ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO CUỐI KỲ

MÔN HỌC: PHÂN TÍCH DỮ LIỆU WEB

Hệ thống Gợi ý việc làm tại Việt Nam

Ứng dụng kép Lọc theo Nội dung và Xử lý Ngôn ngữ tự nhiên

GVHD: Tiến Sĩ Nguyễn Thôn Dã

Nhóm: 2

Thành viên nhóm:

Stt	Mã số sinh viên	Họ và tên	Chức danh
1	K224131589	Phạm Thanh Hưng	Thành viên
2	K224111452	Nguyễn Hoàng Kim	Thành viên
3	K224131604	Phạm Trúc Phương	Thành viên
4	K224111469	Hồ Song Tín	Nhóm trưởng
5	K224111475	Huỳnh Huệ Trúc	Thành viên
6	K224111474	Huỳnh Hiếu Trung	Thành viên

LỜI CẢM ƠN CỦA NHÓM

Trước hết, nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến Tiến sĩ, Giảng viên Nguyễn Thôn Dã – người đã tận tình giảng dạy và truyền đạt những kiến thức quý báu về Phân tích Dữ liệu Web trong suốt học kỳ vừa qua. Sự tận tâm và nhiệt huyết của thầy không chỉ giúp chúng em nắm vững các kiến thức chuyên môn, mà còn khơi dậy tinh thần nghiên cứu và tư duy phản biện trong quá trình học tập.

Chúng em xin chân thành cảm ơn thầy Nguyễn Thôn Dã vì đã luôn kiên nhẫn hỗ trợ, giải đáp các thắc mắc, cũng như đưa ra những nhận xét mang tính xây dựng, giúp nhóm hoàn thiện bài báo cáo một cách tốt nhất. Những góp ý chuyên sâu và thực tiễn từ thầy đã giúp chúng em có thêm góc nhìn toàn diện trong việc phân tích dữ liệu và ứng dụng công cụ phù hợp cho bài nghiên cứu.

Một lần nữa, chúng em xin gửi lời tri ân sâu sắc đến thầy Nguyễn Thôn Dã vì đã đồng hành, định hướng và tạo điều kiện thuận lợi cho nhóm trong suốt quá trình thực hiện đề tài. Sự hướng dẫn tận tình của thầy là nguồn động lực to lớn giúp chúng em hoàn thành báo cáo này với sự nghiêm túc, cầu thị và tinh thần học hỏi không ngừng.

Nhóm 2

LỜI CAM KẾT

Chúng em xin cam kết rằng toàn bộ nội dung trong báo cáo này là do chính nhóm chúng em thực hiện, dựa trên kiến thức đã học, quá trình thu thập và phân tích dữ liệu một cách trung thực và khách quan. Mọi tài liệu tham khảo, số liệu và hình ảnh sử dụng trong báo cáo đều được trích dẫn rõ ràng, đầy đủ. Nhóm hoàn toàn chịu trách nhiệm trước nhà trường và giảng viên hướng dẫn nếu có bất kỳ vi phạm nào liên quan đến đạo đức học thuật.

Nhóm 2

MỤC LỤC

DANH MỤC BẢNG

DANH MỤC HÌNH ẢNH

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	**Ý nghĩa**
NLP	Natural Language Processing
TF-IDF	Term Frequency - Inverse Document Frequency
JD	Job Description
CV	Curriculum Vitae
PCA	Principal Component Analysis
CBF	Content-Based Filtering
NER	Named Entity Recognition
ETNLP	Evaluation and Visualization Toolkit for Natural Language Processing
MAP	Mean Average Precision at K
nDCG	Normalized Discounted Cumulative Gain at K
URL	Uniform Resource Locator

PHẦN 1. NỘI DUNG ĐỒ ÁN

Tóm tắt

Thị trường lao động Việt Nam hiện tồn tại khoảng cách lớn giữa người tìm việc và nhà tuyển dụng do lỗ hổng ngôn ngữ nghề nghiệp, định kiến hệ thống và sự mất cân đối thông tin. Tình trạng này đặt ra yêu cầu cấp thiết về một hệ thống gợi ý việc làm thông minh và phù hợp với ngữ cảnh tiếng Việt. Báo cáo này trình bày một hệ thống đề xuất việc làm dựa trên phương pháp Lọc theo nội dung (Content-Based Filtering) kết hợp với các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) tiên tiến như TF-IDF, Pho2Vec và PhoBERT. Tập dữ liệu gồm hơn 46.000 hồ sơ ứng viên và 21.000 tin tuyển dụng được thu thập từ các nền tảng tuyển dụng lớn tại Việt Nam. Sau quá trình tiền xử lý và biểu diễn dữ liệu, quá trình thực nghiệm cho thấy mô hình PhoBERT kết hợp PCA đạt hiệu suất vượt trội với **Precision@30 lên tới 93%**, **nDCG@30 đạt 97%** và **MAP@30 đạt 94%**. Kết quả nghiên cứu thể hiện tiềm năng áp dụng thực tiễn cao, giúp tăng cường kết nối cung-cầu lao động và nâng cao hiệu quả thị trường lao động Việt Nam.

1. Giới thiệu đề tài

1.1. Bối cảnh và vấn đề cần nghiên cứu

Thị trường lao động Việt Nam hiện đang đối mặt với một nghịch lý đáng lo ngại: dù có vô vàn thông tin tuyển dụng, nhưng cả người tìm việc và nhà tuyển dụng đều đang lạc lối trong hệ sinh thái của chính họ.Theo báo cáo Tổng cục Thống kê (Quý 1/2024)**:** Số liệu từ Tổng cục Thống kê thường xuyên chỉ ra rằng, dù tỷ lệ thất nghiệp chung có thể thấp, nhưng tỷ lệ thiếu việc làm (làm việc dưới 35 giờ/tuần hoặc không tối ưu năng lực) vẫn còn đáng kể. Đồng thời, nhiều doanh nghiệp vẫn phản ánh khó khăn trong việc tìm kiếm nhân sự chất lượng cao hoặc lao động có kỹ năng chuyên biệt. Sự bùng nổ của các nền tảng số đã khiến dữ liệu trở nên hỗn loạn, và điều này vô tình làm gia tăng khoảng cách về sự thấu hiểu thực chất giữa hai bên. Thực trạng này được thể hiện rõ qua ba "vết nứt vô hình" đang chia cắt thị trường lao động.

Lỗ hổng ngôn ngữ nghề nghiệp là một trong những rảo cản lớn nhất. Ứng viên, đặc biệt là những người trẻ, thường chuẩn bị hồ sơ xin việc (CV) một cách sơ sài, thiếu đi sự trau chuốt và ngôn ngữ chuyên môn cần thiết. Trong khi đó, bản mô tả công việc của các doanh nghiệp lại thường xuyên sử dụng những thuật ngữ chuyên ngành phức tạp, đôi khi khó hiểu đối với những người chưa có kinh nghiệm. Minh chứng là, nhiều trang tuyển dụng như TopCV, CareerBuilder, VietnamWorks thường xuyên đưa ra lời khuyên về cách viết CV hiệu quả, nhấn mạnh việc tránh chung chung, liệt kê và thay vào đó là định lượng hóa thành tích, sử dụng ngôn ngữ chuyên nghiệp. Điều này ngụ ý rằng các ứng viên thường có tình trạng viết CV sơ sài, thiếu thông tin cụ thể là phổ biến. Ngược lại, các hướng dẫn cho nhà tuyển dụng cũng thường khuyến nghị viết JD rõ ràng, tránh thuật ngữ quá chuyên môn nếu không cần thiết để tiếp cận được nhiều ứng viên hơn.. Điều này tạo ra một "cuộc đối thoại giữa hai thế giới song song", nơi thông tin không thể được truyền tải và thấu hiểu một cách hiệu quả, dẫn đến việc bỏ lỡ nhiều cơ hội phù hợp cho cả hai phía.

Thêm vào đó, định kiến hệ thống đang loại trừ các nhóm lao động phi truyền thống. Các thuật toán tuyển dụng truyền thống thường được thiết kế dựa trên các tiêu chí và khuôn mẫu cố định, vô tình "vô hình hóa" những đối tượng đặc biệt như người chuyển giới hay lao động trên 50 tuổi. Họ có thể sở hữu kỹ năng, kinh nghiệm và sự tận tâm đáng quý, nhưng lại gặp rào cản từ chính hệ thống tuyển dụng thiếu linh hoạt và không có khả năng nhận diện giá trị tiềm ẩn của họ. Điều này không chỉ gây thiệt thòi cho cá nhân mà còn làm mất đi một nguồn lực lao động tiềm năng cho nền kinh tế. Và nó cũng vô tình vi phạm quy định cấm phân biệt đối xử trong Bộ luật Lao động (Điều 8: Các hành vi bị nghiêm cấm trong lĩnh vực lao động) và Luật Bình đẳng giới (năm 2006) ngụ ý rằng các hành vi phân biệt đối xử dựa trên giới tính, độ tuổi, tôn giáo, v.v., là có thực và cần được ngăn chặn.

Theo báo cáo tình hình nhân lực tại các khu công nghiệp, thị trường lao động đang tồn tại một bản đồ cơ hội méo mó. Các khu công nghiệp trọng điểm như Bình Dương, Hải Phòng, Bắc Ninh, nơi đang có nhu cầu lớn về nhân lực và cơ hội việc làm phong phú, lại thường xuyên rơi vào tình trạng "đói nhân tài". Ngược lại, các thành phố lớn lại đang tràn ngập những ứng viên thất nghiệp hoặc làm những công việc không đúng với chuyên môn, thiếu đi chất lượng và sự ổn định. Sự mất cân bằng này cho thấy sự thiếu kết nối giữa cung và cầu lao động, cũng như sự thiếu hụt thông tin về cơ hội việc làm chất lượng tại các khu vực đang phát triển.

Những "vết nứt" này không chỉ đơn thuần là thất bại về mặt công nghệ hay hệ thống thông tin. Thay vào đó, chúng phản ánh một sự sụp đổ của niềm tin trong một hệ thống thiếu cơ chế giải mã những tín hiệu nghề nghiệp tinh tế. Để giải quyết nghịch lý này, cần có những giải pháp mang tính chiến lược, không chỉ tập trung vào việc cải thiện công nghệ mà còn phải chú trọng đến việc xây dựng lại cầu nối giao tiếp, phá bỏ định kiến và cung cấp cái nhìn toàn diện hơn về thị trường lao động. Liệu việc phát triển các nền tảng thông minh hơn có thể giúp giải quyết triệt để những vấn đề này, hay chúng ta cần một sự thay đổi sâu rộng hơn về tư duy và chính sách? Vì vậy, nhóm chúng tôi quyết định nghiên cứu để giải quyết vấn đề này bằng cách xây dựng lại hệ thống quản lý việc làm cho ứng viên lẫn doanh nghiệp.

1.2. Mục tiêu và tầm quan trọng của nghiên cứu

1.2.1. Mục tiêu nghiên cứu

Nghiên cứu này tập trung vào việc giải quyết bài toán nâng cao độ chính xác trong gợi ý việc làm, đặc biệt là trong bối cảnh thị trường lao động Việt Nam với những đặc thù riêng về ngôn ngữ và dữ liệu. Để đạt được điều đó, đề tài đặt ra các mục tiêu cụ thể như sau:

Mục tiêu tổng quát: Xây dựng và đánh giá một mô hình gợi ý việc làm ứng dụng phương pháp Lọc dựa trên nội dung (Content-Based Filtering) có tích hợp các kỹ thuật Xử lý Ngôn ngữ Tự nhiên (NLP) tiên tiến nhằm cải thiện độ chính xác và mức độ phù hợp của các gợi ý so với các hệ thống truyền thống.

Các mục tiêu cụ thể:

1. Nghiên cứu và phân tích: Nghiên cứu các phương pháp Lọc dựa trên nội dung, các kỹ thuật NLP phù hợp cho tiếng Việt (như PhoBERT, Pho2Vec), và phân tích đặc điểm của dữ liệu mô tả công việc (Job Description - JD) và hồ sơ ứng viên tại Việt Nam.
2. Xây dựng mô hình xử lý dữ liệu: Phát triển các module có khả năng tiền xử lý, trích xuất đặc trưng và vector hóa thông tin từ văn bản tiếng Việt chuyển đổi dữ liệu phi cấu trúc thành dạng vector có cấu trúc mà máy tính có thể hiểu và so sánh được.
3. Thiết kế và triển khai thuật toán gợi ý: Xây dựng thuật toán Lọc dựa trên nội dung, sử dụng các vector đặc trưng đã được trích xuất để tính toán độ tương đồng giữa hồ sơ của ứng viên và các mô tả công việc, từ đó đưa ra danh sách gợi ý được xếp hạng.
4. Kiểm thử và đánh giá: Thực nghiệm mô hình trên tập dữ liệu thực tế từ các nền tảng tuyển dụng tại Việt Nam. Xây dựng các tiêu chí đánh giá để đo lường hiệu quả và so sánh độ chính xác của mô hình đề xuất với các phương pháp cơ bản khác

1.2.2. Tầm quan trọng của nghiên cứu

Nghiên cứu được thực hiện không chỉ mang ý nghĩa về mặt học thuật mà còn có giá trị ứng dụng thực tiễn sâu sắc, tác động đến nhiều đối tượng trong thị trường lao động.

Về phương diện khoa học, đề tài góp phần giải quyết thách thức trong việc áp dụng Xử lý Ngôn ngữ Tự nhiên cho tiếng Việt – một ngôn ngữ có đặc thù thanh điệu và cấu trúc phức tạp – vào lĩnh vực hệ thống gợi ý. Kết quả nghiên cứu sẽ cung cấp một mô hình lai (hybrid) hiệu quả, kết hợp giữa Lọc dựa trên nội dung và NLP sâu, làm phong phú thêm cơ sở tri thức về việc xây dựng các hệ thống gợi ý thông minh trong ngành nhân sự. Hơn nữa, việc đánh giá thực nghiệm trên dữ liệu thực tế tại Việt Nam sẽ mang lại những số liệu tham khảo giá trị cho các nghiên cứu liên quan trong tương lai.

Về mặt thực tiễn, giải pháp đề xuất hứa hẹn mang lại lợi ích toàn diện. Đối với người tìm việc, hệ thống giúp họ nhanh chóng tiếp cận những cơ hội phù hợp nhất, tiết kiệm thời gian và công sức trước sự quá tải thông tin tuyển dụng. Đối với nhà tuyển dụng, công cụ này là một bộ lọc hiệu quả để tìm ra các ứng viên tiềm năng, tối ưu hóa chi phí và quy trình. Đồng thời, các nền tảng việc làm tích hợp một hệ thống gợi ý chính xác cao sẽ nâng cao được trải nghiệm người dùng, từ đó tạo ra lợi thế cạnh tranh bền vững.

Trên bình diện xã hội rộng hơn, nghiên cứu góp phần thúc đẩy sự vận hành hiệu quả của thị trường lao động, giúp kết nối tối ưu giữa cung và cầu. Điều này không chỉ hỗ trợ giải quyết bài toán việc làm mà còn thúc đẩy sự phát triển nguồn nhân lực chất lượng cao cho đất nước.

1.3. Câu hỏi nghiên cứu

Để định hướng cho việc đạt được các mục tiêu đã đề ra, nghiên cứu sẽ tập trung trả lời câu hỏi chính sau: “Làm thế nào để ứng dụng kết hợp Lọc dựa trên Nội dung và

Xử lý Ngôn ngữ Tự nhiên có thể nâng cao một cách hiệu quả độ chính xác của hệ thống gợi ý việc làm trong bối cảnh đặc thù của thị trường Việt Nam?”

Để làm rõ câu hỏi chính, nghiên cứu sẽ giải quyết các câu hỏi phụ cụ thể sau:

- Đâu là phương pháp hiệu quả nhất để biểu diễn và trích xuất đặc trưng ngữ nghĩa từ dữ liệu văn bản tiếng Việt không cấu trúc (hồ sơ ứng viên và mô tả công việc) nhằm phục vụ cho bài toán so khớp?
- Mức độ cải thiện về độ chính xác của mô hình Lọc dựa trên Nội dung khi tích hợp các kỹ thuật nhúng từ sâu (deep embeddings) từ NLP so với các phương pháp truyền thống (như TF-IDF hay so khớp từ khóa) là như thế nào?
- Hiệu suất của mô hình đề xuất, khi được đánh giá trên tập dữ liệu thực tế tại Việt Nam thông qua các chỉ số đo lường chuẩn (Precision@K, Recall@K, nDCG@K), có đáp ứng được yêu cầu về một hệ thống gợi ý chất lượng cao hay không?

2. Tổng quan về bộ dữ liệu

2.1. Nguồn gốc bộ dữ liệu

Bộ dữ liệu sử dụng trong nghiên cứu bao gồm hai phần: dữ liệu hồ sơ ứng viên và dữ liệu tin tuyển dụng. Hai tập dữ liệu này được nhóm trực tiếp thu thập thông qua quá trình xây dựng trình thu thập dữ liệu từ các nền tảng tuyển dụng phổ biến tại 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. Dữ liệu được thu thập bằng script Python sử dụng các thư viện như requests, BeautifulSoup, và được lưu dưới dạng .csv để phục vụ xử lý tiếp theo. Kết quả thu được bộ dữ liệu với tổng cộng 46.124 hồ sơ ứng viên, Mỗi bản ghi ứng viên chứa thông tin cá nhân, kỳ vọng nghề nghiệp và trình độ chuyên môn như tên ứng viên, kinh nghiệm làm việc, công việc mong muốn, nơi làm việc mong muốn, kỹ năng, mục tiêu nghề nghiệp, tình trạng hôn nhân, giới tính, độ tuổi và mức lương kỳ vọng.

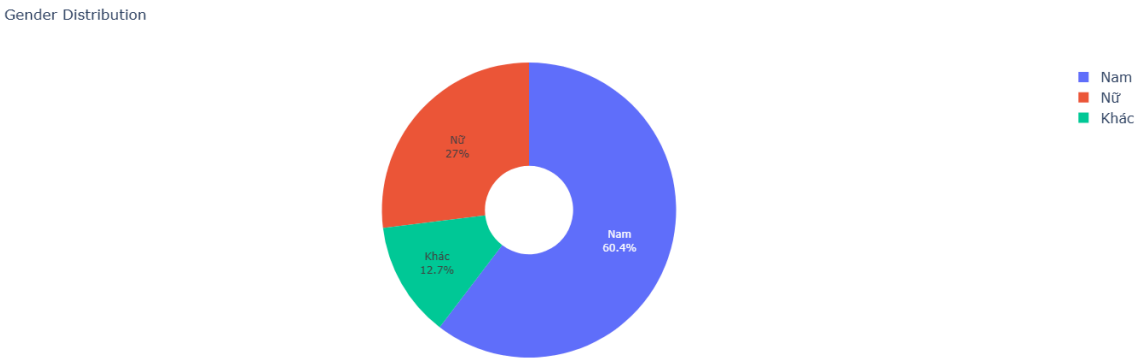
Và, dữ liệu tin tuyển dụng được nhóm crawler từ trang *CareerViet.vn* – một phiên bản nội địa hóa của CareerBuilder Việt Nam. Đây là một trong những nền tảng tuyển dụng lớn, chuyên cung cấp các vị trí việc làm có chất lượng, đặc biệt trong lĩnh vực văn phòng, kỹ thuật, công nghệ và quản lý. Từ trang danh sách công việc của CareerViet.vn, nhóm xây dựng một crawler có khả năng phân trang, thu thập liên kết từng tin tuyển dụng, và trích xuất các thông tin chi tiết như: tên công việc, tên công ty, địa điểm làm việc, mức lương, loại hình công việc, mô tả công việc, yêu cầu kinh nghiệm, kỹ năng cần thiết và thời gian đăng tuyển.

Để tiến hành thu thập dữ liệu, nhóm sử dụng ngôn ngữ lập trình Python, kết hợp với ba công cụ hỗ trợ chính là Selenium, BeautifulSoup và Playwright. Cụ thể, với dữ liệu công việc, nhóm xây dựng quy trình duyệt và trích xuất thông tin từ hàng chục nghìn trang tuyển dụng trên CareerViet.vn, thu về tổng cộng 21.861 mẫu tin, mỗi tin tuyển dụng bao gồm các trường quan trọng như tên công việc, ngành nghề, mức lương, địa điểm làm việc, yêu cầu công việc, mô tả chi tiết, phúc lợi, thời gian ứng tuyển và loại hình công việc.

Dữ liệu được thu thập trong khoảng thời gian từ tháng 6 năm 2025, với mục đích hoàn toàn phục vụ cho học thuật và nghiên cứu trong khuôn khổ môn học Phân tích dữ liệu web tại trường Đại học Kinh tế - Luật. Tập dữ liệu sau khi thu thập được xử lý sơ bộ để loại bỏ các bản ghi trùng lặp, bản ghi thiếu thông tin cần thiết và chuẩn hóa định dạng để sẵn sàng cho các bước tiền xử lý văn bản và mô hình hóa sau này.

Việc sử dụng hai nguồn dữ liệu từ hai phía khác nhau – ứng viên và nhà tuyển dụng – cho phép nhóm nghiên cứu xây dựng một hệ thống gợi ý việc làm theo hướng lọc theo nội dung (Content-Based Filtering) với tính cá nhân hóa cao. Thông tin từ hồ sơ ứng viên được dùng để xác định mức độ tương đồng với từng công việc, từ đó đưa ra các gợi ý sát thực và đa dạng hơn, phản ánh đúng nhu cầu và năng lực của người lao động trong thực tế. Bộ dữ liệu này không chỉ mang giá trị học thuật mà còn có tiềm năng cao trong ứng dụng thực tiễn nếu được cập nhật và mở rộng thêm trong tương lai.

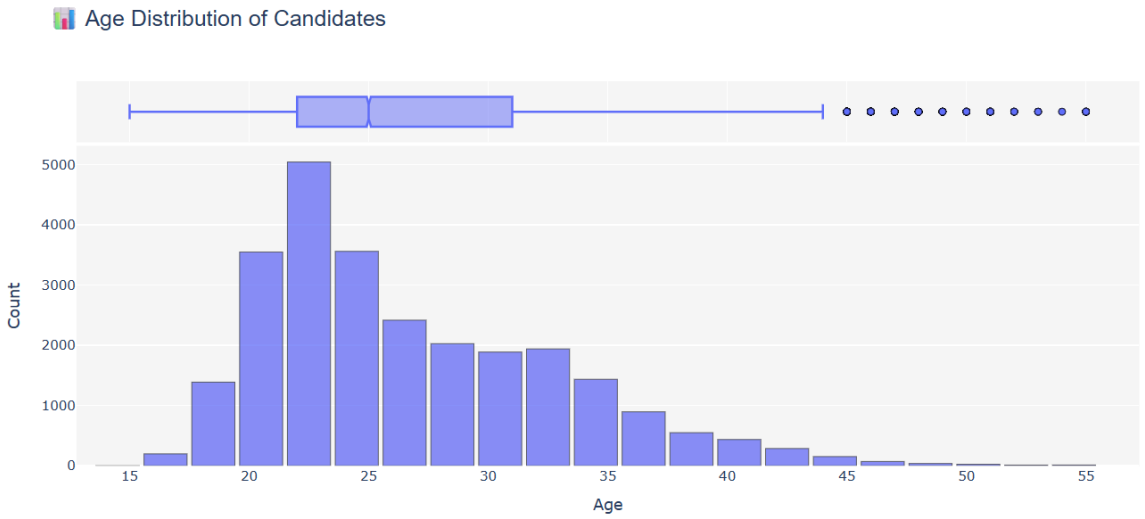
2.2. Bộ dữ liệu ứng viên



Hình 2.1 : Biểu đồ phân bố giới tính

Biểu đồ tròn cho thấy sự phân bố giới tính của các ứng viên trong tập dữ liệu hiện tại. Tỷ lệ ứng viên nam chiếm ưu thế rõ rệt với 60.4%, trong khi nữ chỉ chiếm 27%. Điều này cho thấy lực lượng ứng viên có xu hướng thiên về nam giới, có thể là do đặc thù lĩnh vực, ngành nghề, hoặc do xu hướng thị trường lao động vẫn đang phần nào thiên lệch về giới trong tuyển dụng.

Bên cạnh hai giới tính chính, biểu đồ cũng ghi nhận tỷ lệ nhỏ các giới tính khác. Nhóm được phân loại là "Khác" chiếm khoảng 12.4%. Điều này phản ánh một mức độ đa dạng nhất định trong dữ liệu, cho thấy xã hội và thị trường lao động đang dần ghi nhận và tạo cơ hội cho các nhóm không thuộc hai giới tính truyền thống.

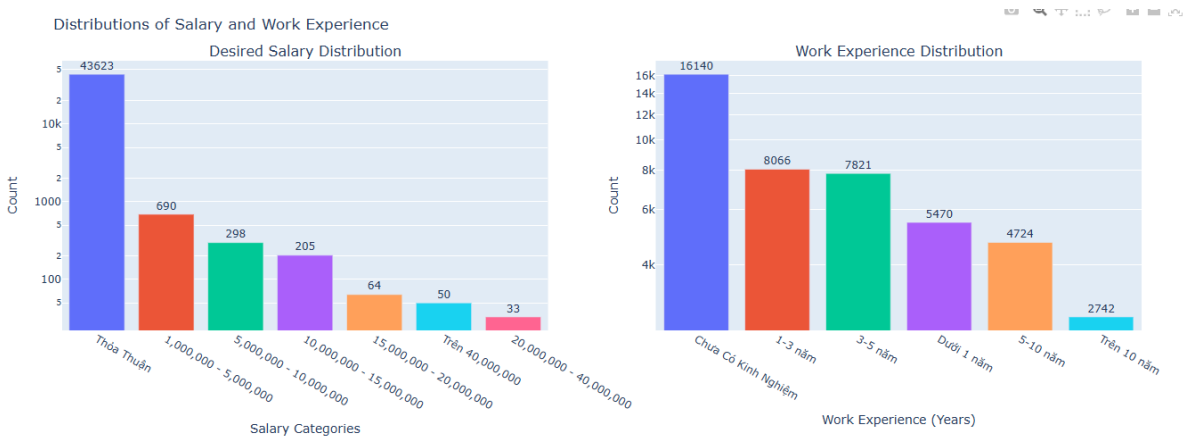


Hình 2.2 : Biểu đồ phân tán tuổi ứng viên

Biểu đồ kết hợp giữa histogram và boxplot cho thấy số lượng ứng viên tập trung chủ yếu ở độ tuổi từ 20 đến 26, trong đó đỉnh của phân phối rơi vào khoảng 23 tuổi. Đây là lứa tuổi điển hình của sinh viên mới tốt nghiệp hoặc vừa bắt đầu đi làm trong một vài năm đầu tiên.

Phần boxplot phía trên histogram cung cấp thêm thông tin về sự phân tán và độ lệch của dữ liệu. Đa số ứng viên có độ tuổi nằm trong khoảng từ đầu 20 đến khoảng 30 tuổi. Phía ngoài khoảng tứ phân vị xuất hiện nhiều outliers – tức là những ứng viên có độ tuổi cao hơn đáng kể, lên đến 55 tuổi. Tuy nhiên, số lượng những ứng viên này rất nhỏ, cho thấy đây chỉ là thiểu số.

Dạng phân phối này là đặc trưng của thị trường lao động trẻ, phần lớn các ứng viên đang trong giai đoạn đầu sự nghiệp, hoặc thậm chí mới tiếp cận thị trường lao động lần đầu tiên.

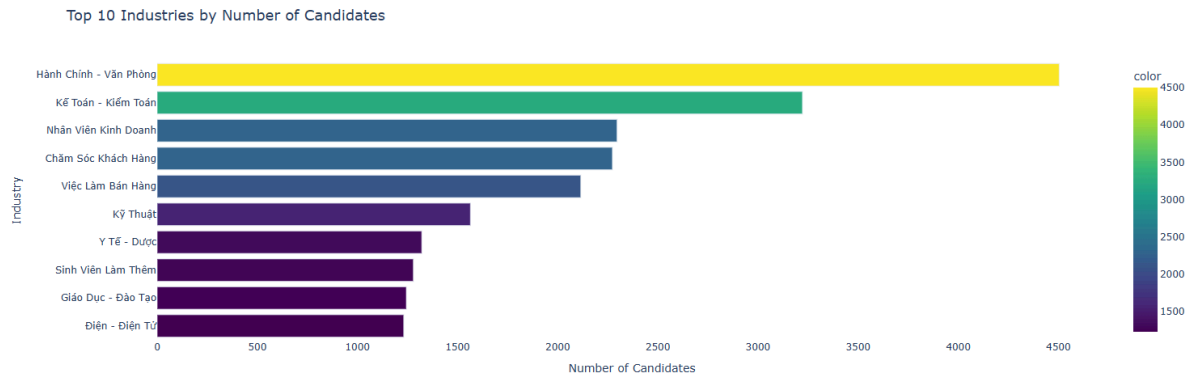


Hình 2.3: Các biểu đồ phân bố lương và kinh nghiệm làm việc ứng viên

Phía bên trái của biểu đồ là phân bố mức lương kỳ vọng của ứng viên. Một điểm nổi bật rất đáng chú ý là số lượng ứng viên chọn "Thỏa thuận" gần như áp đảo hoàn toàn, với hơn 43,000 người. Điều này phản ánh rằng đa phần ứng viên không nêu rõ mức lương cụ thể mà họ mong muốn. Có thể lý giải hiện tượng này là do ứng viên chưa có đủ thông tin để định giá bản thân, hoặc muốn linh hoạt để thương lượng với nhà tuyển dụng.

Mức lương cụ thể được kỳ vọng chủ yếu nằm ở khoảng từ 1 đến 10 triệu đồng, với số lượng giảm dần ở các nhóm lương cao hơn. Số ứng viên kỳ vọng lương từ 15 triệu đồng trở lên rất ít, đặc biệt là nhóm trên 40 triệu gần như không đáng kể. Điều này khớp với đặc điểm của một lực lượng lao động trẻ, ít kinh nghiệm, và phù hợp với các vị trí đầu vào hoặc chưa quản lý.

Phía bên phải là biểu đồ kinh nghiệm làm việc. Số liệu này cho thấy gần 16,140 ứng viên chưa có kinh nghiệm. Đây là nhóm đông nhất trong toàn bộ dữ liệu. Hai nhóm tiếp theo là ứng viên có từ 1 đến 3 năm và từ 3 đến 5 năm kinh nghiệm, lần lượt chiếm khoảng 8,000 người mỗi nhóm. Những ứng viên có từ 5 đến 10 năm và trên 10 năm kinh nghiệm ít hơn đáng kể.

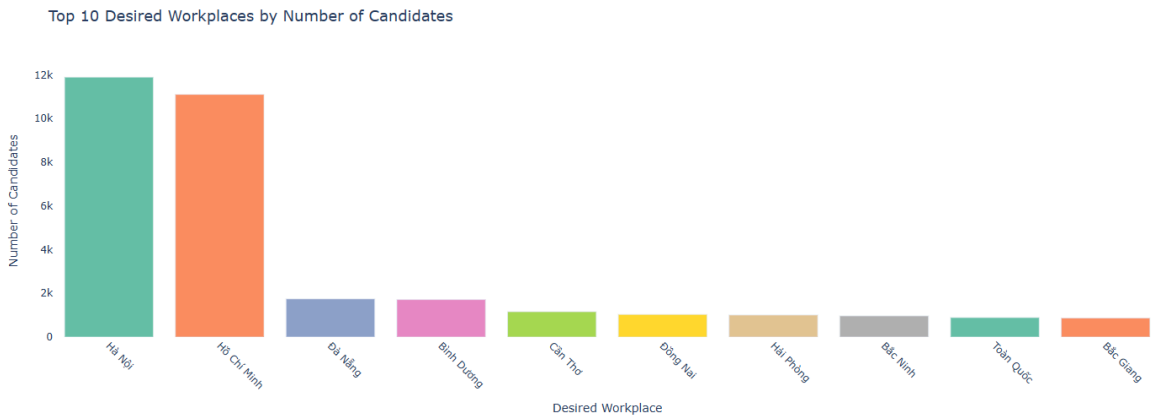


Hình 2.4: Biểu đồ phân bố số lượng ứng viên cho các ngành nghề

Biểu đồ thanh ngang thể hiện top 10 ngành nghề có số lượng ứng viên nhiều nhất. Ngành "Hành chính – Văn phòng" đứng đầu, với số lượng gần 4,500 ứng viên. Đây là lĩnh vực phổ thông, yêu cầu kỹ năng mềm cơ bản, phù hợp với sinh viên mới ra trường và ứng viên chưa có kinh nghiệm.

Xếp sau đó là các ngành như "Kế toán – Kiểm toán", "Nhân viên kinh doanh" và "Chăm sóc khách hàng". Đây đều là những ngành có nhu cầu tuyển dụng cao trên thị trường, thường xuyên tuyển số lượng lớn và yêu cầu đầu vào không quá khắt khe về kinh nghiệm hoặc chuyên môn cao.

Những ngành như "Y tế – Dược", "Kỹ thuật", "Giáo dục – Đào tạo", và "Điện – Điện tử" có số lượng ứng viên thấp hơn đáng kể. Lý do có thể là vì những ngành này yêu cầu bằng cấp chuyên môn hoặc kỹ năng kỹ thuật cao, nên không phải ứng viên nào cũng có thể tham gia. Ngoài ra, sinh viên khối kỹ thuật thường ít hơn các ngành kinh tế – xã hội, dẫn đến số lượng hồ sơ cũng giảm theo.

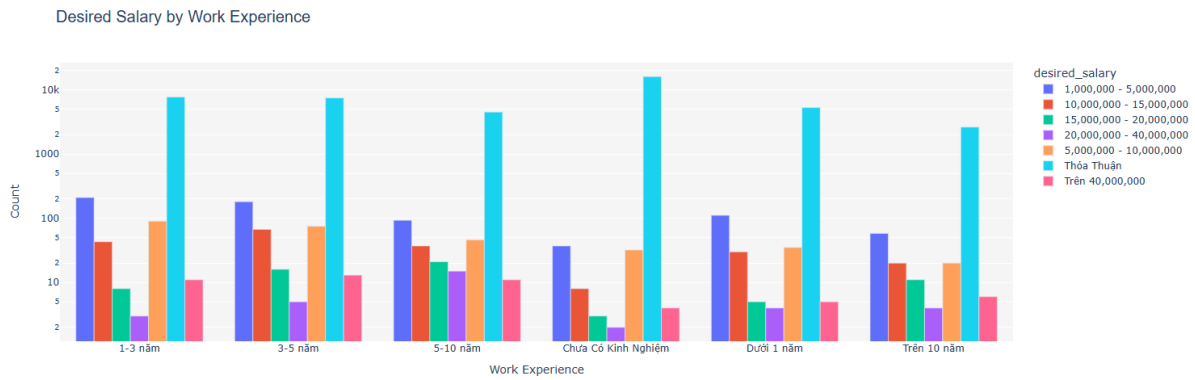


Hình 2.5: Biểu đồ top 10 nơi làm việc được các ứng viên yêu thích

Hai địa điểm có số lượng ứng viên mong muốn làm việc nhiều nhất là **Hà Nội** và **TP. Hồ Chí Minh**, với số lượng lần lượt là gần 12,000 và 11,000 người. Đây là hai trung tâm kinh tế lớn nhất cả nước, nơi tập trung nhiều cơ hội việc làm, mức thu nhập cao và hệ thống doanh nghiệp phong phú. Không quá bất ngờ khi đây là lựa chọn hàng đầu của ứng viên.

Từ vị trí thứ ba trở đi, số lượng ứng viên giảm mạnh. Các tỉnh như Đà Nẵng, Bình Dương, Cần Thơ, Đồng Nai… mỗi nơi chỉ thu hút vài nghìn ứng viên. Những địa phương như Hải Phòng, Bắc Ninh, Bắc Giang – dù có khu công nghiệp lớn – nhưng lại không phải ưu tiên hàng đầu với ứng viên. Ngoài ra, lựa chọn "Toàn Quốc" – tức là sẵn sàng làm việc ở nhiều nơi – cũng chiếm tỷ lệ nhất định nhưng không đáng kể.

Dữ liệu này cho thấy sự mất cân đối trong định hướng tìm việc của ứng viên. Thị trường lao động hiện vẫn tập trung quá nhiều vào các thành phố lớn, trong khi cơ hội việc làm đang dần mở rộng ra các khu vực vệ tinh và tỉnh thành khác.



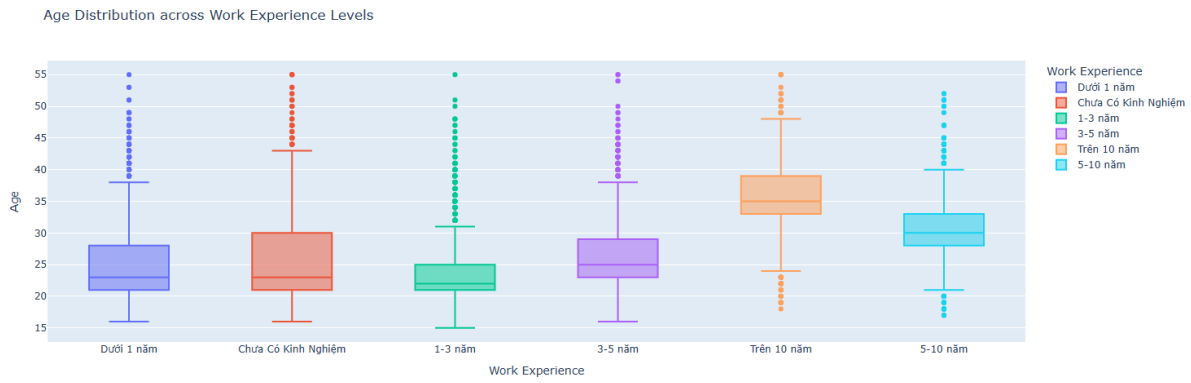
Hình 2.6: Biểu đồ phân bố kinh nghiệm làm việc ứng viên theo mức lương mong muốn

Biểu đồ này thể hiện mối liên hệ giữa mức kinh nghiệm và kỳ vọng lương của ứng viên. Nhìn tổng thể, có thể nhận thấy rằng mức lương được nhiều người lựa chọn nhất ở mọi cấp độ kinh nghiệm chính là "Thỏa thuận". Điều này cho thấy một bộ phận lớn ứng viên chưa xác định rõ được mức lương kỳ vọng hoặc muốn để ngỏ để thương lượng với nhà tuyển dụng.

Đối với nhóm chưa có kinh nghiệm và dưới 1 năm, lựa chọn "Thỏa thuận" chiếm ưu thế rõ rệt, cho thấy sự dè dặt và linh hoạt từ phía ứng viên mới ra trường. Mức lương từ 5 đến 10 triệu đồng là lựa chọn phổ biến thứ hai, phù hợp với thị trường tuyển dụng cho người mới bắt đầu.

Khi ứng viên có từ 1 đến 5 năm kinh nghiệm, xu hướng lựa chọn mức lương cao hơn bắt đầu xuất hiện. Tuy nhiên, "Thỏa thuận" vẫn là lựa chọn phổ biến nhất, điều này cho thấy nhiều người có kinh nghiệm cũng chưa đặt ra một mức lương cụ thể, có thể do thị trường biến động hoặc mong muốn đánh giá mức độ phù hợp sau phỏng vấn.

Ở nhóm có từ 5 đến 10 năm kinh nghiệm và trên 10 năm, mức lương kỳ vọng bắt đầu đa dạng và có phần dịch chuyển lên các khoảng cao hơn như 15 – 20 triệu và trên 40 triệu. Tuy nhiên, số người chọn các mức này vẫn thấp so với những người chọn "Thỏa thuận" hoặc mức trung bình. Có thể hiểu rằng, mặc dù có kinh nghiệm lâu năm, nhiều người vẫn giữ sự linh hoạt trong đàm phán, hoặc thị trường vẫn chưa đáp ứng tương xứng với kỳ vọng của họ.



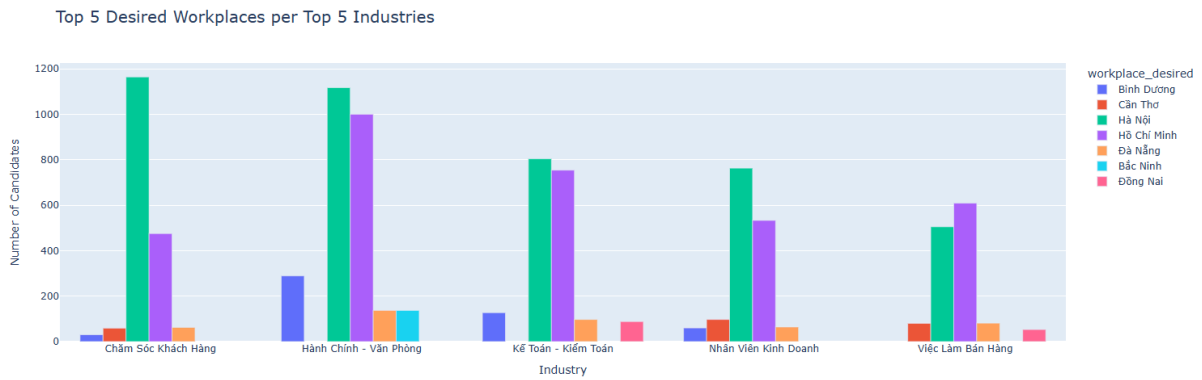
Hình 2.7: Biểu đồ phân bố kinh nghiệm làm việc theo độ tuổi các ứng viên

Biểu đồ này mô tả số lượng ứng viên mong muốn làm việc tại các thành phố lớn đối với năm ngành nghề phổ biến nhất. Có thể thấy, Hà Nội và TP. Hồ Chí Minh là hai địa phương được lựa chọn nhiều nhất ở tất cả các ngành. Trong ngành Chăm sóc khách hàng và Hành chính - Văn phòng, TP. Hồ Chí Minh chiếm ưu thế, theo sau là Hà Nội. Điều này phản ánh sự tập trung cơ hội việc làm ở các đô thị lớn, đồng thời thể hiện xu hướng dịch chuyển lao động về các trung tâm kinh tế.

Riêng ngành Hành chính - Văn phòng có một điểm đáng chú ý: ngoài Hà Nội và TP.HCM, Bình Dương cũng ghi nhận một lượng lớn ứng viên mong muốn làm việc. Điều này có thể cho thấy sự phát triển mạnh về hành chính – văn phòng ở các tỉnh công nghiệp vệ tinh, nơi có nhiều khu công nghiệp và văn phòng hành chính.

Ngành Kế toán – Kiểm toán và Nhân viên Kinh doanh cũng cho thấy sự quan tâm lớn đến hai thành phố lớn, song có sự phân tán nhẹ sang các địa phương khác như Đà Nẵng, Bắc Ninh và Đồng Nai. Ở ngành Việc làm Bán hàng, tuy số lượng thấp hơn, TP. Hồ Chí Minh vẫn là nơi có sức hút mạnh nhất.

Nhìn chung, kết quả biểu đồ phản ánh thực tế rằng các ứng viên có xu hướng lựa chọn nơi làm việc tại các trung tâm kinh tế lớn, nơi có nhiều cơ hội phát triển nghề nghiệp và mức lương cao hơn.



Hình 2.8: Biểu đồ phân bố top 5 nơi làm việc đối với 5 ngành nghề nổi bật nhất

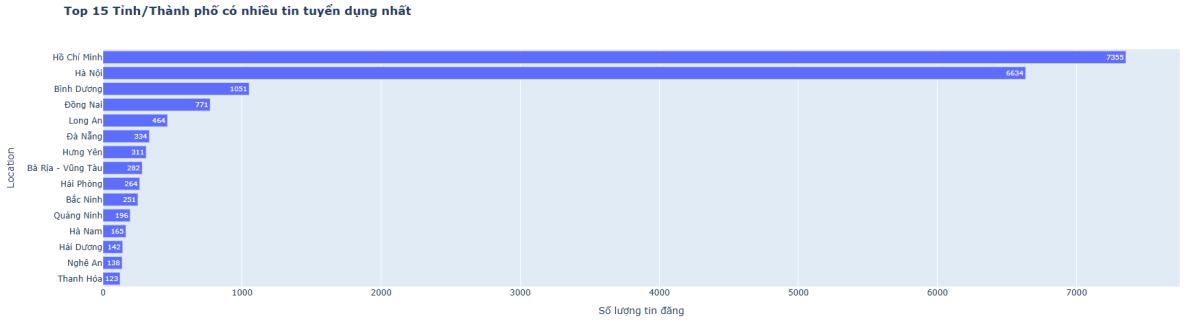
Biểu đồ này là dạng boxplot mô tả sự phân bố độ tuổi của ứng viên theo từng nhóm kinh nghiệm. Ở nhóm chưa có kinh nghiệm hoặc dưới 1 năm, độ tuổi chủ yếu dao động từ 18 đến khoảng 27 tuổi, phần lớn là sinh viên mới tốt nghiệp hoặc mới đi làm. Tuy nhiên, cũng có một số ít ứng viên ở độ tuổi ngoài 30 nhưng vẫn chưa có kinh nghiệm hoặc mới bắt đầu làm việc, có thể là người chuyển ngành hoặc từng tạm dừng công việc.

Những người có từ 1 đến 5 năm kinh nghiệm thường nằm trong khoảng 22 đến 32 tuổi, với median (giá trị trung vị) tăng dần theo kinh nghiệm. Điều này phù hợp với thực tế vì đa số người ra trường khoảng 22 tuổi và bắt đầu tích lũy kinh nghiệm từ đó.

Nhóm có từ 5 đến 10 năm kinh nghiệm có độ tuổi dao động rộng hơn, phổ biến trong khoảng 27 đến 40 tuổi. Một số người có thể đã có sự ổn định trong công việc hoặc đang tìm kiếm cơ hội phát triển mới.

Đặc biệt, nhóm có hơn 10 năm kinh nghiệm thường tập trung ở độ tuổi từ 35 đến ngoài 45, với một số trường hợp vượt 50 tuổi. Nhóm này có thể đang ở giai đoạn quản lý hoặc chuyên gia trong lĩnh vực của mình.

2.3. Bộ dữ liệu công việc

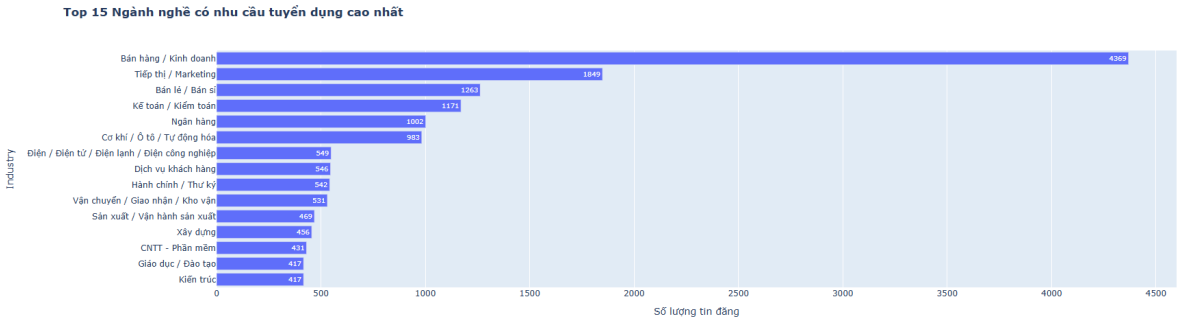


Hình 2.9: Biểu đồ top 15 vùng có nhiều tin tuyển dụng nhất

Biểu đồ thể hiện số lượng tin tuyển dụng tại 15 địa phương đứng đầu cả nước. Rõ ràng, TP. Hồ Chí Minh và Hà Nội là hai trung tâm có số lượng tin tuyển dụng áp đảo so với phần còn lại, với lần lượt 7.355 và 6.634 tin. Điều này cho thấy vai trò đầu tàu kinh tế của hai đô thị lớn, đồng thời phản ánh sự tập trung lớn của các doanh nghiệp, tập đoàn, và tổ chức tuyển dụng tại đây.

Khoảng cách giữa Hà Nội/TP.HCM và các tỉnh còn lại là rất lớn. Vị trí thứ ba, Bình Dương, chỉ ghi nhận khoảng 1.051 tin tuyển dụng – chưa bằng một phần sáu so với TP.HCM. Điều này phần nào phản ánh cơ cấu phát triển không đều giữa các vùng miền. Tuy nhiên, sự xuất hiện của các tỉnh công nghiệp như Đồng Nai, Long An, Hưng Yên, Bắc Ninh cho thấy rằng các khu công nghiệp và vùng ven đô đang dần trở thành điểm đến tiềm năng cho người lao động.

Các địa phương khác như Đà Nẵng, Hải Phòng, Quảng Ninh – đại diện cho miền Trung và vùng ven biển – tuy số lượng tin tuyển dụng không cao bằng, nhưng vẫn thể hiện vai trò trong quá trình chuyển dịch sản xuất và mở rộng đô thị hóa tại Việt Nam.



Hình 2.10: Biểu đồ top 15 ngành nghề có nhu cầu tuyển dụng cao nhất

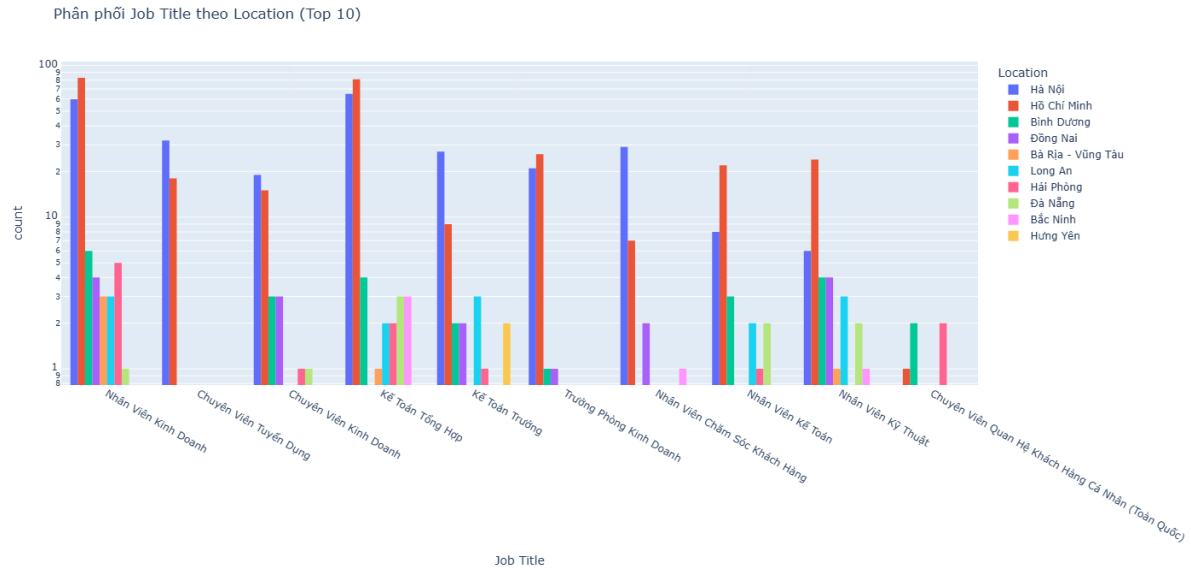
Biểu đồ thứ hai cho thấy nhu cầu tuyển dụng phân bổ theo ngành nghề. Trong đó, ngành Bán hàng / Kinh doanh dẫn đầu với số lượng tin tuyển dụng vượt trội (4.369 tin), gần gấp đôi ngành đứng thứ hai là Tiếp thị / Marketing với 1.849 tin. Điều

này không gây bất ngờ bởi Bán hàng là bộ phận cốt lõi trong hầu hết doanh nghiệp, đặc biệt là các công ty thương mại, dịch vụ, và bán lẻ.

Tiếp theo là các ngành cũng rất phổ biến và cần nguồn nhân lực lớn như Bán lẻ / Bán sỉ, Kế toán – Kiểm toán, và Ngân hàng. Những lĩnh vực này có đặc điểm là cần đội ngũ nhân sự đông đảo, từ vị trí entry-level đến chuyên viên cao cấp. Vì vậy, tần suất tuyển dụng cao là điều dễ hiểu.

Ngoài ra, các ngành mang tính kỹ thuật và sản xuất như Cơ khí – Ô tô – Tự động hóa, Điện – Điện lạnh – Điện công nghiệp, và Sản xuất – Vận hành cũng chiếm tỷ trọng đáng kể, cho thấy nhu cầu về lao động kỹ thuật hiện vẫn rất lớn, nhất là ở các khu công nghiệp.

Từ dữ liệu biểu đồ, có thể thấy rằng xu hướng tuyển dụng hiện tại nghiêng nhiều về các ngành mang tính vận hành – bán hàng – kinh doanh – kỹ thuật, phản ánh nhu cầu phát triển lực lượng lao động nhằm hỗ trợ hoạt động sản xuất và thương mại trên cả nước.



Hình 2.11: Biểu đồ phân phối ngành nghề theo vùng

Ta có thể thấy rất rõ ràng rằng hai thành phố lớn là Hồ Chí Minh và Hà Nội chiếm phần lớn số lượng tuyển dụng ở hầu hết các chức danh công việc. Trong biểu

đồ, các cột màu đỏ đại diện cho Hồ Chí Minh và màu xanh dương cho Hà Nội xuất hiện thường xuyên và có chiều cao vượt trội so với các địa phương khác, điều này phản ánh rõ vai trò trung tâm kinh tế và thị trường lao động sôi động tại hai thành phố này.

Vị trí "Nhân Viên Kinh Doanh" là công việc phổ biến nhất trong toàn bộ tập dữ liệu, đặc biệt nổi bật tại Hồ Chí Minh và Hà Nội, với số lượng bài đăng vượt trội so với các địa phương khác. Các vị trí như "Chuyên Viên Kinh Doanh", "Kế Toán Tổng Hợp" và "Trưởng Phòng Kinh Doanh" cũng xuất hiện thường xuyên, chủ yếu tập trung tại hai trung tâm kinh tế lớn là Hà Nội và Hồ Chí Minh, nhưng mức độ chênh lệch giữa hai thành phố không quá lớn. Ví dụ, "Chuyên Viên Kinh Doanh" ở Hồ Chí Minh chiếm khoảng 80 lượt đăng, trong khi Hà Nội gần 65, cho thấy mức độ cạnh tranh nhân lực giữa hai khu vực.

Một số vị trí khác như "Kế Toán Trưởng", "Nhân Viên Kế Toán", "Chuyên Viên Tuyển Dụng", hay "Nhân Viên Chăm Sóc Khách Hàng" có mặt tại nhiều tỉnh, nhưng với quy mô nhỏ hơn đáng kể. Điều này phản ánh rằng các công việc hành chính - kế toán tuy có nhu cầu rộng khắp, nhưng phần lớn vẫn tập trung tuyển ở các đô thị lớn.

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 một phương pháp thống kê dùng để đánh giá độ quan trọng của một từ trong một văn bản cụ thể so với toàn bộ văn bản. 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 tiêu chuẩn của TF-IDF được định nghĩa như sau:

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

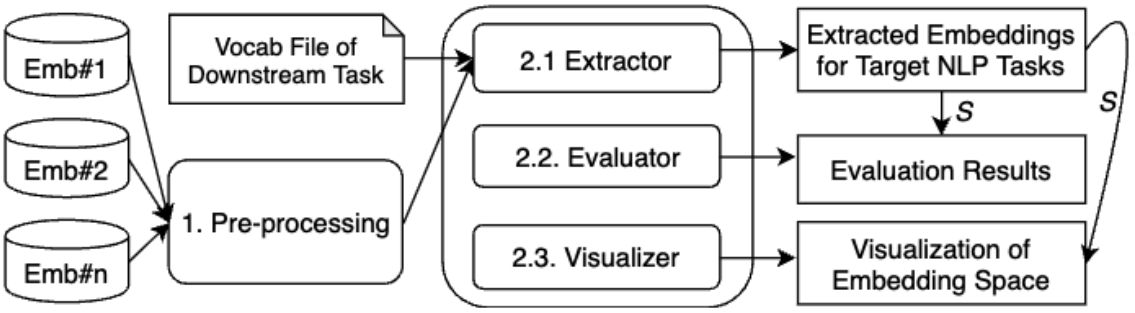
Trong đó $TF(t, d)$ là tần suất của từ t trong tài liệu d, N là tổng số tài liệu và $df(t)$ là số tài liệu chứa từ t. Phép nhân giữa TF và IDF giúp nhấn mạnh những từ đặc trưng (ít xuất hiện trên toàn bộ tập văn bản nhưng lặp lại nhiều trong một tài liệu cụ thể), đồng thời loại bỏ các từ phổ biến mang tính nhiễu. Theo Aizawa (2003), phương pháp này có nền tảng chặt chẽ về lý thuyết thông tin, gắn liền với entropy và tính bất định thông tin trong tài liệu. Trong lĩnh vực tuyển dụng, TF-IDF đã được ứng dụng để trích xuất đặc trưng từ CV và mô tả công việc (JD), qua đó đo độ tương đồng giữa ứng viên và vị trí cần tuyển (Apaza et al., 2021). Luthfi và Lhaksamana (2020) cho thấy việc kết hợp TF-IDF và các mô hình phân loại như SVM giúp tăng độ chính xác trong việc phân loại ứng viên phù hợp, đạt tới 86,3%. Ngoài ra, TF-IDF còn là thành phần quan trọng trong các hệ thống đề xuất việc làm dạng content-based filtering, đóng vai trò đại diện văn bản đầu vào.

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). Mô hình này hỗ trợ hai dạng biểu diễn: word-level (ở cấp độ từ) và syllable-level (ở cấp độ âm tiết), phản ánh đặc trưng ngôn ngữ tiếng Việt vốn mang tính phân đoạn cao theo âm tiết. Pho2Vec được huấn luyện trên tập dữ liệu tiếng Việt có quy mô lớn (trên 20 GB), bao gồm các nguồn như Wikipedia, báo chí, văn bản chính thống và mạng xã hội. Nhờ đó, các từ tiếng Việt được ánh xạ vào không gian vector liên tục có khả năng biểu diễn ngữ nghĩa sâu sắc và giữ mối quan hệ tương đồng giữa các từ.

Một đóng góp đáng chú ý trong phát triển Pho2Vec là quá trình đánh giá chọn lọc embedding thông qua hệ thống ETNLP (Evaluation and Visualization Toolkit for NLP). ETNLP cho phép so sánh, trực quan hóa và chọn ra embedding hiệu quả nhất

dựa trên các tiêu chí về hiệu suất thực nghiệm trên các tác vụ hạ nguồn (downstream tasks). Cụ thể, Pho2Vec đã được đánh giá trên các nhiệm vụ như Named Entity Recognition (NER), Word Analogy và Text-to-SQL và cho thấy kết quả vượt trội hơn so với các embedding không chuyên biệt cho tiếng Việt. Hình dưới đây minh họa quy trình tổng thể của ETNLP trong việc trích xuất, đánh giá và trực quan hóa các embedding trước khi áp dụng vào bài toán NLP cụ thể:



Hình 3.1: Quy trình tổng quát của ETNLP, trong đó S là tập các embedding được trích xuất để đánh giá và trực quan hóa cho một tác vụ NLP. (Nguồn: Nguyen, D. Q. & Nguyen, A. T., 2020)

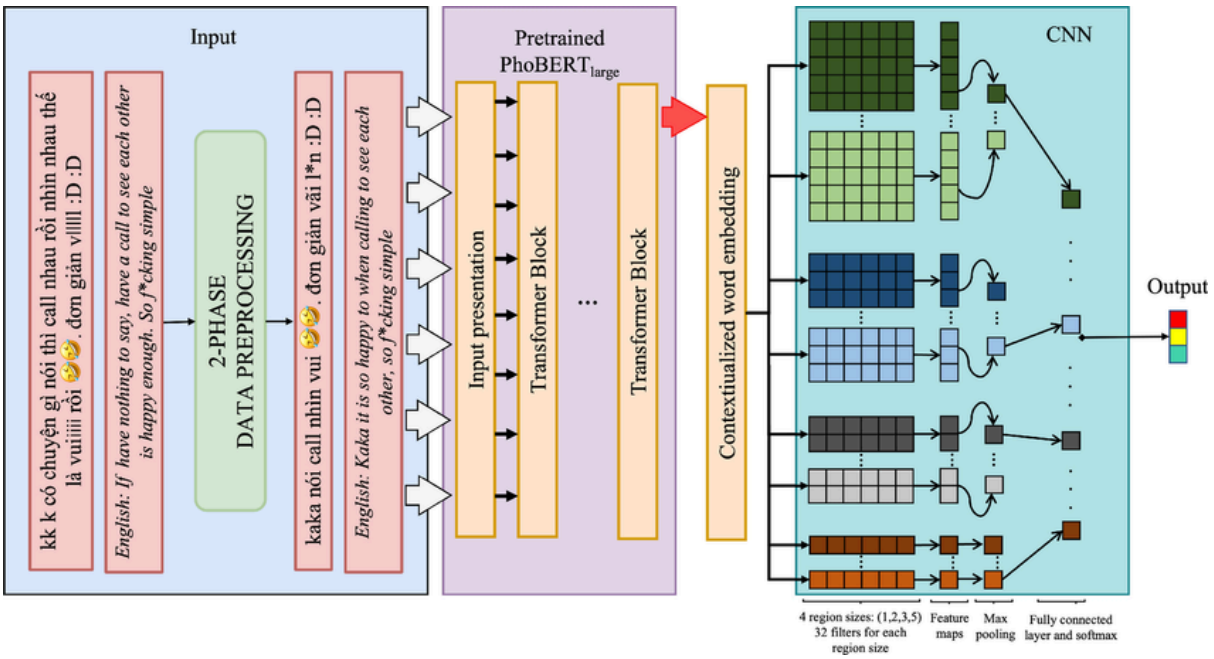
Trong các hệ thống gợi ý việc làm, Pho2Vec có thể được sử dụng để chuyển đổi văn bản mô tả công việc (Job Description – JD) và thông tin ứng viên (CV) sang dạng vector số. Từ đó, mô hình học máy có thể tính toán độ tương đồng giữa JD và CV thông qua khoảng cách cosine trong không gian vector. So với các kỹ thuật thống kê như TF-IDF, Pho2Vec có khả năng biểu diễn mối quan hệ ngữ nghĩa mềm mại và tổng quát hơn giữa các từ đồng nghĩa hoặc tương cận, từ đó tăng hiệu quả trong việc xác định mức độ phù hợp giữa ứng viên và công việc. Nghiên cứu của Apaza et al. (2018) cũng khẳng định rằng việc kết hợp Word2Vec và cosine similarity trong hệ thống đề xuất việc làm giúp cải thiện độ chính xác đề xuất so với phương pháp truyền thống.

3.1.3. PhoBERT

PhoBERT là mô hình ngôn ngữ đầu tiên được tiền huấn luyện (pre-trained) dành riêng cho tiếng Việt, dựa trên kiến trúc RoBERTa. Không giống như Pho2Vec vốn hoạt động dựa trên Skip-gram và cho embedding độc lập ngữ cảnh, PhoBERT sử

dụng cơ chế self-attention của transformer để học các biểu diễn ngữ cảnh (contextualized embeddings). Điều này có nghĩa là biểu diễn vector của một từ trong PhoBERT sẽ thay đổi tùy thuộc vào vị trí và ngữ nghĩa của từ đó trong câu. Mô hình được huấn luyện trên tập dữ liệu lớn Vietnamese Treebank, có hai phiên bản: PhoBERT-base (135 triệu tham số) và PhoBERT-large (370 triệu tham số). Theo Nguyen và Nguyen (2020), PhoBERT đạt hiệu suất vượt trội so với các mô hình đa ngôn như XLM-R trên nhiều tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt như phân tích cú pháp, phân loại văn bản, và nhận diện thực thể có tên (NER).

Trong các nghiên cứu gần đây về hệ thống đề xuất việc làm, PhoBERT thường được sử dụng để tạo embedding ngữ cảnh cho đoạn văn bản như mô tả công việc (JD) hoặc hồ sơ ứng viên (CV). Embedding được trích xuất từ token đặc biệt [CLS], đại diện cho toàn bộ chuỗi văn bản. Phương pháp này cho phép hệ thống gợi ý hiểu sâu hơn về mục tiêu nghề nghiệp, kỹ năng, hoặc điều kiện tuyển dụng được thể hiện một cách ẩn ý trong ngôn ngữ tự nhiên. Điển hình, nghiên cứu của Tran et al. (2022) cho thấy việc kết hợp PhoBERT và CNN giúp cải thiện đáng kể độ chính xác trong các tác vụ phân loại tiếng Việt. Ngoài ra, embedding từ PhoBERT còn được dùng làm đầu vào cho các hệ thống content-based filtering khi tính toán độ tương đồng giữa văn bản CV và JD bằng khoảng cách cosine.



Hình 3.2: Tổng quan mô hình PhoBERT và ứng dụng trong pipeline NLP (Source: Tran, K. Q., Nguyen, D. T., Nguyen, T. H., & Bui, T. (2022))

PhoBERT đóng vai trò quan trọng trong việc nâng cao năng lực hiểu ngữ cảnh và sắc thái ngôn ngữ tiếng Việt, từ đó hỗ trợ hiệu quả hơn cho các hệ thống gợi ý việc làm có tính ngữ nghĩa cao.

3.1.4. PCA

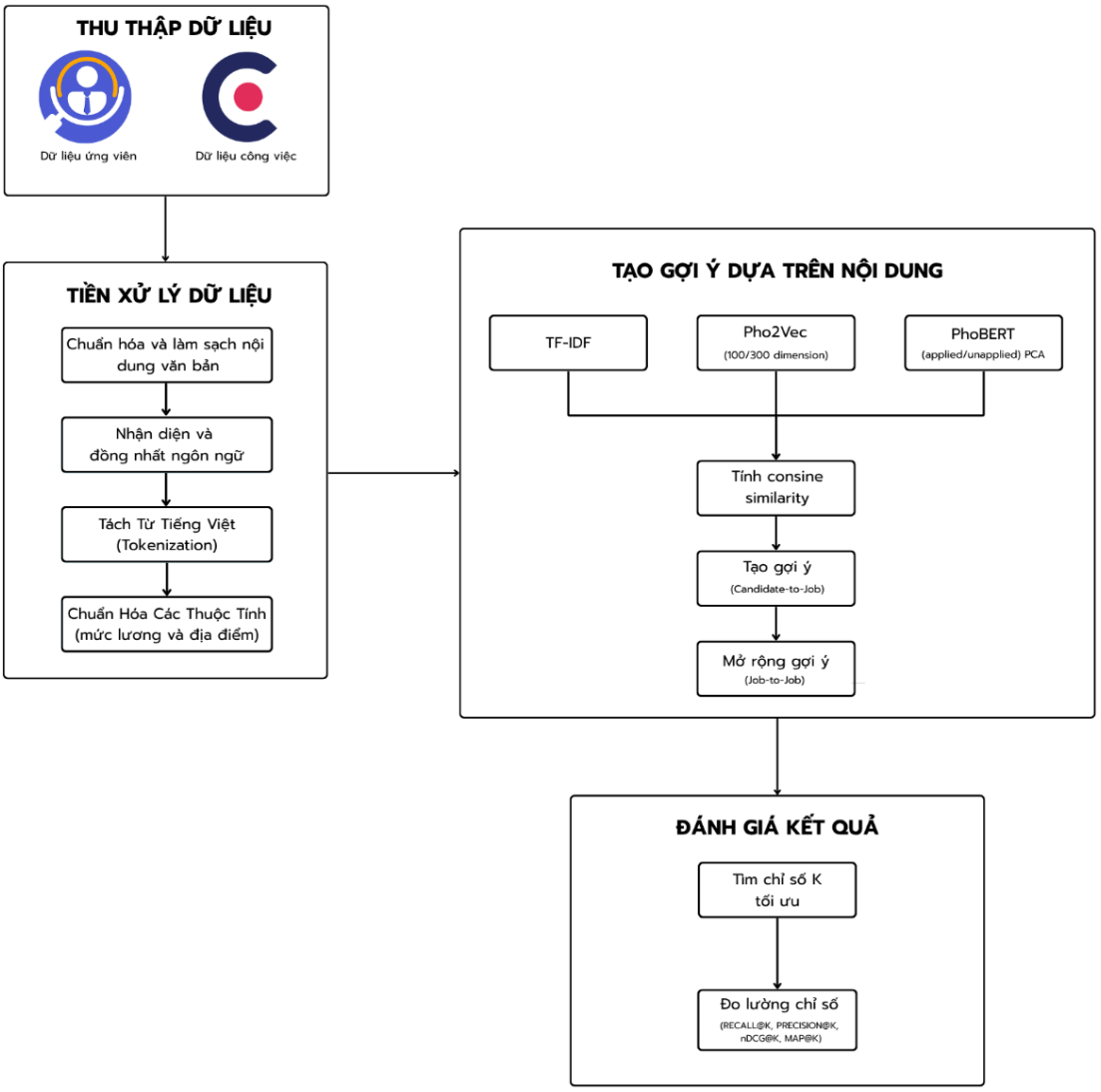
PCA (Principal Component Analysis) là một phương pháp giảm chiều dữ liệu tuyến tính được sử dụng rộng rãi trong học máy và khai phá dữ liệu. Mục tiêu của PCA là tìm ra một hệ trục tọa độ mới mà tại đó các thành phần chính (principal components) giữ lại phần lớn phương sai của dữ liệu, đồng thời không bị tương quan tuyến tính với nhau. Các thành phần chính được xác định thông qua phân tích giá trị riêng (eigenvalues) và vector riêng (eigenvectors) của ma trận hiệp phương sai. Kỹ thuật này giúp chuyển dữ liệu từ không gian gốc nhiều chiều về không gian mới có ít chiều hơn nhưng vẫn giữ lại được những đặc trưng quan trọng nhất.

PCA thường được áp dụng sau khi trích xuất đặc trưng từ các kỹ thuật như TF-IDF, Word2Vec, Pho2Vec hoặc PhoBERT, nhằm giảm số chiều vector đầu vào cho mô hình học máy. Việc này mang lại nhiều lợi ích như tăng tốc quá trình huấn luyện,

giảm nguy cơ overfitting, đồng thời tiết kiệm bộ nhớ và tài nguyên xử lý. Theo Berry et al. (1995), PCA được xem là một phương pháp nền tảng trong phân tích dữ liệu vì khả năng giữ lại cấu trúc chính của dữ liệu ban đầu trong khi loại bỏ các nhiễu không cần thiết.

Trong các hệ thống đề xuất việc làm (job recommendation), PCA đặc biệt hiệu quả khi cần tích hợp nhiều loại đặc trưng đầu vào như embedding của mô tả công việc (JD), embedding của hồ sơ ứng viên (CV), embedding kinh nghiệm làm việc hoặc mức lương mong muốn. Việc giảm số chiều vector xuống mức 100-300 sau khi xử lý bằng PhoBERT giúp duy trì độ chính xác của mô hình trong khi cải thiện đáng kể hiệu suất tính toán.

3.2 Quy trình thực hiện nghiên cứu

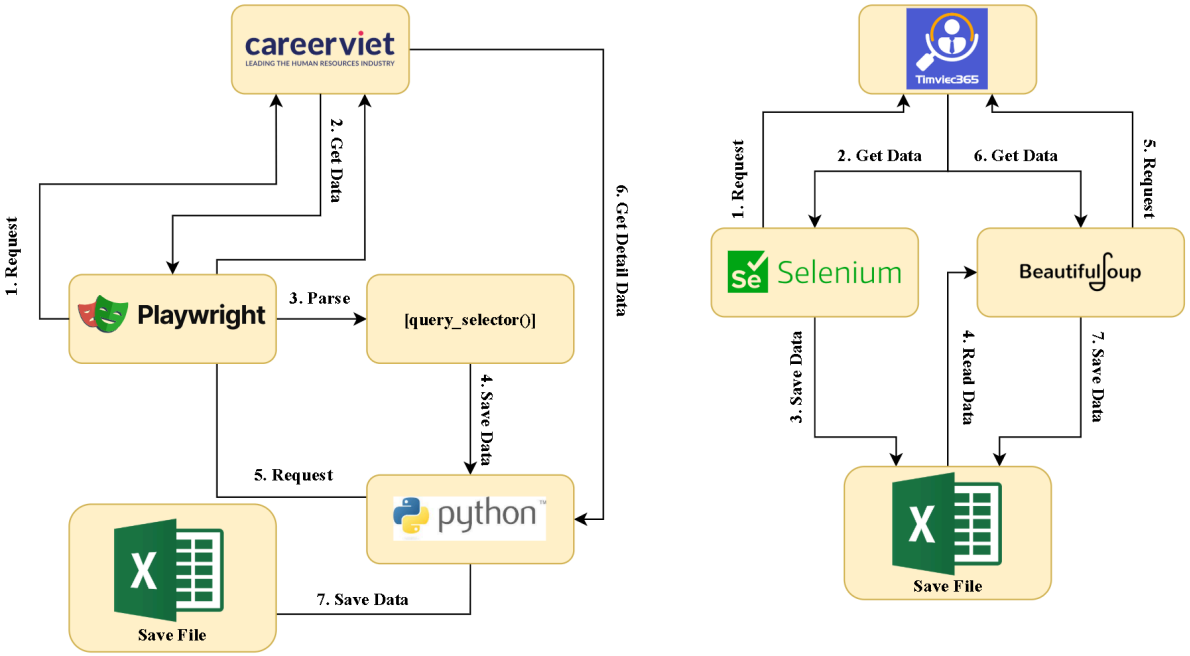


Hình 3.3: Mô hình nghiên cứu

3.2.1 Thu thập dữ liệu

Bộ dữ liệu sử dụng trong báo cáo này có tên là Vietnamese Jobs Dataset được nhóm nghiên cứu thu thập từ các trang web tìm việc trực tuyến. Sử dụng ngôn ngữ lập trình Python, kết hợp với ba framework được hỗ trợ cho việc cào dữ liệu là Selenium, BeautifulSoup và Playwright, để thu thập thông tin về công việc và ứng viên. Chi tiết về quy trình thu thập dữ liệu được trình bày ở Hình X. Bộ dữ liệu được thu thập gồm hai bộ: dữ liệu công việc (jobs) và dữ liệu ứng viên (users). Bộ dữ liệu công việc được nhóm nghiên cứu thu thập trên trang web *careerviet.vn* gồm 21.861 công việc với 10

thuộc tính. Bộ dữ liệu ứng viên được thu thập trên trang web *timviec365.vn* gồm 46,124 ứng viên với 17 thuộc tính. Thông tin chi tiết về các thuộc tính của bộ dữ liệu được thể hiện tại Bảng X.



Hình 3.4: Quy Trình Crawl Data.

Bảng 1: Thông tin chi tiết của các thuộc tính.

Bộ dữ liệu công việc		**Bộ dữ liệu ứng viên**	
Thuộc tính	**Ý nghĩa**	**Thuộc tính**	**Ý nghĩa**
Industry	Ngành nghề	Industry	Ngành nghề
Job Name	Tiêu đề tuyển dụng	Desired Job	Công việc mong muốn
Salary	Mức lương	Desired Salary	Mức lương mong muốn
Location	Địa chỉ làm việc	Workplace Desired	Nơi làm việc mong muốn
Type	Hình thức công việc	Type	Hình thức công việc

Time	Hạn nộp hồ sơ	User URL	URL của ứng viên
Benefit	Phúc lợi	User ID	Mã số ứng viên
Description	Mô tả công việc	User Name	Tên ứng viên
Requirement	Yêu cầu công việc	Current Place of Residence	Nơi ở hiện tại
URL	URL của công việc	Province	Tên tỉnh/TP
		District	Tên Quận/Huyện
		Gender	Giới tính
		Marriage	Tình trạng hôn nhân
		Age	Tuổi
		Work Experience	Kinh nghiệm làm việc
		Career Goals	Mục tiêu trong sự nghiệp
		Skills	Kĩ năng

Bộ dữ liệu được thu thập từ tháng 6/2025 với mục đích học và nghiên cứu môn học Phân Tích Dữ Liệu WEB. Hình X thể hiện ví dụ về các điểm dữ liệu thô trong bộ dữ liệu.

	user_url	userid	user_name	industry	workplace_desired	desired_salary	gender	marriage	age	work_experience	desired_job_translated
0	https://timviec365.vn/uvtv2/dao-huu-tai-uv1111...	1111684868	Đào Hữu Tài	Kd Bất Động Sản	Hà Nội	Thỏa Thuận	Nam	Độc Thân	38.0	Dưới 1 năm	Chuyên Viên Kinh Doanh Bất Động Sản
1	https://timviec365.vn/uvtv2/ho-vinh-duc-uv199803	199803	Hồ Vĩnh Đức	Xây Dựng	Hồ Chí Minh	Thỏa Thuận	Nam	Khác	37.0	Không yêu cầu kinh nghiệm	Kỹ Sư Kinh Tế Xây Dựng
2	https://timviec365.vn/uvtv2/ngo-thi-bich-tram-...	1271594	Ngô Thị Bích Trâm	Khách Sạn - Nhà Hàng	Khánh Hòa	Thỏa Thuận	Nam	Độc Thân	NaN	Chưa Có Kinh Nghiệm	Nhân Viên Nhà Hàng
3	https://timviec365.vn/uvtv2/vo-ba-loi-uv160413	160413	Võ Bá Lợi	Giao Thông Vận Tải - Thủy Lợi - Cầu Đường	Hồ Chí Minh	Thỏa Thuận	Nam	Độc Thân	NaN	Chưa Có Kinh Nghiệm	Đầu Bếp
4	https://timviec365.vn/uvtv2/vo-thi-thuy-van-uv...	1111646039	Võ Thị Thuỳ Vân	Quản Trị Kinh Doanh	Đồng Nai	Thỏa Thuận	Nữ	Độc Thân	21.0	Chưa Có Kinh Nghiệm	Thực Tập Sinh Kinh Doanh

Hình 3.5: 5 điểm dữ liệu thô đầu tiên của bộ dữ liệu ứng viên.

	Job_Name	Company_name	Location	Salary	Time	Benefit	Description	Requirement	URL	Industry	Type
0	Chuyên Viên Hỗ Trợ Kỹ Thuật (ô tô thương mại)...	CÔNG TY CP TM DV AN SƯƠNG	Hà Nội	Cạnh tranh	10/31/2025	Chế độ bảo hiểm, Du Lịch, Đồng phục, Chế độ th...	• Hỗ trợ kỹ thuật & hướng dẫn sửa chữa cho hệ ...	• Tốt nghiệp đại học ngành Cơ khí ô tô. • Tối ...	https://careerviet.vn/vi/tim-viec-lam/chuyen-v...	Cơ khí / Ô tô / Tự động hóa	Nhân viên chính thức
1	CHUYÊN VIÊN PHỤ TÙNG (MỚI)	CÔNG TY CP TM DV AN SƯƠNG	Hồ Chí Minh	5,000,000 - 10,000,000	9/30/2025	Chế độ bảo hiểm, Du Lịch, Đồng phục, Chế độ th...	- Xây dựng chính sách, kế hoạch và KPI kinh do...	-Tốt nghiệp ĐH ngành Cơ khí ô tô, Logistics, K...	https://careerviet.vn/vi/tim-viec-lam/chuyen-v...	Tiếp thị / Marketing	Nhân viên chính thức
2	PHÓ TRƯỞNG BỘ PHẬN PHÁP LÝ – ĐĂNG KIỂM XE CBU/...	CÔNG TY CP TM DV AN SƯƠNG	Hà Nội	Cạnh tranh	8/31/2025	Chế độ bảo hiểm, Du Lịch, Đồng phục, Chế độ th...	Mô tả công việc: - Quản lý hồ sơ pháp lý – kỹ ...	- Tốt nghiệp kỹ thuật ô tô, cơ khí, luật hoặc ...	https://careerviet.vn/vi/tim-viec-lam/pho-truo...	Luật / Pháp lý	Nhân viên chính thức
3	Tài Xế Lái Xe Cho Giám Đốc (MỚI)	CÔNG TY CP TM DV AN SƯƠNG	Hồ Chí Minh	5,000,000 - 10,000,000	6/30/2025	Chế độ bảo hiểm, Du Lịch, Đồng phục, Chế độ th...	Mô tả công việc Phụ trách đưa đón, chuyên chở ...	Nam, sức khỏe tốt, không bị cận; GPLX hạng B2 ...	https://careerviet.vn/vi/tim-viec-lam/tai-xe-l...	Vận chuyển / Giao nhận / Kho vận	Nhân viên chính thức
4	Phiên dịch viên tiếng Trung Quốc (MỚI)	CÔNG TY CP TM DV AN SƯƠNG	Hà Nội	15,000,000 - 20,000,000	8/31/2025	Chế độ bảo hiểm, Du Lịch, Chế độ thưởng, Chăm ...	- Phiên dịch, biên dịch tài liệu kỹ thuật và n...	- Thành thạo tiếng Trung - Ưu tiên ứng viên có...	https://careerviet.vn/vi/tim-viec-lam/phien-di...	Hành chính / Thư ký	Nhân viên chính thức

Hình 3.6: 5 điểm dữ liệu thô đầu tiên của bộ dữ liệu công việc.

3.2.2 Tiền xử lý dữ liệu

Trong các bài toán xử lý ngôn ngữ tự nhiên, giai đoạn tiền xử lý dữ liệu đóng vai trò nền tảng, quyết định trực tiếp đến hiệu quả của mô hình phân tích. Mục tiêu của giai đoạn này là chuyển đổi dữ liệu văn bản thô, vốn không đồng nhất và chứa nhiều nhiễu, về một định dạng sạch sẽ, cấu trúc và sẵn sàng cho việc phân tích. Dưới đây là mô tả chi tiết quy trình các bước chúng tôi thực hiện việc xử lý văn bản để chuẩn bị cho việc phân tích và xây dựng mô hình gợi ý công việc.

Bước 1: Chuẩn Hóa và Làm Sạch Nội Dung Văn Bản

Bước đầu tiên tập trung vào việc làm sạch và chuẩn hóa dữ liệu đầu vào. Quy trình bắt đầu bằng việc kiểm tra và đảm bảo toàn bộ dữ liệu đều ở định dạng chuỗi (string), những định dạng khác sẽ được giữ nguyên.

Tiếp theo, để loại bỏ các yếu tố không mang giá trị ngữ nghĩa, hệ thống tiến hành làm sạch sâu nội dung. Các thẻ HTML, thường xuất hiện trong dữ liệu thu thập từ web, sẽ được gỡ bỏ hoàn toàn thông qua việc sử dụng biểu thức chính quy. Song song đó, các ký tự đặc biệt như ký tự xuống dòng, tab, cùng các loại dấu câu và ký hiệu (!, @, #, _, *, v.v.) cũng được loại bỏ để văn bản trở nên tinh gọn.

Sau khi làm sạch, văn bản được chuẩn hóa về mặt định dạng. Các khoảng trắng thừa được thay thế bằng một dấu cách duy nhất. Một bước quan trọng trong xử lý văn bản tiếng Việt là loại bỏ các stopwords – những từ xuất hiện với tần suất cao nhưng

không đóng góp nhiều vào ý nghĩa của câu (ví dụ: "và", "là", "của", "tại"). Việc này được thực hiện dựa trên một danh sách stopwords được xây dựng sẵn, giúp mô hình tập trung vào các từ khóa cốt lõi.

Cuối cùng, để đảm bảo tính nhất quán, toàn bộ văn bản được chuẩn hóa về định dạng chữ viết. Cụ thể, văn bản được chuyển về chữ thường, sau đó viết hoa ký tự đầu của mỗi từ. Thao tác này giúp đồng bộ hóa dữ liệu, tránh trường hợp mô hình nhận diện sai một từ chỉ vì sự khác biệt trong cách viết hoa.

Bước 2: Nhận Diện và Đồng Nhất Ngôn Ngữ

Để đảm bảo toàn bộ không gian dữ liệu được xử lý thống nhất, quy trình tiến hành nhận diện và đồng nhất hóa ngôn ngữ. Hệ thống sử dụng kết hợp các thư viện chuyên dụng là *langdetect* để tự động xác định ngôn ngữ của từng đoạn văn bản. Trong trường hợp phát hiện văn bản không phải là tiếng Việt, mô-đun dịch tự động *googletrans* được sử dụng để chuyển ngữ nội dung sang tiếng Việt. Bước này có ý nghĩa quan trọng trong việc tạo ra một tập dữ liệu đồng nhất về mặt ngôn ngữ, là tiền đề cho các bước phân tích sâu hơn.

Bước 3: Tách Từ Tiếng Việt (Tokenization)

Tách từ là bước cuối cùng và mang tính đặc thù cao trong xử lý ngôn ngữ tự nhiên tiếng Việt. Do đặc điểm của tiếng Việt là ngôn ngữ đơn âm tiết nhưng lại có rất nhiều từ ghép (ví dụ: "mục tiêu nghề nghiệp", "khoa học dữ liệu"), việc tách từ đơn thuần dựa trên khoảng trắng sẽ phá vỡ cấu trúc và ý nghĩa của các cụm từ quan trọng.

Để giải quyết thách thức này, chúng tôi sử dụng công cụ tách từ chuyên sâu dành riêng cho tiếng Việt là *ViTokenizer*. Công cụ này có khả năng nhận diện chính xác ranh giới của các từ đơn và từ ghép, qua đó phân tách chuỗi văn bản thành một danh sách các "token" (từ) hoàn chỉnh và có nghĩa. Sau khi tách, các token này được kết hợp lại thành một chuỗi duy nhất, với mỗi token được ngăn cách bởi một dấu khoảng trắng. Kết quả của bước này là một bộ dữ liệu có cấu trúc, sẵn sàng để đưa

vào các mô hình học máy cho những tác vụ như phân loại văn bản, phân tích cảm xúc hay nhận dạng thực thể.

Bước 4: Chuẩn Hóa Các Thuộc Tính "Mức Lương Mong Muốn" và "Kinh Nghiệm Làm Việc"

Hai thuộc tính định tính quan trọng là **mức lương** và **kinh nghiệm làm việc** thường xuất hiện với nhiều định dạng khác nhau giữa dữ liệu ứng viên và dữ liệu công việc. Để đảm bảo tính nhất quán và thuận tiện cho phân tích, các giá trị trong hai thuộc tính này được chuẩn hóa theo các bước sau:

Bảng 2: Chuẩn hóa dữ liệu lương và kinh nghiệm

Ứng viên		**Công việc**
desired_salary	*work_experience*	*Salary*
Thỏa thuận	Chưa có kinh nghiệm	Thỏa thuận / Cạnh tranh
1,000,000 - 5,000,000	Dưới 1 năm	Dưới 5,000,000
5,000,000 - 10,000,000	1-3 năm	5,000,000 - 10,000,000
10,000,000 - 15,000,000	3-5 năm	10,000,000 - 15,000,000
15,000,000 - 20,000,000	5-10 năm	15,000,000 - 20,000,000
20,000,000 - 40,000,000	Trên 10 năm	20,000,000 - 40,000,000
Trên 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 đầu 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 xử lý dưới dạng dữ liệu phân loại rời rạc. Cột "work_experience" ban đầu bao gồm bảy giá trị: "Dưới 1

năm", "1-3 năm", "3-5 năm", "5-10 năm", "Trên 10 năm", "Không yêu cầu kinh nghiệm" và "Chưa Có Kinh Nghiệm". Các giá trị này được mã hóa bằng Label Encoding thành các số nguyên từ 0 đến 6. Tương tự, cột "desired_salary" chứa bảy mức lương: "Thỏa Thuận", "1,000,000 - 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" và "Trên 40,000,000", cũng được mã hóa từ 0 đến 6.

Sau bước mã hóa, hai cột này được đưa vào mô hình đơn giản sử dụng các lớp Input, Embedding, Flatten và Dense trong Keras để học vector biểu diễn liên tục cho từng giá trị phân loại. Cột "work_experience" được ánh xạ thành vector có chiều dài 3, còn "desired_salary" được ánh xạ thành vector dài 5. Kết quả là hai embedding vector có tổng cộng 8 chiều cho mỗi ứng viên, phản ánh thông tin về kinh nghiệm làm việc và kỳ vọng thu nhập.

Vector 8 chiều này sau đó được kết hợp trực tiếp với đặc trưng văn bản đã xử lý thô (chẳng hạn như TF-IDF, Pho2Vec hoặc PhoBERT), tạo thành một biểu diễn đầu vào thống nhất cho mỗi ứng viên. Việc ghép nối này đảm bảo rằng thông tin định lượng và định tính từ cả văn bản và dữ liệu phi văn bản được tích hợp đầy đủ trong pipeline gợi ý.

Bước 2: Biểu diễn văn bản bằng các kỹ thuật NLP

Ở phía ứng viên, các trường "proc_desired_job", "proc_industry" và "proc_workplace" sau khi tiền xử lý được ghép nối cùng với hai đặc trưng phân loại đã mã hóa là "work_experience_enc" và "desired_salary_enc" để tạo thành một chuỗi văn bản duy nhất gọi là combined_text_candidate. Ở phía công việc, combined_text_job được xây dựng từ các trường "Job_Name_processed", "Industry_processed", "Location_processed", "Requirement_processed", "Description_processed" và "Salary_enc". Hai chuỗi văn bản này lần lượt đại diện cho hồ sơ ứng viên và thông tin tuyển dụng, phục vụ cho bước biểu diễn đặc trưng bằng kỹ thuật xử lý ngôn ngữ tự nhiên.

TF-IDF được áp dụng trực tiếp lên các chuỗi văn bản đã ghép sau khi được xử lý thô (bao gồm chuẩn hóa chữ thường, loại bỏ ký tự đặc biệt và stopword). Với tham số max_features=1000, mỗi văn bản được chuyển thành vector thưa có 1.000 chiều.

Pho2Vec được sử dụng ở hai mô hình: 100 chiều và 300 chiều. Sau khi token hóa, mỗi từ trong chuỗi được ánh xạ sang vector tiền huấn luyện tương ứng. Biểu diễn cuối cùng của một văn bản được tính bằng trung bình cộng các vector từ, cho ra vector đặc trưng dense 100 hoặc 300 chiều tùy phiên bản sử dụng.

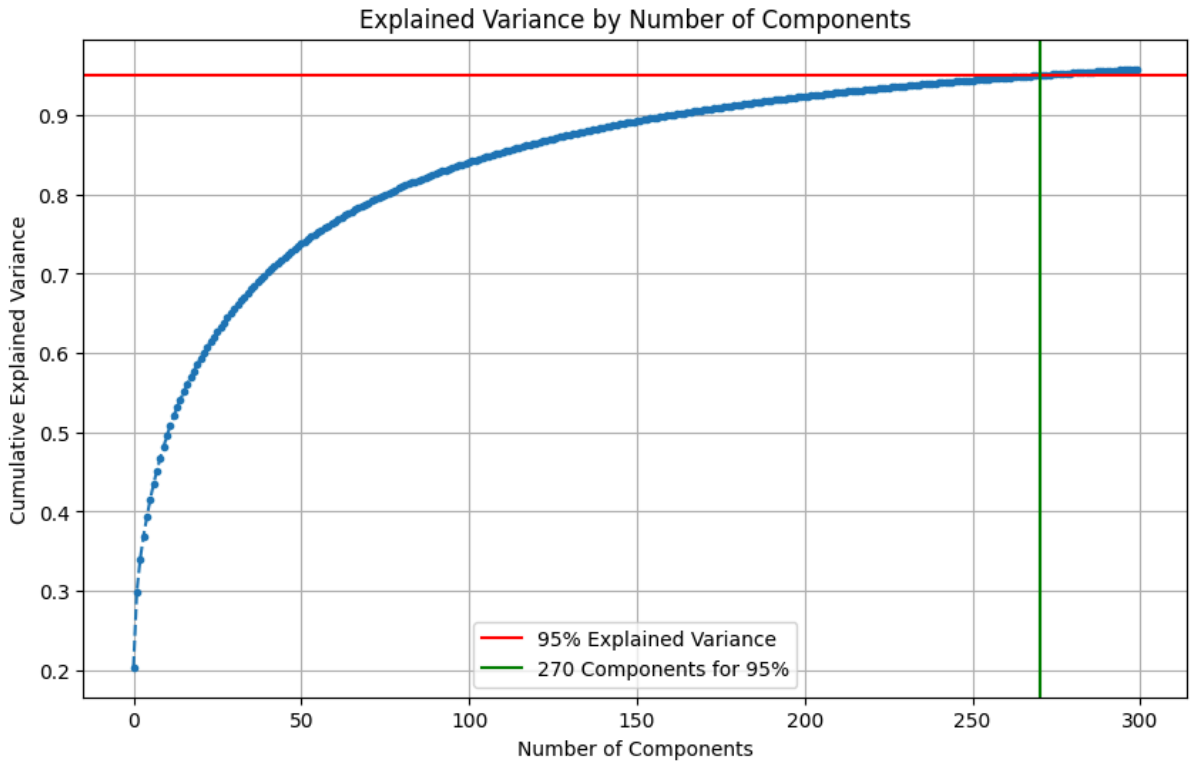
Mô hình PhoBERT được áp dụng với tokenizer và kiến trúc vinai/phobert-base, trong đó vector biểu diễn đặc trưng tương ứng với token [CLS] có độ dài ban đầu là 768 chiều. Để chuẩn hóa cấu trúc dữ liệu đầu vào và tối ưu hóa hiệu quả xử lý trong quá trình huấn luyện mô hình, nhóm nghiên cứu tiến hành giảm chiều vector này xuống còn 300 thông qua kỹ thuật Phân tích Thành phần Chính (Principal Component Analysis – PCA). Việc áp dụng PCA trong giai đoạn tiền xử lý dữ liệu được thực hiện với hai mục tiêu chính:

1. **Giảm chiều dữ liệu nhằm tối ưu hóa tài nguyên tính toán**

Các vector đầu ra của PhoBERT có số chiều lớn, dẫn đến chi phí xử lý cao và thời gian huấn luyện kéo dài. Việc giảm số chiều từ 768 xuống còn 300 thông qua PCA giúp giảm tải đáng kể cho hệ thống, cải thiện tốc độ huấn luyện và tiết kiệm tài nguyên tính toán mà không làm mất nhiều thông tin quan trọng trong dữ liệu.

2. **Loại bỏ nhiễu và giữ lại các thành phần thông tin quan trọng**

PCA cho phép xác định các thành phần chính thể hiện phần lớn phương sai của dữ liệu, qua đó loại bỏ những chiều không mang nhiều ý nghĩa về mặt thông tin hoặc có thể chứa nhiễu. Nhờ đó, dữ liệu đầu vào của mô hình trở nên tinh gọn và hiệu quả hơn trong việc học biểu diễn ngữ nghĩa.



Hình 3.7: Mối quan hệ giữa số lượng thành phần chính và phần phương sai tích lũy

Từ biểu đồ có thể nhận thấy, đường cong màu xanh lam biểu thị phần phương sai tích lũy tăng dần theo số lượng thành phần chính, phản ánh mức độ thông tin được bảo toàn trong quá trình giảm chiều bằng phương pháp PCA. Đường màu đỏ đánh dấu ngưỡng 95% phương sai – một chuẩn phổ biến trong các nghiên cứu thực nghiệm nhằm đảm bảo duy trì phần lớn thông tin có ý nghĩa từ tập dữ liệu gốc. Việc ngưỡng này được đạt tại thành phần thứ 270 cho thấy rằng việc giảm số chiều từ 768 xuống 270 không gây thất thoát đáng kể về mặt thông tin, đồng thời giúp loại bỏ các chiều dư thừa hoặc chứa nhiễu không cần thiết. Tuy nhiên, để tăng cường khả năng biểu diễn đặc trưng và đảm bảo mức độ bao phủ thông tin cao hơn trong các giai đoạn huấn luyện tiếp theo, nghiên cứu quyết định giữ lại 300 thành phần chính. Việc lựa chọn này không chỉ tạo điều kiện thuận lợi cho quá trình huấn luyện mà còn hỗ trợ việc so sánh hiệu quả giữa các mô hình học máy khác nhau. Do đó, cần thiết tiến hành thêm các thử nghiệm thực nghiệm với nhiều mô hình khác nhau để xác định mô hình tối ưu nhất cho bài toán đặt ra.

Các biểu diễn vector tạo thành từ TF-IDF, Pho2Vec và PhoBERT cho cả combined_text_candidate và combined_text_job sẽ là đầu vào cho bước tính toán mức độ tương đồng giữa ứng viên và công việc. Những biểu diễn này, khi kết hợp với vector đặc trưng phi văn bản ở Bước 1, tạo nên tập đặc trưng đầy đủ, phản ánh toàn diện thông tin từ hai phía trong hệ thống gợi ý việc làm.

Bảng 3. Kích thước vector đầu ra theo từng kỹ thuật biểu diễn văn bản

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

Quá trình thử nghiệm được thiết kế nhằm kiểm chứng độ chính xác của mô hình gợi ý dựa trên nội dung văn bản, sử dụng phương pháp content-based filtering (CBF). Hệ thống hoạt động theo hai hướng chính: đầu tiên là gợi ý trực tiếp công việc

từ hồ sơ ứng viên (candidate-to-job), sau đó là mở rộng danh sách công việc bằng cách tìm các công việc tương tự (job-to-job expansion) từ các gợi ý ban đầu.

Ở hướng thứ nhất, mỗi ứng viên được biểu diễn bằng một vector đặc trưng. Tương tự, các công việc cũng được biểu diễn bằng vector có cùng cấu trúc. Sự tương đồng giữa ứng viên và công việc được tính bằng chỉ số cosine similarity, cho phép hệ thống đo lường độ gần nhau về mặt nội dung giữa hai vector. Dựa trên chỉ số này, hệ thống tiến hành sắp xếp các công việc theo mức độ phù hợp và chọn ra Top-10 công việc có điểm số cao nhất cho mỗi ứng viên.

Sau bước gợi ý ban đầu, nhóm tiếp tục mở rộng danh sách công việc bằng phương pháp tìm kiếm các công việc có nội dung tương tự với những công việc đã được chọn. Phương pháp mở rộng job-to-job này sử dụng lại không gian biểu diễn vector ban đầu, áp dụng phép tính cosine similarity giữa các công việc với nhau để tìm ra thêm 5 công việc tương đương về yêu cầu kỹ năng, ngành nghề và mô tả công việc ứng với mỗi công việc đã được đề xuất. Bằng cách kết hợp cả hai hướng, hệ thống không chỉ dừng lại ở những công việc khớp trực tiếp với ứng viên mà còn mở rộng cơ hội bằng cách tìm kiếm những công việc "gần đúng" trong ngữ cảnh nghề nghiệp.

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. Nhãn này đóng vai trò như một thước đo tham chiếu giúp đánh giá khách quan kết quả mô hình.

4. Kết quả thực nghiệm

4.1. Các chỉ số đánh giá

Để đánh giá hiệu quả của hệ thống gợi ý công việc, nhóm nghiên cứu sử dụng bốn chỉ số chính: Precision@K, Recall@K, nDCG@K và MAP@K, trong đó K là siêu tham số đại diện cho số lượng gợi ý hàng đầu mà người dùng nhận được. Việc lựa chọn giá trị K phù hợp giúp phản ánh mức độ chính xác và hữu ích của hệ thống từ

góc nhìn người dùng, vốn thường chỉ quan tâm đến một số gợi ý đầu tiên. Tất cả các chỉ số trên đều phụ thuộc vào giá trị của K. Do đó, để lựa chọn giá trị K tối ưu, nhóm nghiên cứu sẽ thử nghiệm với các giá trị K khác nhau: [5, 10, 20, 30, 40, 50]. Mục tiêu là tìm ra giá trị K phù hợp nhất, giúp ứng viên nhận được các gợi ý công việc liên quan nhất. Cụ thể:

Precision@K: đo lường tỷ lệ các mục được gợi ý trong Top-K thực sự là phù hợp (tức là đúng với kỳ vọng của người dùng), tức đo lường tỷ lệ các công việc trong Top-K gợi ý thực sự phù hợp với hồ sơ ứng viên, hệ thống có gợi đúng hay không

$$Precision@K = \frac{Số\ lượng\ công\ việc\ phù\ hợp\ trong\ Top-K}{K}$$

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.

$$Recall@K = \frac{Số\ lượng\ công\ việc\ phù\ hợp\ trong\ Top-K}{Tổng\ số\ công\ việc\ phù\ hợp\ thực\ tế}$$

nDCG@K (Normalized Discounted Cumulative Gain): đo lường chất lượng xếp hạng trong Top-K gợi ý, tức là công việc phù hợp không chỉ xuất hiện mà còn đứng ở vị trí cao trong danh sách.

$$nDCG@K = \frac{DCG@K}{IDCG@K}$$

Trong đó:

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{log_2(i+1)}$$

MAP@K (Mean Average Precision at K) là chỉ số tổng hợp cho biết độ chính xác trung bình của các đề xuất Top-K trên toàn bộ tập ứng viên, có tính đến cả mức độ phù hợp lẫn vị trí xuất hiện của từng công việc.

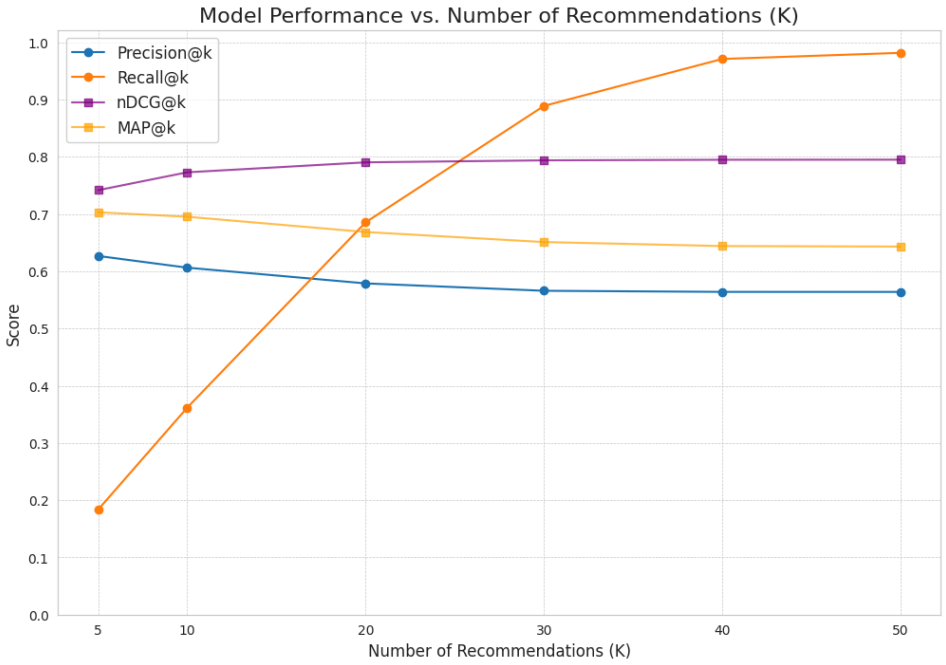
$$MAP@K = \frac{1}{N} \sum_{n=1}^{N} AP@K_n$$

Trong đó:

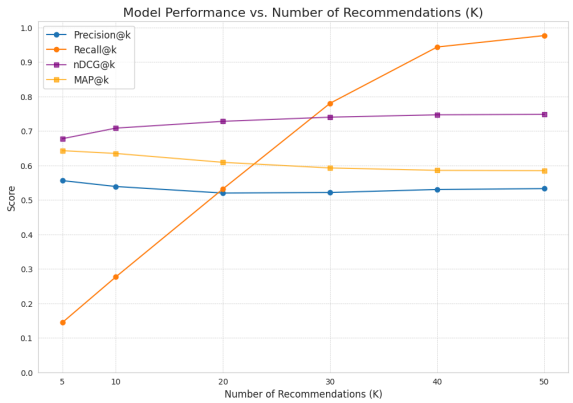
$$AP@K = \frac{1}{min(K,R)} \sum_{i=1}^{K} P(i).rel_i$$

4.2. Kết quả phân tích dữ liệu

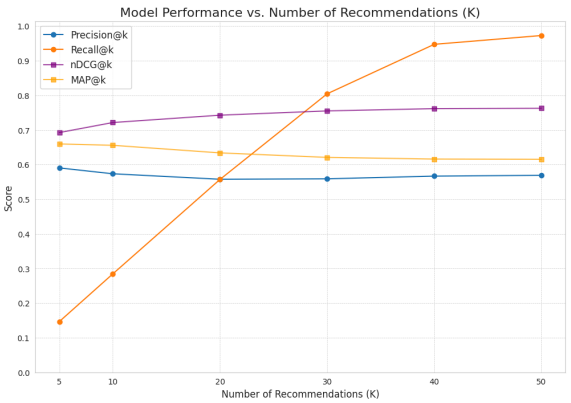
Để lựa chọn giá trị K phù hợp, nhóm nghiên cứu không chỉ hướng đến việc tối đa hóa Recall@K mà còn cân bằng với độ chính xác (Precision@K) và đảm bảo chất lượng xếp hạng qua MAP@K và nDCG@K. Một trong những tiêu chí quan trọng là xác định "Điểm Gãy" của Lợi ích (Point of Diminishing Returns) — tức là điểm mà việc tăng K không còn mang lại cải thiện đáng kể về hiệu suất. Điểm "sweet spot" này thường được nhận biết khi đường cong Recall bắt đầu thoai thoải, cho thấy hệ thống đã bao phủ phần lớn các công việc phù hợp và việc tiếp tục tăng K chỉ mang lại giá trị gia tăng rất nhỏ.



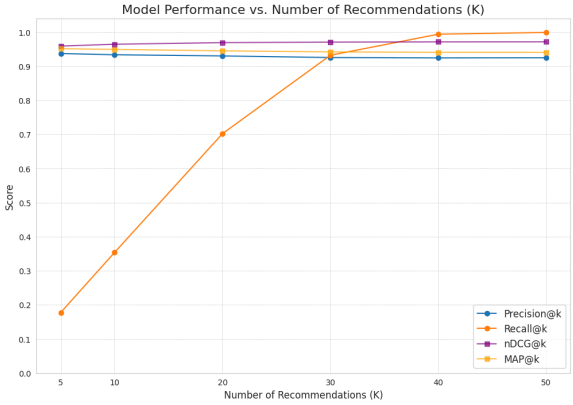
Hình 4.1: Kết quả lựa chọn chỉ số K của TF-IDF



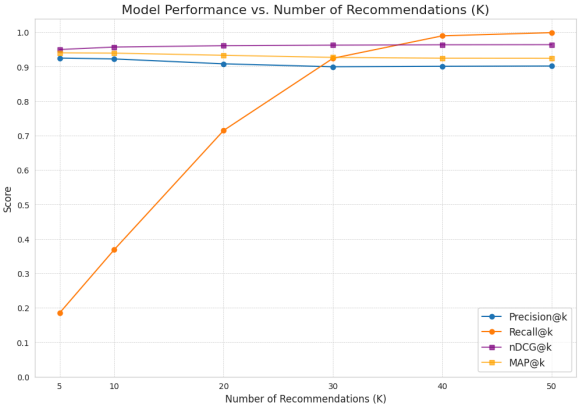
Hình 4.2: Kết quả lựa chọn chỉ số K của Pho2Vec (100)



Hình 4.3: Kết quả lựa chọn chỉ số K của Pho2Vec (300)



Hình 4.4: Kết quả lựa chọn chỉ số K của PhoBERT (PCA)



Hình 4.5: Kết quả lựa chọn chỉ số K của PhoBERT

Dựa trên kết quả thực nghiệm từ tất cả các mô hình được triển khai trong nghiên cứu, nhóm lựa chọn K = 30 làm giá trị ngưỡng chính thức để đánh giá các chỉ số như Precision@K, Recall@K, nDCG@K và MAP@K. Quyết định này được đưa ra dựa trên sự quan sát đồng nhất về hiệu suất của các mô hình.

Bảng 4: Kết quả phân tích dữ liệu

Mô hình	Precision@30	Recall@30	nDCG@30	MAP@30
TF-IDF	0.57	0.89	0.80	0.65
Pho2Vec100	0.52	0.78	0.74	0.59
Pho2Vec300	0.56	0.80	0.76	0.62
PhoBERT	0.90	0.92	0.96	0.93
PhoBERT (PCA)	0.93	0.93	0.97	0.94

Trong số các mô hình được thử nghiệm, TF-IDF tuy là một phương pháp truyền thống nhưng vẫn mang lại kết quả khá ấn tượng. Với Precision@30 đạt khoảng 56% và Recall@30 lên tới gần 89%, TF-IDF cho thấy khả năng bao phủ tốt các công việc phù hợp với hồ sơ ứng viên, tức mô hình có thể nhận diện tương đối hiệu quả các từ khóa quan trọng, đặc biệt khi dữ liệu có sự trùng lặp cao giữa nội dung hồ sơ và mô tả công việc. Tuy nhiên, do TF-IDF chỉ dựa vào tần suất từ và không hiểu ngữ cảnh, nên độ chính xác trong việc xếp hạng (thể hiện qua nDCG và MAP) vẫn còn hạn chế so với các mô hình hiện đại.

Với Pho2Vec, hai phiên bản sử dụng vector 100 chiều và 300 chiều lần lượt đạt Precision@30 ở mức xấp xỉ 52% và 56%. Khi chuyển sang phiên bản 300 chiều, hiệu suất có phần cải thiện, cho thấy số chiều cao hơn giúp biểu diễn thông tin phong phú hơn. Tuy vậy, phương pháp trung bình vector (mean pooling) vẫn là một điểm yếu cố hữu, bởi việc tổng hợp đặc trưng từ các từ đơn lẻ mà không quan tâm đến ngữ cảnh khiến mô hình dễ bỏ sót các mối liên hệ ngữ nghĩa phức tạp, đặc biệt trong các đoạn mô tả dài hoặc nhiều lớp nghĩa.

Ở phía ngược lại, PhoBERT đạt hiệu năng vượt trội nhất trong tất cả các mô hình. Precision@30 chạm mốc gần 90%, Recall@30 vượt 92%, và MAP@30 lên tới 92.7%. Khác với TF-IDF hay Pho2Vec, PhoBERT dựa trên kiến trúc transformer nên có khả năng nắm bắt tốt các biểu hiện ngữ cảnh, giúp việc so khớp giữa CV và JD diễn ra sâu sắc và chính xác hơn. Khi kết hợp với kỹ thuật PCA để giảm chiều vector

từ 768 còn 300, mô hình PhoBERT (PCA) tiếp tục đạt được kết quả tối ưu nhất, với MAP@30 đạt 94.2%. PCA không chỉ giúp rút gọn không gian biểu diễn mà còn góp phần loại bỏ nhiễu, từ đó nâng cao chất lượng tổng thể của hệ thống gợi ý.

Từ các kết quả trên, có thể thấy rằng sự chênh lệch giữa các mô hình truyền thống và hiện đại là rất rõ rệt. TF-IDF vẫn có vai trò nhất định nhờ khả năng đơn giản, hiệu quả trong trường hợp từ khóa mang tính mô tả rõ ràng. Tuy nhiên, các mô hình ngữ nghĩa sâu như PhoBERT – đặc biệt khi kết hợp với PCA – đã cho thấy ưu thế rõ ràng trong việc nắm bắt mối quan hệ ngữ nghĩa phức tạp, giúp đưa ra các gợi ý việc làm sát thực và chính xác hơn.

Ngoài ra, nhóm cũng đã tiến hành chạy mô hình với dữ liệu mới hoàn toàn, bao gồm 1983 dữ liệu ứng viên và 4200 dữ liệu công việc mới. Dưới đây là kết quả thực nghiệm nhóm đã ghi nhận (với k = 30).

Bảng 5: Kết quả phân tích dữ liệu

Mô hình	**Precision@30**	**Recall@30**	**nDCG@30**	**MAP@30**
TF-IDF	0.57	0.60	0.59	0.37
PhoBERT (PCA)	0.25	0.76	0.52	0.29
Pho2Vec100	0.32	0.72	0.54	0.38
Pho2Vec300	0.33	0.72	0.55	0.40
PhoBERT	0.27	063	0.47	0.30

Mặc dù các mô hình đề xuất, đặc biệt là PhoBERT và PhoBERT (PCA), đã đạt được hiệu suất rất cao trên tập dữ liệu đánh giá ban đầu, một hạn chế lớn và quan trọng của nghiên cứu này đã xuất hiện khi kiểm thử trên một tập dữ liệu hoàn toàn mới (bao gồm các ứng viên và tin tuyển dụng chưa từng thấy). Hiệu suất của các mô hình trên thực tế đã sụt giảm nghiêm trọng, cho thấy một khoảng cách đáng kể giữa kết quả trong môi trường thí nghiệm và khả năng ứng dụng trong thế giới thực.

Sự khác biệt này là cực kỳ rõ rệt đối với các mô hình PhoBERT. Mặc dù ban đầu đạt các chỉ số gần như hoàn hảo (MAP > 0.93), hiệu suất của chúng đã giảm mạnh xuống chỉ còn khoảng 0.30 trên dữ liệu mới. Đây là một dấu hiệu rõ ràng cho thấy mô hình đã không tổng quát hóa tốt như kỳ vọng.

Hiện tượng này cho thấy một vấn đề kinh điển trong học máy: overfitting (quá khớp). Có thể xuất phát từ hai nguyên nhân sau:

Đầu tiên, quá khớp với tập dữ liệu ban đầu. Các mô hình, đặc biệt là các mô hình phức tạp như PhoBERT, dường như đã "học thuộc lòng" các đặc điểm, mẫu câu và từ vựng đặc trưng của tập dữ liệu huấn luyện và đánh giá ban đầu. Các chỉ số hiệu suất cao chót vót (trên 90%) chính là một dấu hiệu cảnh báo cho khả năng này. Khi gặp dữ liệu mới với các đặc điểm khác biệt, mô hình không thể áp dụng những gì đã "học thuộc" một cách hiệu quả.

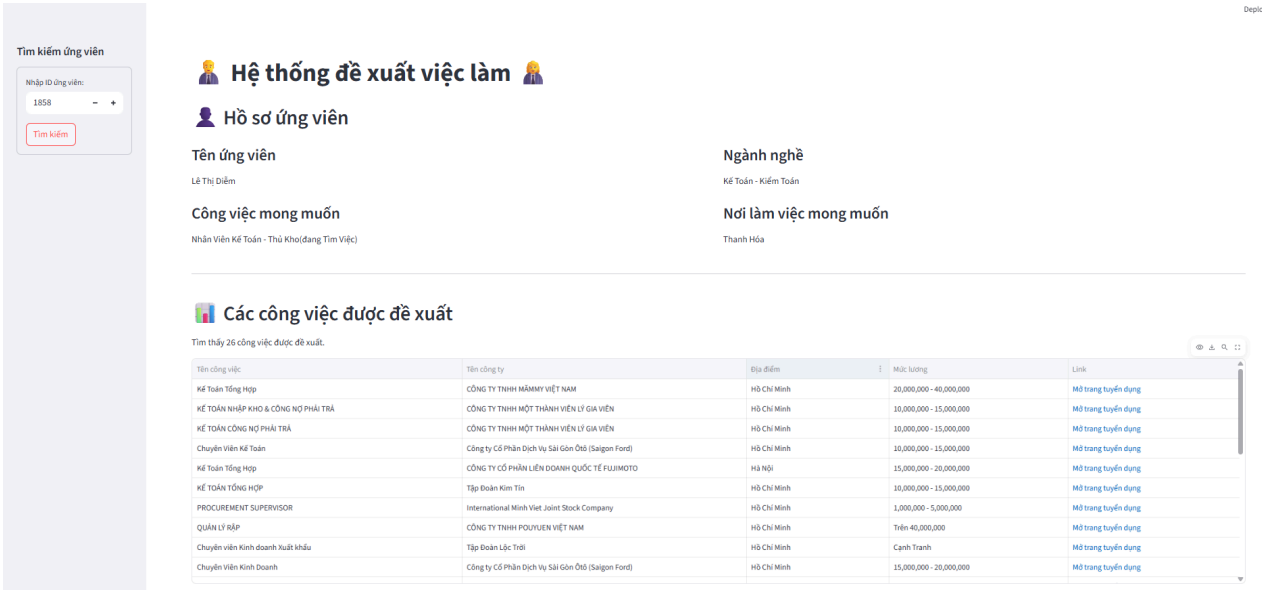
Thứ hai, có sự khác biệt về phân phối dữ liệu (Data Distribution Shift). Đây là nguyên nhân sâu xa hơn. Tập dữ liệu mới có thể có những khác biệt căn bản so với tập dữ liệu gốc, ví dụ:

- **Từ vựng và Thuật ngữ:** Các tin tuyển dụng và CV mới có thể đến từ các ngành nghề khác (ví dụ: marketing, tài chính thay vì chỉ IT), sử dụng bộ từ vựng, thuật ngữ chuyên ngành và cách diễn đạt khác.
- **Cấu trúc và Định dạng:** Cách viết mô tả công việc (JD) hoặc trình bày hồ sơ (CV) có thể khác biệt, khiến mô hình khó trích xuất thông tin tương đồng.
- **Sự thay đổi của dữ liệu theo thời gian (Temporal Drift):** Các kỹ năng và yêu cầu công việc thay đổi theo thời gian. Dữ liệu mới có thể phản ánh các xu hướng tuyển dụng gần đây hơn mà mô hình chưa được học.

Với phát hiện này cho thấy rằng mặc dù các mô hình có tiềm năng lớn, chúng chưa sẵn để triển khai trong môi trường thực tế mà không có các bước cải tiến thêm. Kết quả của nghiên cứu nhấn mạnh tầm quan trọng của việc đánh giá mô hình trên nhiều tập dữ liệu "out-of-distribution" (dữ liệu có phân phối khác) để có cái nhìn trung thực về khả năng tổng quát hóa của chúng.

4.3 Mô phỏng giao diện hiển thị và kết quả đề xuất

Để minh họa trực quan và kiểm chứng khả năng ứng dụng của mô hình gợi ý việc làm trong thực tiễn, nhóm nghiên cứu đã phát triển một trang web demo hệ thống gợi ý việc làm với mô hình PhoBERT kết hợp PCA.



Hình 4.6: Web demo hệ thống gợi ý công việc

Phần hiển thị chính của giao diện được thiết kế với tiêu đề "Hệ thống đề xuất việc làm", thể hiện rõ ràng mục đích và chức năng của hệ thống nhằm hỗ trợ người dùng trong việc tìm kiếm và lựa chọn công việc phù hợp. Ngay bên dưới tiêu đề là khu vực thông tin hồ sơ ứng viên, cung cấp đầy đủ các trường dữ liệu cơ bản gồm: Tên ứng viên, Ngành nghề, Công việc mong muốn, và Nơi làm việc mong muốn. Ví dụ minh họa: ứng viên Lê Thị Diễm, thuộc ngành nghề Kế toán - Kiểm toán, có nguyện vọng công việc là Nhân viên Kế toán - Thủ kho, và địa điểm làm việc mong muốn tại Thanh Hóa.Tiếp theo là khu vực trình bày danh sách công việc được hệ thống đề xuất. Hệ thống thông báo tổng số công việc phù hợp đã tìm thấy (ví dụ: 26 công việc), đồng thời hiển thị kết quả dưới dạng bảng dữ liệu. Bảng này bao gồm các cột thông tin chính: Tên công việc, Tên công ty, Địa điểm làm việc, Mức lương, và Liên kết tới trang tuyển dụng. Thiết kế này nhằm đảm bảo tính trực quan, giúp người

dùng dễ dàng theo dõi, so sánh và truy cập chi tiết thông tin tuyển dụng một cách thuận tiện và hiệu quả.

5. Kết luận

5.1 Đóng Góp Chính của Nghiên Cứu

Nghiên cứu đã xây dựng thành công một hệ thống gợi ý việc làm hiệu suất cao dành riêng cho tiếng Việt, dựa trên phương pháp Lọc Dựa trên Nội dung (Content-Based Filtering) kết hợp với các kỹ thuật Xử lý Ngôn ngữ Tự nhiên (NLP) hiện đại, bao gồm TF-IDF, Pho2Vec và PhoBERT.

Thông qua quá trình thực nghiệm, nhóm nghiên cứu đã tiến hành so sánh hiệu quả của các mô hình trên tập dữ liệu thực tế, sử dụng các chỉ số đánh giá chuẩn như Precision@30, Recall@30, nDCG@30 và MAP@30. Kết quả cho thấy mô hình PhoBERT kết hợp PCA đạt hiệu năng tối ưu, với **Precision@30 lên tới 93%**, **nDCG@30 đạt 97%** và **MAP@30 đạt 94%**, vượt trội so với các phương pháp truyền thống như TF-IDF (Precision@30 đạt 57%, nDCG@30 đạt 80%). Sự kết hợp giữa PhoBERT và PCA không chỉ giúp nâng cao độ chính xác của hệ thống mà còn góp phần tối ưu hóa tài nguyên tính toán nhờ giảm chiều dữ liệu trong khi vẫn duy trì phần lớn thông tin quan trọng.

Kết quả nghiên cứu đã khẳng định tính khả thi và tiềm năng của việc kết hợp các kỹ thuật NLP ngữ cảnh với phương pháp Lọc Dựa trên Nội dung trong việc giải quyết bài toán gợi ý việc làm, đặc biệt trong bối cảnh dữ liệu và ngôn ngữ tiếng Việt còn nhiều thách thức. Đồng thời, quá trình kiểm thử trên tập dữ liệu mới đã làm rõ hạn chế về khả năng tổng quát hóa của mô hình, qua đó nhấn mạnh nhu cầu cần thiết phải nghiên cứu các giải pháp để nâng cao chất lượng gợi ý.

Nghiên cứu này đã đóng góp một khung tham chiếu thực nghiệm toàn diện cho việc ứng dụng Xử lý Ngôn ngữ Tự nhiên kết hợp với Lọc Dựa trên Nội dung trong hệ thống gợi ý việc làm cho tiếng Việt, góp phần làm phong phú thêm tri thức học thuật và mở ra các định hướng nghiên cứu tiếp theo trong lĩnh vực này.

5.2 Hạn Chế của Nghiên Cứu

Mặc dù nghiên cứu đã đạt được nhiều kết quả tích cực và mang ý nghĩa thực tiễn, hệ thống gợi ý vẫn tồn tại một số hạn chế cần được nhìn nhận rõ ràng để làm cơ sở cho các định hướng cải thiện trong tương lai.

Trước hết, hệ thống hiện tại hoàn toàn dựa trên phương pháp Lọc Dựa trên Nội dung (Content-Based Filtering) mà chưa khai thác được dữ liệu tương tác người dùng. Hệ thống chưa ghi nhận hoặc phân tích các hành vi thực tế như công việc đã xem, đã ứng tuyển, hay thời gian người dùng dành cho mỗi tin tuyển dụng. Sự thiếu vắng các tín hiệu này khiến hệ thống chưa thể ứng dụng các kỹ thuật Lọc Dựa trên Cộng tác (Collaborative Filtering) hoặc mô hình lai (Hybrid Recommender), dẫn đến các gợi ý mặc dù phù hợp về nội dung nhưng chưa chắc đã phản ánh chính xác sở thích tiềm ẩn và hành vi thực tế của người dùng.

Đặc biệt, một hạn chế quan trọng được làm rõ qua thực nghiệm là khả năng tổng quát hóa của mô hình khi áp dụng trên tập dữ liệu hoàn toàn mới. Mặc dù các mô hình, đặc biệt là PhoBERT và PhoBERT (PCA), đã đạt được hiệu suất rất cao trên tập dữ liệu đánh giá ban đầu (Precision@30 lên tới 93% và MAP@30 trên 94%), hiệu suất này đã suy giảm nghiêm trọng khi kiểm thử trên dữ liệu mới. Cụ thể, Precision@30 của PhoBERT (PCA) giảm từ 93% xuống còn khoảng 25%, và MAP@30 giảm từ 94% xuống chỉ còn xấp xỉ 30%. Kết quả này phản ánh rõ rệt vấn đề quá khớp (overfitting), khi mô hình đã học thuộc những đặc trưng riêng của tập dữ liệu huấn luyện mà chưa đủ năng lực khái quát hóa cho dữ liệu thực tế.

Ngoài ra, mô hình còn chịu ảnh hưởng từ chất lượng hồ sơ đầu vào và sự mất cân đối trong phân bố ngành nghề. Các hồ sơ sơ sài, thiếu thông tin hoặc chứa lỗi chính tả làm giảm khả năng mô hình nhận diện chính xác ngữ nghĩa. Đồng thời, sự áp đảo của một số ngành nghề phổ biến như "Bán hàng" hoặc "Hành chính – Văn phòng" dẫn đến thiên lệch trong gợi ý và thu hẹp tính đa dạng nghề nghiệp của hệ thống.

5.3 Hướng Phát Triển Tương Lai

Để khắc phục các hạn chế hiện tại và tiếp tục nâng cao hiệu quả của hệ thống gợi ý việc làm, nghiên cứu hướng đến ba trụ cột phát triển chính trong tương lai: tích hợp dữ liệu hành vi người dùng, tối ưu hóa khả năng bao phủ cơ hội việc làm và đảm bảo tính công bằng trong gợi ý.

Thứ nhất, Phát triển mô hình gợi ý lai (Hybrid Recommendation Model) để cá nhân hóa sâu hơn: Hướng đi quan trọng đầu tiên là xây dựng hệ thống gợi ý lai bằng cách kết hợp phương pháp Lọc cộng tác (Collaborative Filtering) với Lọc dựa trên nội dung (Content-Based Filtering). Bằng cách thu thập và khai thác dữ liệu hành vi người dùng như lượt xem, số lần ứng tuyển, từ khóa tìm kiếm và thời gian tương tác, hệ thống có thể học được sở thích cá nhân và hành vi ngầm của từng người dùng. Nhờ đó, các gợi ý không chỉ đúng về kỹ năng mà còn sát với mong muốn thực tế, mang lại trải nghiệm cá nhân hóa sâu sắc hơn.

Thứ hai, Tăng khả năng bao phủ thông qua hệ thống gợi ý đa tầng: Để cải thiện chỉ số Recall@K, hệ thống sẽ được tiến hành thực nghiệm thêm ở 2 tầng đưa ra gợi ý. Tầng đầu tiên (Candidate Generation) sử dụng các mô hình nhanh và rộng như TF-IDF hoặc Pho2Vec để tạo ra một danh sách rộng gồm hàng trăm công việc tiềm năng. Sau đó, tầng hai (Re-ranking) sẽ dùng mô hình mạnh hơn như PhoBERT để phân tích ngữ nghĩa sâu và xếp hạng lại danh sách này, đẩy các cơ hội "ẩn" nhưng phù hợp lên đầu. Cách tiếp cận này cho phép mở rộng độ bao phủ mà không ảnh hưởng tới độ chính xác của những gợi ý hàng đầu.

Thứ ba, Tối ưu hóa Kỹ thuật Điều chuẩn (Regularization): Kết quả cho thấy các mô hình, đặc biệt là PhoBERT, có độ phức tạp cao và dễ dàng "học thuộc lòng" dữ liệu huấn luyện. Do đó, việc áp dụng các kỹ thuật điều chuẩn mạnh mẽ hơn là một bước đi cần thiết. **Tinh chỉnh các tham số Điều chuẩn hiện có**: Thay vì chỉ sử dụng các giá trị mặc định, cần tiến hành một cuộc tìm kiếm siêu tham số (hyperparameter tuning) có hệ thống cho các kỹ thuật như **Dropout** và **Weight Decay (L2 Regularization)**.

- **Dropout**: Có thể thử nghiệm tăng tỷ lệ dropout trong các tầng cuối của PhoBERT (các tầng fully-connected). Việc này buộc mạng nơ-ron phải học các đặc trưng dự phòng và đa dạng hơn, thay vì phụ thuộc vào một vài nơ-ron cụ thể, từ đó giảm thiểu khả năng ghi nhớ các mẫu nhiễu trong dữ liệu huấn luyện.
- **Weight Decay**: Tăng giá trị của hệ số weight decay sẽ áp đặt một "hình phạt" lớn hơn lên các trọng số có giá trị lớn, khuyến khích mô hình học các trọng số nhỏ hơn và phân tán hơn. Điều này làm cho mô hình trở nên đơn giản hơn và ít nhạy cảm với các biến động nhỏ trong dữ liệu đầu vào.

Áp dụng Early Stopping: Đây là một kỹ thuật điều chuẩn hiệu quả và đơn giản. Trong quá trình huấn luyện, mô hình sẽ được đánh giá định kỳ trên một tập dữ liệu kiểm định (validation set) riêng biệt (không thuộc tập huấn luyện). Quá trình huấn luyện sẽ được dừng lại ngay khi hiệu suất trên tập kiểm định này bắt đầu suy giảm, thay vì tiếp tục cho đến khi đạt số epoch tối đa. Điều này giúp chọn ra được phiên bản mô hình có khả năng tổng quát hóa tốt nhất, trước khi nó bắt đầu quá khớp.

Thứ tư, Tinh chỉnh Mô hình trên Dữ liệu Đa dạng (Fine - Tuning on Diverse Data): Đây là một chiến lược hai giai đoạn nhằm thích ứng hóa mô hình.

Về quy trình: Sau khi đã có mô hình được huấn luyện trên tập dữ liệu gốc (dù đang bị quá khớp), chúng ta sẽ tiếp tục quá trình huấn luyện (fine-tuning) nó trên một tập dữ liệu mới, đa dạng hơn về ngành nghề và nguồn gốc.

Về chiến lược học máy: Trong giai đoạn tinh chỉnh này, điều quan trọng là phải sử dụng một **tốc độ học (learning rate) thấp hơn đáng kể** so với giai đoạn huấn luyện ban đầu. Việc này giúp mô hình từ từ điều chỉnh các trọng số đã học để phù hợp với phân phối dữ liệu mới, mà không phá vỡ các kiến thức ngôn ngữ hữu ích đã học được từ trước (tránh hiện tượng "catastrophic forgetting"). Mục tiêu là để mô hình "thích nghi" chứ không phải "học lại từ đầu".

Thứ năm, Khai thác các kỹ thuật Thích ứng Miền (Domain Adaptation): Đây là hướng tiếp cận tiên tiến nhất, giải quyết trực tiếp vấn đề chênh lệch phân phối dữ liệu (data distribution shift) giữa miền nguồn (dữ liệu gốc) và miền đích (dữ liệu

mới). Mục tiêu là học ra các biểu diễn đặc trưng (feature representations) **bất biến với miền (domain-invariant)**, tức là các đặc trưng này hữu ích cho tác vụ đề xuất bất kể dữ liệu đến từ đâu.

Với phương pháp Huấn luyện Đối nghịch (Adversarial Training), là một kỹ thuật phổ biến trong Domain Adaptation.

- **Kiến trúc**: Xây dựng thêm một mạng nơ-ron phụ gọi là "bộ phân loại miền" (domain classifier). Nhiệm vụ của bộ phân loại này là cố gắng dự đoán xem một biểu diễn đặc trưng (do PhoBERT tạo ra) thuộc về miền dữ liệu gốc hay miền dữ liệu mới.
- **Cơ chế huấn luyện**: Quá trình huấn luyện sẽ có hai mục tiêu đối nghịch nhau:
    1. Mô hình PhoBERT chính (bộ trích xuất đặc trưng) được huấn luyện để vừa tối ưu cho tác vụ đề xuất (matching CV-JD), vừa phải tạo ra các đặc trưng **đánh lừa** được bộ phân loại miền.
    2. Bộ phân loại miền được huấn luyện để ngày càng **giỏi hơn** trong việc phân biệt hai miền.
- **Kết quả**: Khi quá trình huấn luyện hội tụ, bộ trích xuất đặc trưng của PhoBERT sẽ học được cách tạo ra các biểu diễn mà từ đó không thể phân biệt được chúng đến từ miền nào. Đây chính là các đặc trưng tổng quát, bất biến với miền mà chúng ta cần.

Với phương pháp tự huấn luyện (Self-training / Pseudo-labeling). Sử dụng mô hình đã huấn luyện để đưa ra dự đoán trên dữ liệu ở miền đích (dữ liệu mới, không có nhãn). Những cặp (CV, JD) có độ tương đồng cao nhất theo dự đoán của mô hình sẽ được coi là các "nhãn giả" (pseudo-labels) và được thêm vào tập huấn luyện để tiếp tục tinh chỉnh mô hình.

Cuối cùng, Cải thiện chất lượng dữ liệu và tăng cường tính công bằng trong hệ thống: Hệ thống sẽ tích hợp mô-đun “Hỗ trợ và Làm giàu Hồ sơ” (Profile Enrichment) sử dụng NLP để phát hiện lỗi, gợi ý từ khóa, kỹ năng và hoàn thiện thông tin còn thiếu trong CV. Đồng thời, để giảm tình trạng mô hình bị lệch về một số ngành

hoặc khu vực phổ biến (như TP.HCM, Hà Nội; ngành Hành chính – Văn phòng, Bán hàng), nhóm đề xuất áp dụng các kỹ thuật cân bằng dữ liệu như oversampling, gán trọng số nghịch đảo và tái cấu trúc tập huấn luyện theo ngành nghề và vùng địa lý. Điều này sẽ giúp hệ thống đưa ra gợi ý đồng đều và công bằng hơn, tránh bỏ sót cơ hội phù hợp cho các nhóm ngành và địa phương ít dữ liệu.

TÀI LIỆU THAM KHẢO

General Statistics Office. (2024). *Labour and employment report Q1 2024*. Hanoi: Statistical Publishing House.

National Assembly. (2006). *Law on Gender Equality (Law No. 73/2006/QH11)*. Hanoi, Vietnam.

National Assembly. (2019). *Labour Code (Law No. 45/2019/QH14)*. Hanoi, Vietnam.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Apaza, A., Cuadros, F., & Poma, L. (2021). Content-based job recommendation system using NLP and cosine similarity. IEEE Xplore. https://ieeexplore.ieee.org/document/9474226

Apaza, A., Poma, L., & Cuadros, F. (2018). *CV-JD matching using Word2Vec*. In *International Conference on Big Data and Advanced Wireless Technologies*.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. SIAM Review, 37(4), 573–595. https://doi.org/10.1137/1037126

Luthfi, A. H., & Lhaksamana, D. A. (2020). Classification of job applicants using TF-IDF and SVM. Indonesian Journal of Artificial Intelligence and Data Mining, 3(3), 162–167.

Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1037–1042). Association for Computational Linguistics. https://aclanthology.org/2020.findings-emnlp.92/

Tran, K. Q., Nguyen, D. T., Nguyen, T. H., & Bui, T. (2022). Vietnamese hate speech detection using PhoBERT-CNN. Neural Computing and Applications. https://doi.org/10.1007/s00521-022-07023-5

Nguyen, A. T., Dao, M. H., & Nguyen, D. Q. (2020). PhoW2V: Pre-trained Word2Vec embeddings for Vietnamese. In Findings of the Association for Computational Linguistics: EMNLP 2020. https://aclanthology.org/2020.findings-emnlp.126/

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620. https://doi.org/10.1145/361219.361220

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Panchasara, S., Gupta, R. K., & Sharma, A. (2023, August 18–19). AI based job recommendation system using BERT. In *Proceedings of the 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCUBEA58933.2023.10392119

Apaza, H., Vidal, A. A. R. de C., & Saire, J. E. C. (2021). *Job recommendation based on Curriculum Vitae using text mining*. In *Advances in Intelligent Systems and Computing* (Vol. 1363, pp. 1051–1059). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-73100-7_72

Kumar, N., Gupta, M., Sharma, D., & Ofori, I. (2022). Technical job recommendation system using APIs and web crawling. *Computational Intelligence and Neuroscience*, *2022*, Article 7797548. https://doi.org/10.1155/2022/7797548

Narula, R., Kumar, V., Arora, R., & Bhatia, R. (2023, October). Enhancing job recommendations using NLP and machine learning techniques. *Technix International Journal for Engineering Research (TIJER)*, 10(10), a347–a356. ISSN 2349-9249.:https://tijer.org/TIJER/papers/TIJER2310041.pdf

Kwieciński, R., Filipowska, A., Górecki, T., & Dubrov, V. (2023, January 19). *Job recommendations: Benchmarking of collaborative filtering methods for classifieds* (arXiv:2301.07946 [cs.IR]). arXiv. https://doi.org/10.48550/arXiv.2301.07946

Yu, X., Xu, R., Xue, C., Zhang, J., Ma, X., & Yu, Z. (2025, February). *ConFit v2: Improving resume-job matching using hypothetical resume embedding and runner-up hard-negative mining* (arXiv:2502.12361 [cs.CL]). arXiv. https://doi.org/10.48550/arXiv.2502.12361

Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., & Zschech, P. (2025, March 3). *CareerBERT: Matching resumes to ESCO jobs in a shared embedding space for generic job recommendations* (arXiv:2503.02056 [cs.LG]). arXiv. https://doi.org/10.48550/arXiv.2503.02056

Zhao, J., Wang, J., Sigdel, M., Zhang, B., Hoang, P., Liu, M., & Korayem, M. (2021, July 1). *Embedding-based recommender system for job to candidate matching on scale* (arXiv:2107.00221 [cs.IR]). arXiv. https://doi.org/10.48550/arXiv.2107.00221

Deshpande, K. V., Pan, S., & Foulds, J. R. (2020, July). Mitigating demographic bias in AI-based resume filtering. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)* (pp. 1–8). ACM. https://doi.org/10.1145/3386392.3399569

Yang, S., Korayem, M., AlJadda, K., Grainger, T., & Natarajan, S. (2017). Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive statistical relational learning approach. *Knowledge-Based Systems, 136*, 37–45. https://doi.org/10.1016/j.knosys.2017.08.017

Mulay, A., Sutar, S., Patel, J., Chhabria, A., & Mumbaikar, S. (2022). Job recommendation system using hybrid filtering. *ITM Web of Conferences, 44*, 02002. https://doi.org/10.1051/itmconf/20224402002

El-Deeb, R. H., Abdelmoez, W., & El-Bendary, N. (2025). Enhancing e-recruitment recommendations through text summarization techniques. *Information, 16*(4), 333. https://doi.org/10.3390/info16040333

PHẦN 2. PHÂN CÔNG CÔNG VIỆC

TỔNG QUAN QUÁ TRÌNH LÀM VIỆC NHÓM

STT	Tên	MSSV	Vai trò	Công việc	Mức độ hoàn thành
1	Hồ Song Tín	K224111469	Nhóm trưởng	- Thu thập và xử lý dữ liệu ứng viên - Làm slide - Hoàn thiện report báo cáo - Thuyết trình - Giám sát công việc của các thành viên	100%
2	Nguyễn Hoàng Kim	K224111452	Thành viên	- Chuyển đổi văn bản thành vector - Encode metadata - Hoàn thiện report báo cáo - Làm slide - Thuyết trình	100%
4	Huỳnh Hiếu Trung	K224111474	Thành viên	- Thực hiện phân tích khám phá trên dữ liệu ứng viên và công việc - Hoàn thiện report báo cáo - Làm slide - Thuyết trình	100%
5	Huỳnh Huệ Trúc	K224111475	Thành viên	- Thực hiện khớp dữ liệu ứng viên và công việc (candidate-to-job) - Mở rộng thêm danh sách công việc bằng (job-to-job) - Đánh giá mô hình - Làm slide - Hoàn thiện report báo cáo	100%
6	Phạm Thanh Hưng	K224131589	Thành viên	- Thu thập và xử lý dữ liệu công việc - Làm slide	100%

				- Thuyết trình	
7	Phạm Trúc Phương	K224131604	Thành viên	- Hoàn thiện report báo cáo - Format word - Làm slide	60%