ĐOÀN THANH NIÊN CỘNG SẢN HỒ CHÍ MINH  
**BAN CHẤP HÀNH TP. HỒ CHÍ MINH**

|  |
| --- |
|  |

**CÔNG TRÌNH DỰ THI  
GIẢI THƯỞNG SINH VIÊN NGHIÊN CỨU KHOA HỌC EURÉKA  
LẦN THỨ 26 NĂM 2024**

**ĐỊNH DANH NGƯỜI DÙNG THÔNG QUA CÁC TRANG MẠNG XÃ HỘI VÀ ỨNG DỤNG TRONG CÁC HỆ THỐNG KHUYẾN NGHỊ**

LĨNH VỰC NGHIÊN CỨU: CÔNG NGHỆ THÔNG TIN  
CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

Mã số công trình: ……………………………

Thành phố Hồ Chí Minh, tháng 9 năm 2024

# MỤC LỤC

[MỤC LỤC i](#_Toc178105731)

[DANH MỤC CÁC CHỮ VIẾT TẮT iv](#_Toc178105732)

[DANH MỤC BẢNG v](#_Toc178105733)

[DANH MỤC CÁC HÌNH vi](#_Toc178105734)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU 0](#_Toc178105735)

[1.1. Lý do chọn đề tài 0](#_Toc178105736)

[1.2. Mục tiêu nghiên cứu 4](#_Toc178105737)

[1.3. Câu hỏi nghiên cứu 5](#_Toc178105738)

[1.4. Đối tượng và phạm vi nghiêm cứu 6](#_Toc178105739)

[1.5. Phương pháp nghiên cứu 7](#_Toc178105740)

[1.6. Ý nghĩa nghiên cứu 8](#_Toc178105741)

[1.7. Kết cấu nghiên cứu 8](#_Toc178105742)

[CHƯƠNG 2: CƠ SỞ LÝ LUẬN VÀ MÔ HÌNH NGHIÊN CỨU 10](#_Toc178105743)

[2.1. Các khái niệm liên quan 10](#_Toc178105744)

[2.1.1. Định danh người dùng (User Identification) 10](#_Toc178105745)

[2.1.2. Các nền tảng mạng xã hội (Social networking platforms) 10](#_Toc178105746)

[2.2. Các phương pháp sử dụng 11](#_Toc178105747)

[2.2.1. Edit distance 11](#_Toc178105748)

[2.2.2. Nhúng từ (Word embedding) 12](#_Toc178105749)

[2.2.3. Trích xuất từ khóa (Keyword extracting) 17](#_Toc178105750)

[2.3. Các khái niệm liên quan 22](#_Toc178105751)

[2.3.1. Độ tương tự Cosine (Cosine similarity) 22](#_Toc178105752)

[2.3.2. Độ chính xác (Accuracy) 23](#_Toc178105753)

[2.4. Tóm tắt chương II 24](#_Toc178105754)

[CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU 25](#_Toc178105755)

[3.1. Quy trình nghiên cứu 25](#_Toc178105756)

[3.1.1. Tiền xử lý dữ liệu (Data preprocessing) 25](#_Toc178105757)

[3.1.2. Tính khoảng cách Levenshtein (Edit distance) 26](#_Toc178105758)

[3.1.3. Trích xuất từ khóa (Keyword extracting) 27](#_Toc178105759)

[3.1.4. Nhúng văn bản (Word embedding) 27](#_Toc178105760)

[3.1.5. Tính độ tương tự Cosine (Cosine similarity) 28](#_Toc178105761)

[3.2. Phương pháp xây dựng mô hình định danh người dùng 28](#_Toc178105762)

[3.2.1. Mô hình xử lý ngôn ngữ tự nhiên (Natural language processing model) 28](#_Toc178105763)

[3.2.2. Giới thiệu về ngôn ngữ Python và công cụ Jupyter Notebook 29](#_Toc178105764)

[3.3. Tóm tắt chương III 29](#_Toc178105765)

[CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU 31](#_Toc178105766)

[4.1. Mẫu nghiên cứu 31](#_Toc178105767)

[4.2. Nhập thư viện 31](#_Toc178105768)

[4.2.1. NLTK (Natural Language Toolkit) 31](#_Toc178105769)

[4.2.2. Sentence-Transformers và Cosine Similarity 32](#_Toc178105770)

[4.2.3. PKE (Python Keyword Extraction) 33](#_Toc178105771)

[4.2.4. Vai trò của các thư viện trong nghiên cứu 33](#_Toc178105772)

[4.3. Kết quả về độ chính xác của thuật toán 34](#_Toc178105773)

[4.3.1. Độ tương tự giữa tên người dùng của các tài khoản 34](#_Toc178105774)

[4.3.2. Độ tương tự giữa nội dung các bài viết 34](#_Toc178105775)

[4.3.3. Độ tương tự giữa các chủ đề người dùng quan tâm 34](#_Toc178105776)

[4.4. Tóm tắt chương IV 34](#_Toc178105777)

[CHƯƠNG 5: KẾT LUẬN VÀ ỨNG DỤNG 36](#_Toc178105778)

[5.1. Kết luận 36](#_Toc178105779)

[5.2. Ứng dụng 37](#_Toc178105780)

[5.2.1. Đề xuất việc làm 38](#_Toc178105781)

[5.2.2. Đề xuất sản phẩm 40](#_Toc178105782)

[5.2.3. Triển khai ứng dụng 41](#_Toc178105783)

[5.3. Hạn chế và hướng nghiên cứu trong tương lai 41](#_Toc178105784)

[5.3.1. Hạn chế 42](#_Toc178105785)

[5.3.2. Hướng nghiên cứu trong tương lai 42](#_Toc178105786)

[5.4. Tóm tắt chương V 43](#_Toc178105787)

[TÀI LIỆU THAM KHẢO 45](#_Toc178105788)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Nghĩa đầy đủ** |
| ID | Identification |
| AI | Artificial Intelligence |
| CV | Curriculum Vitae |
| NLP | Natural language processing |
| HTML | Hyper Text Markup Language |
| ROUGE | Recall-oriented Understudy for Gisting Evaluation |
| CRFs | Conditional Random Fields |
| SVM | Support Vector Machines |
| URL | Uniform Resource Locato |
| NLTK | Natural Language Toolkit |
| PKE | Python Keyword Extraction |
| YAKE | Yet Another Keyword Extractor |

# DANH MỤC BẢNG

[Bảng 4.3.1: Kết quả về độ chính xác dựa trên độ tương tự giữa tên người dùng của các tài khoản 34](#_Toc178105789)

[Bảng 4.3.2: Kết quả về độ chính xác dựa trên độ tương tự giữa nội dung các bài viết. 34](#_Toc178105790)

[Bảng 4.3.3: Kết quả về độ chính xác dựa trên độ tương tự giữa chủ đề mà người dùng quan tâm 34](#_Toc178105791)

[Bảng 4.4.1: Tổng hợp độ chính xác của các thuật toán sử dụng 35](#_Toc178105792)

# DANH MỤC CÁC HÌNH

[Hình 1‑1: Ứng dụng thực tiễn: hồ sơ chuyên gia 1](#_Toc178105793)

[Hình 1‑2: Ứng dụng thực tiễn: Trực quan hoá sở thích người dùng và gợi ý sản phẩm 2](#_Toc178105794)

[Hình 1‑3: Ứng dụng thực tiễn: Phân cụm người dùng 3](#_Toc178105795)

[Hình 2‑1: Thống kê dữ liệu pre-training 16](#_Toc178105796)

[Hình 2‑2:Mô tả mô hình trích xuất từ khóa 18](#_Toc178105797)

[Hình 3‑1: Quy trình nghiên cứu 29](#_Toc178105798)

[Hình 5‑1: Hệ thống đề xuất việc làm. 39](#_Toc178105799)

[Hình 5‑2: Hệ thống đề xuất sách 40](#_Toc178105800)

[Hình 5‑3: Mã QR triển khai ứng dụng 41](#_Toc178105801)



# GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU

## Lý do chọn đề tài

Trong thời đại số hóa hiện nay, mạng xã hội đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của con người. Đặc biệt sau khoảng thời gian giãn cách xã hội đợt đại dịch COVID-19 làm cho việc sử dụng mạng xã hội ngày càng rộng rãi [1]. Các nền tảng như Facebook và Instagram không chỉ là nơi người dùng chia sẻ thông tin cá nhân, hình ảnh, và nội dung mà còn là kho dữ liệu quý giá giúp các tổ chức, doanh nghiệp thu thập và phân tích hành vi của người dùng. Tuy nhiên, việc một người có thể sở hữu nhiều tài khoản trên các nền tảng khác nhau dẫn đến thách thức lớn trong việc định danh và xác định rằng các tài khoản này có thuộc về cùng một cá nhân hay không. Việc xác định được các tài khoản trên hai nền tảng này thuộc cùng một cá nhân sẽ giúp ta xây dựng được hồ sơ người dùng (user profile). Hồ sơ người dùng có nhiều ứng dụng thực tế trong nhiều lĩnh vực khác nhau như hồ sơ chuyên gia (expert profiles), gợi ý sản phẩm (item recommendation) trên các nền tảng thương mại điện tử (Shopee, Lazada, Tiki, Netflix), tiếp thị trực tuyến (online marketing), gom cụm người dùng (user clustering) và kết nối việc làm (job matching) trên các nền tảng như LinkedIn. Những ứng dụng này sẽ giúp tăng hiệu quả tìm kiếm chuyên gia trong các mạng lưới chuyên nghiệp về một lĩnh vực cụ thể, nâng cao gợi ý sản phẩm trong các hệ thống thương mại điện tử, tối ưu hóa chiến lược tiếp thị đúng mục tiêu, tinh chỉnh phân khúc người dùng và cách mạng hóa quy trình kết nối việc làm. Chi tiết của một số ứng dụng vừa nên có thể được tóm lược như sau.

**Hồ sơ chuyên gia (expert profiles)**: Ứng dụng này liên quan đến việc xây dựng hồ sơ cho các nhà nghiên cứu chuyên về một lĩnh vực cụ thể. Hệ thống ArnetMining là một ví dụ. Một yêu cầu cơ bản trong hệ thống này là xây dựng hồ sơ cho từng chuyên gia (nhà nghiên cứu). Mỗi hồ sơ thường chứa thông tin cơ bản ( ảnh đại diện, tên cơ quan, vị trí và liên hệ), quá trình học tập (trường đại học, sau đại học, chuyên ngành), lĩnh vực nghiên cứu và các ấn phẩm khoa học chính đã được công bố. Mỗi hồ sơ có thể được tạo ra bằng cách sử dụng thông tin một phần được trích xuất từ nhiều nguồn trên môi trường internet. Chẳng hạn, thông tin cơ bản và quá trình học tập có thể đến từ trang chủ của chuyên gia/nhà nghiên cứu hoặc từ một trang web giới thiệu về chuyên gia đó, thông tin về các ấn phẩm khoa học có thể được tích hợp từ các thư viện số trực tuyến (DBLP, IEEE hoặc ACM), và lĩnh vực nghiên cứu có thể được suy ra từ cơ sở dữ liệu thu thập được. Hình 1‑1 minh họa ví dụ về hồ sơ của một nhà nghiên cứu. Phần trên bên trái hiển thị trang chủ của nhà nghiên cứu, phần dưới bên trái hiển thị trang DBLP/ACM liệt kê các ấn phẩm khoa học, và phần bên phải hiển thị hồ sơ “được suy luận” đầy đủ của nhà nghiên cứu đó, bao gồm lĩnh vực nghiên cứu được khai thác từ các ấn phẩm xuất bản. Nếu tất cả các hồ sơ nhà nghiên cứu được xây dựng chính xác, chúng ta sẽ có một tập hợp lớn các cơ sở dữ liệu cấu trúc tốt về các nhà nghiên cứu trên khắp thế giới. Cơ sở dữ liệu này sau đó có thể phục vụ cho các nhiệm vụ A diagram of a document

Description automatically generated with medium confidencethực tiễn khác nhau, chẳng hạn như tìm kiếm chuyên gia về một chủ đề hoặc dự án cụ thể.

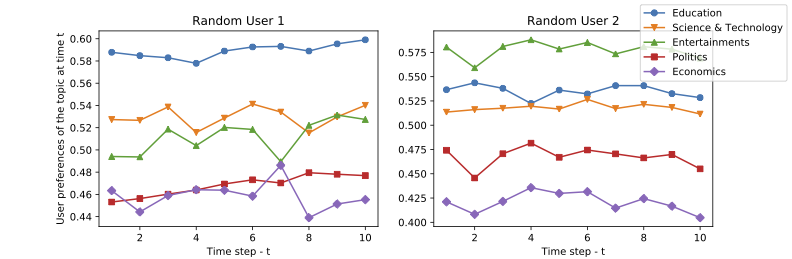
Hình 1‑1: Ứng dụng thực tiễn: hồ sơ chuyên gia

(nguồn: <https://dspace.jaist.ac.jp/dspace/bitstream/10119/17781/14/paper.pdf>)

**Gợi ý sản phẩm (item recommendation):** Người dùng trên các hệ thống thương mại điện tử thường thay đổi sở thích của họ về các loại sản phẩm khác nhau theo thời gian. Một yêu cầu quan trọng trong các hệ thống này là nắm bắt chính xác sự thay đổi này để gợi ý các sản phẩm phù hợp tại một [khoảng] thời điểm cụ thể. Suy luận sở thích của người dùng và trực quan hóa kết quả có thể nâng cao hiệu quả của quá trình gợi ý sản phẩm. Hình 1‑2 minh họa kết quả trực quan hóa mức độ “quan tâm” của hai người dùng Twitter đối với năm chủ đề (Giáo dục, Kinh tế, Khoa học & Công nghệ, Giải trí và Chính trị) theo thời gian. Một số thông tin hữu ích có thể được được kết xuất từ quá trình trực quan hoá này như sau:

* Người dùng 1 quan tâm nhất đến Giáo dục, và
* Người dùng 2 quan tâm nhất đến Giải trí, và
* Sự quan tâm của người dùng 1 đối với Chính trị đang tăng lên theo thời gian.

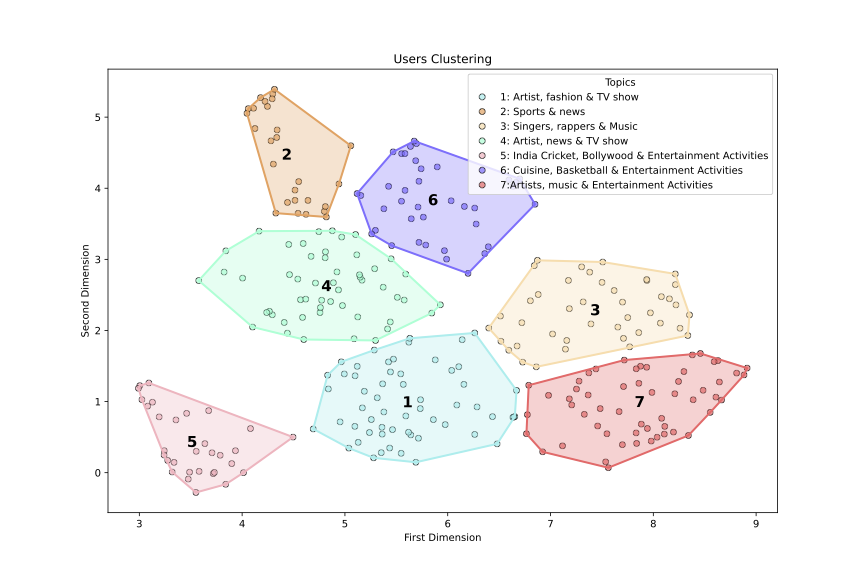
Những quan sát này rất hữu ích cho việc gợi ý sản phẩm. Ví dụ, chúng ta có thể đề xuất một phiếu giảm giá cho các khóa học trực tuyến cho người dùng 1, hoặc gợi ý vé xem phim hoặc thiết bị thể thao cho người dùng 2.



Hình 1‑2: Ứng dụng thực tiễn: Trực quan hoá sở thích người dùng và gợi ý sản phẩm

(nguồn: <https://dspace.jaist.ac.jp/dspace/bitstream/10119/17781/14/paper.pdf>)

**Phân cụm người dùng (user clustering):** Một nhóm người dùng “tương tự” có thể được phân cụm động vào cùng một cụm dựa trên sở thích chung về một chủ đề cụ thể. Vấn đề nghiên cứu này trở nên phức tạp hơn khi dữ liệu đầu vào là các văn bản ngắn và bài toán được xét trong ngữ cảnh động. Phương pháp của ứng viên đề xuất có thể cung cấp những cách tiếp cận hữu ích để kết xuất thông tin giá trị trong bài toán này. Ví dụ, đo lường ngữ nghĩa giữa các cặp từ khóa trong sở thích của người dùng có thể tiết lộ rằng những người dùng có mối quan tâm “tương tự” trong một khoảng thời gian cụ thể. Thông tin này cho phép hệ thống gợi ý đề xuất một sản phẩm phù hợp không chỉ cho cá nhân mà còn cho cả nhóm khách hàng trong hệ thống. Ngoài ra, những thông tin này có thể giúp hệ thống xác định sở thích của những người dùng “có vẻ” khác biệt đáng kể trong ngữ cảnh tĩnh nhưng lại cùng chung một mối quan tâm (hội tụ) về một sản phẩm/chủ đề nhất định tại một số thời điểm trong ngữ cảnh động. Ví dụ, chủ đề quan tâm của một chính trị gia và một vận động viên có thể khác nhau, nhưng họ có thể chia sẻ sở thích chung về điện ảnh, ẩm thực hoặc du lịch trong một khoảng thời gian cụ thể. Kết quả trực quan của quá trình phân cụm người dùng được minh họa trong Hình 1‑3.



Hình 1‑3: Ứng dụng thực tiễn: Phân cụm người dùng

(nguồn: <https://dspace.jaist.ac.jp/dspace/bitstream/10119/17781/14/paper.pdf>)

Nhận thấy tầm quan trọng và tính thực tiễn của việc xây dựng hồ sơ người dùng, nhóm chúng tôi quyết định thực hiện đề tài này để xây dựng thuật toán định danh người dùng, từ đó trích xuất và tổng hợp thông tin của họ từ các nền tảng khác nhau. Bên cạnh đó nhóm chúng tôi cũng phát triển ứng dụng đề xuất việc làm và sách dựa trên những chủ đề mà người dùng quan tâm.

## Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là xây dựng và phát triển một phương pháp hiệu quả nhằm định danh người dùng trên các nền tảng mạng xã hội khác nhau, cụ thể là Facebook và Instagram. Từ đó, nghiên cứu sẽ:

* **Xác định mức độ tương đồng:** giữa các tài khoản mạng xã hội, một người dùng có thể có sở hữu hoặc dùng đồng thời cả hai hoặc nhiều mạng xã hội. Vì thế mà việc giúp nhận diện liệu các tài khoản trên hai nền tảng này có thuộc về cùng một cá nhân hay không là rất quan trọng, giúp ta xác định được nội dung có thể được gợi ý cho người dùng ấy.
* **Ứng dụng các thuật toán và phương pháp phân tích dữ liệu**: Chúng tôi sẽ sử dụng những thuật toán phổ biến như khoảng cách Levenshtein (Edit Distance), nhúng văn bản (Word Embedding), và khoảng cách Cosine (Cosine Similarity) để so sánh tên tài khoản và nội dung của người dùng trên hai nền tảng. Từ đó xác định được liệu những nội dung kia có cùng đến từ một người dùng hay là không, qua đó có thể gợi ý đến những nội dung liên quan đến các nội dung trước đó mà người dùng đã đăng trên cùng nhiều nền tảng mạng xã hội (ở đây để dễ so sánh, ta có thể nói đến việc người dùng đăng ở hai nền tảng phổ biến như là Facebook hay là Instagram
* **Xây dựng một hệ thống khuyến nghị**: Về nội dung chính, ở đây chúng tôi sẽ sử dụng các thuật toán để xây dựng 2 hệ thống chính đó là: đề xuất việc làm (Job Recommendation) và đề xuất sách (Book Recommendation). Bằng cách xem xét mà nội dung người dùng đã đăng, từ đó đưa ra những gợi ý về quyển sách có nội dung giống với bài đăng người dùng, đó là Book Recommendation. Về đề xuất việc làm, ta có thể sử dụng thông tin CV mà người dùng đã gửi cho các công ty qua các bài thảo luận ở các nền tảng mạng xã hội, xem xét các yếu tố, kinh nghiệm, kĩ năng người gửi, từ đó gợi ý công việc phù hợp cho họ.
* **Đánh giá hiệu quả của mô hình**: Để có thể đánh giá hiệu suất mô hình hoạt động ta có thể triển khai bằng cách so sánh các kết quả từ các mô hình khác nhau, tối ưu hóa và cải tiến mô hình dựa trên các chỉ số hiệu suất. Nếu hiệu suất đạt kết quả mong muốn (Ở đây, ta lấy việc so sánh 3 thuật toán làm ví dụ, sẽ được nói chi tiết ở phương pháp nghiên cứu), ta có thể triển khai mô hình ổn định và tránh được việc xuất hiện các bug (lỗi hệ thống) trong quá trình triển khai mô hình.

Nghiên cứu này hy vọng mang đến những cải tiến thực tế trong việc định danh người dùng trên mạng xã hội, giải quyết các vấn đề liên quan đến dữ liệu trùng lặp, tài khoản giả mạo, và tăng cường sự đồng bộ trong các hệ thống mạng xã hội.

## Câu hỏi nghiên cứu

Nghiên cứu sẽ tập trung vào việc trả lời các câu hỏi chính sau:

1. Làm thế nào để nhận diện được người dùng trên Facebook và Instagram với độ chính xác cao?
2. Các phương pháp xử lý ngôn ngữ tự nhiên (NLP) như nhúng văn bản (Word Embedding và khoảng cách Cosine (Cosine Similarity) có hiệu quả trong việc so sánh tên và nội dung tài khoản trên các nền tảng khác nhau không?
3. Khoảng cách Levenshtein (Edit Distance) có phải là một công cụ tốt để đo lường sự tương đồng giữa tên người dùng trên các nền tảng mạng xã hội khác nhau hay không?
4. Sự khác biệt giữa các dữ liệu thu thập từ Facebook và Instagram có gây khó khăn trong việc định danh người dùng không? Nếu có, làm cách nào để giải quyết các khó khăn này?
5. Làm thế nào để mô hình nhận diện người dùng có thể được áp dụng trong các ứng dụng thực tế, bao gồm bảo mật và đồng bộ dữ liệu?

## Đối tượng và phạm vi nghiêm cứu

* **Đối tượng nghiên cứu**: Người dùng mạng xã hội trên hai nền tảng Facebook và Instagram, với các thông tin liên quan như tên tài khoản, bài đăng, nội dung hình ảnh, và các hoạt động khác.
* **Phạm vi nghiên cứu**:
  + Về dữ liệu: Dữ liệu sẽ bao gồm các thông tin công khai từ tài khoản mạng xã hội Facebook và Instagram. Ở đây chúng tôi sẽ lấy khoảng 300 người dùng và xấp xỉ 9000 bài đăng mà họ đã đăng trên cùng 2 nền tảng
  + Về kỹ thuật: Nghiên cứu sẽ tập trung vào các thuật toán xử lý ngôn ngữ tự nhiên và các phương pháp phân tích dữ liệu như Edit Distance, Word Embedding, Cosine Similarity, và các kỹ thuật trích xuất từ khóa. Việc xác định từ khóa rất quan trọng, nó sẽ giúp ta xác định liệu các bài đăng ở nhiều nền tảng có thuộc về cùng 1 người dùng hay không.
  + Về không gian: Nghiên cứu sẽ chỉ giới hạn trên hai nền tảng Facebook và Instagram mà không mở rộng sang các mạng xã hội khác để tránh việc quá tải dữ liệu cũng như hiện tượng trùng lặp quá nhiều dữ liệu, ảnh hưởng đến quá trình xây dựng dữ liệu cũng như phân tích dữ liệu
  + Về thời gian: Dữ liệu sẽ được thu thập trong một khoảng thời gian cụ thể để đảm bảo tính đồng nhất và giảm thiểu các yếu tố thay đổi có thể gây ảnh hưởng đến quá trình phân tích.

## Phương pháp nghiên cứu

Để thực hiện nghiên cứu, chúng tôi sử dụng các phương pháp sau:

* **Phương pháp thu thập dữ liệu**: Dữ liệu sẽ được thu thập từ các tài khoản công khai trên Facebook và Instagram, bao gồm tên tài khoản, nội dung bài đăng, hình ảnh và các tương tác người dùng khác. Dữ liệu sẽ được thu thập tự động qua API hoặc các công cụ hỗ trợ khác để đảm bảo tính chính xác và đầy đủ. Ngoài ra dữ liệu đã được gắn nhãn thủ công để phục vụ cho việc phân tích và đánh giá mô hình.
* **Phương pháp làm sạch dữ liệu**: Trước khi tiến hành phân tích, dữ liệu sẽ được làm sạch để loại bỏ các yếu tố nhiễu như thông tin không liên quan, các ký tự đặc biệt, và những từ dừng (stopwords). Quá trình này cũng bao gồm việc chuẩn hóa văn bản, chuyển đổi các từ về dạng gốc (stemming) và loại bỏ các từ không mang ý nghĩa ngữ nghĩa (emoji).
* **Phương pháp phân tích dữ liệu**:
  + **Khoảng cách Levenshtein (Edit Distance)** sẽ được sử dụng để so sánh mức độ tương đồng giữa các tên tài khoản trên Facebook và Instagram.
  + **Trích xuất từ khóa (Keyword Extraction)**: PKE, YAKE! và PositionRank sẽ được áp dụng để xác định các từ khóa chính trong nội dung của người dùng.
  + **Nhúng văn bản (Word Embedding)**: Các mô hình nhúng văn bản sẽ chuyển đổi các đoạn văn bản thành các vector số để phân tích sâu hơn mối quan hệ ngữ nghĩa.
  + **Khoảng cách Cosine (Cosine Similarity)** sẽ đo lường độ tương đồng giữa các vector văn bản để so sánh nội dung bài đăng.
* **Phương pháp đánh giá mô hình**: Sử dụng chỉ số đo lường là độ chính xác để đánh giá hiệu quả của các thuật toán cũng như độ tin cậy của những thuật toán trên cũng như mô hình đã sử dụng.

## Ý nghĩa nghiên cứu

Nghiên cứu này mang lại nhiều ý nghĩa cả về mặt lý thuyết và thực tiễn:

* **Về lý thuyết**: Nghiên cứu đóng góp vào việc phát triển các thuật toán định danh người dùng trên mạng xã hội, đồng thời mở rộng kiến thức về xử lý ngôn ngữ tự nhiên (NLP) trong việc phân tích dữ liệu từ các nền tảng mạng xã hội.
* **Về thực tiễn**: Nghiên cứu này sẽ hỗ trợ các tổ chức và doanh nghiệp trong việc định danh người dùng trên nhiều nền tảng mạng xã hội, từ đó tăng cường khả năng quản lý các tài khoản đa nền tảng một cách hiệu quả. Đồng thời, nghiên cứu còn góp phần nâng cao trải nghiệm người dùng thông qua việc đồng bộ hóa dữ liệu, cải thiện độ chính xác của các hệ thống gợi ý và quảng cáo, giúp cá nhân hóa nội dung tốt hơn và tối ưu hóa các chiến lược tiếp thị trực tuyến.

## Kết cấu nghiên cứu

Nội dung nghiên cứu sẽ được chia thành các chương như sau:

* **CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI NGHIÊN CỨU** – Trình bày bối cảnh, lý do chọn đề tài, mục tiêu, câu hỏi, phạm vi và phương pháp nghiên cứu.
* **CHƯƠNG 2: CƠ SỞ LÝ LUẬN VÀ MÔ HÌNH NGHIÊN CỨU** – Giới thiệu về các nền tảng mạng xã hội như Facebook và Instagram, các phương pháp định danh người dùng đã có, và lý thuyết về xử lý ngôn ngữ tự nhiên.
* **CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU** – Trình bày chi tiết về quy trình thu thập, làm sạch dữ liệu, các thuật toán và mô hình được áp dụng trong nghiên cứu.
* **CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU** – Mô tả quá trình thực nghiệm và trình bày các kết quả thu được từ việc áp dụng các mô hình định danh người dùng.
* **CHƯƠNG 5: KẾT LUẬN VÀ ỨNG DỤNG** – Tổng kết các kết quả nghiên cứu, những hạn chế và gợi ý ứng dụng lý thuyết /thực tiễn cho các nghiên cứu tiếp theo.

# CƠ SỞ LÝ LUẬN VÀ MÔ HÌNH NGHIÊN CỨU

## Các khái niệm liên quan

### Định danh người dùng (User Identification)

**ID người dùng (User ID)** là một phương thức xác thực phổ biến trong các hệ thống máy tính, mạng lưới và trên Internet. Mỗi người dùng đều có một mã định danh duy nhất để phân biệt với những người dùng khác, giúp đảm bảo tính bảo mật và chính xác trong việc quản lý tài khoản.

**Mô hình Định danh người dùng (User Identification Model)** được phát triển nhằm mục tiêu xác định và đối chiếu danh tính người dùng thông qua các thông tin hồ sơ công khai trên mạng xã hội. Đối với mỗi tài khoản cá nhân trên các nền tảng như Facebook, mô hình sẽ phân tích các thông tin như tên người dùng, tên hiển thị và các trường văn bản khác. Sau đó, mô hình so sánh và tìm kiếm sự tương đồng giữa tài khoản này với tài khoản trên các nền tảng khác như Instagram. Bằng cách trích xuất các đặc trưng (features) từ các trường văn bản, mô hình có thể đối chiếu các thông tin trên nhiều nền tảng mạng xã hội khác nhau. Điều này giúp xác định danh tính người dùng một cách hiệu quả, từ đó hỗ trợ các ứng dụng trong việc quản lý và phân tích thông tin người dùng [2].

### Các nền tảng mạng xã hội (Social networking platforms)

Mạng xã hội là một hệ sinh thái trực tuyến đa dạng, các trang web và ứng kết nối hàng tỷ người trên toàn cầu. Các nền tảng này không chỉ đơn thuần là nơi để kết nối với bạn bè và gia đình mà còn là công cụ mạnh mẽ để chia sẻ thông tin, tạo lập cộng đồng, mối quan hệ và thúc đẩy các hoạt động xã hội. Mọi người có thể kết nối với những người khác trong cùng khu vực, gia đình, bạn bè và những người có cùng sở thích. Mạng xã hội đã trở thành một trong những ứng dụng quan trọng nhất của Internet ngày nay và là một phần không thể thiếu của cuộc sống hiện đại. Và các nền tảng mạng xã hội ngày càng phổ biến hơn trong thời gian gần đây, đặc biệt do sau đại dịch COVID-19 [1].

Sự phát triển vượt bậc của các mạng xã hội trực tuyến, ví dụ: Facebook, Twitter, Google Talk, Wechat, v.v., đã cách mạng hóa một cách mới để mọi người kết nối, thể hiện bản thân và chia sẻ nội dung với những người khác trong xã hội mạng ngày nay [3]. Các trang web này cũng cho phép các nhà tiếp thị tiếp cận đúng đối tượng mục tiêu của họ. Dựa và nội dung mà người dùng chia sẻ trên mạng xã hội, các doanh nghiệp có thể dựa vào đó để phân thích hành vi cũng như sở thích của họ, từ đó xây dựng các hệ thống gợi ý sản phẩm tự động. Tuy nhiên các doanh nghiệp hay các nhà nghiên cứu có thể gặp một thách thức lớn là chỉ với tập hợp các bài viết ngắn của mỗi người dùng trên mạng xã hội, nhưng cần phải trích xuất một tập hợp các từ ngữ phản ánh sở thích của họ về các chủ đề khác nhau [4].

## Các phương pháp sử dụng

### Edit distance

Phương pháp này nhằm xác định người dùng sử dụng hồ sơ công khai của chính họ. Đối với tài khoản của một cá nhân trên nền tảng xã hội (ví dụ: Facebook, Instagram,…). Chúng tôi rút trích user\_id của các tài khoản người dùng từ dữ liệu thu thập được của người dùng và cố gắng để khớp hai tài khoản của chung một người dùng. Chúng tôi sử dụng độ đo Levenshtein distance (edit distance) để tính khoảng cách giữa trường văn bản userid thuộc tài khoản của cá nhân trên các nền tảng khác nhau, cụ thể là Facebook và Instagram. Đối với bộ dữ liệu này, mỗi người dùng sẽ có 1 user\_id riêng và userid của mỗi người sẽ giống nhau ở tất cả các nền tảng. Mục đích của việc gán nhãn này giúp kiểm tra được tài khoản Facebook nào ứng với tài khoản Instagram là của chung 1 người dùng. Sau đây, chúng tôi sẽ trình bày phương pháp đo độ tương tự được sử dụng trong mô hình.

Levenshtein distance (edit distance) là một chỉ số dùng để do độ tương tự giữa 2 chuỗi ký tự. Cụ thể, đó là số lần chỉnh sửa tối thiểu của một ký tự như thêm, xóa hoặc thay thế để biến đổi một từ thành một từ khác. Đối với hai chuỗi a và b, điều này được định nghĩa:

Trong đó: bằng 0 khi và bằng 1 khi , và sẽ hiển thị khoảng cách giữa i kí tự đầu tiên của chuỗi a và j kí tự đầu tiên của chuỗi b [5].

Khoảng cách Levenshtein càng nhỏ thì các chuỗi càng giống nhau:

* Nếu xét đến ký tự cuối cùng cả hai chuỗi giống nhau, thì edit distance của chúng sẽ bằng nhau tính từ ký tự thứ hai đến ký tự cuối cùng của chúng.
* Nếu ký tự cuối của chúng khác nhau, thì edit distance sẽ là giá trị nhỏ nhất của việc chèn, xóa và thay thế ký tự cuối của chuỗi.

Kết quả của độ đo khoảng cách Levenshtein distance (edit distance) giúp xác định được những userid nào có sự tương đồng lớn nhất. Do đó, nhận diện được tài khoản Facebook nào thật sự là chính chủ ứng với tài khoản Instagram của chính họ.

### Nhúng từ (Word embedding)

Mô hình nhận diện người dùng thông qua nội dung các bài đăng sử dụng Embedding Vectors.

Mục đích của phương pháp này nhằm xác định các mẫu bài đăng của người dùng. Giả sử, nếu hai người dùng đang đăng các bài viết có các thuộc tính: nội dung văn bản, hình ảnh, địa điểm… giống nhau hoặc tương tự về mặt ngữ nghĩa trên hai tài khoản của họ, thì hai tài khoản này có thể thuộc về cùng một người sở hữu. Dữ liệu các bài đăng ghi lại nội dung các bài đăng của họ và điều này quan trọng để chúng tôi áp dụng phương pháp word embedding để chuyển đổi văn bản thành vector và sau đó so sánh độ tương tự bằng các hàm khoảng cách để tìm ra những tài khoản tương đồng với nhau trên hai nền tảng Facebook và Instagram.

Chúng tôi tập trung vào việc tổng hợp dữ liệu các bài đăng của người dùng, điều này khả quan khi một người dùng sở hữu hai tài khoản Facebook và Instagram. Một nguồn thông tin về một sự kiện mà họ đăng tải trên hai nền tảng có thể giống hoàn toàn về toàn bộ nội dung, hoặc sử dụng các từ có ý nghĩa tương tự. Ví dụ: Khi đăng tải về trạng thái của bản thân, nội dung đăng tải trên Instagram: “Cảm thấy hạnh phúc khi ở bên gia đình”, trong khi đó nội dung đăng tải trên Facebook: “Hạnh phúc là khi ở bên gia đình”. Trường hợp các thông tin đăng tải giống nhau hoàn toàn là khách quan, chiếm tỷ lệ nhỏ nhưng điều này giúp đánh giá được độ hiệu quả của phương pháp khi áp dụng với một lượng dữ liệu khá phức tạp như thế này. Do đó, để sử dụng phương pháp này, chúng tôi tiến hành kết hợp lại dữ liệu các bài đăng của một tài khoản người dùng và sau đó lược bỏ các ký tự không quan trọng (ví dụ: emoji, dấu chấm, dấu phẩy, từ nối…) để dữ liệu sau khi chuyển đổi sẽ cho ra kết quả độ đo tương tự chính xác hơn. Tiếp theo, sử dụng mô hình embedding để chuyển đổi sang vector, sử dụng độ đo cosine để tính khoảng cách vector và dựa vào kết quả độ đo cosine để xác định tài khoản người dùng trên hai nền tảng. Khi độ đo cosine càng lớn, thì cho thấy được độ tương tự giữa hai vector càng lớn, suy ra được hai tài khoản này có độ tương đồng lớn về nội dung các bài đăng. Lựa chọn các cặp vector cho kết quả độ đo cosin lớn nhất sẽ nhận diện được tài khoản của người dùng.

Sau đây, chúng tôi sẽ trình bày về khái niệm của một số kỹ thuật sử dụng trong mô hình.

* **Embeding Vectors**

Word embeddings là các mô hình toán học giúp mã hóa mối quan hệ giữa các từ trong một không gian vectơ. Chúng được tạo ra thông qua quá trình huấn luyện không giám sát (an unsupervised training process) dựa trên thông tin về mặt ngữ nghĩa, ngữ pháp hay sự tương đồng về từ ngữ của chúng trong các chuỗi văn bản. Các mối quan hệ được mã hóa bao gồm các thuộc tính ngữ nghĩa và cú pháp của từ.

Ví dụ, biểu diễn vectơ từ ngữ đã được chứng minh có khả năng tiết lộ các tương đồng ngữ nghĩa và các nhóm từ có liên quan về mặt ngữ nghĩa. Biểu diễn vectơ từ ngữ có thể giúp xác định các từ có ý nghĩa tương đồng (ví dụ: "vua" và "hoàng hậu"). Biểu diễn vectơ từ ngữ có thể giúp nhóm các từ có liên quan về mặt ngữ nghĩa (ví dụ: "hoa", "lá"…)

Trong topolpgy (Tô pô học) biểu diễn (embedding) được hiểu là một ánh xạ bảo toàn cấu trúc đơn ánh f: X → Y, giữa hai cấu trúc toán học X và Y. Thuật toán biểu diễn vectơ từ ngữ (f) học một không gian tích nội (Y) để bảo toàn cấu trúc ngôn ngữ trong một kho văn bản tham chiếu D (X), dựa trên một tập từ vựng V. Cấu trúc trong D được phân tích dựa trên mối quan hệ giữa các từ được suy ra từ sự đồng xuất hiện của chúng, theo một định nghĩa ngữ cảnh nhất định.

* **E5 – small – v2**

E5-small-v2 là một trong những phiên bản của mô hình ngôn ngữ dạng transformer, được tối ưu hóa cho hiệu quả và tốc độ xử lý nhanh hơn mà vẫn giữ được độ chính xác trong các tác vụ nhất định.

Trong word embedding, "E5-small-v2" đề cập đến một mô hình với kiến trúc nhỏ gọn (small) dùng để biểu diễn từ vựng (word embedding) dưới dạng các vector số. Các mô hình này học cách mã hóa các từ (hoặc cụm từ, câu) thành các vector sao cho các từ có ngữ nghĩa gần nhau sẽ có các vector gần nhau trong không gian vector.

E5-Small-v2 sử dụng kiến trúc Transformer với một số cải tiến để giảm kích thước và tăng hiệu suất. Các cải tiến này bao gồm việc sử dụng:

* Bộ mã hóa Sparse Attention: Giảm số lượng phép toán cần thiết để tính toán Attention.
* Mạng nơ-ron Feed-forward có kích thước nhỏ hơn: Giảm kích thước của mô hình.
* Kỹ thuật (Knowledge Distillation): Chuyển giao kiến thức từ một mô hình lớn sang một mô hình nhỏ hơn.

E5-Small-v2 có hiệu suất cạnh tranh bộ ngôn ngữ tiếng Anh trong bộ đánh giá MTEB [6], và biến thể được tinh chỉnh theo hướng dẫn để vượt qua các mô hình chỉ dành cho tiếng Anh có kích thước tương đương. Để thể hiện khả năng đa ngôn ngữ của các mô hình, mô hình cũng được đánh giá hiệu suất của chúng trên bộ đánh giá truy xuất thông tin đa ngôn ngữ MIRACL [7] với 16 ngôn ngữ và trên nhiệm vụ khai thác song ngữ [8] [9] với hơn 100 ngôn ngữ.

Bằng cách công khai trọng số mô hình, các nhà thực hành có thể tận dụng các mô hình này cho các nhiệm vụ như truy xuất thông tin, đo lường sự tương đồng ngữ nghĩa và phân cụm trên nhiều ngôn ngữ khác nhau. Ở đây, chúng tôi sẽ sử dụng mô hình E5-Small-v2 cho việc chuyển đổi các bài viết sang vector để phục vụ cho việc định danh người dùng thông qua các bài đăng của họ trên các nền tảng xã hội.

* **GTE-base**

Các mô hình GTE chủ yếu dựa trên khung BERT [10]và hiện tại cung cấp ba kích thước mô hình khác nhau, bao gồm GTE-large, GTE-base, và GTE-small. Các mô hình GTE được huấn luyện trên một tập dữ liệu lớn gồm các cặp văn bản có liên quan, bao phủ nhiều lĩnh vực và kịch bản khác nhau. Điều này cho phép các mô hình GTE được áp dụng vào nhiều tác vụ hạ nguồn của embedding văn bản, bao gồm tìm kiếm thông tin, tương đồng ngữ nghĩa giữa văn bản, xếp hạng lại văn bản, và các ứng dụng khác.

Nó bắt đầu với nền tảng tương tự như BERT, một mô hình nổi tiếng trong xử lý ngôn ngữ tự nhiên (NLP), nhưng sau đó trải qua một quá trình huấn luyện độc đáo gồm hai giai đoạn:

**Huấn luyện không giám sát** (Unsupervised Pre-training): Lấy cảm hứng từ công trình trước đây [11] [12] [13], mô hình ban đầu được huấn luyện trước trên các cặp văn bản tự nhiên được trích xuất từ nhiều nguồn khác nhau. GTE được "nuôi dưỡng" bởi một lượng khổng lồ gồm 800 triệu cặp văn bản, được thu thập từ nhiều nguồn khác nhau trên internet trang web, mạng xã hội, tài liệu học thuật, và nhiều nguồn khác mà không cần sử dụng bất kỳ lời nhắc đặc thù nào cho từng tác vụ [14]. Dữ liệu sau khi được huấn luyện của bài viết gốc được minh họa trong bảng dưới đây.

A table with numbers and text

Description automatically generated

Hình ‑: Thống kê dữ liệu pre-training

**Tinh chỉnh có giám sát** (Supervised Fine-tuning): Trong giai đoạn tinh chỉnh có giám sát, chúng tôi sử dụng các tập dữ liệu nhỏ hơn với chú thích của con người về mức độ liên quan giữa hai đoạn văn bản, kết hợp với các tiêu cực cứng (hard negatives) tùy chọn được khai thác bởi một bộ truy xuất bổ sung để tạo thành các bộ ba văn bản. Quá trình này giúp xử lý các nhiệm vụ đối xứng (semantic textual similarity) và không đối xứng (passage retrieval). Để đảm bảo độ bao phủ đa dạng, chúng tôi đã thu thập dữ liệu từ nhiều nhiệm vụ và lĩnh vực khác nhau như tìm kiếm web, suy luận ngữ nghĩa tự nhiên, kiểm chứng thông tin và các cặp câu đồng nghĩa. [14]

Tổng cộng, khoảng 3 triệu cặp dữ liệu đã được sử dụng để tinh chỉnh mô hình, nhằm đạt được hiệu suất cao hơn cho cả hai loại nhiệm vụ đối xứng và không đối xứng. Sau đó, mô hình rèn luyện kỹ năng của mình trên 3 triệu bộ ba văn bản chất lượng cao từ các tác vụ khác nhau như truy vấn tìm kiếm, hỏi đáp, và diễn giải lại.

Qua quá trình này, GTE học cách phân biệt giữa các văn bản có ngữ nghĩa tương đồng và không tương đồng. Dữ liệu mà bài báo gốc dùng để huấn luyện được sử dụng bởi các nghiên cứu trước đó [15] [16].

### Trích xuất từ khóa (Keyword extracting)

Mô hình này tương tự với mô hình sử dụng phương pháp embedding vector. Điểm khác là chúng tôi sử dụng thêm phương pháp trích xuất từ khóa – keyphrase. Thay vì như ở mô hình trước, dữ liệu các bài đăng được gom và kết hợp lại, sau đó loại bỏ các phần tử không quan trọng bằng phương pháp thủ công. Mô hình này chúng tôi sử dụng thư viện PKE để tiến hành rút trích các từ khóa quan trọng trong văn bản, giúp tóm tắt được nội dung văn bản và nhận biết được chủ đề chính trong chính văn bản đó. Sau đó, sẽ tiến hành chuyển đổi dữ liệu văn bản sang vector và tính độ đo cosine để xem xét kết quả nhằm xác định tài khoản của người dùng.

Sau đây, chúng tôi sẽ trình bày về một số kỹ thuật sử dụng trong mô hình này.

**Keyphrase**

Keyphrase - Trích xuất từ khóa (hay còn gọi là phát hiện từ khóa hoặc phân tích từ khóa) là một kỹ thuật phân tích văn bản tự động trích xuất các từ và cụm từ được sử dụng nhiều nhất và quan trọng nhất từ một văn bản. Nó giúp tóm tắt nội dung của văn bản và nhận dạng các chủ đề chính được thảo luận.

Trích xuất từ khóa sử dụng học máy (machine learning) - trí tuệ nhân tạo (AI) kết hợp với xử lý ngôn ngữ tự nhiên (NLP) để phân tích ngôn ngữ của con người, giúp máy tính có thể hiểu và phân tích được. Nó được sử dụng để tìm từ khóa từ tất cả các dạng văn bản: tài liệu thông thường và báo cáo kinh doanh, bình luận trên mạng xã hội, diễn đàn trực tuyến và đánh giá, tin tức, v.v.

Ví dụ: Chúng ta cần trong đánh giá sau khi mua hàng của khách hàng có nhắc tới những gì liên quan đến sản phẩm cũng như họ thích gì và không thích gì keyphrase sẽ giúp chúng ta làm điều đó:

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động

Hình 2‑2:Mô tả mô hình trích xuất từ khóa

Keyphrase hoạt động dựa trên các phương pháp thống kê đơn giản (Simple Statistical Approaches), Linguistic Approaches, Graph-based Approaches (sử dụng đồ thị), các phương pháp học máy:

* Simple Statistical Approaches – có nhiều phương pháp thống kê có thể áp dụng như word frequency, word collocations và co-occurrences, TF-IDF (short for term frequency–inverse document frequency), và RAKE (Rapid Automatic Keyword Extraction).
* Linguistic Approaches – trích xuất keyword dự trên thông tin về cú pháp, hình thái từ đó trong văn bản (ví dụ: cụm danh từ hoặc danh từ - do nó có thể chứa nhiều thông tin) hoặc thông tin về nghĩa của các từ
* Graph-based Approaches - Phương pháp phổ biến nhất dựa trên đồ thị là mô hình TextRank. Đồ thị có thể được định nghĩa là một tập các đỉnh (vertex) với các kết nối giữa chúng.
* Machine learning – có nhiều thuật toán học máy có thể dùng trong trích xuất keyword như Support Vector Machines (SVM), Conditional Random Fields (CRFs), ROUGE (recall-oriented understudy for gisting evaluation).
* **YAKE!**

YAKE! là một phương pháp không giám sát để trích xuất từ khóa, không yêu cầu sử dụng từ điển hay huấn luyện trước trên tập dữ liệu cụ thể. Nó sử dụng các đặc trưng thống kê được trích xuất từ văn bản để xác định các từ khóa quan trọng từ các tài liệu đơn lẻ, hỗ trợ đa ngôn ngữ và phù hợp với nhiều loại văn bản khác nhau.

**Cách thức hoạt động**

* Bước 1: Bước quan trọng nhất để cải thiện hiệu quả của thuật toán, tài liệu được tiền xử lý thành định dạng máy có thể đọc được để xác định các từ ứng viên tiềm năng.

Tại đây, văn bản được làm sạch và chuyển đổi sang định dạng có thể đọc được bằng máy: các phần quan trọng được xác định và các phần không có thông tin hoặc gây nhiễu được loại bỏ như dấu câu, email, URL và xác định các phần quan trọng.

* Bước 2: Các từ được đại diện dưới dạng một tập hợp các đặc trưng thống kê.

Thuật toán được xây dựng tính toán các chỉ số thống kê liên quan đến các thuật ngữ (terms) trong văn bản, bao gồm tần suất xuất hiện của chúng và các mối quan hệ đồng xuất hiện giữa các từ trong cửa sổ trượt 𝑤 như: Tính tần suất từ (TF), vị trí xuất hiện, xử lý viết tắt và viết hoa, ma trận đồng xuất hiện. Sau khi thuật toán thực thi, chúng ta thu được danh sách các thuật ngữ với các chỉ số thống kê (tần suất, vị trí, viết tắt, viết hoa), cùng với ma trận đồng xuất hiện ghi nhận mối quan hệ giữa các từ

* Bước 3: Tính toán số diểm cho term

Công thức này tính điểm số của mỗi thuật ngữ bằng cách kết hợp các đặc trưng đã tính toán trước đó. Cụ thể, và đại diện cho vị trí và mức độ liên quan của từ, trong khi , , và phản ánh cách mà thuật ngữ xuất hiện trong văn bản.

Công thức này giúp cân bằng giữa các đặc trưng khác nhau để xác định mức độ quan trọng của mỗi thuật ngữ trong văn bản. Kết quả của thuật toán là danh sách các thuật ngữ với điểm số của chúng, giúp cho quá trình trích xuất từ khóa và sắp xếp từ khóa theo mức độ quan trọng.

* Bước 4: Tạo n-gram và tính điểm cho các cụm từ khóa ứng viên

Tại bước này này tạo ra các cụm từ khóa ứng viên (n-grams) từ các câu trong văn bản, với độ dài tối đa n. Thuật toán chỉ chọn những từ quan trọng (như danh từ, từ viết hoa) và loại bỏ các cụm từ bắt đầu hoặc kết thúc bằng từ dừng (stopwords). Mỗi khi cụm từ khóa ứng viên được tìm thấy, trọng số của nó sẽ được tăng lên. Kết quả là danh sách các cụm từ khóa ứng viên với trọng số tương ứng, giúp xác định các cụm từ quan trọng trong văn bản. Tiếp theo thuật toán sẽ tính điểm Skw (score keyword) cho từng từ khóa ứng viên. Nếu một từ không phải là từ dừng, điểm của nó được cộng và nhân vào tổng điểm của cụm. Nếu là từ dừng, thuật toán tính xác suất bigram để điều chỉnh trọng số. Sau khi tính toán, các từ khóa được sắp xếp theo điểm Skw, giúp xác định mức độ quan trọng của từng từ khóa trong văn bản. Kết quả là danh sách các từ khóa đã được sắp xếp theo mức độ quan trọng.

* Bước 5: Loại bỏ trùng lặp và xếp hạng

Thuật toán này kiểm tra từng từ khóa ứng viên và so sánh độ tương tự giữa từ khóa đó với các từ khóa đã được chọn. Nếu độ tương tự (DistanceSimilarity) giữa hai từ khóa lớn hơn ngưỡng cho trước, từ khóa mới sẽ bị bỏ qua, tránh sự trùng lặp trong danh sách từ khóa cuối cùng. Cuối cùng, trả về danh sách các từ khóa không trùng lặp, sắp xếp theo điểm số Skw của chúng. [17]

**Ứng dụng vào bài nghiên cứu:**

Đối với bài nghiên cứu của chúng tôi, dữ liệu người dùng thu thập được chủ yếu là các bài đăng có nội dung ngắn mà, thuật toán YAKE hoạt động hiệu quả trên các văn bản ngắn, điều này làm cho nó lý tưởng cho các loại tài liệu ngắn gọn hoặc dữ liệu rời rạc như tin tức, bài viết trên blog, hoặc các bài đăng trên mạng xã hội [17], vì vậy YAKE thích hợp với mô hình của cúng tôi. Tuy vậy YAKE có một vài điểm bất lợi là thiếu ngữ cảnh sâu và không phù hợp với văn bản có cấu trúc phức tạp do YAKE! chỉ dựa trên các đặc trưng cục bộ như tần suất và đồng xuất hiện, điều này có thể dẫn đến việc bỏ sót các yếu tố ngữ nghĩa sâu hơn hoặc mối liên hệ dài hạn giữa các thuật ngữ.

* **PositionRank**

PositionRank: Một phương pháp không giám sát để trích xuất từ khóa từ các tài liệu học thuật. Thuật toán này cải tiến PageRank truyền thống bằng cách tích hợp thông tin về vị trí xuất hiện của các từ trong tài liệu để cải thiện hiệu suất trích xuất từ khóa.

**Cách thức hoạt động của thuật toán:**

* Bước 1: Xây dựng đồ thị ở cấp độ từ.

Đầu tiên, thuật áp dụng bộ lọc loại từ (part-of-speech filter) bằng cách sử dụng bộ công cụ NLP Stanford, chỉ chọn danh từ và tính từ làm từ ứng cử viên. Sau đó, tạo một đồ thị từ, trong đó mỗi từ được chọn tương ứng với một nút. Hai nút được kết nối với nhau nếu từ tương ứng của chúng cùng xuất hiện trong một cửa sổ gồm **w** từ liên tiếp trong nội dung của tài liệu. Trọng số của cạnh giữa hai nút được tính dựa trên số lần hai từ đó xuất hiện cùng nhau trong cửa sổ **w** từ. Đồ thị này có thể được xây dựng dưới dạng có hướng hoặc không có hướng, nhưng loại đồ thị không ảnh hưởng đáng kể đến hiệu suất trích xuất cụm từ khóa [18]. Vì vậy, nghiên cứu này sử dụng đồ thị không có hướng.

* Bước 2: Thiết kế thuật toán PageRank với trọng số theo vị trí.

Ban đầu, mỗi từ trong đồ thị có điểm PageRank bằng nhau. Sau đó thuật toán tính toán đệ quy để cập nhật điểm của từng từ dựa trên trọng số vị trí và điểm của các từ liên kết với nó. Với các từ xuất hiện đầu tiên và nhiều lần trong chuỗi văn bản thì nó sẽ được đánh trọng số cao hơn. Việc tổng hợp trọng số vị trí cho một từ nhất định nhằm mục đích mang lại độ tin cậy cao hơn cho các từ xuất hiện thường xuyên bằng cách tính đến trọng số vị trí của mỗi lần xuất hiện.

* Bước 3: Tạo thành các cụm từ thành viên.

Các từ ứng viên có vị trí liền kề nhau trong tài liệu sẽ được kết hợp thành cụm từ và được đánh giá dựa trên tổng điểm của các từ riêng lẻ cấu thành cụm từ đó, có độ dài tối đa là ba từ. Cuối cùng, các cụm từ được tính điểm bằng cách sử dụng tổng điểm của các từ riêng lẻ trong cụm từ đó [19]. Các cụm từ có điểm cao nhất được xuất ra dưới dạng dự đoán (tức là cụm từ khóa được dự đoán cho tài liệu) [20].

**Ứng dụng trong bài nghiên cứu:**

PositionRank tích hợp thông tin về vị trí của từ trong tài liệu, giúp các từ xuất hiện sớm có trọng số cao hơn, bên cạnh đó còn xây dựng đồ thị từ tài liệu nhưng ưu tiên vị trí từ, giúp nó nổi bật trong một số ngữ cảnh nhất định [20]. Vì vậy PositionRank sẽ giúp chúng tôi trích suất được các từ khóa thuộc các chủ đề mà người dùng quan tâm từ đó tăng khả năng định danh người dùng.

## Các khái niệm liên quan

### Độ tương tự Cosine (Cosine similarity)

Độ tương đồng cosin (Cosine Similarity) là một phép đo được sử dụng để đánh giá mức độ giống nhau giữa hai vectơ trong không gian tích vô hướng. Nó được tính bằng cosin của góc giữa hai vector, cho biết hướng của hai vectơ có gần nhau hay không. Độ tương đồng cosin thường được sử dụng để đo lường mức độ giống nhau giữa các văn bản trong phân tích văn bản. Góc giữa hai mũi tên này sẽ phản ánh mức độ giống nhau của chúng. Góc càng nhỏ (gần 0 độ), cosin của góc càng lớn (gần 1), cho thấy hai vector càng giống nhau. Ngược lại, góc càng lớn (gần 90 độ), cosin của góc càng nhỏ (gần 0), cho thấy hai vector càng khác nhau. Giả sử chúng ta có hai vector x và y đại diện cho nội dung của hai văn bản. Bằng cách sử dụng độ tương đồng cosin như một hàm tương đồng, công thức được tính như sau:

Trong đó:

‖x‖: Độ dài (độ lớn) của vector x

‖y‖: Độ dài (độ lớn) của vector y

### Độ chính xác (Accuracy)

Trong đề tài này, chúng tôi sử dụng Accuracy để đánh giá độ hiệu quả của các mô hình mà chúng tôi đã đề xuất. Từ đó lựa chọn được phương pháp tối ưu để áp dụng vào việc nhận diện người dùng thông qua các nền tảng mạng xã hội.

Accuracy (độ chính xác) chỉ đơn giản đánh giá mô hình thường xuyên dự đoán đúng đến mức nào. Độ chính xác là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu.

Được tính theo công thức:

Tuy nhiên, một mô hình có độ chính xác cao chưa hẳn đã tốt. Accuracy lộ rõ hạn chế khi được sử dụng trên bộ dữ liệu không cân bằng (imbalanced dataset). Chỉ số accurracy phụ thuộc vào yếu tố khách quan của mô hình, do đó chúng tôi sẽ lựa chọn những mô hình phù hợp để đánh giá đúng các đặc điểm của dữ liệu..

## Tóm tắt chương II

Đầu tiên, phương pháp định danh người dùng (User Identification) dựa trên ID người dùng được sử dụng để đối chiếu tài khoản trên các nền tảng như Facebook và Instagram. Tiếp theo, phương pháp Levenshtein distance (Edit Distance) tính toán số lần chỉnh sửa cần thiết giữa các chuỗi ký tự để đo độ tương đồng giữa các tài khoản. Để tăng độ chính xác, nhúng từ (Word Embedding) được sử dụng để chuyển đổi văn bản thành vector, cho phép so sánh độ tương đồng giữa các bài đăng. Ngoài ra, phương pháp trích xuất từ khóa (Keyword Extraction) giúp tóm tắt nội dung và xác định các chủ đề chính. Cuối cùng, độ tương đồng cosine (Cosine Similarity) được áp dụng để đo lường mức độ giống nhau giữa các vector, hỗ trợ việc phân tích nội dung và xác định tài khoản người dùng trên các nền tảng mạng xã hội khác nhau.

# PHƯƠNG PHÁP NGHIÊN CỨU

## Quy trình nghiên cứu

### Tiền xử lý dữ liệu (Data preprocessing)

Tiền xử lý dữ liệu là một bước không thể thiếu trong quy trình phân tích và khai thác dữ liệu, đặc biệt khi dữ liệu được thu thập từ các nguồn khác nhau, như từ các tài khoản người dùng mạng xã hội lớn như Facebook và Instagram. Dữ liệu thô từ các nền tảng này thường chứa nhiều yếu tố gây nhiễu, chẳng hạn như ký tự đặc biệt, thẻ HTML, và các từ dừng (stopwords) hay các emoji mà họ dùng để biểu đạt cảm xúc, lỗi chính tả trong lúc soạn. Những yếu tố này không chỉ ảnh hưởng đến độ chính xác của mô hình phân tích mà còn có thể làm sai lệch kết quả, dẫn đến các quyết định không chính xác. Chính vì vậy, bước tiền xử lý dữ liệu tập trung vào việc làm sạch và chuẩn hóa dữ liệu nhằm chuẩn bị cho các giai đoạn phân tích tiếp theo. Các bước cụ thể của quy trình này bao gồm:

* **Chuyển đổi văn bản thành chữ thường**: Việc đồng nhất định dạng chữ hoa và chữ thường giúp loại bỏ sự khác biệt không cần thiết giữa các từ, từ đó đảm bảo tính nhất quán trong quá trình phân tích.
* **Loại bỏ ký tự trong dấu ngoặc vuông**: Dữ liệu mạng xã hội thường chứa các thông tin không liên quan được đặt trong các dấu ngoặc vuông. Việc loại bỏ những ký tự này giúp loại bỏ nhiễu và tập trung vào các thông tin có giá trị.
* **Loại bỏ đường dẫn web và thẻ HTML**: Các liên kết URL và thẻ HTML thường không mang lại thông tin ngữ nghĩa có ích cho phân tích nội dung. Chúng gây nhiễu và làm giảm hiệu suất của các mô hình phân tích ngôn ngữ, do đó việc loại bỏ là cần thiết.
* **Loại bỏ dấu câu và ký tự xuống dòng**: Các ký tự này thường không có giá trị về mặt ngữ nghĩa và chỉ làm phân tán dữ liệu. Việc loại bỏ giúp đảm bảo văn bản trở nên liền mạch, dễ dàng cho việc xử lý sau này.
* **Loại bỏ các từ chứa số và emoji**: Những ký tự không chứa ý nghĩa ngữ nghĩa hoặc chỉ mang tính chất biểu tượng cảm xúc sẽ được loại bỏ để giảm thiểu nhiễu trong quá trình phân tích.
* **Loại bỏ từ dừng (Stopwords)**: Các từ dừng như "là" (as, be, like), "và" (and), "nhưng" (but) thường ít mang thông tin hữu ích khi phân tích ngữ nghĩa và nội dung chính. Việc loại bỏ những từ này giúp tập trung vào các từ khóa quan trọng hơn.
* **Sử dụng thuật toán stemming:** Các thuật toán này giúp rút gọn từ về dạng gốc, giảm thiểu sự phân tán do biến thể ngữ pháp hoặc ngữ nghĩa của từ vựng. Thuật toán này đảm bảo rằng các biến thể khác nhau của cùng một từ sẽ được hiểu như nhau trong quá trình phân tích, hạn chế được việc xuất hiện những trùng lặp không cần thiết

Kết quả của bước tiền xử lý dữ liệu là chúng ta có một tập dữ liệu được làm sạch, tinh gọn, và sẵn sàng cho các bước phân tích tiếp theo. Tập dữ liệu này sẽ cung cấp nền tảng mạnh mẽ cho việc xây dựng và triển khai các mô hình phân tích, đảm bảo rằng các mô hình không bị ảnh hưởng bởi các yếu tố gây nhiễu

### Tính khoảng cách Levenshtein (Edit distance)

Khoảng cách Levenshtein, hay còn gọi là Edit Distance, là một phương pháp hữu ích trong việc đo lường sự tương đồng giữa hai chuỗi ký tự. Đặc biệt, khi so sánh tên tài khoản trên các nền tảng mạng xã hội như Facebook và Instagram, phương pháp này giúp đánh giá số lượng thao tác tối thiểu cần thực hiện để chuyển đổi một chuỗi ký tự thành chuỗi khác. Các thao tác này bao gồm việc chèn, xóa hoặc thay thế ký tự.

Trong lúc chúng ta định danh người dùng, việc một cá nhân sử dụng nhiều tài khoản trên các nền tảng khác nhau là điều phổ biến. Các tên tài khoản có thể chỉ khác nhau đôi chút, do lỗi đánh máy hoặc do họ cố ý tạo ra các biến thể của tên mình để cách điệu theo xu hướng của nền tảng đó. Khoảng cách Levenshtein có thể phát hiện ra các tài khoản với tên tương tự, giúp nhận diện được liệu chúng có thuộc về cùng một cá nhân hay không.

Phương pháp này đặc biệt hữu ích trong các trường hợp phát hiện tài khoản trùng lặp hoặc các tài khoản giả mạo, mang lại hiệu quả cao trong quá trình phân tích và nhận diện người dùng trên mạng xã hội.

### Trích xuất từ khóa (Keyword extracting)

Trích xuất từ khóa là một trong những bước quan trọng trong việc phân tích nội dung văn bản, đặc biệt là khi cần xác định các chủ đề chính hoặc mối quan tâm của người dùng trên mạng xã hội. Quá trình này giúp lọc ra những từ hoặc cụm từ mang tính đại diện và có giá trị ngữ nghĩa cao, giúp làm nổi bật các thông tin cốt lõi từ nội dung bài đăng của người dùng.

Chúng tôi đã áp dụng các mô hình trích xuất từ khóa tiên tiến như YAKE! và PositionRank từ thư viện PKE (Python Keyphrase Extraction). Các mô hình này có khả năng tự động xác định và xếp hạng các từ khóa quan trọng dựa trên tần suất xuất hiện, vị trí trong văn bản và các đặc trưng ngữ nghĩa khác. Mỗi mô hình đều có những ưu điểm riêng biệt, giúp phân tích và trích xuất từ khóa một cách chính xác và hiệu quả.

Việc trích xuất từ khóa không chỉ giúp nhận diện các chủ đề mà người dùng quan tâm, mà còn hỗ trợ trong việc phân loại nội dung và tìm ra những điểm tương đồng giữa các tài khoản khác nhau. Điều này rất hữu ích trong việc nhận diện các tài khoản có nội dung tương tự hoặc được quản lý bởi cùng một cá nhân.

### Nhúng văn bản (Word embedding)

Nhúng văn bản (Word Embedding) là một kỹ thuật quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), giúp chuyển đổi văn bản thành các vector số trong không gian đa chiều. Thay vì chỉ xem xét các từ dưới dạng chuỗi ký tự, phương pháp này giúp biểu diễn các từ và cụm từ theo ngữ nghĩa, từ đó cho phép các mô hình máy học hiểu và phân tích nội dung một cách chính xác hơn.

Trong nghiên cứu này, chúng tôi đã sử dụng hai mô hình nhúng văn bản phổ biến là E5-small-v2 và GIST. Các mô hình này đã được kiểm chứng trong nhiều nghiên cứu trước đây và cho thấy hiệu quả cao trong việc nhận diện ngữ nghĩa của văn bản. Khi áp dụng vào nội dung mạng xã hội, nhúng văn bản giúp phân tích ngữ nghĩa của các bài đăng một cách sâu sắc, cho phép mô hình hiểu được mối quan hệ giữa các từ trong câu, từ đó cải thiện đáng kể độ chính xác của mô hình nhận diện người dùng.

### Tính độ tương tự Cosine (Cosine similarity)

Độ tương tự Cosine là một phương pháp hữu hiệu để đo lường sự tương tự giữa hai vector trong không gian đa chiều, đặc biệt là khi phân tích nội dung văn bản. Khi được áp dụng trong bối cảnh mạng xã hội, khoảng cách Cosine giúp so sánh mức độ tương đồng giữa nội dung bài đăng hoặc hồ sơ người dùng dựa trên các vector đã được chuyển đổi từ văn bản.

Độ tương tự Cosine đo lường cosine của góc giữa hai vector, và giá trị càng gần 1 thì độ tương đồng giữa các văn bản càng cao. Phương pháp này đặc biệt hữu ích trong việc phát hiện các tài khoản có nội dung tương tự, từ đó giúp nhận diện các tài khoản có khả năng thuộc về cùng một cá nhân.

## Phương pháp xây dựng mô hình định danh người dùng

### Mô hình xử lý ngôn ngữ tự nhiên (Natural language processing model)

Các mô hình xử lý ngôn ngữ tự nhiên (NLP) đóng vai trò cốt lõi trong việc phân tích và hiểu nội dung từ các bài đăng trên mạng xã hội. Trong nghiên cứu này, chúng tôi đã sử dụng nhiều mô hình NLP khác nhau, từ những phương pháp đơn giản như khoảng cách Levenshtein đến các phương pháp tiên tiến hơn như nhúng văn bản và tính độ tương tự Cosine.

Sự kết hợp của các mô hình NLP này giúp chúng tôi xây dựng được một hệ thống nhận diện người dùng chính xác và hiệu quả. Việc áp dụng các mô hình hiện đại không chỉ tăng độ chính xác trong việc phân tích nội dung mà còn cho phép nhận diện người dùng một cách thông minh hơn, ngay cả khi họ sử dụng các biến thể ngôn ngữ khác nhau trên nhiều nền tảng.

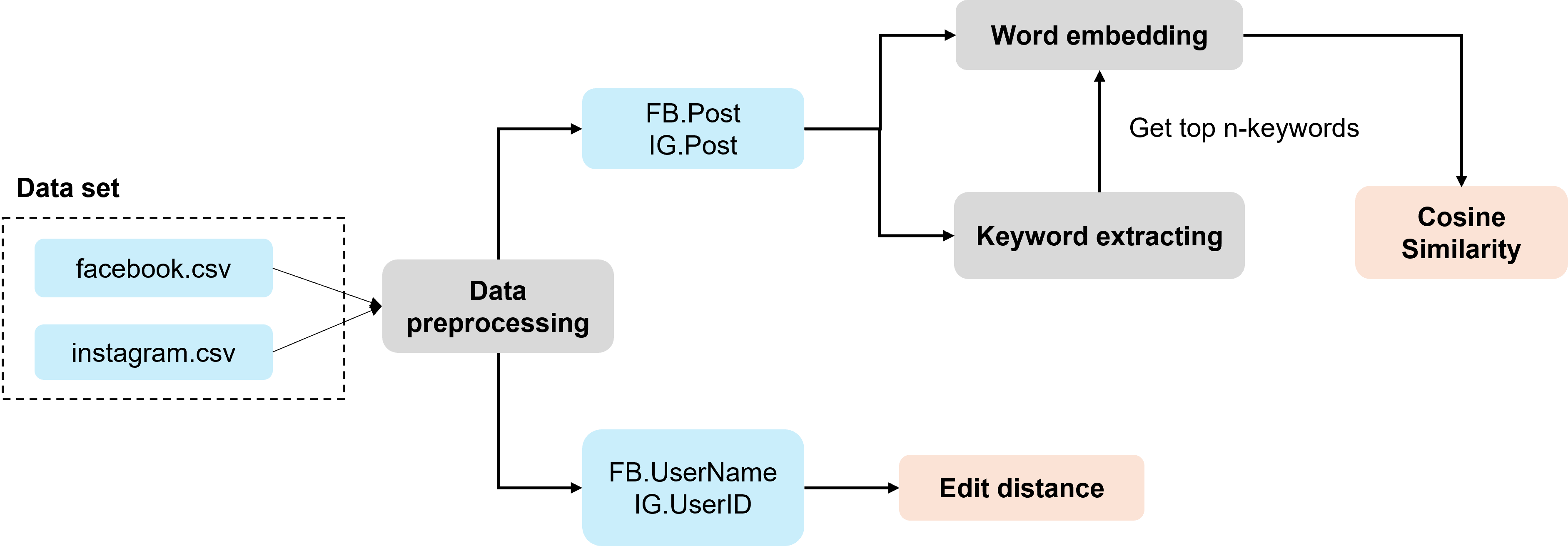
### Giới thiệu về ngôn ngữ Python và công cụ Jupyter Notebook

Python là ngôn ngữ lập trình phổ biến và mạnh mẽ, đặc biệt phù hợp cho các nhiệm vụ phân tích dữ liệu và xử lý ngôn ngữ tự nhiên. Trong nghiên cứu này, chúng tôi sử dụng Python kết hợp với Jupyter Notebook - một công cụ tương tác mạnh mẽ, giúp thực hiện các phép tính, thử nghiệm mô hình, và trực quan hóa dữ liệu một cách dễ dàng và hiệu quả.

Python cung cấp nhiều thư viện mạnh mẽ như numpy, pandas, nltk, sklearn, và PKE. Những thư viện này hỗ trợ toàn bộ quy trình từ thu thập, xử lý dữ liệu đến xây dựng và đánh giá mô hình, giúp đảm bảo rằng nghiên cứu đạt được kết quả tốt nhất.

## Tóm tắt chương III

Chương III đã trình bày chi tiết quy trình nghiên cứu và các phương pháp chính được sử dụng trong việc xây dựng mô hình nhận diện người dùng trên mạng xã hội.



Hình 3‑1: Quy trình nghiên cứu

Các bước quan trọng như tiền xử lý dữ liệu, tính khoảng cách Levenshtein, nhúng văn bản và tính độ tương tự Cosine đã được thảo luận kỹ lưỡng. Ngoài ra, việc trích xuất từ khóa cũng đóng vai trò quan trọng trong việc hiểu và phân tích nội dung của người dùng, từ đó ta có thể xây dựng được các mô hình gợi ý hoặc quảng cáo những sản phẩm có nội dung tương tự bài đăng của họ.

Sự kết hợp của các phương pháp xử lý ngôn ngữ tự nhiên (NLP) và các công cụ mạnh mẽ như Python và Jupyter Notebook đã giúp tạo ra một nền tảng phân tích dữ liệu vững chắc, đảm bảo cho việc phát triển và tối ưu hóa việc nhận dạng tài khoản đa nền tảng và xây dựng mô hình gợi ý sản phẩm (Recommendation System).

# KẾT QUẢ NGHIÊN CỨU

## Mẫu nghiên cứu

Mẫu nghiên cứu của chúng tôi bao gồm dữ liệu từ hai nền tảng mạng xã hội phổ biến, cụ thể là Instagram và Facebook. Mẫu dữ liệu bao gồm các thông tin chính sau:

* User\_id: Mã định danh duy nhất của người dùng trên mỗi nền tảng, giúp nhận diện từng cá nhân trong hệ thống trên Instagram.
* Username: Tên người dùng được hiển thị trên Facebook.
* Post: Nội dung bài đăng của người dùng, bao gồm văn bản thể hiện các suy nghĩ, sở thích, quan điểm hoặc thông tin mà họ chia sẻ công khai trên mạng xã hội trên cả hai nền tảng.

Mẫu dữ liệu chứa các bài đăng từ nhiều người dùng khác nhau trên cả hai nền tảng, cho phép chúng tôi thực hiện phân tích đa chiều về hành vi, sở thích và thói quen trực tuyến của họ. Mẫu này sẽ được sử dụng để phân tích và so sánh, từ đó đưa ra các đề xuất cá nhân hóa cho người dùng như cơ hội việc làm và sách dựa trên nội dung bài đăng của họ.

## Nhập thư viện

Trong đề tài này, chúng tôi sử dụng một số thư viện và mô hình mạnh mẽ nhằm phục vụ việc phân tích và xử lý ngôn ngữ tự nhiên (NLP) để thực hiện các nhiệm vụ chính như định danh người dùng, embedding dữ liệu và trích xuất từ khóa từ các bài đăng trên mạng xã hội. Dưới đây là giới thiệu về các thư viện chính được sử dụng:

### NLTK (Natural Language Toolkit)

**Mục đích sử dụng:** Thư viện NLTK được sử dụng cho nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên. Trong đề tài này, chúng tôi sử dụng NLTK để tính toán Edit Distance giữa các tên người dùng (Username) nhằm đo lường độ tương đồng giữa các tên người dùng trên các nền tảng mạng xã hội khác nhau.

**Tính năng nổi bật:**

* Edit Distance (hay Levenshtein Distance) là phép đo tính khoảng cách giữa hai chuỗi bằng cách tính số lần thay thế, xóa hoặc thêm ký tự để chuyển từ chuỗi này sang chuỗi khác.
* NLTK cung cấp một phương pháp tính toán đơn giản và hiệu quả cho nhiệm vụ này, giúp chúng tôi định danh được người dùng trên nhiều nền tảng khác nhau.

### Sentence-Transformers và Cosine Similarity

**Mục đích sử dụng:** Trong nghiên cứu này, chúng tôi sử dụng Sentence-Transformers để tạo ra các embedding cho văn bản, và sau đó sử dụng Cosine Similarity để đo lường độ tương đồng ngữ nghĩa giữa các bài viết của người dùng trên các nền tảng khác nhau.

**Tính năng nổi bật:**

* Sentence-Transformers là một thư viện giúp tạo ra các vector ngữ nghĩa từ các đoạn văn bản (sentence embedding) bằng cách sử dụng các mô hình như GTE-base và e5-small-v2. Quá trình này cho phép chúng tôi chuyển đổi văn bản từ các bài đăng thành các vector số đại diện cho ý nghĩa của văn bản.
* Sau khi đã tạo được các embedding từ văn bản, chúng tôi sử dụng Cosine Similarity để tính toán độ tương đồng giữa các bài đăng. Cosine Similarity là một phép đo phổ biến để so sánh hai vector, giúp xác định mức độ giống nhau về mặt ngữ nghĩa của các bài viết trên nhiều nền tảng xã hội.
* Giá trị Cosine Similarity gần với 1 cho thấy hai bài viết có ngữ nghĩa tương đồng cao, giúp hệ thống hiểu rõ hơn về các chủ đề mà người dùng quan tâm và cá nhân hóa các đề xuất phù hợp hơn.

### PKE (Python Keyword Extraction)

**Mục đích sử dụng:** Thư viện PKE được sử dụng để trích xuất từ khóa từ các bài đăng của người dùng trên mạng xã hội, từ đó nắm bắt các chủ đề chính và mối quan tâm của họ.

**Các phương pháp sử dụng:**

* PositionRank: Là một thuật toán xếp hạng từ khóa dựa trên vị trí của các từ quan trọng trong văn bản. PositionRank hoạt động dựa trên việc xác định các từ quan trọng nhất trong văn bản và thứ tự xuất hiện của chúng, từ đó trích xuất các từ khóa chính có ý nghĩa.
* YAKE! (Yet Another Keyword Extractor):là một thuật toán trích xuất từ khóa không cần dữ liệu huấn luyện trước. Nó hoạt động bằng cách đánh giá tầm quan trọng của các từ trong văn bản dựa trên tần suất xuất hiện và ngữ cảnh của chúng.

### Vai trò của các thư viện trong nghiên cứu

**NLTK** giúp định danh người dùng thông qua việc so sánh tên người dùng từ các nền tảng khác nhau, hỗ trợ việc tìm kiếm và khớp danh tính.

**Sentence-Transformers** giúp chuyển các bài viết thành các vector có ý nghĩa, còn Cosine Similarity cho phép đo lường mức độ tương đồng giữa chúng. Kết hợp hai công cụ này, chúng tôi có thể phân tích và so sánh các bài đăng của người dùng trên nhiều nền tảng khác nhau một cách chính xác.

**PKE** hỗ trợ trích xuất từ khóa và xác định các chủ đề chính trong bài đăng của người dùng, Bằng cách trích xuất từ khóa chính từ các bài viết của người dùng trên các nền tảng mạng xã hội khác nhau, hệ thống có thể so sánh và xác định các từ khóa chung hoặc tương đồng mà một người dùng có xu hướng sử dụng nhiều lần.

## Kết quả về độ chính xác của thuật toán

### Độ tương tự giữa tên người dùng của các tài khoản

|  |  |
| --- | --- |
|  | **Edit distance** |
| Accuracy | 40% |

Bảng 4.3.1: Kết quả về độ chính xác dựa trên độ tương tự giữa tên người dùng của các tài khoản

### Độ tương tự giữa nội dung các bài viết

|  |  |  |
| --- | --- | --- |
|  | **Gte base** | **E5 small v2** |
| Accuracy | 92% | 91% |

Bảng 4.3.2: Kết quả về độ chính xác dựa trên độ tương tự giữa nội dung các bài viết.

### Độ tương tự giữa các chủ đề người dùng quan tâm

|  |  |  |
| --- | --- | --- |
|  | **Position Rank** | **YAKE** |
| Accuracy | 92% | 91% |

Bảng 4.3.3: Kết quả về độ chính xác dựa trên độ tương tự giữa chủ đề mà người dùng quan tâm

## Tóm tắt chương IV

Sau khi thực hiện các phương pháp để định danh người dùng, hầy hết các phương pháp đều có độ chính xác khoảng 90%. Tuy nhiên với phương pháp sử dụng Edit Distance thì chỉ nhận được 40%. Sau đây là bảng tổng hợp kết quả của các phương pháp mà chúng tôi sử dụng:

|  |  |  |
| --- | --- | --- |
| Phương pháp | Thuật toán/mô hình | Độ chính xác (Accuracy) |
| **Edit Distance** | Levenshtein distance | 40% |
| **Post Embedding** | Gte-base | 92% |
| E5-small-v2 | 91% |
| **Keyword Extracting** | YAKE | 90.67% |
| Position Rank | 92.33% |

Bảng 4.4.1: Tổng hợp độ chính xác của các thuật toán sử dụng

Dựa vào kết quả trên, có thể thấy các phương pháp sử dụng nội dung người dùng đăng tải như Post Embedding và Keyword Extracting mang lại hiệu suất khá cao. Điều này cho thấy tập người dùng mà chúng tôi nghiên cứu tuy họ sử dụng các nền tảng khác nhau nhưng độ tương tự về các chủ đề mà họ quan tâm cũng như cách dùng từ ngữ đều không có sự khác biệt rõ rệt.

# KẾT LUẬN VÀ ỨNG DỤNG

## Kết luận

Trong thời đại công nghệ số và mạng xã hội phát triển mạnh mẽ như hiện nay, việc định danh người dùng trên các nền tảng trực tuyến đã trở thành một nhu cầu thiết yếu. Từ những tài khoản cá nhân trên Facebook, Instagram cho đến các hồ sơ trực tuyến khác, người dùng không chỉ để lại dấu vết về tên tuổi mà còn chia sẻ những sở thích, thói quen và hành vi của mình thông qua các bài đăng, bình luận và nội dung tương tác hàng ngày. Hiểu được danh tính và hành vi người dùng không chỉ giúp các doanh nghiệp tiếp cận đúng đối tượng khách hàng mà còn giúp cá nhân hóa trải nghiệm của họ, từ việc tìm kiếm cơ hội nghề nghiệp đến việc nhận các gợi ý sản phẩm phù hợp.

Để đáp ứng nhu cầu này, chúng tôi đã phát triển và đề xuất ba mô hình định danh người dùng dựa trên những dữ liệu cơ bản và hành vi trực tuyến: (i) mô hình edit distance dựa trên tên người dùng, (ii) mô hình độ tương tự giữa các bài viết, và (iii) mô hình độ tương tự giữa từ khóa hay chủ đề người dùng thường đề cập. Những mô hình này không chỉ tập trung vào việc xác định danh tính người dùng mà còn mở ra nhiều cơ hội để cá nhân hóa và tối ưu hóa các gợi ý, từ việc làm cho đến các sản phẩm tiêu dùng. Kết quả cho thấy mô hình edit distance đạt độ chính xác khoảng 40%, khẳng định rằng tên người dùng không ảnh hưởng quá nhiều đến độ nhận diện danh tính trên nhiều nền tảng (**Q3**). Mô hình độ tương tự giữa các bài viết đạt độ chính xác khoảng 90%, cung cấp cái nhìn sâu sắc về nội dung người dùng chia sẻ, trong khi mô hình độ tương tự giữa từ khóa đạt 91%, giúp hệ thống hiểu thêm về sở thích và quan điểm cá nhân của người dùng. Với kết quả trên việc định danh người dùng sử dụng phương pháp phân tích nội dung các bài đăng mang lại hiệu quả tốt hơn. Có thể thấy người dùng thường sẽ đăng tải các bài đăng theo sở thích cũng như các chủ đề mà họ quan tâm (**Q1**, **Q2**). Tuy nhiên, bộ dữ liệu của chúng tôi là dữ liệu đã qua chọn lọc, thu thập trong khoảng thời gian ngắn nhất định nên sự khác biệt về nội dung giữa các nền tảng là không quá nhiều. Nếu một bộ dữ liệu có người dùng khi đăng tải lên hai tài khoản với nội dung hoàn toàn khác nhau có thể sẽ làm thuật toán khó khăn trong việc định danh. Tuy nhiên vì hai tài khoản đều thuộc sở hữu bởi cùng một người nên văn phong, cách trình bày cũng như các chủ đề mà họ quan tâm cũng không có quá nhiều sự khác biệt. Để có thể giải quyết vấn đề này có thể sẽ phải thu thập nhiều dữ liệu hơn về vị trí địa lý, thời gian đăng bài. Bên cạnh đó, việc theo dõi một người dùng trong thời gian dài để nắm bắt được sự thay đổi về nhu cầu, sở thích cũng có thể tăng độ chính xác cho thuận toán (**Q4**).

Với các kết quả từ những mô hình này, việc nhận diện người dùng không chỉ giúp phân biệt danh tính giữa các tài khoản mà còn mở ra khả năng ứng dụng trong các lĩnh vực **gợi ý công việc** và **gợi ý sản phẩm** (recommendation system). Hành vi và sở thích cá nhân được phân tích từ tên, bài viết, và các chủ đề thường đề cập giúp chúng tôi có thể cá nhân hóa các gợi ý một cách hiệu quả hơn, đặc biệt trong bối cảnh việc làm và thương mại điện tử ngày càng dựa trên trải nghiệm cá nhân hóa. Ngoài ra dựa vào các hành vi của người dùng có thể nắm bắt được đâu là những tài khoản có hành vi lừa đảo từ đó phát hiện kịp thời tránh sự mất mát xảy ra (**Q5**).

## Ứng dụng

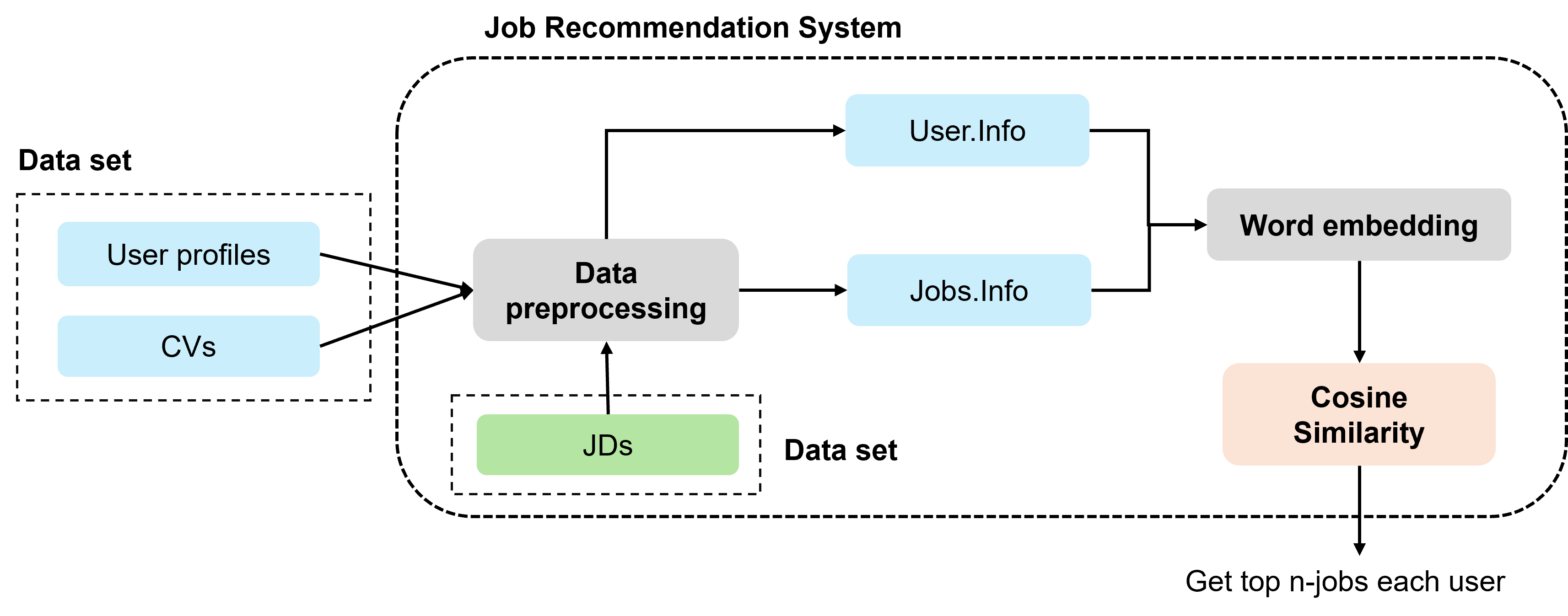
Trong thời đại dữ liệu lớn, các kênh thu thập thông tin người dùng ngày càng phong phú. Việc xác định người dùng trên mạng xã hội có thể mang lại lợi ích cho an ninh mạng, truy xuất thông tin và xây dựng các hệ thống đề xuất [21]. Bên cạnh đó với nhu cầu tìm kiếm công việc phù hợp và nâng cao tri thức ngày càng tăng, nhóm chúng tôi đã phát triển ứng dụng đề xuất việc làm và đề xuất sách với mục tiêu hỗ trợ người dùng trong việc khám phá các cơ hội nghề nghiệp và tài liệu học tập phù hợp nhất. Ứng dụng này tận dụng khả năng phân tích từ các hồ sơ mạng xã hội để đưa ra các gợi ý chính xác và được cá nhân hóa tối đa, mang lại lợi ích thiết thực cho người dùng.

Ứng dụng đề xuất công việc và sách của chúng tôi đã tạo ra một bước đột phá trong việc kết hợp công nghệ AI với dữ liệu cá nhân từ các hồ sơ trực tuyến và mạng xã hội để mang lại giá trị thiết thực cho người dùng. Thông qua việc phân tích CV và các bài đăng trên mạng xã hội như Facebook và Instagram, hệ thống của chúng tôi đã có khả năng nhận diện sở thích, kỹ năng và nhu cầu cá nhân của người dùng một cách chính xác hơn, từ đó cung cấp các đề xuất việc làm phù hợp với kỹ năng và định hướng nghề nghiệp của họ.

Giả sử chúng tôi đã thu thập đầy đủ thông tin từ hồ sơ cá nhân của người dùng trên nền tảng mạng xã hội như Facebook và Instagram, hệ thống sẽ có khả năng phân tích toàn diện hơn về người dùng, từ đó cải thiện đáng kể độ chính xác trong việc đề xuất công việc và tài liệu học tập. Việc có cái nhìn toàn diện về hành vi, sở thích, mối quan tâm, và các tương tác xã hội của người dùng sẽ cho phép hệ thống cung cấp những gợi ý chi tiết và cá nhân hóa đến mức tối đa. Nhờ có dữ liệu phong phú từ nhiều nền tảng, hệ thống có thể đề xuất các cơ hội việc làm không chỉ dựa trên kỹ năng hiện tại của người dùng, mà còn theo các xu hướng mà họ có thể quan tâm trong tương lai. Điều này đặc biệt hữu ích trong những ngành công nghiệp thay đổi nhanh chóng, nơi việc theo kịp các xu hướng và kỹ năng mới là yếu tố sống còn để thành công.

### Đề xuất việc làm

Ứng dụng đề xuất việc làm của chúng tôi dựa vào việc phân tích chi tiết CV cùng với hồ sơ mà chúng tôi thu thập được từ người dùng để cung cấp các gợi ý việc làm phù hợp nhất. Hệ thống tự động trích xuất và đánh giá các yếu tố như kỹ năng chuyên môn, kinh nghiệm làm việc, trình độ học vấn, và mục tiêu nghề nghiệp. Từ đó, hệ thống gợi ý những cơ hội việc làm tương thích với năng lực và mong muốn phát triển của người dùng.



Hình ‑: Hệ thống đề xuất việc làm.

Hình 5‑1 mô tả quy trình hoạt động của hệ thống đề xuất việc làm. Đầu tiên dữ liệu đầu vào của hệ thống là hồ sơ người dùng (user profile) hoặc CV của người dùng. Bên cạnh đó hệ tống cũng lưu trữ các thông tin việc làm mà chúng tôi đã thu thập được từ các công ty (JDs). Đến bước Data preprocessing hệ thống sẽ tiến hành xử lý các dữ liệu đầu vào và cho ra hai tập dữ liệu đó là dữ liệu về người dùng (user.info) và dữ liệu về công việc (jobs). Hai tập dữ liệu này sẽ là đầu vào cho mô hình nhúng từ. Tại đây các thông tin về người dùng cũng như vị trí việc làm sẽ được biến đổi sang vector. Qua đó, nhờ thuật toán tính độ tương tự Cosine (Cosine similarity) sẽ tính độ tương tự của người dùng với từng vị trí công việc khác nhau. Cuối cùng đầu ra của hệ thống sẽ là n công việc có độ tương tự Cosine cao nhất, tức hồ sơ hay CV của người dùng có độ tương hợp cao với các vị trí công việc này.

Ví dụ, nếu một người dùng có kinh nghiệm làm kế toán và đã liệt kê trong CV các kỹ năng như quản lý tài chính, phân tích dữ liệu tài chính, hệ thống sẽ đề xuất các cơ hội việc làm phù hợp như kế toán trưởng hoặc chuyên gia phân tích tài chính. Nếu mục tiêu nghề nghiệp của người dùng là phát triển thêm về kỹ năng quản lý, hệ thống có thể gợi ý các vị trí cao hơn như giám đốc tài chính (CFO), thậm chí cả những công việc kết hợp giữa quản lý tài chính và công nghệ tài chính (FinTech), giúp họ phát triển sự nghiệp theo đúng định hướng.

Với khả năng cá nhân hóa cao, ứng dụng mang lại một trải nghiệm tìm việc hiệu quả, tiết kiệm thời gian và tăng tỷ lệ thành công khi ứng tuyển vào các vị trí phù hợp. Hệ thống không chỉ tìm kiếm công việc hiện tại mà còn hỗ trợ người dùng đạt được các mục tiêu phát triển dài hạn, đảm bảo họ luôn tiếp cận với những cơ hội phù hợp nhất

### Đề xuất sản phẩm

Ứng dụng đề xuất sản phẩm của chúng tôi sử dụng công nghệ AI để phân tích các bài đăng của người dùng trên mạng xã hội và đưa ra gợi ý về **sách** phù hợp với sở thích và nhu cầu của họ. Bằng cách đọc và phân tích nội dung các bài viết, các từ khóa thường được nhắc đến, cũng như chủ đề mà người dùng quan tâm, hệ thống có thể cung cấp những đầu sách chính xác, giúp người dùng tiếp cận kiến thức và tài liệu phù hợp nhất.

Hình 5‑2 mô tả quy trình hoạt động của hệ thống đề xuất sách. Quy trình hoạt động của hệ thống đề xuất sách cũng khá giống với hệ thống đề xuất việc làm. Bao gồm quá trình tiền xử lý dữ liệu (Data preprocessing), nhúng từ (Word embedding) và tính độ tương tự Cosine (Cosine similarity). Tuy nhiên sự khác biệt là chúng tôi chỉ dựa vào hồ sơ của người dùng (user profile) để trích xuất các chủ đề mà người dùng quan tâm. Bên cạnh đó dữ liệu được lưu trong hệ thống là bản tóm tắt hay mô tả của những cuốn sách mà nhóm thu thập được. Và cuối cùng hệ thống sẽ đề xuất những đầu sách mà người dùng có thể quan tâm dựa trên thông tin mà chúng tôi đã thu thập được từ người dùng.

A screenshot of a computer

Description automatically generated

Hình ‑: Hệ thống đề xuất sách

Ví dụ, nếu người dùng thường xuyên chia sẻ các bài viết liên quan đến phát triển kỹ năng cá nhân hoặc thảo luận về các xu hướng công nghệ, hệ thống sẽ gợi ý các cuốn sách về phát triển bản thân hoặc sách về công nghệ mới như trí tuệ nhân tạo, blockchain. Điều này giúp cá nhân hóa trải nghiệm người dùng, mang lại các đề xuất sách có giá trị và liên quan trực tiếp đến mối quan tâm của họ.

Với khả năng phân tích nội dung từ các bài đăng, hệ thống không chỉ dừng lại ở việc gợi ý sách dựa trên sở thích hiện tại mà còn có thể mở rộng các đề xuất, giúp người dùng khám phá những lĩnh vực mới mà họ có thể chưa nghĩ đến. Điều này mang lại một cách tiếp cận cá nhân hóa hơn trong việc học tập và phát triển kiến thức

### Triển khai ứng dụng

Sau đây là QR code triển khai ứng dụng của chúng tôi trên website:

A qr code with black squares

Description automatically generated

Hình 5‑3: Mã QR triển khai ứng dụng

## Hạn chế và hướng nghiên cứu trong tương lai

Mặc dù ứng dụng đề xuất việc làm dựa trên phân tích CV của chúng tôi đã mang lại những kết quả tích cực, vẫn tồn tại một số hạn chế cần được khắc phục và cải thiện trong tương lai.

### Hạn chế

**Phụ thuộc vào chất lượng CV và hồ sơ người dùng:** Hệ thống hiện tại hoàn toàn dựa vào thông tin có trong CV và hồ sơ người dùng để đưa ra gợi ý. Nếu người dùng không cung cấp đầy đủ hoặc chính xác thông tin, các đề xuất có thể thiếu chính xác hoặc không phản ánh đúng khả năng và mục tiêu nghề nghiệp của họ. Bên cạnh đó việc định danh sai các tài khoản không phải của họ dẫn đến đề xuất sai các sản phẩm cũng như vị trí công việc dẫn đến việc giảm chất lượng trải nghiệm của họ trên các trang mạng xã hội.

**Thiếu khả năng cập nhật thông tin liên tục:** Trong khi CV thường chỉ được cập nhật khi có thay đổi lớn trong sự nghiệp, người dùng có thể không thường xuyên bổ sung thông tin mới về kỹ năng, kinh nghiệm. Điều này dẫn đến việc các đề xuất có thể không phản ánh đầy đủ các thay đổi nhỏ trong hành vi và kỹ năng của người dùng theo thời gian.

**Hạn chế về ngữ cảnh:** Hệ thống dựa chủ yếu trên phân tích ngữ nghĩa của CV, do đó có thể bỏ qua những thông tin liên quan đến hoàn cảnh cá nhân hoặc mong muốn thay đổi ngành nghề mà không được thể hiện rõ ràng trong CV.

### Hướng nghiên cứu trong tương lai

**Kết hợp phân tích đa nguồn dữ liệu:** Trong tương lai, hệ thống có thể kết hợp dữ liệu từ các mạng xã hội (Instagram, Facebook) cùng với CV để đưa ra gợi ý chính xác hơn. Việc kết hợp này sẽ giúp cung cấp một cái nhìn toàn diện hơn về kỹ năng và mục tiêu của người dùng, từ đó đề xuất những cơ hội việc làm và phát triển chuyên nghiệp một cách toàn diện.

**Phân tích hành vi và kỹ năng theo thời gian thực**: Hướng đi tiếp theo là phát triển khả năng cập nhật liên tục thông tin về các chủ đề quan tâm, kỹ năng và kinh nghiệm của người dùng. Thay vì chỉ dựa vào CV tĩnh, hệ thống có thể theo dõi các khóa học trực tuyến mà người dùng hoàn thành, những dự án mới tham gia, hoặc các chứng chỉ mới đạt được để cập nhật kịp thời các đề xuất công việc. Bên cạnh đó việc theo dõi người dùng trong khoảng thời gian dài có thể nắm bắt được sự thay đổi các chủ đề mà người dùng yêu thích từ đó có thể đề xuất các sản phẩm cũng như vị trí công việc mà có thể họ sẽ quan tâm trong tương lai.

**Phát triển khả năng hiểu ngữ cảnh cá nhân:** Nghiên cứu thêm về các mô hình AI có khả năng phân tích ngữ cảnh cá nhân của người dùng, bao gồm mong muốn thay đổi ngành nghề, lộ trình phát triển cá nhân, và các yếu tố cá nhân khác để cải thiện sự chính xác và phù hợp của các đề xuất việc làm.

**Cải thiện khả năng cá nhân hóa sâu hơn:** Hệ thống có thể tiếp tục phát triển khả năng cá nhân hóa bằng cách tích hợp thêm các công cụ học máy để dự đoán xu hướng phát triển nghề nghiệp của người dùng, không chỉ dựa trên thông tin hiện tại mà còn dự đoán nhu cầu kỹ năng tương lai.

**Bảo mật và quyền riêng tư:** Trong khi mở rộng khả năng thu thập và phân tích thông tin từ nhiều nguồn khác nhau, bảo mật và quyền riêng tư của người dùng cần được nghiên cứu kỹ lưỡng hơn. Hướng đi này sẽ đảm bảo các dữ liệu cá nhân được bảo vệ chặt chẽ và chỉ sử dụng cho mục đích cải thiện trải nghiệm tìm việc của người dùng.

## Tóm tắt chương V

Hệ thống của chúng tôi sử dụng ba mô hình chính để nhận diện người dùng dựa trên tên, độ tương tự bài viết, và từ khóa, giúp cá nhân hóa các gợi ý và tối ưu hóa trải nghiệm của họ. Mô hình đã đạt những kết quả tích cực, đặc biệt trong lĩnh vực đề xuất công việc và sản phẩm như sách.

Hệ thống phân tích CV để gợi ý việc làm phù hợp dựa trên kỹ năng, kinh nghiệm và mục tiêu nghề nghiệp của người dùng, với khả năng cung cấp các cơ hội việc làm tương thích và giúp người dùng phát triển sự nghiệp dài hạn. Trong khi đó, ứng dụng đề xuất sách được phát triển từ các bài đăng trên mạng xã hội, giúp cá nhân hóa việc lựa chọn sách phù hợp với sở thích và nhu cầu học tập của người dùng.

Tuy nhiên, hệ thống hiện còn một số hạn chế như phụ thuộc vào chất lượng CV, thiếu khả năng cập nhật thời gian thực, và chưa khai thác đầy đủ dữ liệu từ mạng xã hội. Do đó, các hướng nghiên cứu trong tương lai sẽ tập trung vào việc kết hợp phân tích đa nguồn dữ liệu, phát triển khả năng hiểu ngữ cảnh cá nhân và nâng cao tính cá nhân hóa sâu hơn, đồng thời đảm bảo bảo mật và quyền riêng tư cho người dùng.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | A. B. N. S. R. B. R. A. S. Nayyar, “Social Media in the Time of a Pandemic,” *Global Perspectives of COVID-19 Pandemic on Health, Education, and Role,* pp. 289-303, 2023. |
| [2] | P. a. H. L. K. w. a. H. A. Solanki, “User Identification across Social Networking Sites using User Profiles and Posting Patterns,” *International Joint Conference on Neural Networks (IJCNN),* pp. 1-8, 2021. |
| [3] | J. Y. X. Z. C. Z. Daokun Zhanga, “User Profile Preserving Social Network Embedding,” *Proceedings of the Twenty-Sixth International Joint Conference on,* pp. 3378--3384, 2017. |
| [4] | D.-V. a. T. T.-T. a. S. K. a. H. V.-N. Vo, “Deep Generative Networks Coupled With Evidential Reasoning for Dynamic User Preferences Using Short Texts,” *IEEE Transactions on Knowledge and Data Engineering,* tập 35, pp. 6811-6826, 2023. |
| [5] | V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Dokl. Akad. Nauk SSSR,* tập 163, số 4, pp. 845--848, 1965. |
| [6] | N. T. L. M. N. R. Niklas Muennighoff, “MTEB: Massive Text Embedding Benchmark,” 2023. |
| [7] | Z. L. ,. Y. Z. D. L. P. Xie, “Retrieval, Challenging Decoder helps in Masked Auto-Encoder Pre-training for Dense Passage,” 2023. |
| [8] | S. S. a. R. R. Pierre Zweigenbaum, “Proceedings of 11th Workshop on Building and Using Comparable Corpora,” pp. 39-42, 2018. |
| [9] | H. S. Mikel Artetxe, “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond,” *Transactions of the Association for Computational Linguistics,* 2019. |
| [10] | M.-W. C. K. L. K. T. Jacob Devlin, “BERT: Pre-training of deep bidirectional transformers for language understanding,” trong *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. |
| [11] | G. H. A. N. C. J. M. K. H. D. C. a. Y. Y. Jianmo Ni, “Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models,” *In Findings of the Association for Computational Linguistics: ACL 2022,* pp. 1864-1874, 2022a. |
| [12] | C. Q. J. L. Z. D. G. H. A. J. M. V. Z. Y. L. K. Jianmo Ni, “Large dual encoders are generalizable retrievers,” *Proceedings of the 2022 Conference on Empirical,* p. 9844–9855, 2022b. |
| [13] | T. X. R. P. A. R. J. M. H. J. T. Q. Arvind Neelakantan, “Text and code embeddings by contrastive pre-training,” 2022. |
| [14] | Z. L. X. Z. Y. Z. D. L. P. X. M. Zhang, “Towards General Text Embeddings with Multi-stage Contrastive Learning,” p. https://arxiv.org/abs/2308.03281, 2023. |
| [15] | L. G. a. J. Callan, “Retrieval, Unsupervised Corpus Aware Language Model Pre-training for Dense Passage,” 2021. |
| [16] | Z. L. Y. Z. D. L. P. Xie, “Challenging Decoder helps in Masked Auto-Encoder Pre-training for Dense Passage Retrieval,” 2023. |
| [17] | V. M. A. P. A. J. C. N. ,. A. J. Ricardo Campos, “YAKE! Keyword extraction from single documents using,” *Information Sciences,* tập 509, pp. 257-289, 2020. |
| [18] | P. T. Rada Mihalcea, “Textrank: Bringing order into text,” trong *Conference on Empirical Methods in Natural Language*, 2004. |
| [19] | J. X. Xiaojun Wan, “Single document keyphrase extraction using neighborhood knowledge,” trong *American Association for Artificial Intelligence*, 2008. |
| [20] | C. C. Corina Florescu, “PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents,” trong *55th Annual Meeting of the Association for Computational Linguistics*, 2017. |
| [21] | M. H. W. H. Z. K. D. K. Qu Y, “A Multiple Salient Features-Based User Identification across Social Media,” *Entropy. 2022; 24(4):495.,* p. 495, 2022. |
| [22] | C. O. Q. V. L. Andrew M. Dai, “Document Embedding with Paragraph Vectors,” 07 2015. |
| [23] | K. Kenyon-Dean, “Word Embedding Algorithms as Generalized Low Rank Models and their Canonical Form,” 2019. |
| [24] | E. N. J. C. K. C. Kian Kenyon-Dean, “Deconstructing word embedding algorithms,” 2020. |
| [25] | J. B. a. H. Chabanne., “Embedding edit distance to enable private,,” *Human-centric Computing and Information Sciences,* 2012. |
| [26] | D. a. S.-C. J. a. T. T. a. W. L. a. V. S. a. K. I. Chatzakou, “User Identity Linkage in Social Media Using Linguistic and Social Interaction Features,” trong *Proceedings of the 12th ACM Conference on Web Science*, New York, NY: Association for Computing Machinery, 2020, p. 295–304. |
| [27] | R. &. L. H. Zafarani, “Connecting Corresponding Identities across Communities,” *Proceedings of the International AAAI Conference on Web and Social Media,* pp. 354-357, 2019. |
| [28] | D.-V. a. H. A. a. H. V.-N. Vo, “An integrated framework of learning and evidential reasoning for user profiling using short texts,” *Information Fusion,* pp. 27-42, 2020. |
| [29] | K. N. T. S. I. S. Alec Radford, “Improving Language Understanding by Generative Pre-Training,” 2018. |
| [30] | H. S. W. S. J. K. Y. W. Y. H. M. O. W.-t. Y. N. A. S. L. Z. T. Yu, “One Embedder, Any Task: Instruction-Finetuned Text Embeddings,” 2023. |