "r", "n",  
"i", "n", "g"

**Character-Based**

## 4. Vector hoá dữ liệu sử dụng phoBERT

→ Mục tiêu: Bước đầu để thực hiện các tính toán:

- Chuyển văn bản thành số
- Đếm tần xuất xuất hiện....

# 4. Vector hoá dữ liệu sử dụng phoBERT

Ví dụ: “Nhóm mình đang tìm hiểu phoBERT”

Luồng xử lý chính:

B1: Xếp câu theo độ dài, Chỉ định thiết bị sử dụng (gpu or cpu)

B2: **Lấy embedding**

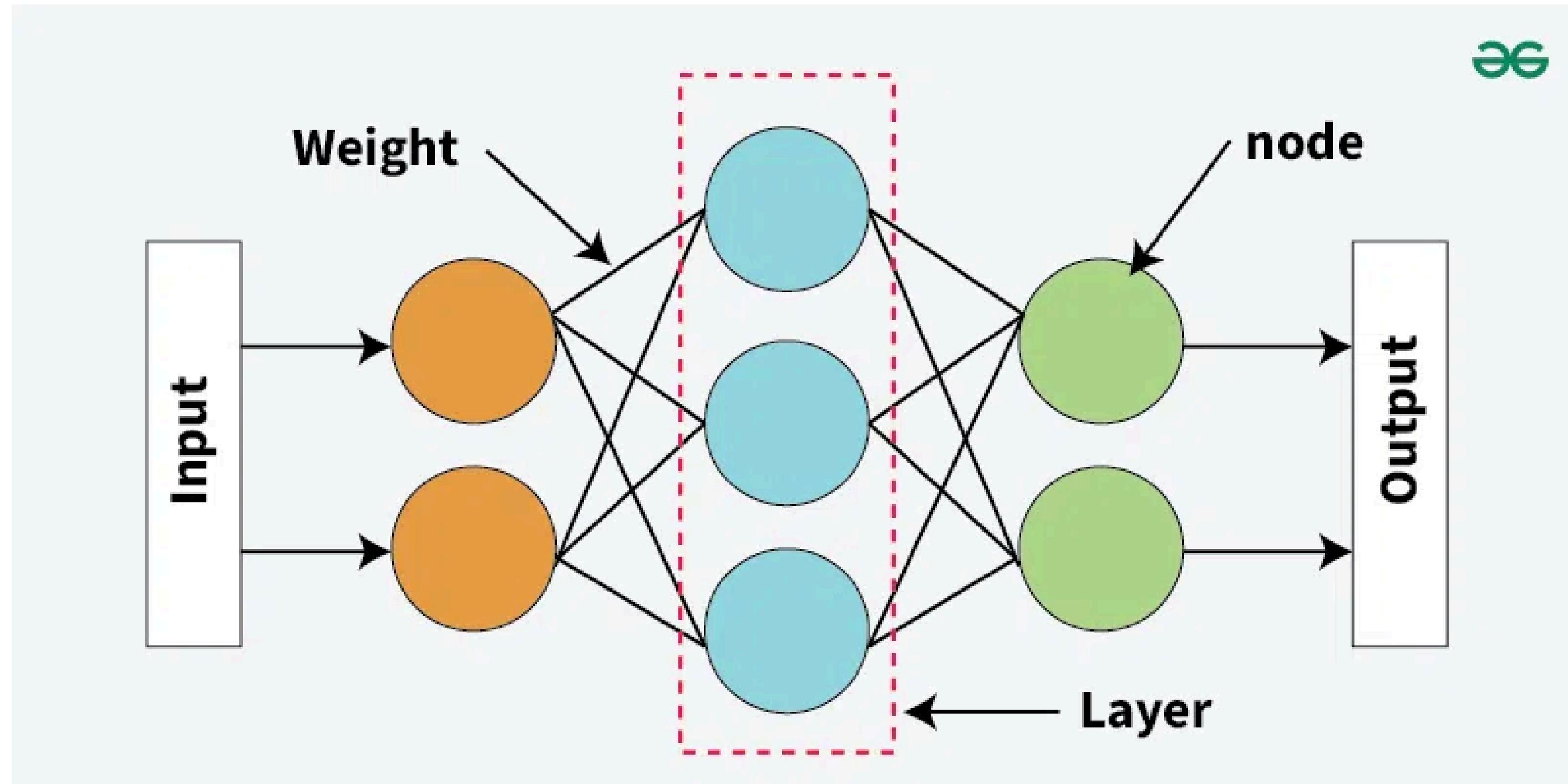
b2.1: Chuyển **token** thành **token ids**

b2.2: Đẩy **token ids** vào model:

*(Truyền token ids vào các layer bên trong của model  
trong file model.safetensors)*

B3: **Trả về vector** dạng np array, tensor, list tensor tùy conf.

## 4. Vector hoá dữ liệu sử dụng phoBERT



src: <https://www.geeksforgeeks.org/deep-learning/neural-network-node/>

# 4. Vector hoá dữ liệu sử dụng phoBERT

```
▶ ▾
from sentence_transformers import SentenceTransformer
model = SentenceTransformer("vinai/phobert-base")

embedding = model.encode('Nhóm mình đang tìm hiểu phoBERT', normalize_embeddings=True)
embedding.shape # => Số chiều bằng (768,)
embedding

[2] ✓ 4.6s

...
No sentence-transformers model found with name vinai/phobert-base. Creating a new one with mean pooling.

...
array([-7.51986634e-03,  1.70325972e-02, -3.80394422e-02, -1.36641487e-02,
       -3.85071523e-02,  3.87815982e-02, -6.43990040e-02, -1.01122754e-02,
       -3.95898968e-02, -3.98599245e-02,  1.71731878e-02,  1.05369471e-01,
       6.86495507e-04, -1.24018453e-02,  7.97104381e-04,  7.17359362e-03,
       2.98525710e-02,  2.29311883e-02, -1.71285346e-02, -3.32339108e-02,
       4.87042591e-03, -2.75367852e-02, -2.85343118e-02,  5.13379574e-02,
      -3.36058525e-04, -9.00713913e-03,  8.39892924e-02,  2.58513466e-02,
       3.41088474e-02,  4.03110646e-02,  3.46182212e-02,  2.62455344e-02,
       1.55008892e-02, -9.68195591e-03,  4.17049266e-02,  4.64267842e-02,
       1.27256718e-02, -1.14048803e-02,  1.13658356e-02,  8.13027695e-02,
       4.21655774e-02,  8.70336592e-03,  2.13540774e-02,  2.81511806e-02,
```

# 4. Vector hoá dữ liệu sử dụng phoBERT

	Câu 1	Câu 2	CountVectorizer	TF-IDF	PhoBERT	
0	Tôi thích uống cà phê vào buổi sáng.	Buổi sáng, tôi hay nhâm nhi ly coffee.		0.3750	0.2735	0.6888
1	Trời hôm nay thật đẹp và mát mẻ.	Thời tiết hôm nay dễ chịu và trong lành.		0.3536	0.2556	0.6814
2	Anh ấy đã rời công ty từ tháng trước.	Tháng rồi, anh ta đã nghỉ việc ở chỗ làm.		0.3333	0.2748	0.7316
3	Tôi cảm thấy rất buồn vì kết quả này.	Kết quả này khiến tôi thất vọng và chán nản.		0.4216	0.3331	0.6787
4	Cô ấy là một giáo viên tận tâm và nhiệt huyết.	Cô ta luôn hết lòng với nghề dạy học.		0.1005	0.0791	0.6439

Đánh giá nhanh: Với 2 câu có ý nghĩa giống nhau thì  
**PhoBERT cho ra kết quả tương đồng tốt hơn**

## 4. Vector hoá dữ liệu sử dụng phoBERT

**Kết luận là:**

Việc dùng mô hình LLM (phoBERT) sẽ cho ra các vector với độ chính xác về ngữ nghĩa cao hơn các kỹ thuật đã được học: tf-idf, word count,...



## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

- Database Qdrant lưu trữ vectors đã tính được ở bước trước.
- Tính toán tương đồng Cosine giữa truy vấn người dùng với các dữ liệu đang lưu
- Đánh giá độ tương đồng sau khi đã so sánh dựa trên vector hóa bằng Phobert bằng phương pháp không gian 2D

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

***Biến truy vấn người dùng thành vectơ:***

query = “Tôi thích ăn bún hoặc phở bò”

d1: Tôi thích ăn phở

d2: Tôi yêu bún bò

d3: Tôi ghét học toán

d4: Con mèo đang ngủ

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

Phương pháp tính tương đồng dựa trên Cosine:

$$\text{Độ tương đồng Cosine}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Trong đó:

- $A \cdot B$  là tích vô hướng giữa hai vector.
- $\|A\| \cdot \|B\|$  là độ dài (chuẩn Euclid) của vector A và B.

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

Giá trị:

Gần 1  $\Rightarrow$  Hai vector rất giống nhau (cùng hướng).

Gần 0  $\Rightarrow$  Hai vector gần như vuông góc  $\Rightarrow$  không liên quan.

Gần -1  $\Rightarrow$  Hai vector ngược hướng.

# 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

**Doc2vec (Vectơ size = 50)**

Document	V1	V2	V3	V4..V50
d1	0.2153	-0.0351	0.1012	0.0501
d2	0.2021	-0.0201	0.1113	0.0432
d3	-0.1123	0.2133	-0.0213	0.0802
d4	-0.0032	-0.0433	0.0012	-0.0045
query	0.2087	-0.0291	0.1045	0.0479

d1: Tôi thích ăn phở => similarity: 0.0637

d2: Tôi yêu bún bò => similarity: -0.0634

d3: Tôi ghét học toán => similarity: 0.1580

d4: Con mèo đang ngủ => similarity: 0.1829

# 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

## Doc2vec (Vectơ size = 4)

Document	V1	V2	V3	V4
d1	0.2153	-0.0351	0.1012	0.0501
d2	0.2021	-0.0201	0.1113	0.0432
d3	-0.1123	0.2133	-0.0213	0.0802
d4	-0.0032	-0.0433	0.0012	-0.0045
query	0.2087	-0.0291	0.1045	0.0479

- d1: Tôi thích ăn phở => similarity: 0.6230  
d2: Tôi yêu bún bò => similarity: 0.0752  
d3: Tôi ghét học toán => similarity: -0.0719  
d4: Con mèo đang ngủ => similarity: -0.4749



## Doc2vec

- Hiểu ngữ nghĩa, ngữ cảnh
- Biết được từ đồng nghĩa
- Nếu tập dữ liệu huấn luyện không nhiều sẽ không hiểu 2 ý trên

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

**PhoBERT**

	0	1	2	3	4	5	6	7	8	9
d1	0.19234	0.10234	-0.0234	0.05812	...					
d2	0.17542	0.12012	-0.0112	0.05099	...					
d3	-0.0231	0.03429	0.11234	0.00812	...					
d4	0.00123	-0.0192	0.09212	-0.0456	...					
query	0.18412	0.11043	-0.0156	0.05587	...					

d1: Tôi thích ăn phở => similarity: 0.8527  
d2: Tôi yêu bún bò => similarity: 0.8056  
d3: Tôi ghét học toán => similarity: 0.6633  
d4: Con mèo đang ngủ => similarity: 0.4450



**PhoBERT**

- Hiểu ngữ nghĩa, ngữ cảnh
- Biết được từ đồng nghĩa
- Tập dữ liệu huấn luyện không cần nhiều

## **5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn**

**Đánh giá độ tương đồng bằng phương pháp không gian 2D  
(Trực quan hóa bằng t-SNE)**

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

### Database lưu trữ Qdrant



Không gian 2D

Vì sao cần đưa về không gian 2D?

- Vector PhoBERT có đến 768 chiều  
→ không thể trực quan hóa được bằng mắt.
- Do đó cần phương pháp giảm chiều.
- Phân cụm vector
- Các chấm càng gần nhau mang ý nghĩa tương đồng gần giống nhau

## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

### Database lưu trữ Qdrant

Giản lược t-SNE (t-Distributed Stochastic Neighbor Embedding)

- Là một kỹ thuật giảm chiều (dimensionality reduction) tập trung vào duy trì sự tương đồng giữa các điểm dữ liệu trong không gian nhiều chiều khi biểu diễn chúng ở không gian thấp hơn (thường là 2D hoặc 3D để trực quan hóa).

Cách hoạt động (giản lược)

- Tính toán xác suất tương đồng giữa các điểm trong không gian cao (high-dimension).
- Ánh xạ các điểm vào không gian thấp (2D hoặc 3D) sao cho các xác suất tương đồng giữa các điểm trong không gian thấp gần giống với không gian cao.
- Tối ưu bằng thuật toán gradient descent để giảm sai lệch giữa hai phân bố (KL divergence).

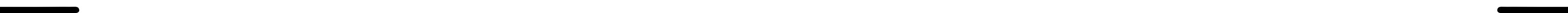
## 5 - Tính toán tương đồng và chọn tài liệu có liên quan với truy vấn

### Database lưu trữ Qdrant



Ứng dụng:

- Vector Expansion Graph
- Khám phá các mối liên hệ ngữ nghĩa giữa các document.



# DEMO

# 6 - Đánh giá thực nghiệm và hướng phát triển

The screenshot displays a search interface with two main search bars at the top, both labeled "Tim kiem với PhoBERT AI".

**Search Results for "Đội tuyển futsal Việt Nam tập trung với HLV từng vô địch thế giới":**

- thanhnien.vn • 03:36:59 1/8/2022  
**Đội tuyển futsal Việt Nam tập trung với HLV từng vô địch thế giới**  
đội tuyển futsal việt nam sẽ tập trung tại nhà thi đấu thái son nam (quận 8, tp hcm), nơi có cơ sở vật chất rất tốt giúp các cầu thủ tập luyện một cách tốt nhất ở lần tập
- vov.vn • 22:57:30 31/7/2022  
**HLV từng vô địch World Cup sẽ cùng ĐT Futsal Việt Nam du đấu Thái Lan**  
đt futsal việt nam đã chính thức công bố danh sách triệu tập 23 cầu thủ và sẽ hội quân vào ngày mai (1/8) đây sẽ là đợt tập trung đầu tiên của đội tuyển dưới thời tân
- thethaovanhoa • 09:04:43 1/8/2022  
**U18 Việt Nam tính 'bài tử' đấu Myanmar tại bán kết giải Đông Nam Á**  
đội tuyển u18 nữ việt nam sau khi vượt qua thái lan để chiếm ngôi đầu bảng giành quyền vào bán kết giải vô địch đông nam á đang có những tính toán phù hợp nhất
- qndn.vn • 22:10:59 31/7/2022  
**Tuyển futsal Việt Nam được dắt bởi huấn luyện viên từng vô địch thế giới**  
sau đó, đội sẽ sang thái lan tham dự giải futsal giao hữu quốc tế từ ngày 9-9 đến 17-9, gặp các đội bóng mạnh như iran, morocco, phản lan, angola và chủ nhà thái la
- dantri • 03:50:00 1/8/2022  
**HLV từng vô địch thế giới chính thức dẫn dắt đội tuyển futsal Việt Nam**  
theo đó, đội tuyển futsal việt nam sẽ tập trung với 23 cầu thủ, bắt đầu chuẩn bị cho vòng chung kết (vck) giải futsal vô địch châu á năm 2022, diễn ra ở kuwait ở đợt t
- zingnews • 05:39:05 1/8/2022  
**HLV từng vô địch World Cup dẫn dắt tuyển futsal Việt Nam**

**Search Results for "chiến tranh ở Ukraine":**

- vietnamnet.vn • 08:04:07 1/8/2022  
**Ukraine tin Nga không có cơ hội thắng, vợ tổng thống Ireland gây tranh cãi**  
ông vadatursky được tạp chí forbes xếp hạng là người giàu thứ 24 ở ukraine với tổng trị giá tài sản là 430 triệu usd doanh nhân này từng được trao tặng giải thưởng danh giá "anh hùng của ukraine" tổng...
- vietnamnet.vn • 09:00:26 1/8/2022  
**Ukraine tin Nga không có cơ hội thắng**  
ông vadatursky được tạp chí forbes xếp hạng là người giàu thứ 24 ở ukraine với tổng trị giá tài sản là 430 triệu usd doanh nhân này từng được trao tặng giải thưởng danh giá "anh hùng của ukraine" tổng...
- vnexpress • 06:37:33 1/8/2022  
**Nga tố Ukraine tấn công Hạm đội Biển Đen, Kiev bác bỏ - VnExpress**  
thứ trưởng quốc phòng ukraine volodymyr havrylov hôm 19/7 nói rằng kiev đang chuẩn bị cuộc tấn công nhằm vào hạm đội biển đen của nga ông havrylov cũng cho biết ukraine lên kế hoạch giành lại crimea t...
- danviet • 23:45:41 31/7/2022  
**Lịch thi đấu giải U16 Đông Nam Á 2022 (ngày 31/7)**  
vào lúc 15h chiều nay, u16 việt nam sẽ bắt đầu chiến dịch u16 đón nam á 2022 bằng trận đấu với u16 singapore đây cũng là trận đấu đầu tiên tại giải đấu u16 việt nam đương nhiên là đội được đánh giá ca...
- thanhnien.vn • 23:35:05 31/7/2022  
**UAV Ukraine tấn công trụ sở Hạm đội biển Đen, 6 người bị thương**  
thứ trưởng quốc phòng ukraine volodymyr havrylov hôm 19/7 nói rằng kyiv đang chuẩn bị cuộc tấn công nhằm vào hạm đội biển đen của nga ông havrylov cũng cho biết ukraine lên kế hoạch giành lại crimea t...
- laodong • 00:04:42 1/8/2022  
**Cuộc đua Kpop tháng 8 "như chảo lửa" khi Blackpink TWICE BTS đều nón mít**

Thử nghiệm với một số query - cho ra kết quả tốt

# 6 - Đánh giá thực nghiệm và hướng phát triển

The screenshot displays a search interface with two main search bars at the top, both labeled "Tim kiem với PhoBERT AI".

The left search bar contains the query "Đội tuyển futsal Việt Nam tập trung với HLV từng vô địch thế giới". Below it, several news articles are listed:

- thanhnien.vn • 03:36:59 1/8/2022: **Đội tuyển futsal Việt Nam tập trung với HLV từng vô địch thế giới**  
đội tuyển futsal việt nam sẽ tập trung tại nhà thi đấu thái son nam (quận 8, tp hcm), nơi có cơ sở vật chất rất tốt giúp các cầu thủ tập luyện một cách tốt nhất ở lần tập
- vov.vn • 22:57:30 31/7/2022: **HLV từng vô địch World Cup sẽ cùng ĐT Futsal Việt Nam du đấu Thái Lan**  
đt futsal việt nam đã chính thức công bố danh sách triệu tập 23 cầu thủ và sẽ hội quân vào ngày mai (1/8) đây sẽ là đợt tập trung đầu tiên của đội tuyển dưới thời tân
- thethaovanhoa • 09:04:43 1/8/2022: **U18 Việt Nam tính 'bài tử' đấu Myanmar tại bán kết giải Đông Nam Á**  
đội tuyển u18 nữ việt nam sau khi vượt qua thái lan để chiếm ngôi đầu bảng giành quyền vào bán kết giải vô địch đông nam á đang có những tính toán phù hợp nhất
- qndn.vn • 22:10:59 31/7/2022: **Tuyển futsal Việt Nam được dắt dìu bởi huấn luyện viên từng vô địch thế giới**  
sau đó, đội sẽ sang thái lan tham dự giải futsal giao hữu quốc tế từ ngày 9-9 đến 17-9, gặp các đội bóng mạnh như iran, morocco, phản lan, angola và chủ nhà thái la
- dantri • 03:50:00 1/8/2022: **HLV từng vô địch thế giới chính thức dẫn dắt đội tuyển futsal Việt Nam**  
theo đó, đội tuyển futsal việt nam sẽ tập trung với 23 cầu thủ, bắt đầu chuẩn bị cho vòng chung kết (vck) giải futsal vô địch châu á năm 2022, diễn ra ở kuwait ở đợt t
- zingnews • 05:39:05 1/8/2022: **HLV từng vô địch World Cup dẫn dắt tuyển futsal Việt Nam**

The right search bar contains the query "chiến tranh ở Ukraine". Below it, several news articles are listed:

- vietnamnet.vn • 08:04:07 1/8/2022: **Ukraine tin Nga không có cơ hội thắng, vợ tổng thống Ireland gây tranh cãi**  
ông vadatursky được tạp chí forbes xếp hạng là người giàu thứ 24 ở ukraine với tổng trị giá tài sản là 430 triệu usd doanh nhân này từng được trao tặng giải thưởng danh giá "anh hùng của ukraine" tổng...
- vietnamnet.vn • 09:00:26 1/8/2022: **Ukraine tin Nga không có cơ hội thắng**  
ông vadatursky được tạp chí forbes xếp hạng là người giàu thứ 24 ở ukraine với tổng trị giá tài sản là 430 triệu usd doanh nhân này từng được trao tặng giải thưởng danh giá "anh hùng của ukraine" tổng...
- vnexpress • 06:37:33 1/8/2022: **Nga tố Ukraine tấn công Hạm đội Biển Đen, Kiev bác bỏ - VnExpress**  
thứ trưởng quốc phòng ukraine volodymyr havrylov hôm 19/7 nói rằng kiev đang chuẩn bị cuộc tấn công nhằm vào hạm đội biển đen của nga ông havrylov cũng cho biết ukraine lên kế hoạch giành lại crimea t...
- danviet • 23:45:41 31/7/2022: **Lịch thi đấu giải U16 Đông Nam Á 2022 (ngày 31/7)**  
vào lúc 15h chiều nay, u16 việt nam sẽ bắt đầu chiến dịch u16 đón nam á 2022 bằng trận đấu với u16 singapore đây cũng là trận đấu đầu tiên tại giải đấu u16 việt nam đương nhiên là đội được đánh giá ca...
- thanhnien.vn • 23:35:05 31/7/2022: **UAV Ukraine tấn công trụ sở Hạm đội biển Đen, 6 người bị thương**  
thứ trưởng quốc phòng ukraine volodymyr havrylov hôm 19/7 nói rằng kyiv đang chuẩn bị cuộc tấn công nhằm vào hạm đội biển đen của nga ông havrylov cũng cho biết ukraine lên kế hoạch giành lại crimea t...
- laodong • 00:04:42 1/8/2022: **Cuộc đua Kpop tháng 8 "như chảo lửa" khi Blackpink TWICE BTS đều nón mít**

Thử nghiệm với một số query - cho ra kết quả tốt

# 6 - Đánh giá thực nghiệm và hướng phát triển

The screenshot shows two search results from a PhoBERT AI search interface. The first search query is "Nghe đồn có viên kim cương nào được phát hiện to đùng". The results include news articles from dantri.com.vn, soha.com, 24h.com.vn, and eva.vn. The second search query is "Nghe đồn | Số ca COVID-19 tăng liên tục". The results include news articles from vtv.vn and baoquocte.com. Both queries yield results related to diamond discoveries and COVID-19 cases.

Q Nghe đồn có viên kim cương nào được phát hiện to đùng

Tìm kiếm với PhoBERT AI

dantri • 01:46:40 1/8/2022  
Á hậu Mâu Thủy được bạn trai "bí mật" quỳ gối cầu hôn  
mâu thủy cho rằng đã quen biết lâu đến vậy mà vẫn còn sai những điều giản đơn cho thấy người đó không thấu

soha • 00:55:32 1/8/2022  
Phát hiện viên kim cương hồng cực hiếm, lớn nhất trong 300 năm  
tháng 4 vừa qua, viên kim cương xanh de beers' cullinan đã được sàn đấu giá sotheby's ở hong kong bán với g

24h.com.vn • 23:52:34 31/7/2022  
Ngoài Mona Lisa, bảo tàng Louvre còn rất nhiều kiệt tác đáng chiêm ngưỡng  
được tạc trong giai đoạn 2620 – 2500 trước công nguyên, bức tượng này được phát hiện tại saqqara trong nhữ

eva.vn • 22:33:14 31/7/2022  
Truyện cổ tích: Những câu chuyện cổ tích bằng thơ, bé khám phá thế giới mới hấp dẫn và  
xưa có bà già nghèo chuyên mò cua bắt ốc một hôm bà bắt được một con ốc xinh xinh vỏ nó bieng biếc xanh kt  
...

Q Nghe đồn | Số ca COVID-19 tăng liên tục

Tìm kiếm với PhoBERT AI

vtv.vn • 21:55:25 31/7/2022  
Hơn 50 ngôi làng ở Pakistan bị nhấn chìm trong nước lũ, gây thiệt hại nặng nề  
giao thông trên đường cao tốc quetta - karachi vẫn bị đình chỉ do sập các cây cầu chính và nhiều đoạn đường cao tốc bị nước lũ cuốn trôi \* mời quý độc giả theo dõi các chương trình đã phát sóng của đà...

baoquocte • 22:59:55 31/7/2022  
Covid-19 ở Ấn Độ: Số ca mắc mới giảm sau 3 ngày liên tiếp ghi nhận trên 20.000 ca  
ngày 31/7, bộ y tế ấn độ thông báo, trong 24 giờ qua, số ca mắc mới covid-19 ở nước này đã giảm xuống còn 19 673 ca sau 3 ngày liên tiếp ghi nhận trên 20 000 ca cũng trong 24 giờ qua, ấn độ có thêm 45...

anninhthudo • 22:39:20 31/7/2022  
The Moffatts, 911 và A1 cùng hát tại Hà Nội và TP.HCM  
anh nhớ lại mình từng gặp gỡ trực tiếp the moffatts trong lần nhóm này đến việt nam biểu diễn năm 1999 và cho biết rất nóng lòng tái ngộ các thành viên moffatts để thưởng thức những "if life is so sho...

vtv.vn • 07:02:23 1/8/2022  
Số ca COVID-19 tăng liên tục, các nước châu Á tăng cường biện pháp phòng chống dịch | VTV.VN  
nhà tráng thông báo sẽ triển khai chiến dịch tiêm mũi tăng cường thứ hai vaccine ngừa covid-19 cho người dân trong tháng 9 tới, sử dụng loại vaccine được hiệu chỉnh để nhắm tới các dòng phụ của biến t...

Thử nghiệm với một số query - cho ra kết quả không tốt

## 6 - Đánh giá thực nghiệm và hướng phát triển

### Đánh giá:

- Đối với những truy vấn dài và chính xác có xuất hiện từ trong văn bản, thì kết quả trả về tương đối chính xác.
- Đối với các truy vấn ngắn và không chính xác thì hầu như kết quả trả về không chính xác và sai lệch.

### Hướng phát triển:

- Thực nghiệm các phương pháp xử lý tương đồng khác nhau.
- Áp dụng thêm các chức năng như mở rộng truy vấn và phản hồi có liên quan để giúp đưa ra kết quả tốt hơn cho truy vấn của người dùng.

# **Q & A**