



Nhóm 2

# Ứng dụng Lọc Dựa trên Nội dung kết hợp Xử lý Ngôn ngữ Tự nhiên trong Việc Nâng cao Độ Chính xác của Hệ thống Gợi ý việc làm tại Việt Nam



Presented by  
**Nhóm 2**

Presented to  
**GVHD: TS. Nguyễn Thôn Dã**

# Danh sách thành viên

HỒ SONG TÍN

MSSV: K224111469

(Nhóm Trưởng)

NGUYỄN HOÀNG KIM

MSSV: K224111452

(Thành Viên)

HUỲNH HIẾU TRUNG

MSSV: K224111474

(Thành Viên)

HUỲNH HUỆ TRÚC

MSSV: K224111475

(Thành Viên)

PHẠM THANH HƯNG

MSSV: K224131589

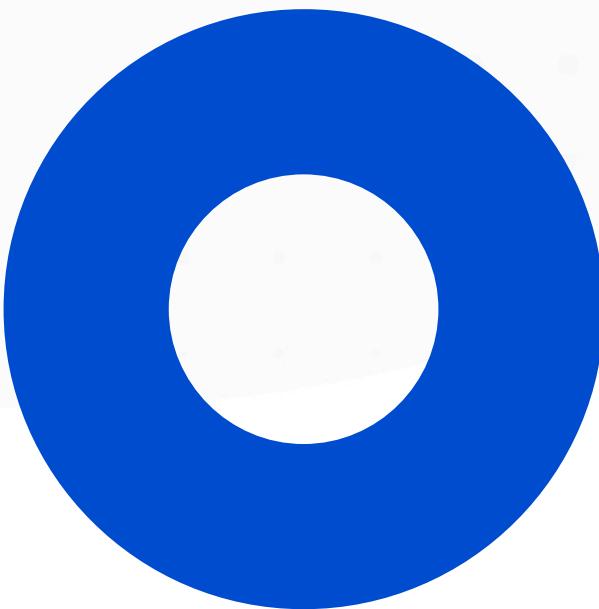
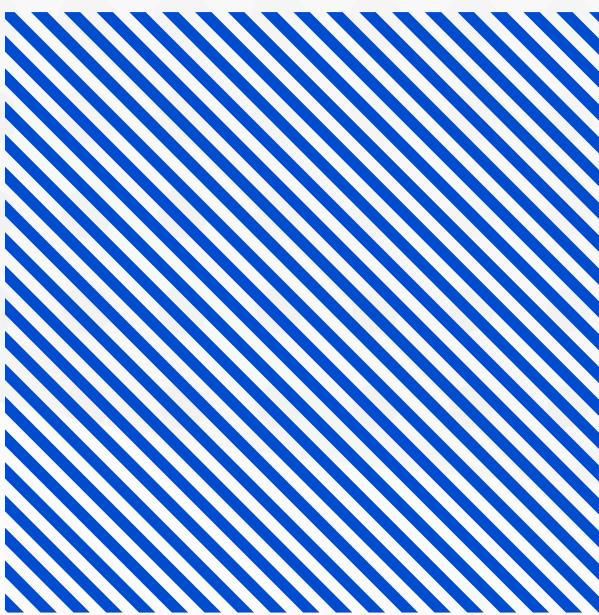
(Thành Viên)

PHẠM TRÚC PHƯƠNG

MSSV: K224131604

(Thành Viên)

## 1. Giới thiệu đề tài nghiên cứu

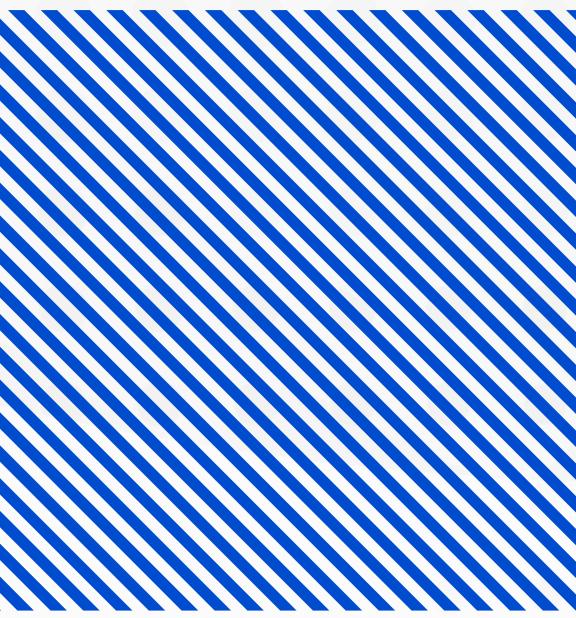


## 1.1. Bối cảnh và vấn đề cần nghiên cứu

**Thị trường lao động Việt Nam hiện đối mặt với nghịch lý "thừa mà thiếu":** Dù có nhiều thông tin tuyển dụng, cả ứng viên và nhà tuyển dụng vẫn gặp khó khăn trong việc kết nối hiệu quả. Ba vấn đề lớn đang chia cắt thị trường:

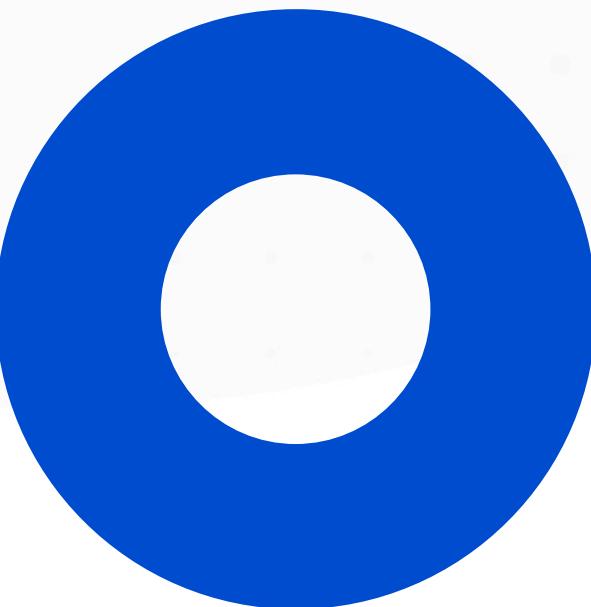
- **Lỗ hổng ngôn ngữ nghề nghiệp:** CV của ứng viên thường sơ sài, thiếu ngôn ngữ chuyên môn, trong khi JD của doanh nghiệp lại quá phức tạp, tạo nên sự thiếu thấu hiểu hai chiều.
- **Định kiến hệ thống:** Các thuật toán tuyển dụng hiện tại thường bỏ qua lao động phi truyền thống như người chuyển giới, lao động lớn tuổi... gây lãng phí nguồn nhân lực tiềm năng và vi phạm quy định về bình đẳng.
- **Bản đồ cơ hội méo mó:** Trong khi các khu công nghiệp thiếu nhân lực nghiêm trọng, thì thành phố lớn lại dư thừa ứng viên không phù hợp chuyên môn.

Những rạn nứt này cho thấy cần **xây dựng lại hệ thống quản lý việc làm**, không chỉ nâng cấp công nghệ mà còn đổi mới tư duy và chính sách.



## 1.2. Mục tiêu nghiên cứu

---

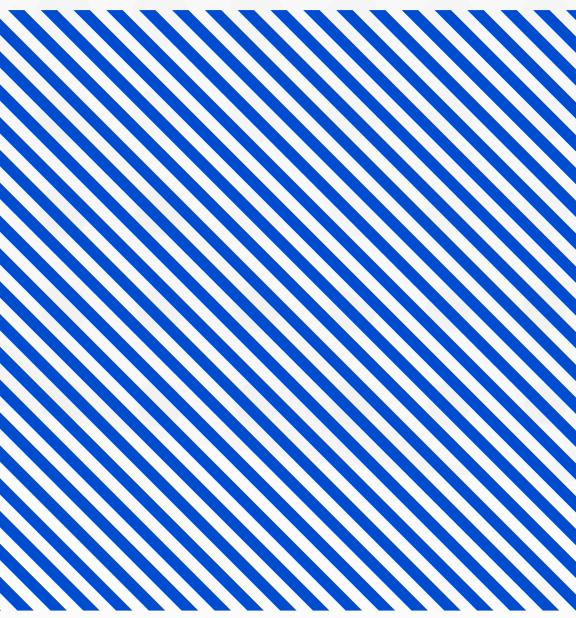


### Mục tiêu tổng quát:

Phát triển mô hình **gợi ý việc làm thông minh** sử dụng phương pháp **Lọc dựa trên nội dung (Content-Based Filtering)** kết hợp các kỹ thuật **Xử lý Ngôn ngữ Tự nhiên (NLP)** hiện đại, nhằm tăng độ chính xác và phù hợp của gợi ý việc làm cho thị trường lao động Việt Nam.

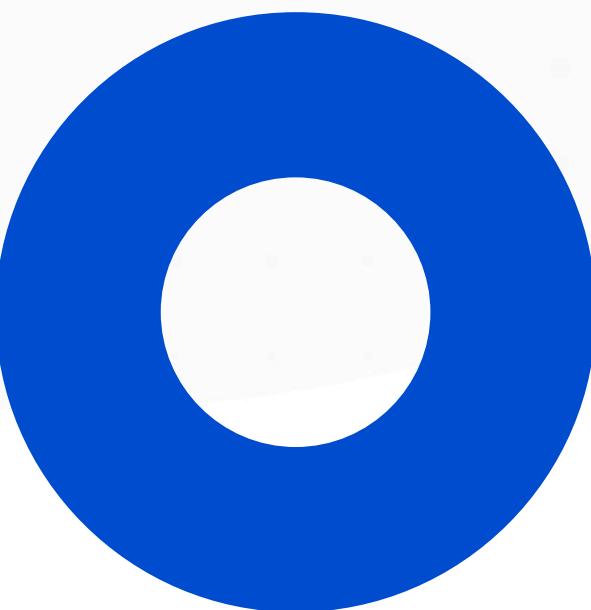
### Mục tiêu cụ thể:

- **Nghiên cứu** các mô hình lọc, kỹ thuật NLP tiếng Việt (PhoBERT, Pho2Vec), và dữ liệu JD & CV.
- **Xây dựng module** xử lý dữ liệu, trích xuất và vector hóa thông tin từ văn bản.
- **Thiết kế thuật toán** gợi ý, tính toán độ tương đồng giữa CV và JD.
- **Kiểm thử & đánh giá** mô hình trên dữ liệu thực tế và so sánh với hệ thống hiện có.



## 1.2. Mục tiêu nghiên cứu

---



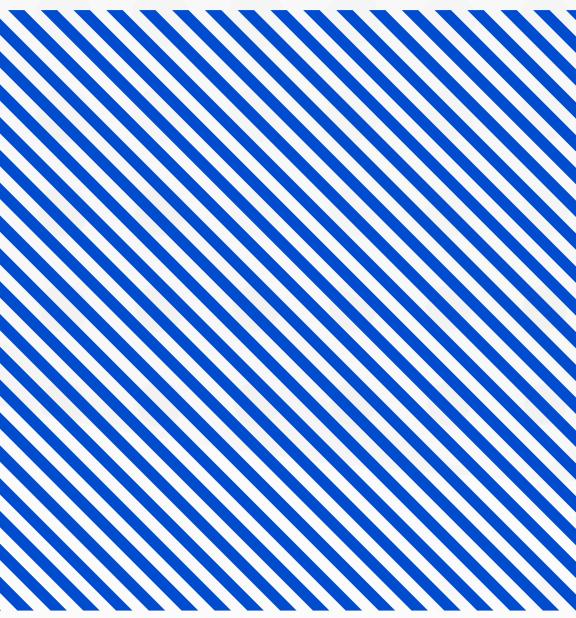
### Về mặt khoa học:

- Góp phần ứng dụng **NLP tiếng Việt** vào hệ thống gợi ý – một lĩnh vực còn nhiều thách thức.
- Đề xuất **mô hình lai (hybrid)** giữa lọc nội dung và NLP sâu, làm giàu thêm tri thức về công nghệ tuyển dụng thông minh.
- Cung cấp **dữ liệu thực nghiệm** có giá trị cho các nghiên cứu tương lai tại Việt Nam.

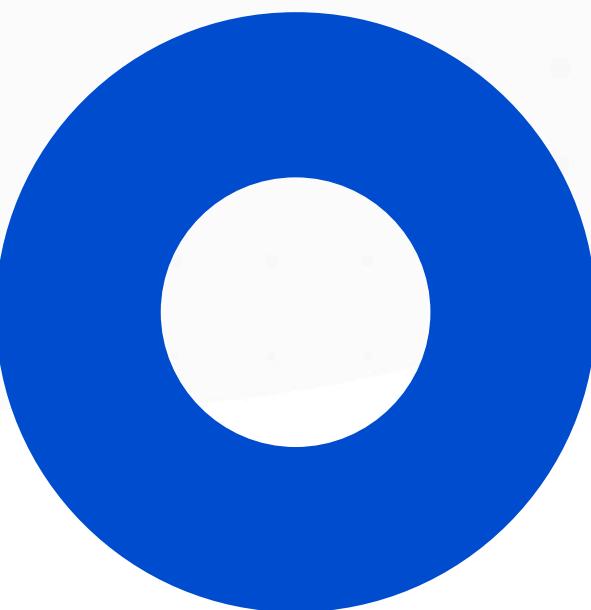
### Về mặt thực tiễn:

- Giúp **ứng viên** tiếp cận cơ hội phù hợp nhanh hơn.
- Hỗ trợ **doanh nghiệp** sàng lọc hiệu quả, tối ưu chi phí tuyển dụng.
- Tăng trải nghiệm người dùng và năng lực cạnh tranh cho các **nền tảng tuyển dụng**.

Tác động xã hội: Thúc đẩy kết nối cung-cầu lao động, nâng cao chất lượng nguồn nhân lực và hiệu quả thị trường lao động Việt Nam.



## 1.3. Câu hỏi nghiên cứu



**Câu hỏi chính:**

**Làm thế nào để kết hợp Lọc dựa trên Nội dung và Xử lý Ngôn ngữ Tự nhiên (NLP) nhằm nâng cao độ chính xác của hệ thống gợi ý việc làm trong bối cảnh đặc thù của thị trường lao động Việt Nam?**

**Các câu hỏi phụ:**

1. Trích xuất đặc trưng:

- Phương pháp nào hiệu quả nhất để biểu diễn ngữ nghĩa từ văn bản tiếng Việt không cấu trúc (CV và JD)?

2. So sánh độ chính xác:

- Tích hợp deep embeddings từ NLP cải thiện ra sao so với TF-IDF hay so khớp từ khóa?

3. Hiệu suất thực tế:

- Mô hình có đạt Precision@K, Recall@K, nDCG@K mong đợi khi thử nghiệm trên dữ liệu thật tại Việt Nam?



# Tổng quan về bộ dữ liệu

2

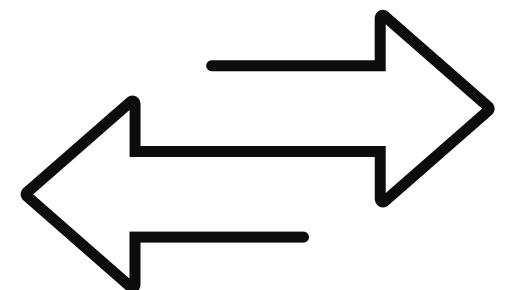
## 2.1 Nguồn gốc bộ dữ liệu

Dữ liệu tin tuyển dụng được nhóm crawler từ trang CareerViet.vn - một phiên bản nội địa hóa của CareerBuilder Việt Nam.

Dữ liệu ứng viên được crawler từ trang Timviec365.vn, nơi công khai hàng chục nghìn hồ sơ ứng viên thuộc nhiều ngành nghề khác nhau.

21.861 mẫu tuyển dụng

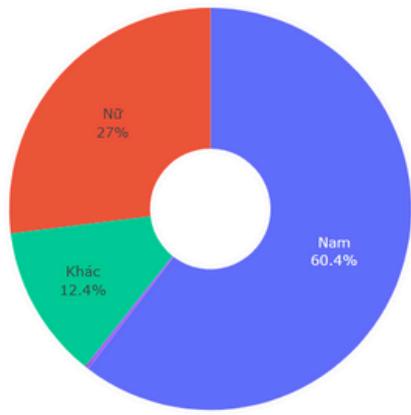
46.124 mẫu hồ sơ ứng viên



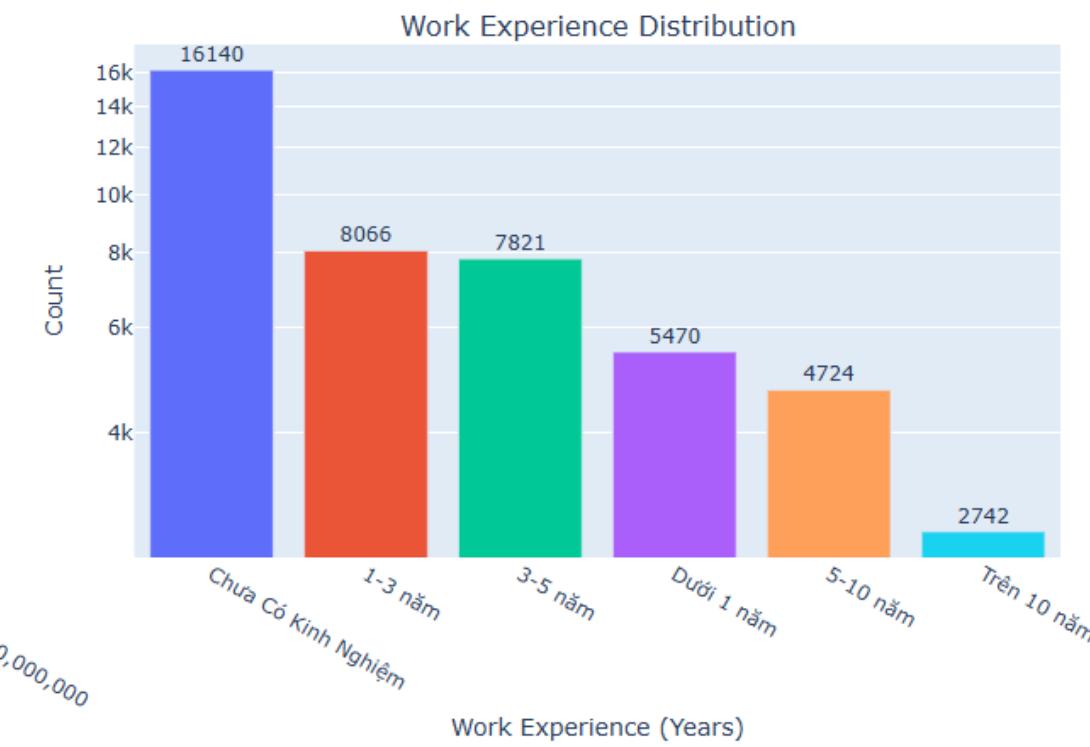
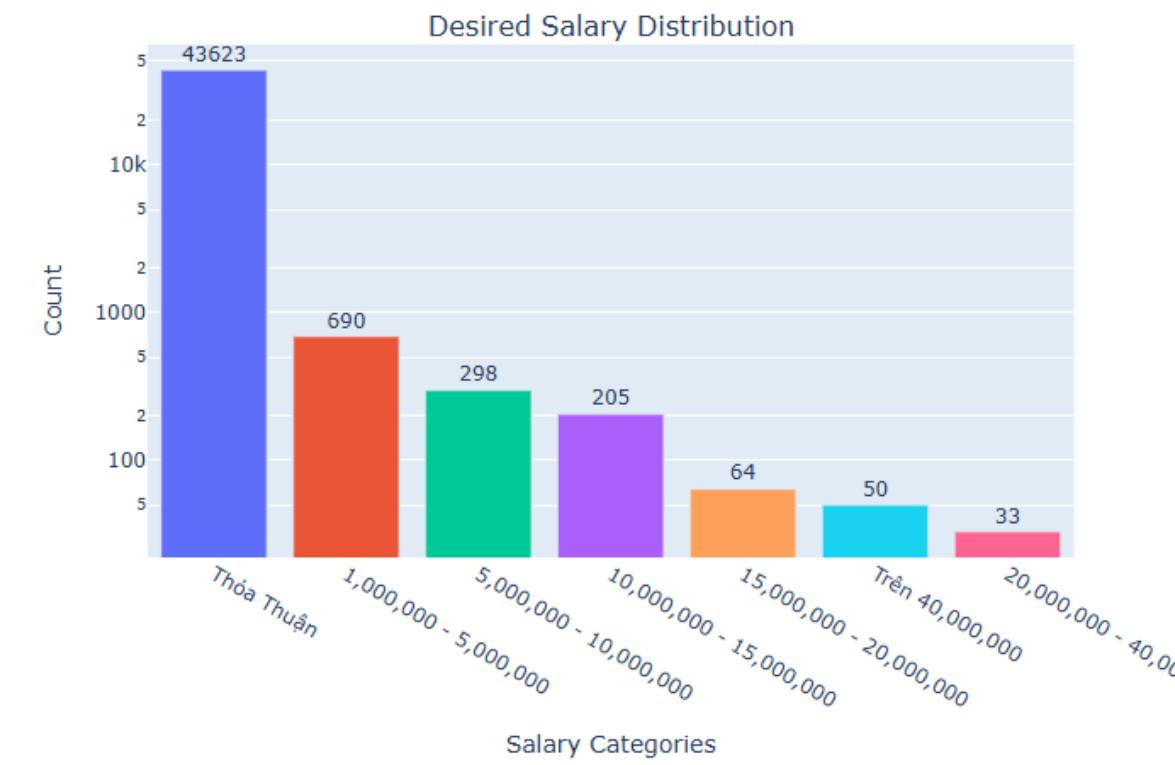
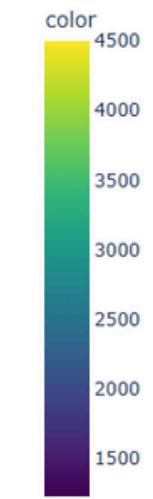
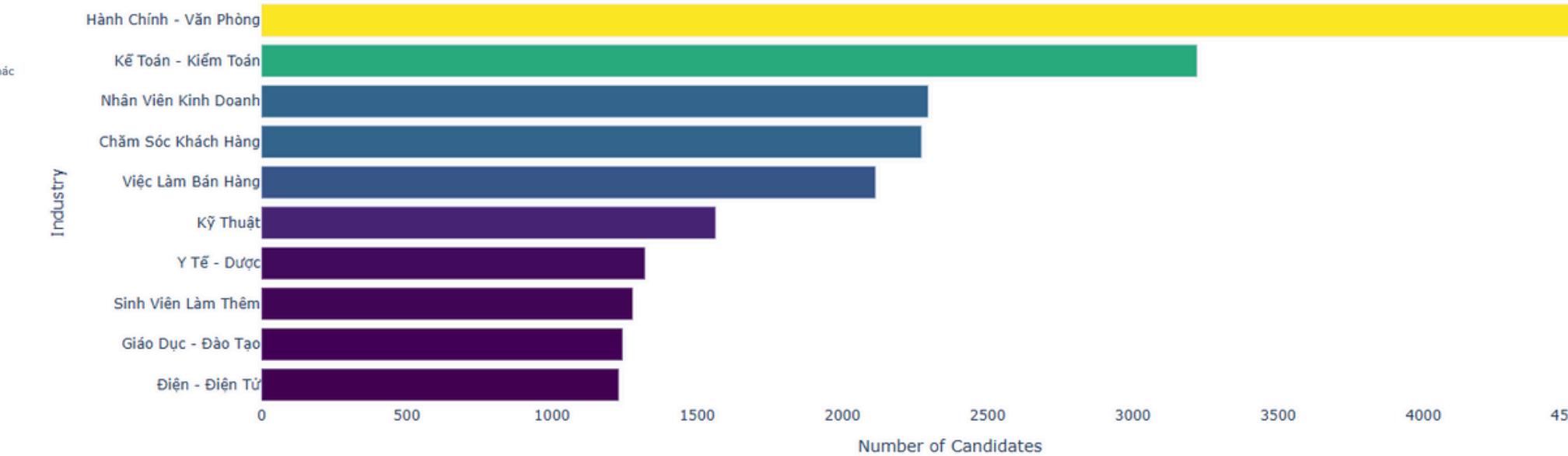
career  
viet

## 2.2 Bộ dữ liệu ứng viên

Gender Distribution

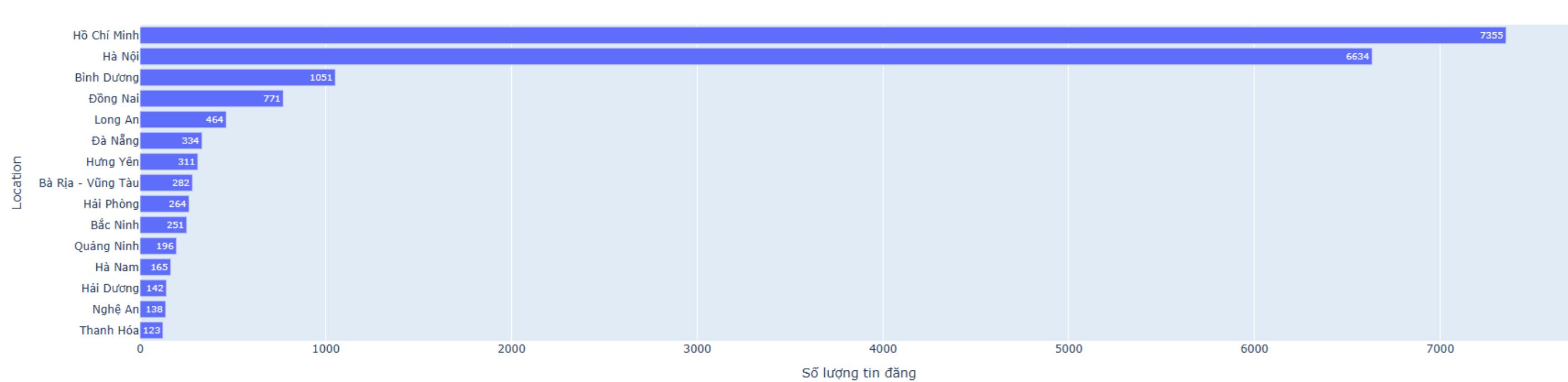


Top 10 Industries by Number of Candidates

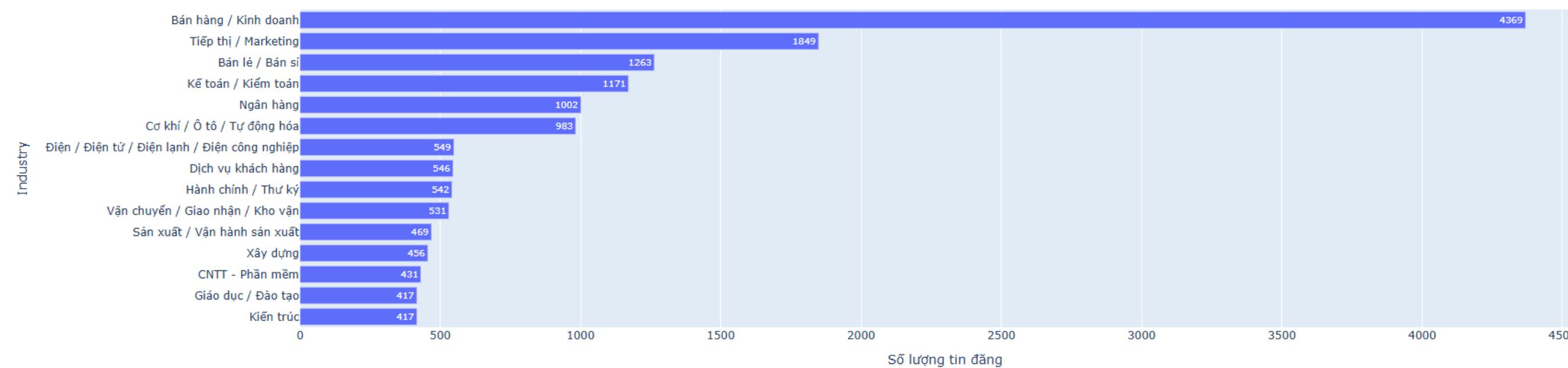


## 2.3 Bộ dữ liệu công việc

Top 15 Tỉnh/Thành phố có nhiều tin tuyển dụng nhất



Top 15 Ngành nghề có nhu cầu tuyển dụng cao nhất



3

# Phương pháp luận nghiên cứu



# 3.1 Cơ sở lý thuyết

## 3.1.1. TF - IDF

**TF-IDF (Term Frequency – Inverse Document Frequency)** là phương pháp thống kê đánh giá độ quan trọng của từ trong một văn bản so với toàn bộ tập tài liệu.

Nó kết hợp hai thành phần:

- **Term Frequency (TF)**: tần suất xuất hiện của từ trong một văn bản.
- **Inverse Document Frequency (IDF)**: đo độ hiếm của từ đó trong toàn bộ tập tài liệu, nhằm giảm trọng số cho các từ phổ biến.

### Công thức

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{df(t)}\right)$$

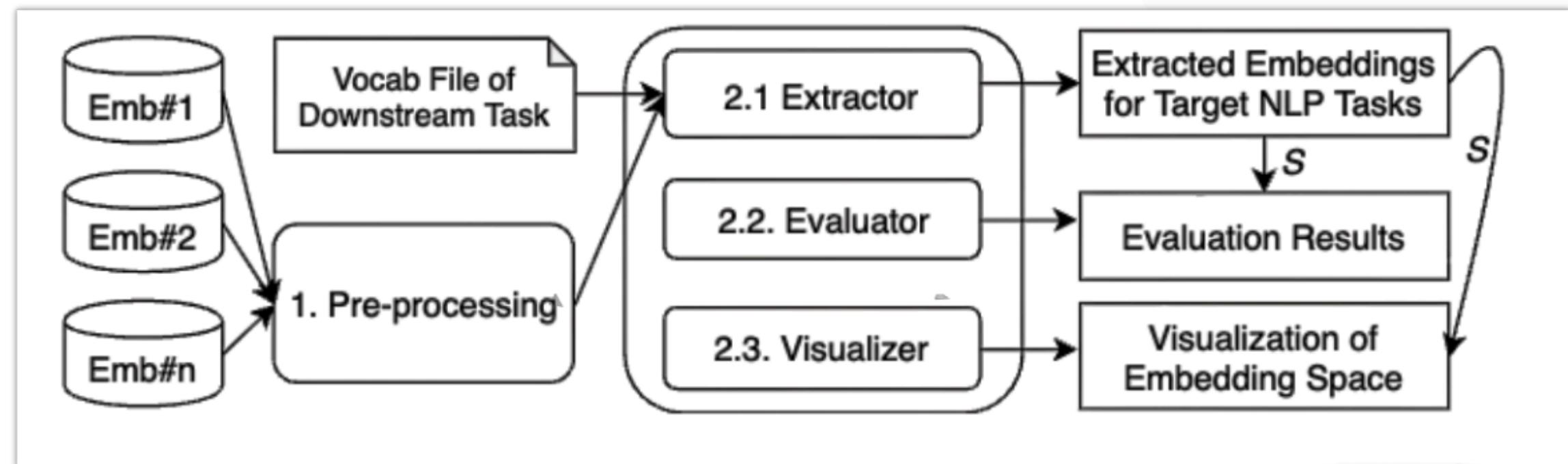


# 3.1 Cơ sở lý thuyết

## 3.1.2 Pho2Vec

**Pho2Vec** (còn gọi là **PhoW2V**) là một mô hình biểu diễn từ (word embedding) tiền huấn luyện được xây dựng dành riêng cho tiếng Việt, dựa trên kiến trúc Word2Vec của Mikolov et al. (2013).

- Hỗ trợ biểu diễn từ ở cấp từ và âm tiết, phản ánh đặc điểm ngôn ngữ tiếng Việt.
- Được huấn luyện trên dữ liệu lớn (Wikipedia, báo chí, mạng xã hội...).
- Cho phép ánh xạ từ vào không gian vector để biểu diễn ngữ nghĩa.
- Được đánh giá và chọn lọc thông qua hệ thống ETNLP (Nguyen & Nguyen, 2020).



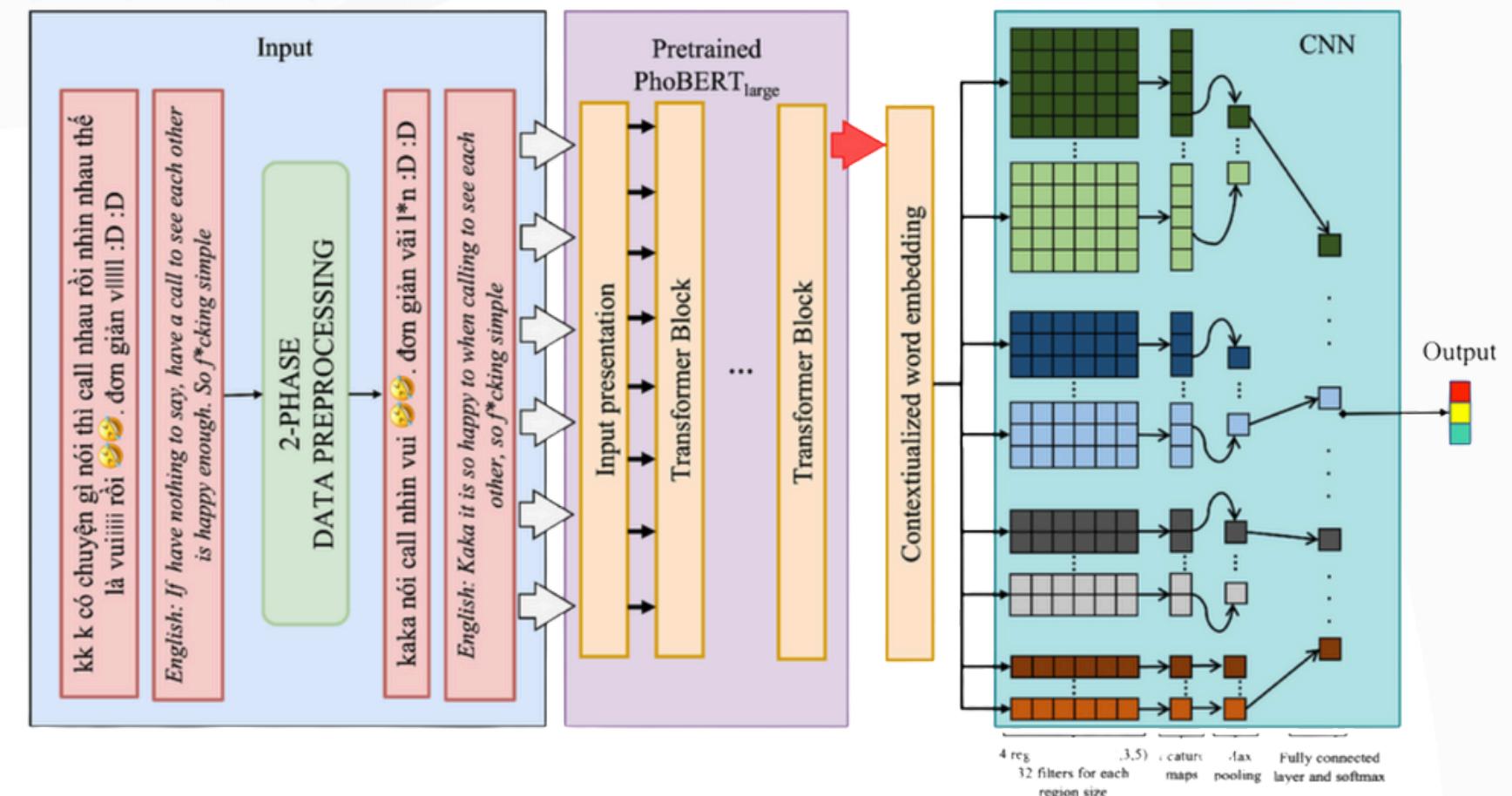
Hệ thống ETNLP (Source: Nguyen & Nguyen, 2020).

# 3.1 Cơ sở lý thuyết

## 3.1.3 PhoBERT

**PhoBERT** là mô hình ngôn ngữ tiếng Việt dựa trên **RoBERTa**, học biểu diễn từ theo **ngữ cảnh**.

- Sử dụng kiến trúc transformer và cơ chế self-attention.
- Embedding của từ phụ thuộc vào vị trí và ngữ nghĩa trong câu.
- Huấn luyện trên Vietnamese Treebank, gồm phiên bản base và large.
- Kết hợp PhoBERT với CNN giúp tăng độ chính xác trong phân loại văn bản (Tran et al., 2022).



Tổng quan mô hình PhoBERT và ứng dụng trong pipeline NLP  
(Source: Tran, K. Q., Nguyen, D. T., Nguyen, T. H., & Bui, T. (2022))

# 3.1 Cơ sở lý thuyết

## 3.1.4 PCA

**PCA (Principal Component Analysis)** là kỹ thuật giảm chiều dữ liệu tuyến tính, giúp chuyển dữ liệu từ không gian nhiều chiều về không gian ít chiều hơn mà vẫn giữ được phần lớn phương sai và đặc trưng quan trọng.

- Các thành phần chính (principal components) được xác định qua **eigenvalues** và **eigenvectors** từ ma trận hiệp phương sai.
- Giúp loại bỏ nhiễu và giảm tương quan tuyến tính giữa các đặc trưng.





## 3.2 Quy trình thực hiện nghiên cứu

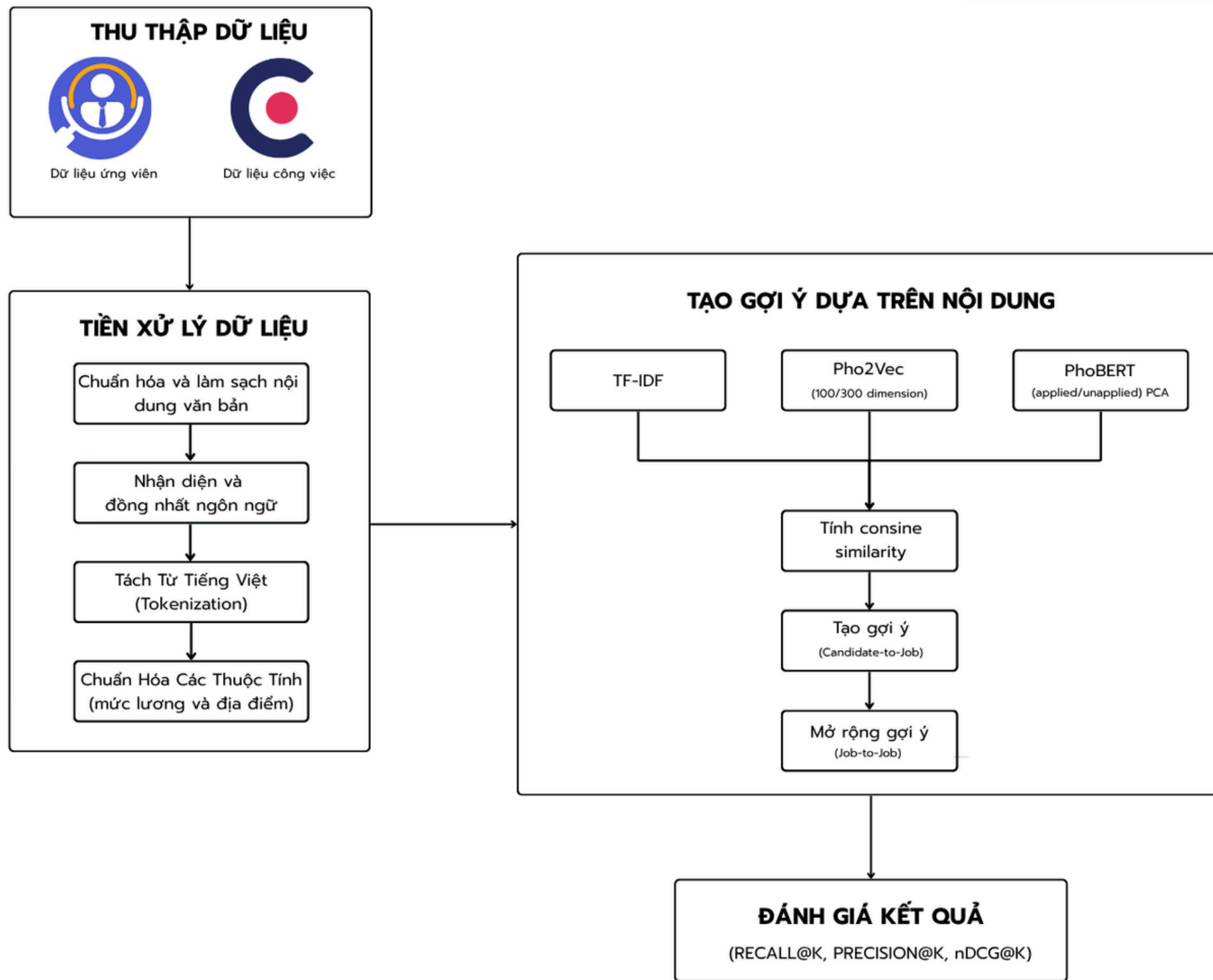
**3.2.1**  
**Thu thập dữ liệu**

**3.2.2**  
**Tiền xử lý dữ liệu**

**3.2.3**  
**Tạo đặc trưng và biểu diễn dữ liệu**

**3.2.4**  
**Tạo gợi ý và đánh giá**

# Mô hình nghiên cứu



### 3.2.1. Thu thập dữ liệu

Bộ dữ liệu được thu thập vào tháng **6/2025**.

Kết quả sau khi thu thập được:

- Dữ liệu công việc:
  - **21.861** hàng dữ liệu
  - 11 thuộc tính
- Dữ liệu ứng viên:
  - **46,124** hàng dữ liệu
  - 17 thuộc tính

Website for Data Crawling

career  
viet

Dữ liệu công việc



Dữ liệu ứng viên



Selenium

Playwright

Playwright

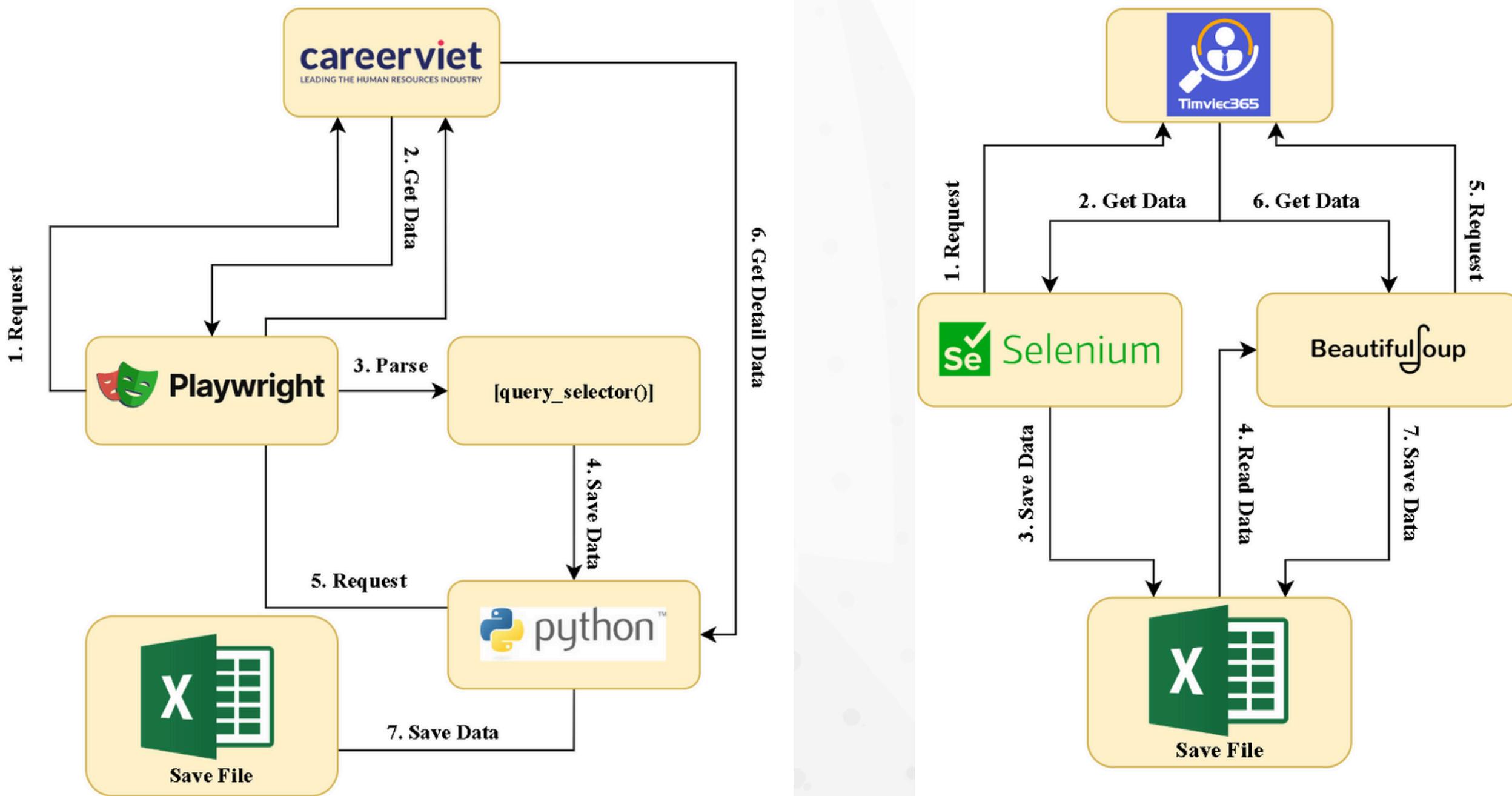
BeautifulSoup

BeautifulSoup

Web Crawling Frameworks

## Sơ Đồ Quy Trình Thu Thập Dữ Liệu

### 3.2.1. Thu thập dữ liệu



## 3.2.2 Tiết kiệm dữ liệu

### Xử lý dữ liệu văn bản

- ✓ Chuẩn Hóa và Làm Sạch Nội Dung Văn Bản
- ✓ Nhận Diện và Đóng Nhất Ngôn Ngữ
- ✓ Tách Từ Tiếng Việt (Tokenization)



## 3.2.2 Tiền xử lý dữ liệu

### Chuẩn hóa mức lương và lương

#### ỨNG VIÊN

desired_salary	work_experience
Thỏa thuận	Chưa có kinh nghiệm
1,000,000 - 5,000,000	Dưới 1 năm
5,000,000 - 10,000,000	1-3 năm
10,000,000 - 15,000,000	3-5 năm
15,000,000 - 20,000,000	5-10 năm
20,000,000 - 40,000,000	Trên 10 năm
Trên 40,000,000	

#### CÔNG VIỆC

Salary
Thỏa thuận / Cạnh tranh
Dưới 5,000,000
5,000,000 - 10,000,000
10,000,000 - 15,000,000
15,000,000 - 20,000,000
20,000,000 - 40,000,000
Trên 40,000,000

### 3.2.3

## Tạo đặc trưng và biểu diễn dữ liệu

Để hệ thống đề xuất có thể xử lý và đưa ra gợi ý phù hợp, tất cả các đặc trưng đều vào từ phía ứng viên và công việc cần được chuẩn hóa và biểu diễn về dạng số học. Quá trình này bao gồm hai hướng chính:

### Bước 1: Xử lý dữ liệu phi văn bản

- Hai trường "work\_experience" và "desired\_salary" được mã hóa bằng Label Encoding
- Dùng Keras Embedding layer để ánh xạ thành vector liên tục
- Ghép với đặc trưng văn bản để tạo đầu vào thống nhất cho mỗi ứng viên.

### 3.2.3

## Tạo đặc trưng và biểu diễn dữ liệu

### Bước 2: Biểu diễn văn bản bằng NLP

Ghép nối các trường văn bản để tạo:

- combined\_text\_candidate
- combined\_text\_job

Ứng dụng 3 kỹ thuật biểu diễn và áp dụng PCA cho kĩ thuật PhoBERT và Pho2Vec, ta được:

Phương pháp	Loại vector	Đầu vào xử lý	Kích thước đầu ra
TF-IDF	Sparse	combined_text_candidate/job	1.000 chiều
Pho2Vec (100d)	Dense	combined_text_candidate/job	100 chiều
Pho2Vec (300d)	Dense	combined_text_candidate/job	300 chiều
PhoBERT + PCA	Dense	combined_text_candidate/job	300 chiều (từ 768)

### 3.2.4

## Tạo gợi ý và đánh giá

- Gợi ý trực tiếp công việc từ hồ sơ ứng viên (candidate-to-job)
- Mở rộng danh sách công việc bằng cách tìm các công việc tương tự (job-to-job expansion) từ các gợi ý ban đầu.

### Sau đó:

Để kiểm tra mức độ chính xác của các đề xuất, nhóm sử dụng một hàm heuristic để gán nhãn ground truth cho từng cặp ứng viên - công việc. Việc gán nhãn dựa trên một tổ hợp các điều kiện thực tế như sự tương thích về ngành nghề, vị trí, địa điểm làm việc, kinh nghiệm và lương.

### Cuối cùng:

Ba chỉ số đánh giá chính được áp dụng là Precision@K, Recall@K và nDCG@K

4

# KẾT QUẢ THỰC NGHIỆM



## 4.1. Các chỉ số đánh giá

Loại chỉ số	Ý nghĩa
Precision@K	Đo lường tỷ lệ các công việc trong Top-K gợi ý thực sự phù hợp với hồ sơ ứng viên
Recall@K	Đo lường tỷ lệ các công việc phù hợp được mô hình gợi ý thành công trong Top-K, so với toàn bộ công việc phù hợp thực tế.
nDCG@K	Đo lường chất lượng xếp hạng trong Top-K gợi ý

## 4.2. Kết quả phân tích dữ liệu

Mô hình	Precision@10	Recall@10	nDCG@10
TF-IDF	0,6144	0,3499	0,7795
Pho2Vec100	0,5406	0,2659	0,7095
Pho2Vec300	0,5748	0,2731	0,7231
PhoBERT	0,9232	0,3548	0,9574
PhoBERT (PCA)	0,9347	0,3413	0,9648

Mô hình **PhoBERT** cho thấy sự vượt trội rõ rệt với Precision@10 đạt 0.9232, Recall@10 là 0.3548 và nDCG@10 lên đến 0.9574 - cao hơn đáng kể so với tất cả các mô hình khác. Khi áp dụng thêm kỹ thuật PCA để giảm chiều vector đầu ra của **PhoBERT (PCA)**, mô hình thậm chí còn đạt được Precision@10 là 0.9347 và nDCG@10 là 0.9648 - **thiết lập mức hiệu quả cao nhất trong toàn bộ bảng kết quả**.

# 5

# KẾT LUẬN



# 5.1 Đóng Góp Chính của Nghiên Cứu



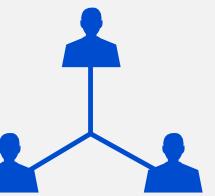
## Mục tiêu nghiên cứu

Xây dựng hệ thống gợi ý việc làm hiệu suất cao dành riêng cho tiếng Việt.



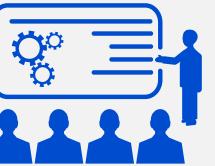
## Biểu diễn và trích xuất đặc trưng

Nghiên cứu sử dụng PhoBERT kết hợp PCA để biểu diễn ngữ nghĩa hiệu quả, đảm bảo độ chính xác cao và tăng tốc độ xử lý.



## Kết quả thực nghiệm nổi bật

PhoBERT + PCA đạt được kết quả vượt trội so với các mô hình khác



## Ý nghĩa và đóng góp

Kết hợp giữa NLP ngữ cảnh (contextual NLP) và Content-Based Filtering là một hướng tiếp cận hiệu quả cho bài toán gợi ý việc làm tiếng Việt.



## 5.2 Hạn Chế của Nghiên Cứu



### Chỉ sử dụng Content-Based Filtering

Mô hình chỉ khai thác nội dung văn bản mà chưa sử dụng dữ liệu tương tác người dùng, khiến không thể áp dụng Collaborative Filtering và làm giảm khả năng phản ánh đúng sở thích hay mục tiêu ngầm của người dùng.



### Chất lượng dữ liệu đầu vào

Hồ sơ kém chất lượng và mất cân bằng dữ liệu ngành nghề khiến mô hình hiểu sai ngữ nghĩa và gợi ý thiên lệch, làm thu hẹp lựa chọn việc làm và giảm trải nghiệm người dùng.

### Recall@K còn thấp

Recall@10 chỉ đạt 35.48%, cho thấy hệ thống còn bỏ sót nhiều công việc tiềm năng, đặc biệt ở các ngành ít phổ biến hoặc JD viết không chuẩn, dẫn đến danh sách gợi ý thiếu đa dạng.



## 5.3 Hướng Phát Triển Tương Lai

### Phát triển mô hình gợi ý lai

Kết hợp Collaborative Filtering và Content-Based Filtering, đồng thời khai thác dữ liệu hành vi người dùng để học sở thích ngầm, từ đó cá nhân hóa gợi ý hiệu quả hơn.

### Tăng khả năng bao phủ bằng hệ thống gợi ý đa tầng

Áp dụng hệ thống gợi ý hai tầng: Tầng 1 dùng mô hình để tạo danh sách mở rộng với Tầng 2 để phân tích sâu và xếp hạng lại, giúp tăng Recall@K mà vẫn đảm bảo độ chính xác.

### Cải thiện chất lượng dữ liệu

Phát triển mô-đun "Hỗ trợ và Làm giàu Hồ sơ" để tự động phát hiện lỗi và gợi ý bổ sung từ khóa, kỹ năng còn thiếu, nhằm nâng cao chất lượng CV, giảm thiên lệch dữ liệu và tăng tính công bằng trong gợi ý.



# Thank You