

# XỬ LÝ NGÔN NGỮ TỰ NHIÊN

## TÊN DỰ ÁN: PHÂN LOẠI VĂN BẢN TIẾNG VIỆT THEO CHỦ ĐỀ

GVHD: TS. Đoàn Thị Hồng Phước

Sinh viên thực hiện:

1. Lê Ngọc Ánh
2. Nguyễn Bá Nhật



# MỤC LỤC:

- I. MỞ ĐẦU.
- II. NỘI DUNG.
- III. XÂY DỰNG MÔ HÌNH.
- IV. ĐÁNH GIÁ.

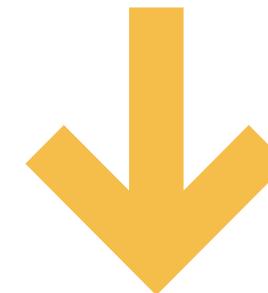


# I. MỞ ĐẦU

## 1. Giới thiệu bài toán:

- Phân loại tự động các văn bản tiếng Việt vào đúng chủ đề tương ứng, giúp việc quản lý, tìm kiếm và xử lý thông tin trở nên dễ dàng hơn.

Văn bản tiếng Việt



Chủ đề văn bản



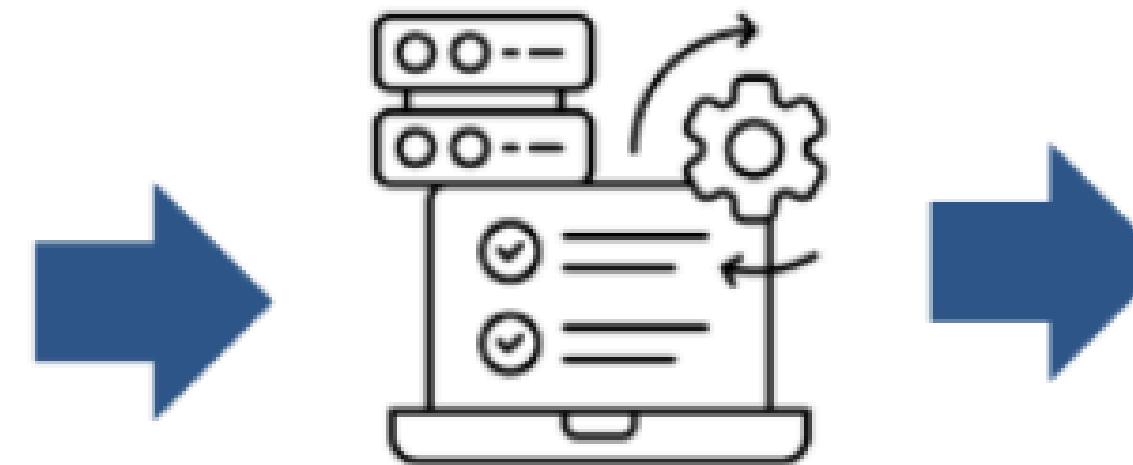
- **Mục tiêu:** tìm hiểu cách triển khai bài toán và triển khai trên mô hình.

## **II. NỘI DUNG:**

# Quy trình tổng thể bài toán (Pipeline)



Văn bản đầu  
vào



Tiền xử lý văn bản



Biểu diễn đặc  
trưng văn bản



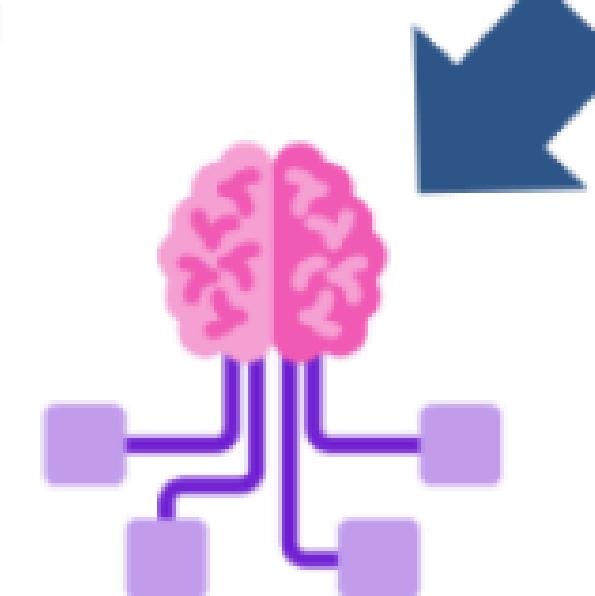
Giảm chiều  
(SVD)



Dự đoán chủ đề



Đánh giá mô hình



Mô hình học  
máy/học sâu

# 1. Tiền xử lý dữ liệu:

## Quy trình tiền xử lí:

### **Chuẩn hóa văn bản:**

xóa bỏ dấu câu, các ký tự đặc biệt, teencode và từ viết tắt.

### **Xóa stopword:**

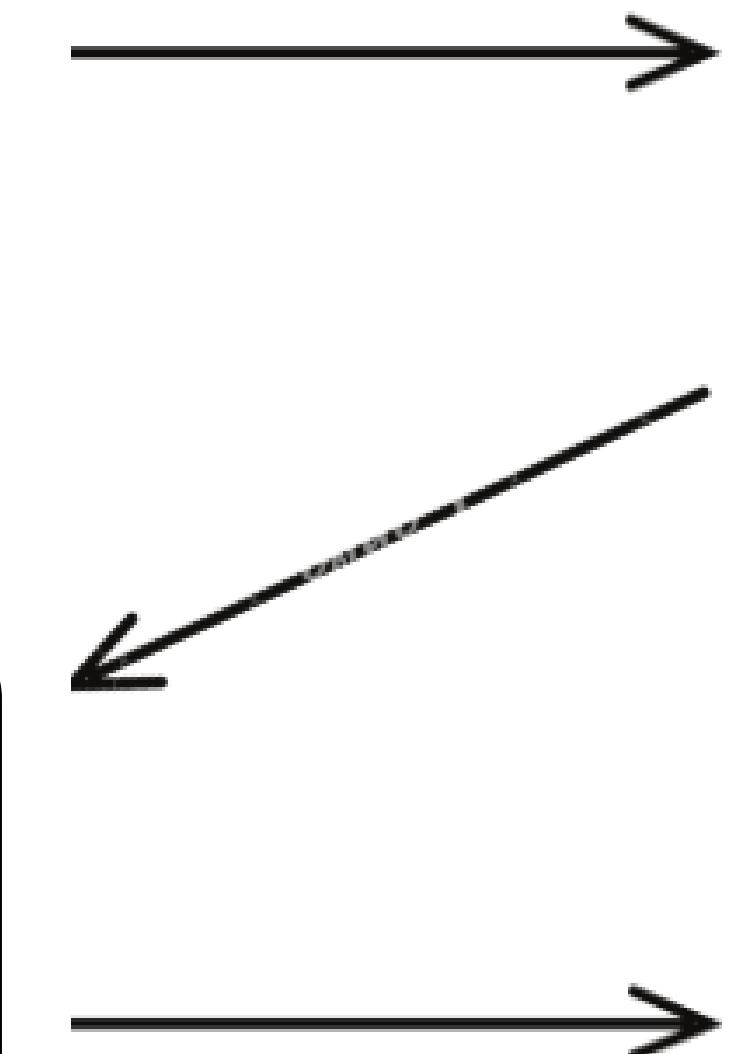
xóa các từ phổ biến nhưng không quan trọng trong nội dung.

### **Tách từ (tokenizer):**

VD: Xử lý ngôn ngữ tự nhiên  
→Xử\_lý\_ngôn\_ngữ\_tự\_nhiên

### **Trích xuất đặc trưng:**

đưa văn bản sau xử lí về dạng vector số hóa.



# Trích xuất đặc trưng:

## Static Embedding:

- Bow, Tf-Idf, Word2Vec.

- Fixed Vector:

- Gán vector cố định cho từ, bất kể ngữ cảnh.

- Pros:

- Tốc độ xử lí nhanh.
- Yêu cầu tài nguyên thấp.
- Phù hợp với các bài toán đơn giản.

- Cons:

- Không thể xử lí đa nghĩa.
- Hiểu biết ngữ cảnh còn hạn chế.
- Không phù hợp với bài toán khó.

## Contextual Embedding:

- BERT, PhoBERT, RoBERTa.

- Dynamic Vector:

- Tạo ra các vector động dựa vào ngữ cảnh xung quanh từ đó.

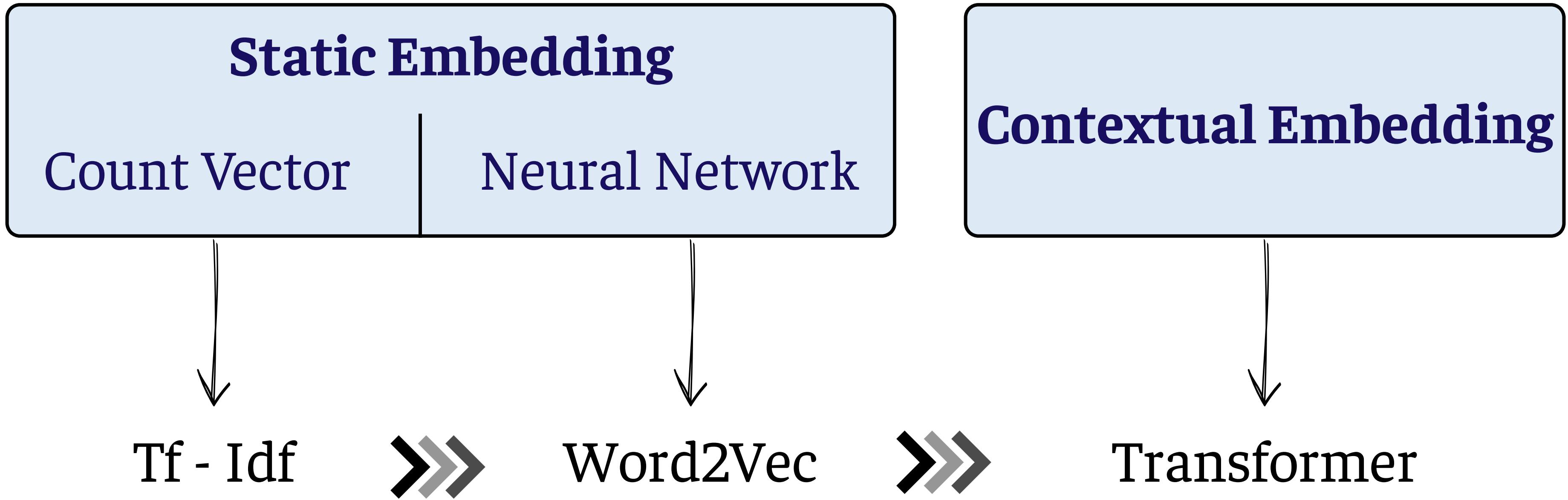
- Pros:

- Xử lí đa nghĩa tốt.
- Học ngữ cảnh + cú pháp.
- Cho kết quả tốt với bài toán phức tạp.

- Cons:

- Tốn tài nguyên.
- Khó triển khai.
- Overkill với bài toán đơn giản.

## Trích xuất đặc trưng:



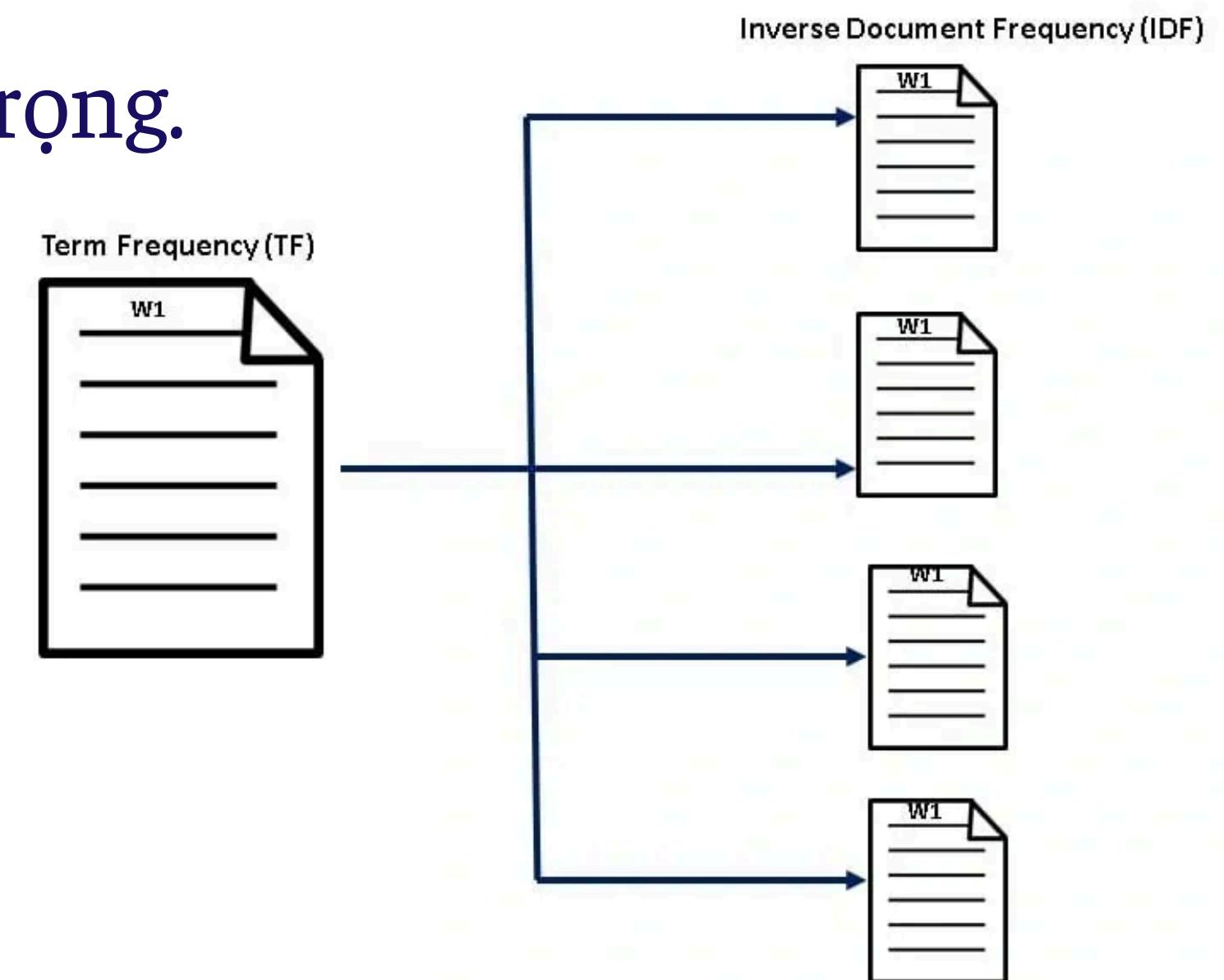
## 1.1 Tf-Idf:

- **Tf**: tần số từ trong văn bản.
- **Idf**: Tần số nghịch của từ trong “tập” văn bản càng nhỏ càng phổ biến, càng lớn càng hiếm.



**Tf \* Idf = Tf-Idf**: càng lớn càng quan trọng.

- **VD**: Mèo và chó đang chạy.
  - **Mèo** và **chó** đang **chạy**.
  - **Mèo** và **chó** đang **chạy**.



-Vector thường được biểu diễn dưới dạng:  $X \in \mathbb{R}^{n \times m}$

		unique word									
		đang	chạy	đua	thắng	cúp	những	đảng	dân	quyền	luật
doc	thể thao	0.16...	0.23...	0.23...	0.23...	0.234...	0.166...	0	0	0	0
	Chính trị	0.16...	0	0	0	0	0.166...	0.234...	0.2342...	0.23...	0.2342...

- **n**: số lượng doc.

- **m**: số lượng unique word.

- **Tf-Idf vector** thường có kích thước lớn dựa theo n và m nên phải kết hợp các thuật toán giảm chiều dữ liệu khác.

- Mô hình phân loại thường sử dụng ML truyền thống: Logistic Regression, SVM.

### **1.3 Hạn chế của TF-IDF:**

**-Không hiểu ngữ nghĩa:**

VD: 1. Chuột (máy tính) = Chuột (động vật).

2. To ≠ Bụ.

**-Không quan tâm trật tự từ:**

VD: Chó cắn người = Người cắn chó.

**-Thường được sử dụng nhiều trong:** + phân loại văn bản.

+ gợi ý sản phẩm.

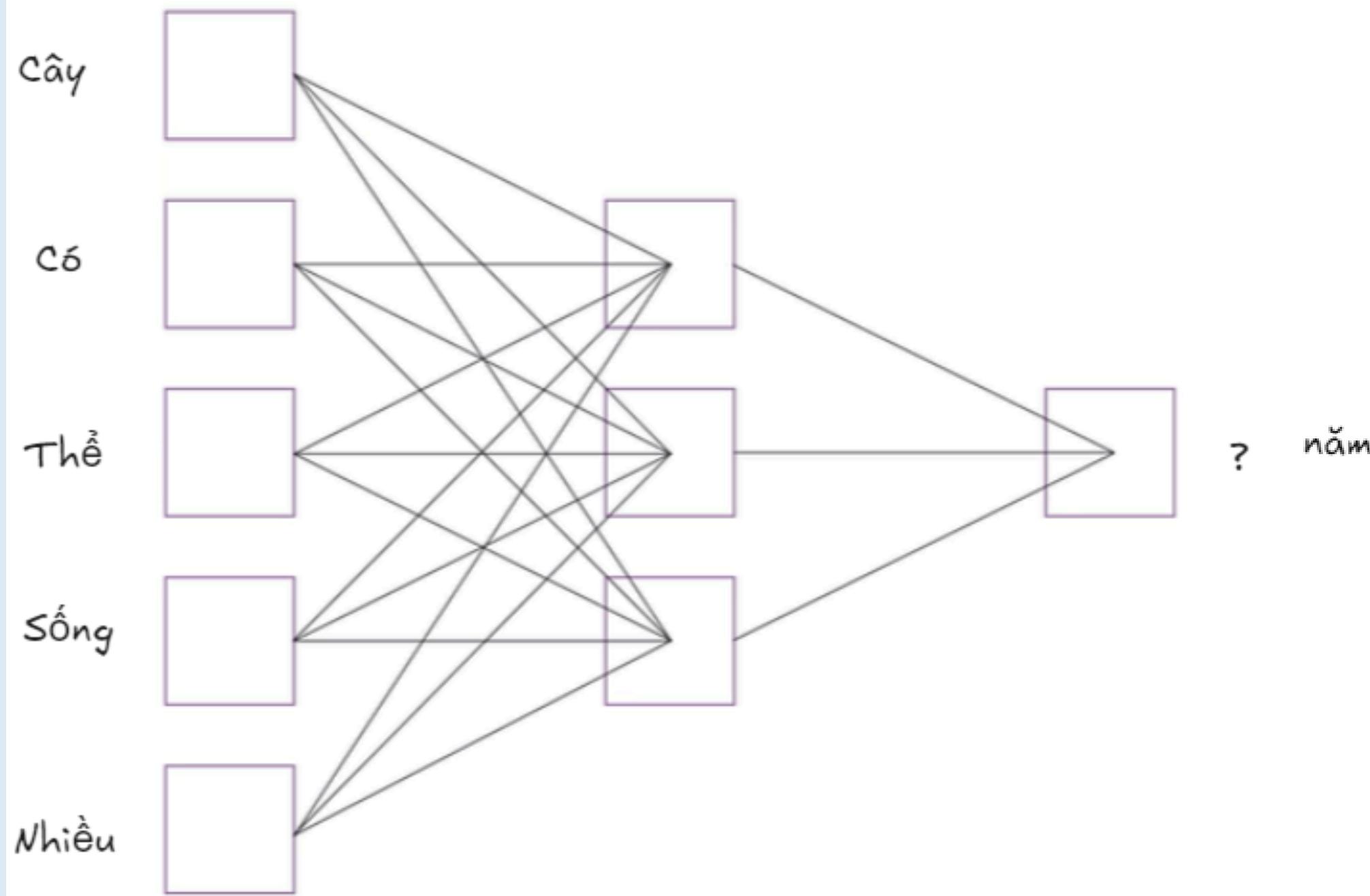
+ xếp hạng tìm kiếm.

+ phát hiện thư rác.

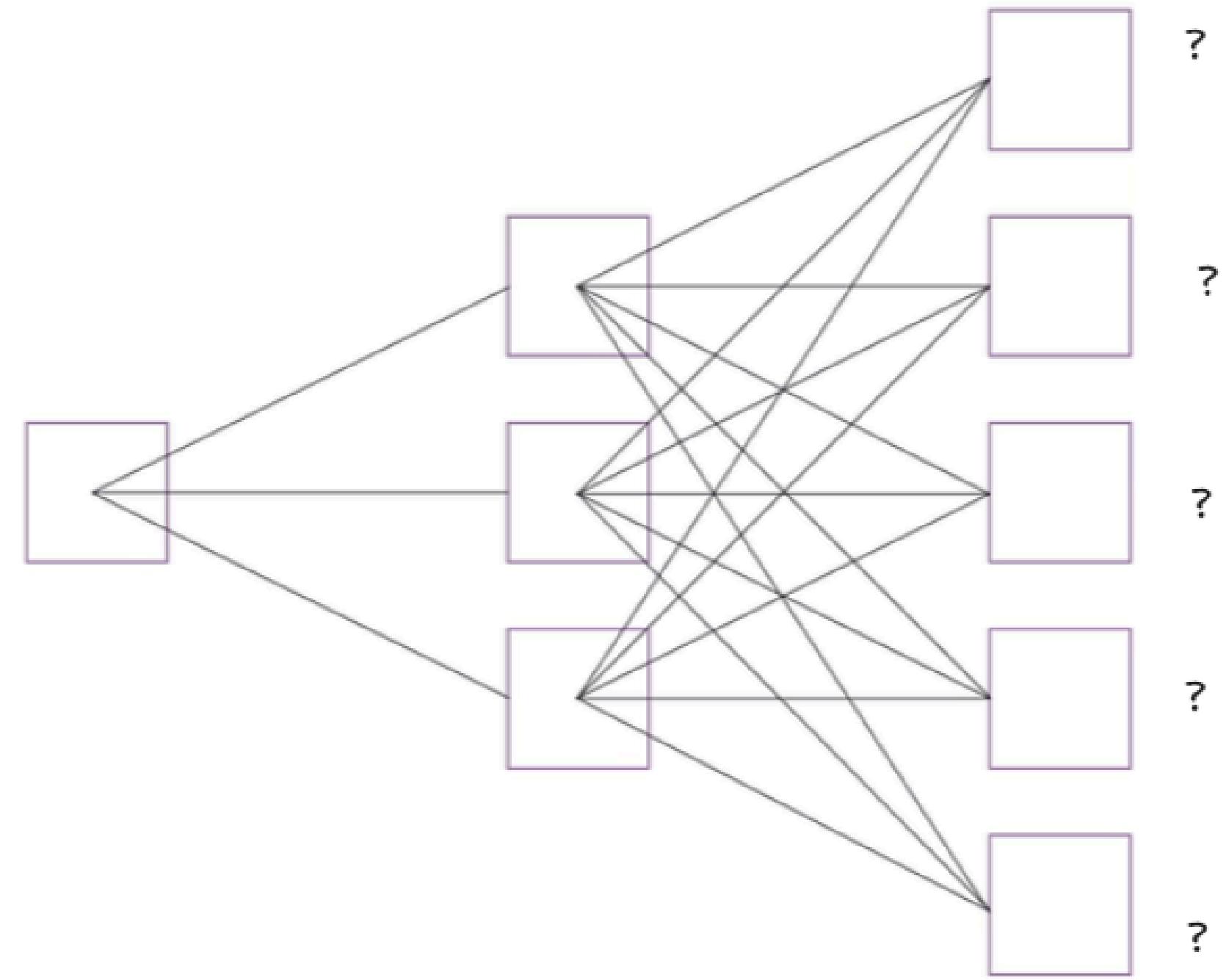


## 1.4 Word2Vec:

CBOW (Continuous Bag of Words)

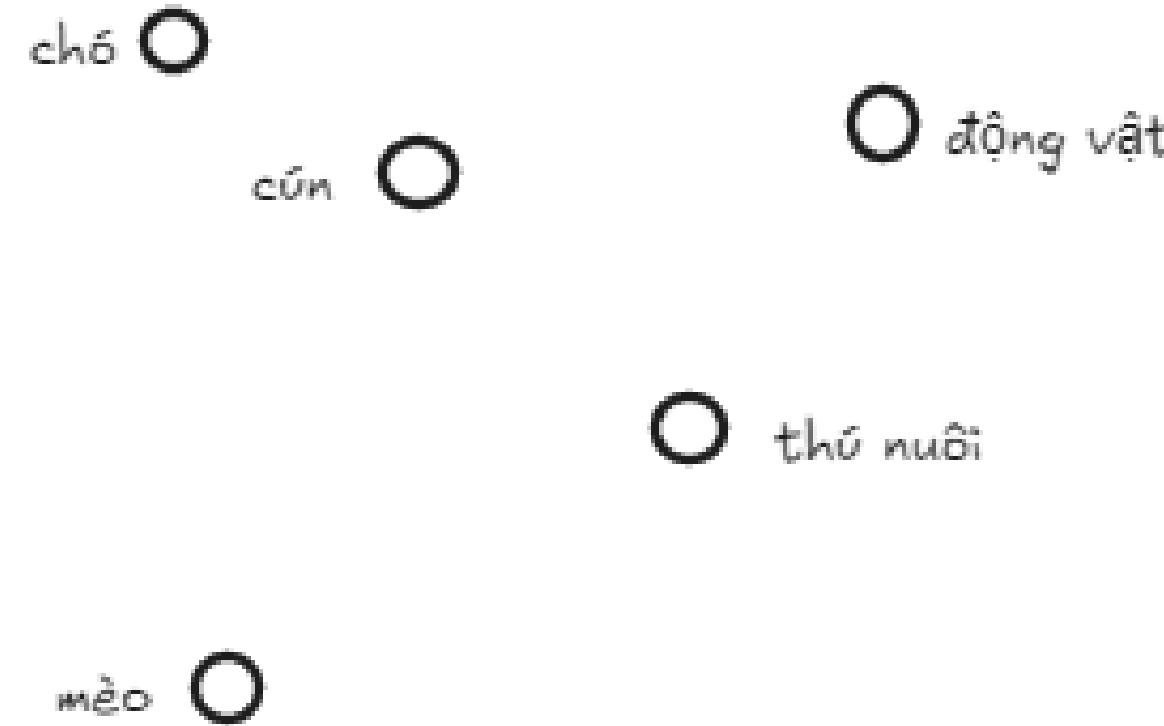


Skip Gram

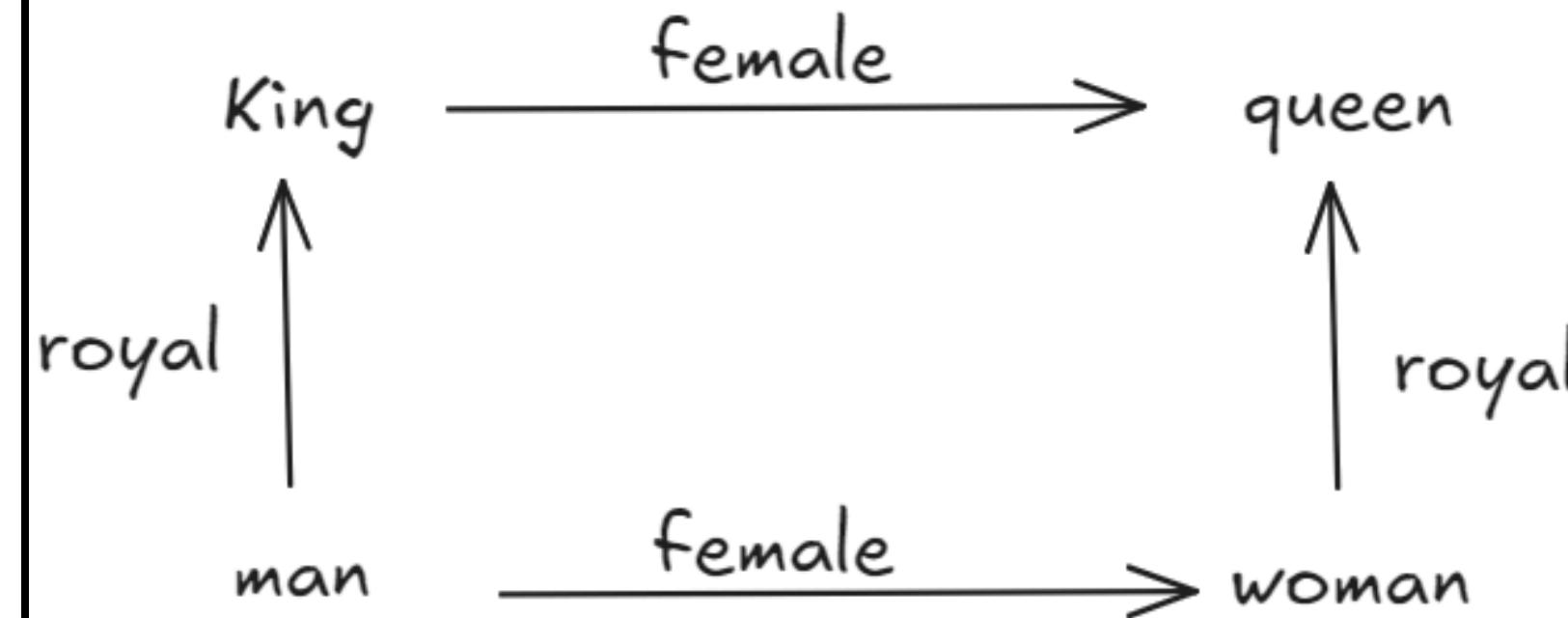


# 1.5 Word2Vec Embedding Space:

Các từ tương tự sẽ ở gần nhau



Cấu trúc tính của từ tương tự

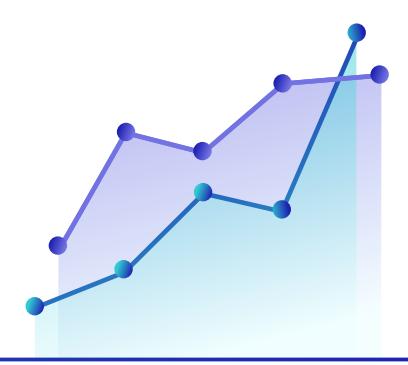


- Học từ ngữ cảnh.
- Hiểu được quan hệ từ vector có chiều thấp, dense.
- Thường sử dụng pretrained embedding.

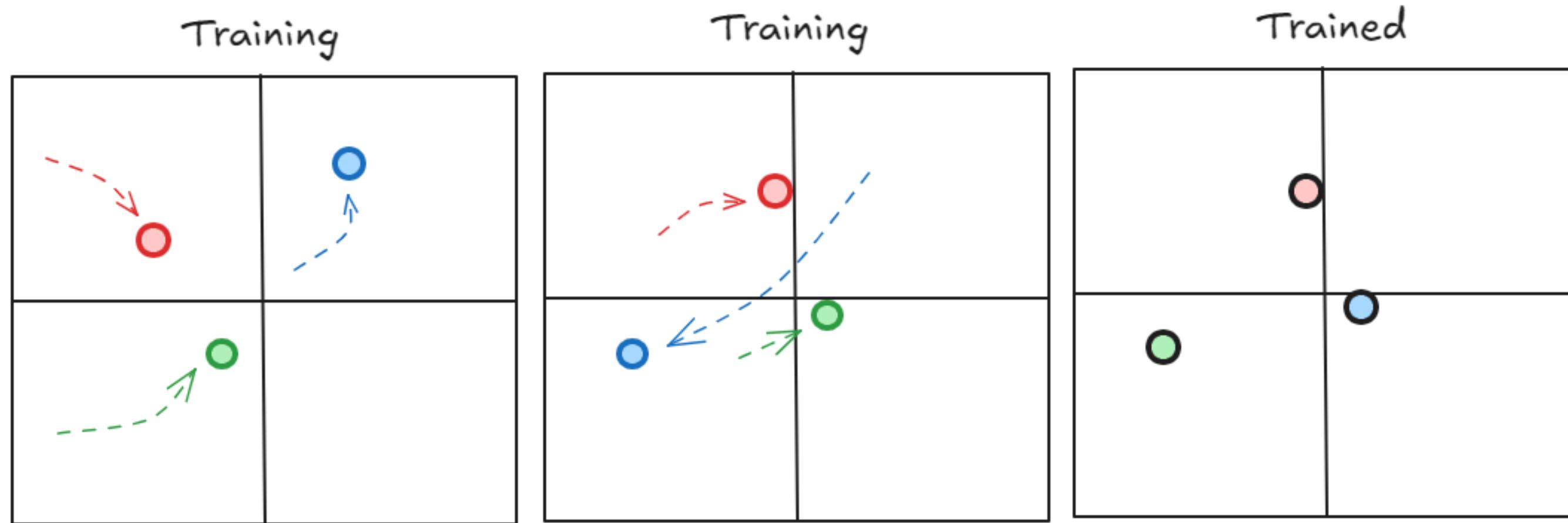
- Thường dùng để:
  - + cải thiện hệ thống gợi ý.
  - + cải thiện dịch ngôn ngữ.

→ Mô hình phân loại:  
Deep learning cho ra kết quả tốt hơn ML truyền thống.

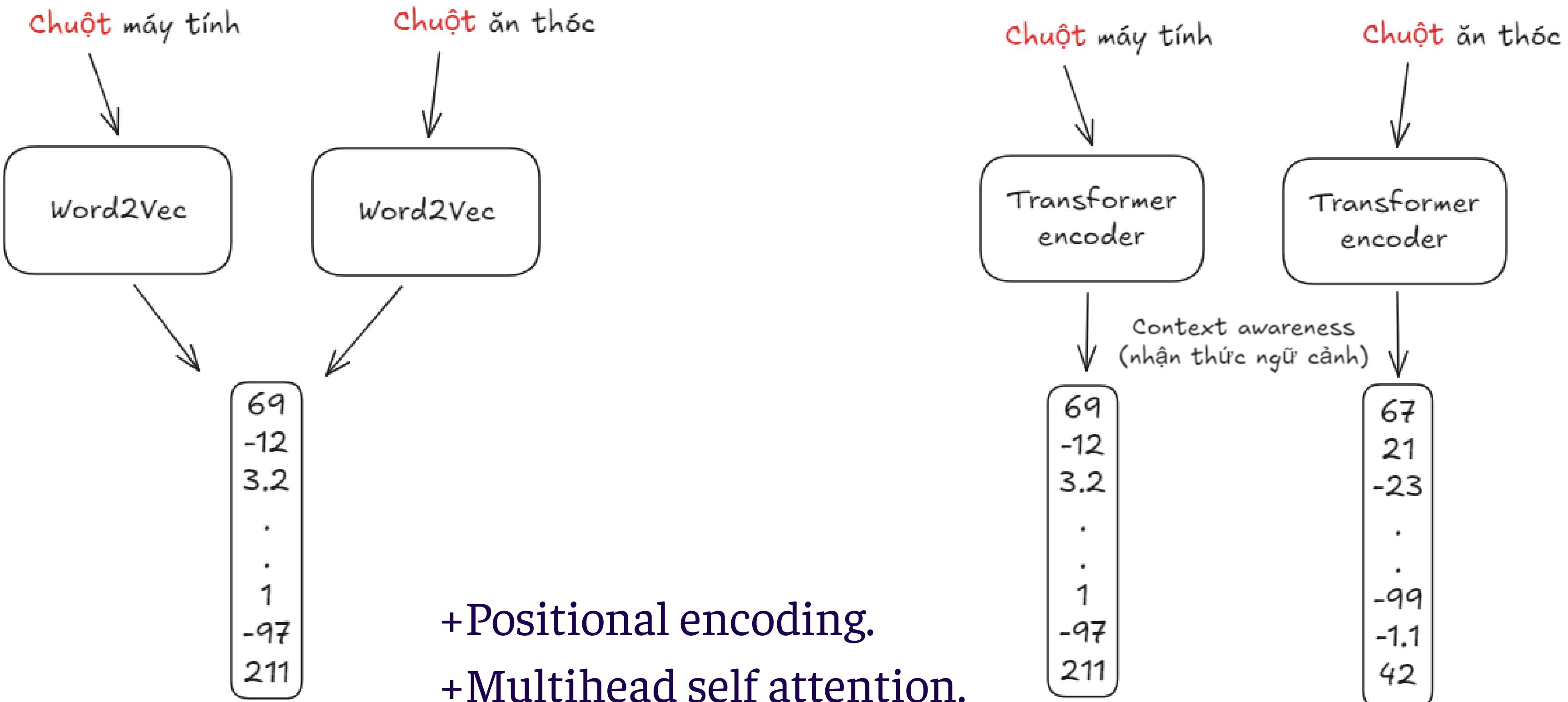
## 1.6 Hạn chế của Word2Vec:



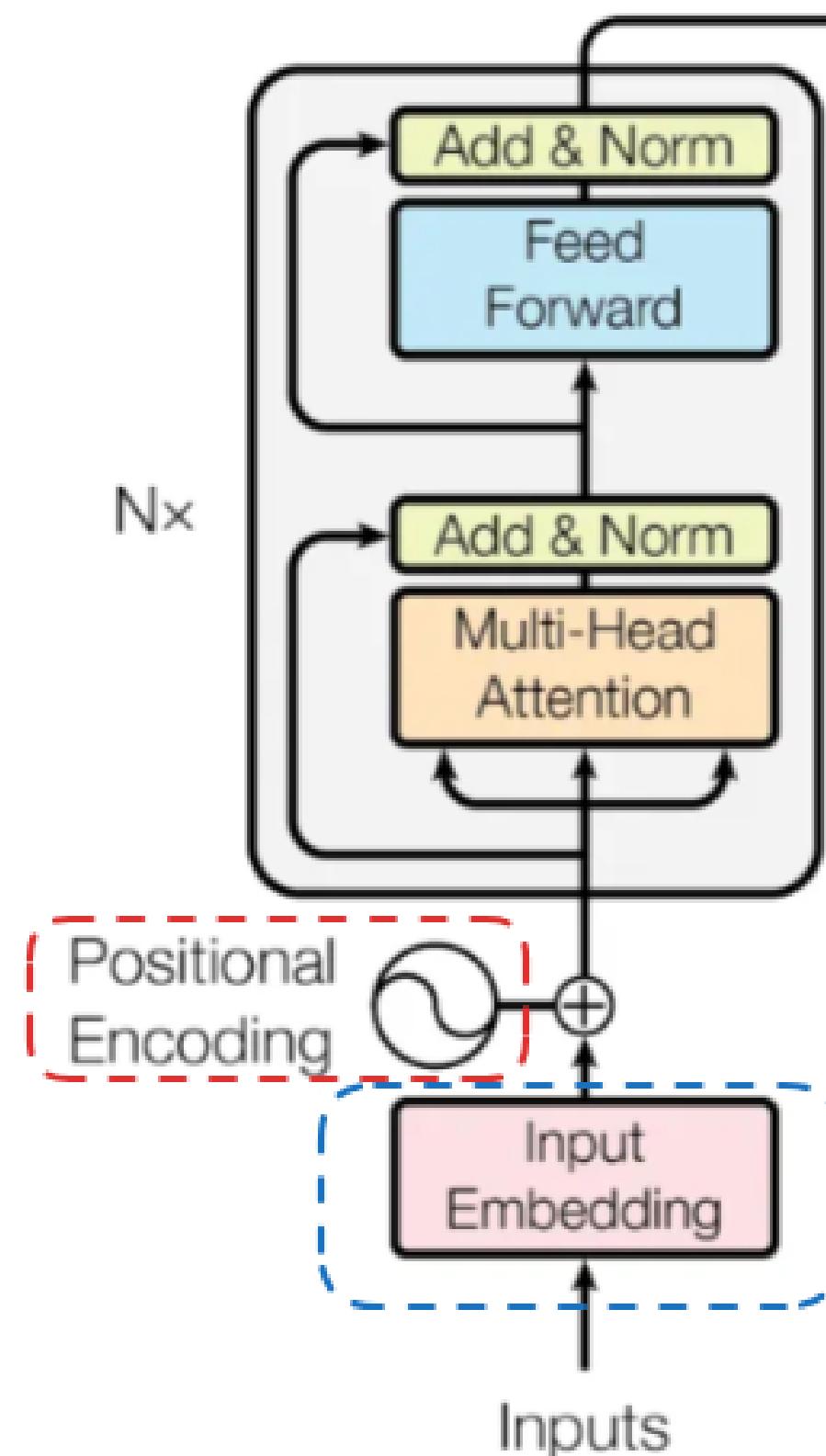
- Vẫn không quan tâm thứ tự từ.
- Tuy hiểu được ngữ nghĩa nhưng vẫn chỉ biểu diễn 1 vector mỗi từ.  
VD: Chuột (máy tính) = Chuột (động vật).
- Khi train: các vector chuyển động để phù hợp với các từ tương tự khác thêm vào.
- Khi train xong: các vector trở nên cố định → khó hiểu được bối cảnh.



# 1.7 Transformer:



## • Positional encoding:



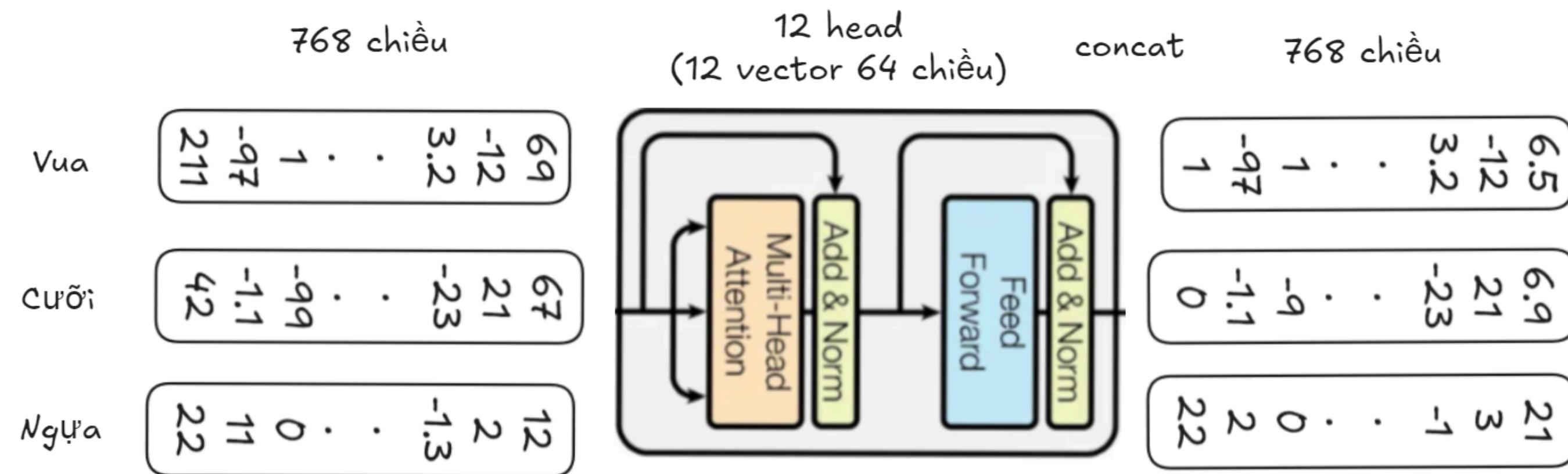
	Vua	Cưỡi	Ngựa
Word embedding	(0 1)	(-12 5.6)	(-53 -22)
Position embedding	(0.2 2)	(0 0.4)	(3 2)
	=	=	=
Word embedding into network	(0.2 3)	(-12 6)	(-50 -20)

- - Hiểu thứ tự từ.  
 - Quan hệ cú pháp câu.  
 - Hiểu câu.

## • Multihead attention:

+ Multihead attention cho phép mô hình nhìn câu từ nhiều gốc độ cùng lúc, mỗi head học một loại quan hệ khác nhau giữa các từ rồi ghép lại để hiểu câu toàn diện:

- Hiểu câu từ nhiều gốc độ.
- Học nhiều loại quan hệ từ phức tạp.
- Thu được ngữ nghĩa sâu và chính xác.





## Kết quả của các bài nghiên cứu trước đây:

- **Contextual embedding (Transformer):**

- Các mô hình sử dụng cấu trúc Transformer đạt hiệu quả vượt trội trên các chỉ số so với các mô hình khác.

- **Count Vector (TF-IDF):**

- Các mô hình ML truyền thống như LR và SVM cũng cho kết quả khá ấn tượng dù không bằng Transformer.

- Tuy nhiên kết quả vẫn rất cao cho thấy mô hình hoàn toàn có thể được sử dụng trong các hệ thống thực tế, đặc biệt khi tài nguyên phần cứng có hạn hoặc cần triển khai với thời gian huấn luyện hạn chế.

- **Neural Network (Word2Vec):**

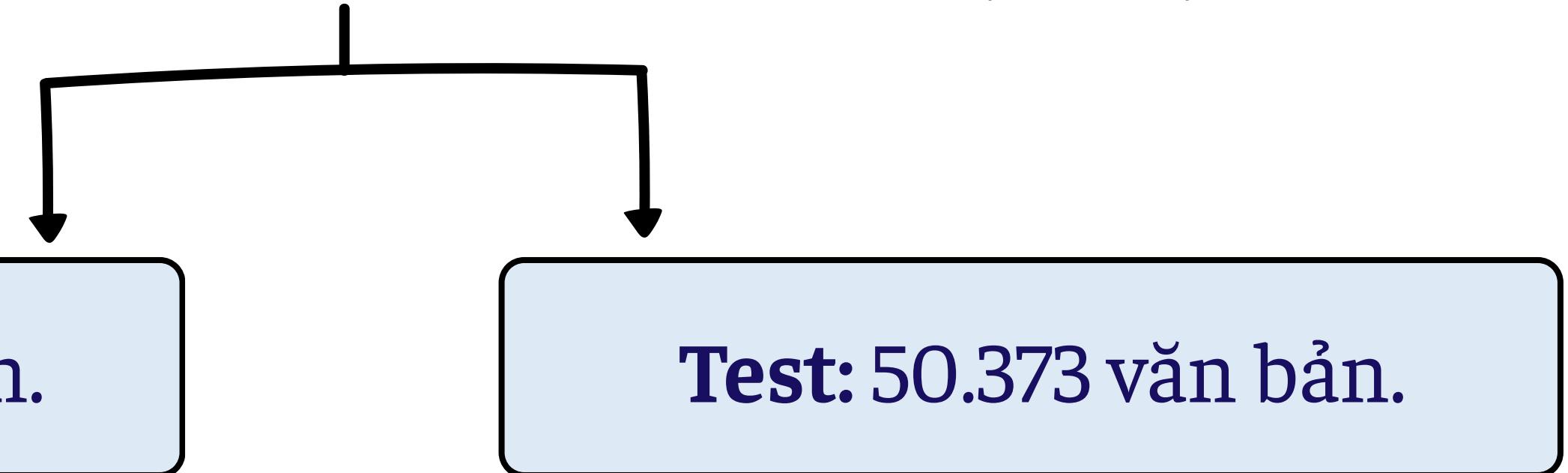
- Với các mô hình Deep Learning chỉ cho ra kết quả tạm ổn.

- Nguyên nhân có thể xuất phát từ việc phân loại bài báo chủ yếu dựa vào đặc điểm từ khóa và cụm từ mang tính chủ đề, ít phụ thuộc vào ngữ cảnh của nó trong câu.



### III. XÂY DỰNG MÔ HÌNH:

- Bộ dữ liệu được sử dụng:
- Bộ dữ liệu được sử dụng trong đề tài là VNTC (Vietnamese News Text Classification). Dữ liệu gồm các bài báo tiếng Việt được thu thập từ nhiều trang tin uy tín như VNExpress, Tuổi Trẻ, Thanh Niên và Người Lao Động. Mỗi bài viết được gán nhãn thuộc một trong 10 chủ đề khác nhau.
- Quy mô dữ liệu tổng cộng gồm 84.132 văn bản tiếng Việt, được chia thành:



\*\*\*Train\*\*\*

Topic	Topic ID	#files
*****		
Chinh tri Xa hoi	XH	5219
Doi song	DS	3159
Khoa hoc	KH	1820
Kinh doanh	KD	2552
Phap luat	PL	3868
Suc khoe	SK	3384
The gioi	TG	2898
The thao	TT	5298
Van hoa	VH	3080
Vi tinh	VT	2481
Total		33759

\*\*\*Test\*\*\*

Chinh tri Xa hoi	XH	7567
Doi song	DS	2036
Khoa hoc	KH	2096
Kinh doanh	KD	5276
Phap luat	PL	3788
Suc khoe	SK	5417
The gioi	TG	6716
The thao	TT	6667
Van hoa	VH	6250
Vi tinh	VT	4560
Total		50373

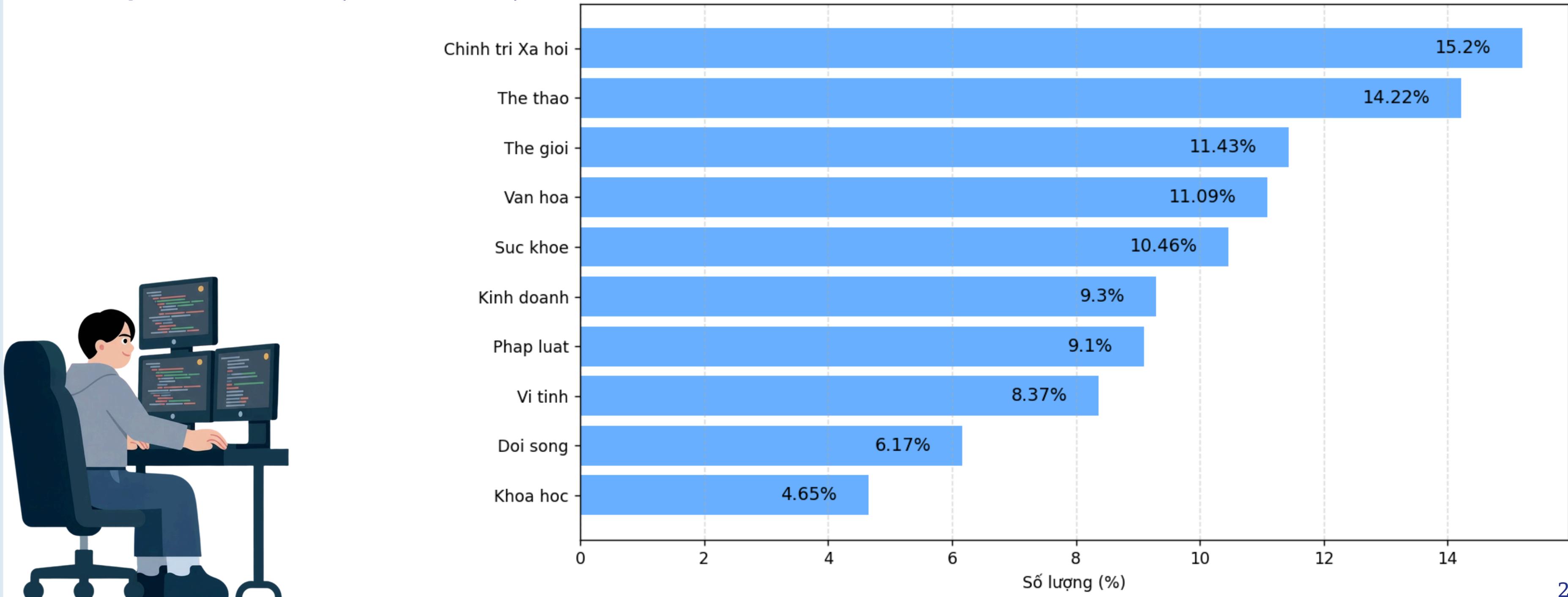
## Cấu trúc mõi văn bản dữ liệu:

- **Văn bản:** nội dung bài báo tiếng Việt (dạng text thuần).

- **Nhãn chủ đề:** một trong mười chủ đề sau: Chính trị - Xã hội; Đời sống; Khoa học; Kinh doanh; Pháp luật; Sức khỏe; Thế giới; Thể thao; Văn hóa và Vi tính.

- **DATASET:** Bộ dữ liệu gồm 10 folder (chủ đề) chứa các file văn bản có độ dài bất kỳ (30 - 50000 ký tự) được chia tỷ lệ 80% train – 20% test bằng phương pháp stratified split nhằm đảm bảo phân bố nhãn đồng đều.
- Trong quá trình huấn luyện, tập validation không được tách cố định mà được thực hiện thông qua Stratified 10-fold Cross-Validation trên tập train để giảm ảnh hưởng của việc chia dữ liệu ngẫu nhiên và đánh giá mô hình một cách ổn định hơn.

Tỷ lệ phân bổ dữ liệu theo từng chủ đề





## Mục tiêu:

### 1. Thủ các mô hình phân loại khác nhau:

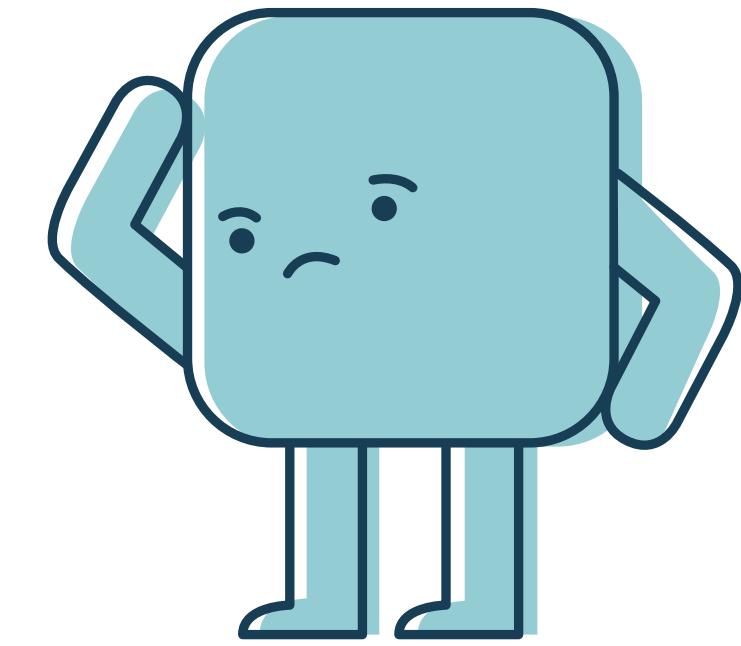
- Machine Learning : LR, SVM, XGBoost
- Deep Learning : DNN, LSTM.

### 2. Kết hợp với các cách trích xuất đặc trưng khác nhau:

- TF - IDF.
- TF - IDF + SVD (để giảm chiều).
- Word2Vec (Pretrained)
- link tải: [https://github.com/sonvx/word2vecVN//vi\\_word2vec.bin](https://github.com/sonvx/word2vecVN//vi_word2vec.bin).

### 3. Để đánh giá tổng quan bài toán với static embedding.

- Cân bằng dữ liệu bằng phương pháp class-weight và sample-weight (XG) trong quá trình huấn luyện mô hình.



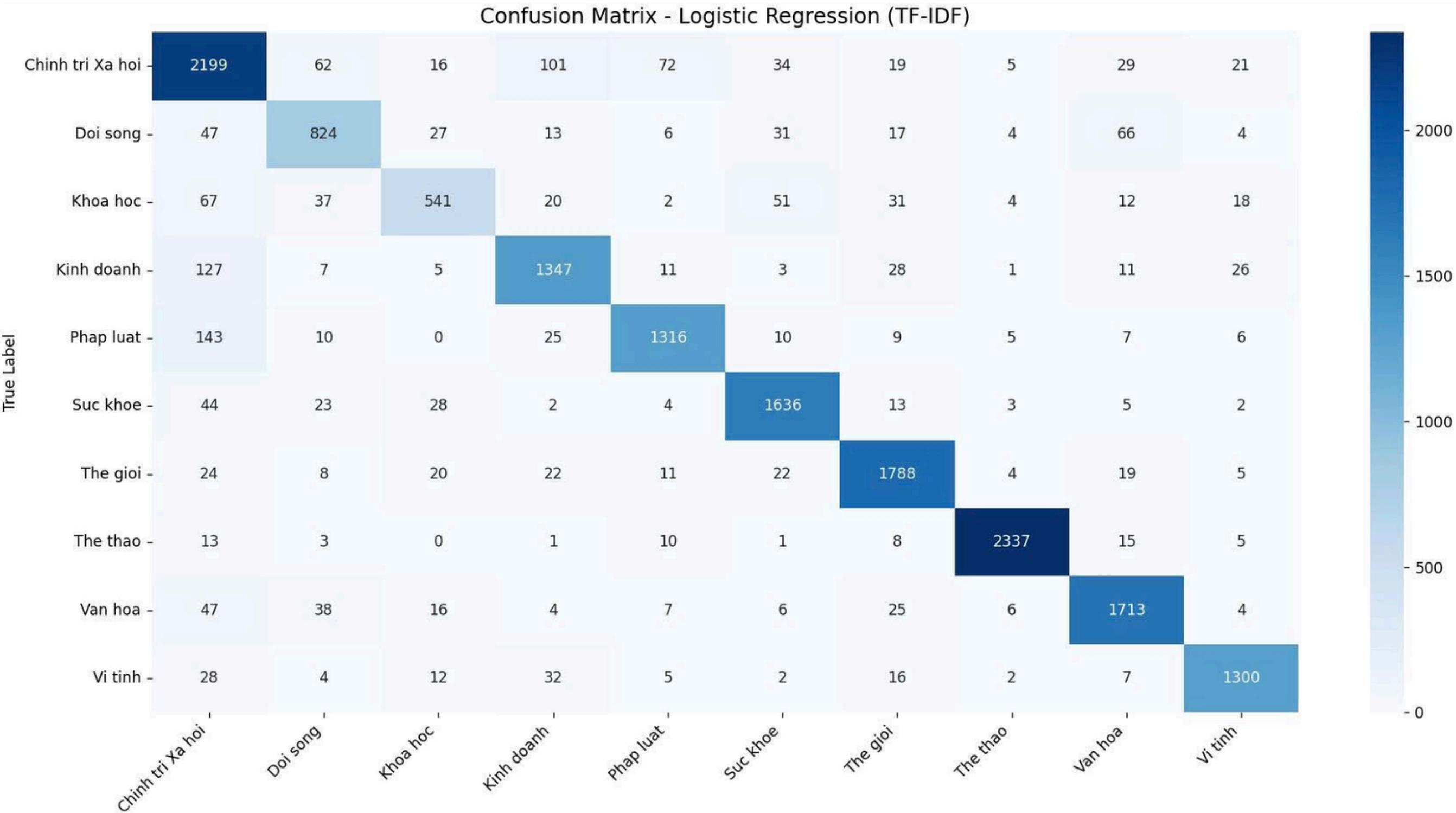
- **Top 5 thực nghiệm cao nhất (accuracy):**

- TF-IDF word kết hợp TF-IDF char giữ lại nhiều đặc trưng quan trọng của văn bản tiếng Việt.
- SVD giảm chiều nhưng làm mất chi tiết tuy nhiên cách biệt không nhiều và vượt trội hơn nếu nhắc đến chiều vector đã được giảm làm mô hình huấn luyện nhanh hơn.
- Word2Vec học ngữ cảnh nhưng không giữ trọng số từ khóa làm kém hiệu quả trong phân loại chủ đề.

➔ **SVM** cho kết quả tốt nhất trong tất cả các mô hình, DNN cũng cho kết quả bất ngờ nhưng cũng có thể hiểu DNN là mô hình DL gần với ML truyền thống nhất.

SUMMARY OF RESULTS	
<hr/>	
<hr/>	
Top performing models:	
1. SVM_TFIDF	: 0.9359
2. LR_TFIDF	: 0.9302
3. SVM_TFIDF_SVD	: 0.9293
4. DNN_TFIDF_SVD	: 0.9196
5. XGB_TFIDF_SVD	: 0.9156

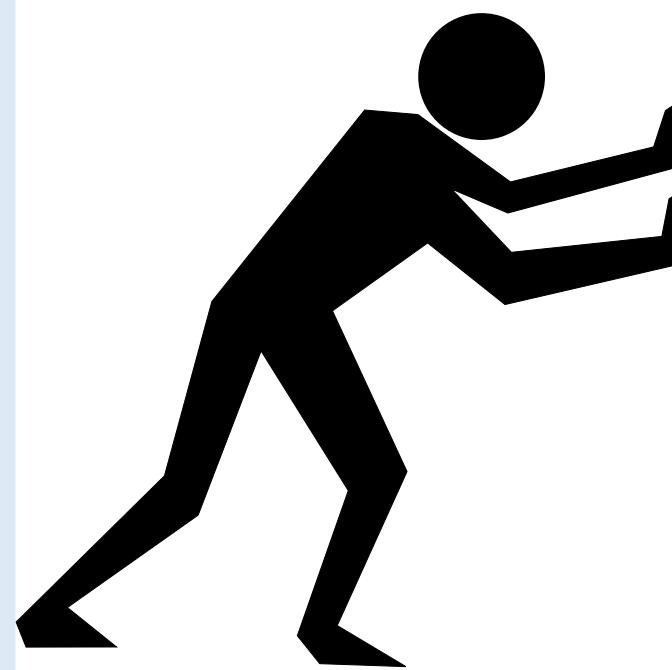
# ma trận nhầm lẫn khi không cân bằng



=> Ma trận nhầm lẫn cho thấy mô hình có xu hướng dự đoán tốt các lớp chiếm đa số, trong khi các lớp có ít dữ liệu vẫn còn bị nhầm lẫn đáng kể, phản ánh ảnh hưởng của sự mất cân bằng dữ liệu.

# IV. ĐÁNH GIÁ:

## Ma trận nhầm lẫn



Confusion Matrix - SVM\_TFIDF (Best Model)

True label	Predicted label									
	Chinh tri Xa hoi -	Doi song -	Khoa hoc -	Kinh doanh -	Phap luat -	Suc khoe -	The gioi -	The thao -	Van hoa -	Vi tinh -
Chinh tri Xa hoi -	2314	43	14	50	51	30	13	5	23	15
Doi song -	34	888	14	7	6	16	21	8	45	0
Khoa hoc -	20	22	670	9	0	31	10	2	3	16
Kinh doanh -	51	6	5	1458	7	2	11	0	3	23
Phap luat -	77	3	0	13	1418	3	5	7	2	3
Suc khoe -	29	13	26	1	3	1677	7	2	0	2
The gioi -	12	9	10	11	4	18	1830	11	13	5
The thao -	7	3	0	1	10	0	3	2353	13	3
Van hoa -	26	19	12	1	4	2	10	7	1784	1
Vi tinh -	8	4	6	14	1	1	10	2	6	1356

## Chỉ số đánh giá

- **Accuracy (Độ chính xác tổng thể):** Đo lường tỷ lệ số lượng dự đoán đúng trên tổng số mẫu trong toàn bộ các lớp.
- **Macro-Precision (Độ chính xác theo lớp – macro average):** Đo lường trung bình tỷ lệ mẫu được dự đoán đúng trên tổng số mẫu được dự đoán là thuộc về mỗi lớp (precision trung bình).
- **Macro-Recall (Khả năng nhận diện theo lớp – macro average):** Đo lường trung bình tỷ lệ mẫu đúng thực sự của mỗi lớp được mô hình nhận diện chính xác.
- **Macro-F1:** Đo lường trung bình độ chính xác tổng hợp giữa Precision và Recall cho tất cả các lớp. F1-score từng lớp được tính bằng trung bình điều hòa giữa Precision và Recall, sau đó lấy trung bình trên toàn bộ các lớp (macro average).



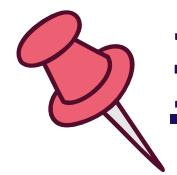
## Đánh giá hiệu quả của mô hình phân loại trên tập test thông qua các chỉ số đo lường:

Accuracy	Precision (macro)	Recall (macro)	F1-Score (macro)
0,94	0.93	0.93	0.93



Mô hình SVM - TF-IDF đạt độ chính xác cao (Accuracy  $\approx 94\%$ ), cho thấy phần lớn văn bản được phân loại đúng.

- **Macro-Precision** = 0.93: Mô hình dự đoán khá chính xác ở hầu hết các lớp, ít bị dự đoán nhầm sang lớp khác.
- **Macro-Recall** = 0.93: Mô hình nhận diện được đa số mẫu thuộc từng lớp, kể cả các lớp có ít dữ liệu.
- **Macro-F1** = 0.93: Mô hình cân bằng tốt giữa Precision và Recall trên toàn bộ 10 lớp.



## Đánh giá chi tiết các chỉ số cho từng lớp phân loại:

```
Model saved: SVM_TFIDF
2025-12-24 01:12:06,491 - INFO - Validation accuracy: 0.9340
2025-12-24 01:12:06,491 - INFO - Test accuracy: 0.9359
2025-12-24 01:12:06,491 - INFO - Training time: 2:05:15.099957

classification Report for SVM_TFIDF:
      precision    recall  f1-score   support
Chinh tri Xa hoi       0.90      0.90      0.90      2558
Doi song             0.88      0.85      0.87      1039
Khoa hoc             0.89      0.86      0.87      783
Kinh doanh            0.93      0.93      0.93     1566
Phap luat             0.94      0.93      0.93     1531
Suc khoe              0.94      0.95      0.95     1760
The gioi              0.95      0.95      0.95     1923
The thao              0.98      0.98      0.98     2393
Van hoa               0.94      0.96      0.95     1866
Vi tinh               0.95      0.96      0.96     1408

accuracy                   0.94     16827
macro avg                 0.93     16827
weighted avg              0.94     16827
```

- Mô hình đạt hiệu năng cao với Accuracy. Phần lớn các lớp đều có Precision, Recall và F1-score cao (khoảng 0.85–0.98), cho thấy mô hình dự đoán ổn định và ít sai lệch.
- Tuy nhiên lớp Đời sống và Khoa học có các chỉ số thấp hơn nhưng đã cải thiện mất cân bằng. Cho thấy mô hình gặp khó khăn khi nhận diện đúng, nguyên nhân có thể do nội dung lớp này có ít dữ liệu nhất.



# Kết quả kiểm thử mô hình với câu thực tế:

Nhập câu (kết thúc bằng #):

Ngọc Thanh và Cẩm Miền đoạt áo vàng (NLĐ)- Hôm qua, 27-12, giải đua xe đạp “Về Phước Long xây chiến thắng” 2004 đã kết thúc sau 3 chặng đua: Thủ Dầu Một-Xoài, Đồng Xoài - Phước Long và Phước Long- Thủ Dầu Một. Áo vàng chung cuộc thuộc về tay đua Lê Ngọc Thanh ở nội dung nam và Cẩm Miền ở nội dung nữ. Đội bộ Thực vật Sài Gòn Dofilm đoạt giải nhất đồng đội nam và đội Cấp thoát nước Bình Dương đoạt giải nhất đồng đội nữ.

#

→ Kết quả: The thao

Nhập câu (kết thúc bằng #):

Chắc bạn đã biết con trai? Trai đã phải chịu đựng đau đớn biết bao nhiêu khi có những hạt cát rơi vào. Nhưng rồi, mặc cho nỗi đau dày vò, trai vẫn can trường bọc lấy những hạt cát để từng ngày tạo nên những viên ngọc đẹp tuyệt vời. Trai là loài động vật có hai mảnh vỏ. Một đôi vợ chồng cũng như con trai vậy. Người vợ và người chồng như hai mảnh vỏ trai gắn kết với nhau để rồi một ngày tạo ra một viên ngọc trai quý báu. Khi có một vật lạ rơi vào bên trong, nếu hai mảnh vỏ trai chỉ làm điều đơn giản là đẩy nó ra ngoài hoặc tách rời nhau và không phối hợp với nhau, sẽ không bao giờ có những viên ngọc trai quý báu. Đầu tiên, chúng ta phải biết chấp nhận những điều khó chịu như những hạt cát, và rồi tận dụng những hạt cát đó để tạo ra một cái gì đó tuyệt vời hơn. Cuộc sống hôn nhân cũng thế.

#

→ Kết quả: Doi song

Nhập câu (kết thúc bằng #):

Bán cả xe hơi của công ty chơi cá độ (NLĐ)- Ngày 27-12, Công an Q.11- TPHCM bắt khẩn cấp Nguyễn Hữu Hiệp (SN 1975, ngụ Bến Tre, tạm trú Q.9). Hiệp là lái xe của Công ty Sacai tại Q.10 - TPHCM. Do nợ nần trong cá độ bóng đá, Hiệp đã lấy ô tô Toyota Zace BS: 52X-1496 của Công ty Sacai đến Thủ Đức cầm lấy 150 triệu đồng. Sau đó, Hiệp bỏ trốn về quê nhưng đã bị công an bắt giữ. Công an Q.11 đã thu hồi xe trả cho Công ty Sacai.

#

→ Kết quả: Phap luat

Nhập câu (kết thúc bằng #):

Các nhà khoa học Viện Sinh thái và Tài nguyên sinh vật Việt Nam và Viện Động vật Saint Petersburg của Nga vừa phát hiện loài thằn lằn mới ở miền Trung Việt Nam và Nam Lào. Loài thằn lằn có tên gọi nhông Natalia, là một họ của loài nhông, phân bố ở Đà Nẵng, Quảng Nam, Gia Lai, Kon Tum và các tỉnh của Lào như Saravani, Xê Kong.

#

→ Kết quả: Khoa hoc

Nhập câu (kết thúc bằng #):

Nhiều nghiên cứu cho thấy chỉ cần duy trì những thói quen đơn giản như ngủ đủ giấc, uống đủ nước, vận động nhẹ 30 phút mỗi ngày và ăn nhiều rau xanh có thể giảm đáng kể nguy cơ mắc bệnh tim mạch, tiểu đường và béo phì. Việc thay đổi từng chút một sẽ giúp cơ thể thích nghi và duy trì lâu dài.

#

→ Kết quả: Suc khoe



**Thank you for your attention!**

**We sincerely appreciate your time and interest in our project.**

## Top performing models:

1.	SVM_TFIDF	:	0.9359
2.	LR_TFIDF	:	0.9302
3.	SVM_TFIDF_SVD	:	0.9293
4.	DNN_TFIDF_SVD	:	0.9196
5.	XGB_TFIDF_SVD	:	0.9156
6.	LR_TFIDF_SVD	:	0.9127
7.	DNN_WORD2VEC	:	0.8775
8.	LSTM_TFIDF_SVD	:	0.8719
9.	DNN_NGRAM_SVD	:	0.8681
10.	XGB_NGRAM_SVD	:	0.8605
11.	SVM_NGRAM_SVD	:	0.8605
12.	SVM_WORD2VEC	:	0.8533
13.	LR_WORD2VEC	:	0.8495
14.	LR_NGRAM_SVD	:	0.8491
15.	LSTM_WORD2VEC	:	0.8439

BEST MODEL: SVM\_TFIDF with accuracy: 0.9359