HỘI SINH VIÊN VIỆT NAM

HỌC VIỆN TÀI CHÍNH **VIỆN TOÁN HỌC**

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC
THAM GIA HỘI THI KHOA HỌC SINH VIÊN TOÀN QUỐC
"OLYMPIC KINH TẾ LƯỢNG VÀ ỨNG DỤNG" LẦN THỨ IX, 2024

ĐỀ TÀI:

GOM CỤM NGƯỜI DÙNG TRÊN CÁC NỀN TẢNG MẠNG XÃ HỘI VÀ ỨNG DỤNG TRONG KINH DOANH

Người hướng dẫn: TS. Võ Đức Vĩnh

Tập thể sinh viên thực hiện:

1. Đỗ Phi Long: DH37KH01, Toán Kinh Tế, Trường đại học Ngân Hàng tp HCM
2. Đinh Quang Toàn: DH37KH01, Toán Kinh Tế, Trường đại học Ngân Hàng tp HCM
3. Nguyễn Thùy Trinh: DH37KH01, Toán Kinh Tế, Trường đại học Ngân Hàng tp HCM
4. Nguyễn Nhật Hoa: DH37KH01, Toán Kinh Tế, Trường đại học Ngân Hàng tp HCM
5. Phan Trịnh Quốc An: DH37KH01, Toán Kinh Tế, Trường đại học Ngân Hàng tp HCM

Hồ Chí Minh - tháng 03 năm 2024

MỤC LỤC

DANH MỤC HÌNH ẢNH
DANH MỤC BẢNG BIỂU
DANH MỤC TỪ VIẾT TẮT
TÓM TẮT

CHƯƠNG 1: GIỚI THIỆU NGHIÊN CỨU 1
1.1. Lý do chọn đề tài 1
1.2. Mục tiêu nghiên cứu 2
1.3. Câu hỏi nghiên cứu 2
1.4. Đối tượng và phạm vi nghiên cứu 3
1.5. Phương pháp nghiên cứu 3
1.6. Đóng góp của nghiên cứu 4
1.7. Cấu trúc của nghiên cứu 4
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4
2.1. Các khái niệm liên quan 4
2.1.1. Gom cụm người dùng trên mạng xã hội. 4
2.1.2. Vector embedding (Vector nhúng) 5
2.1.3. Giảm chiều dữ liệu 5
2.1.4. Trích xuất từ khóa (key word) và mô hình hóa chủ đề (Topic Modeling) 6
2.2. Các lý thuyết liên quan 6
2.2.1. Các phương pháp giảm chiều dữ liệu 6
2.2.2. Các mô hình Vector embedding (Vector nhúng). 13
2.2.3. Phương pháp mô hình hóa chủ đề LDA 16
2.3. Các nghiên cứu liên quan 16
2.3.1. Nghiên cứu trong nước 16
2.3.2. Nghiên cứu ngoài nước 17
CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU 18
3.1. Thu thập dữ liệu nghiên cứu (Data obtaining) 19
3.2. Làm sạch dữ liệu (Data cleaning) 19
3.2.1. Tách từ và chuẩn hóa (Tokenization và Normalization) 20
3.2.2. Loại bỏ từ dừng (Stop words removing) 20
3.2.3. Loại bỏ đường dẫn (Link, URL removing) 21

3.2.4. Chuyển từ về dạng gốc (Word stemming) 21
3.3. Vector nhúng (Vector embedding) 22
3.4. Giảm chiều dữ liệu (Dimensionality reduction) 23
3.5. Gom cụm người dùng (Users clustering) 23
3.6. Mô hình hóa chủ đề (Topic modeling) 23
3.7. Ứng dụng nghiên cứu (Aplications) 24
CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU 24
4.1. Khai báo các thư viện cần thiết 24
4.2. Kết quả gom cụm 25
4.3. Cụm người dùng và chủ đề 28
4.4. Ứng dụng 31
4.4.1. Xây dựng cộng đồng người dùng 31
4.4.2. Đề xuất sản phẩm cho người dùng 33
4.5. So sánh với các nghiên cứu trước 35
CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ 36
5.1. Kết luận 36
5.2. Hạn chế của đề tài 37
5.3. Hướng phát triển đề tài 38

TÀI LIỆU THAM KHẢO

DANH MỤC HÌNH ẢNH

Hình 2.1 Giảm chiều chưa tối ưu 6
Hình 2.2 Giảm chiều tối ưu 7
Hình 2.3 Không gian đa chiều gốc và đường cong phân phối chuẩn.....8
Hình 2.4 Ma trận tương đồng ban đầu.....8
Hình 2.5 Ma trận tương đồng mới.....9
Hình 2.6 Mô hình kết nối cục bộ và tập mở mờ.....10
Hình 2.7 Vùng lân cận và xu hướng các điểm hàng xóm 11
Hình 2.8 Trọng lượng cạnh không đồng ý 12
Hình 2.9 Đồ thị có trọng số cạnh thích hợp 12
Hình 2.10 Kết quả điểm chuẩn MTEB giữa các mô hình 15
Hình 3.1 Tổng quan quá trình nghiên cứu.....19
Hình 3.2 Quá trình làm sạch dữ liệu.....20
Hình 3.3 Quy trình nhúng văn bản thành vector 22
Hình 4.1 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-base và 3 phương pháp giảm chiều dữ liệu) 26
Hình 4.2 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-small và 3 phương pháp giảm chiều dữ liệu) 27
Hình 4.3 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-large và 3 phương pháp giảm chiều dữ liệu) 27
Hình 4.4 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-base và phương pháp giảm chiều dữ liệu Umap) 28
Hình 4.5 Trực quan các cụm người dùng kèm chủ đề tương ứng (Thực hiện bởi mô hình E5-base và phương pháp giảm chiều dữ liệu Umap) 31

DANH MỤC BẢNG BIỂU

Bảng 3.1 Ví dụ minh họa kết quả tách từ và chuẩn hóa dữ liệu.....20
Bảng 3.2 Ví dụ minh hoạ kết quả loại bỏ từ dừng 21
Bảng 3.3 Ví dụ minh họa kết quả loại bỏ đường dẫn.....21
Bảng 3.4 Ví dụ minh họa kết quả chuyển từ về dạng gốc.....22
Bảng 4.1 Bảng kết quả đánh giá các chỉ số gom cụm 27
Bảng 4.2 Mô hình chủ đề cho từng cụm người dùng 29
Bảng 4.3 Chi tiết thông tin người dùng trong các nhóm chủ đề 31
Bảng 4.4 Minh họa sản phẩm được đề xuất tới các cụm chủ đề người dùng.....34

DANH MỤC TỪ VIẾT TẮT

TỪ VIẾT TẮT	NGHĨA TIẾNG ANH	NGHĨA TIẾNG VIỆT
API	Application Programming Interface	Giao diện lập trình ứng dụng
CBOW	Continuous Bag of Words	Túi từ liên tục
CNM	Clauset-Newman-Moore	Thuật toán phân cụm cộng đồng
GUI	Graphical User Interface	Giao diện đồ hoạ người dùng
HTTP	Hypertext Transfer Protocol	Giao thức truyền tải siêu văn bản
ID	Identification	Nhận dạng
INC	Incre-Comm-Extraction	Thuật toán trích xuất cộng đồng tăng dần
JSON	JavaScript Object Notation	Định dạng đối tượng JavaScript
LASSO	Least Absolute Shrinkage and Selection Operator	Toán tử lựa chọn và co rút tuyệt đối nhỏ nhất
LDA	Latent Dirichlet Allocation	Thuật toán phân bổ Dirichlet tiềm ẩn
MTEB	Massive Text Embedding Benchmark	Điểm chuẩn nhúng văn bản lớn
PCA	Principal Component Analysis	Phân tích thành phần chính
STC	Short Text Clustering	Phân cụm văn bản ngắn
STS	Semantic Textual Similarity	Sự tương đồng về mặt ngữ nghĩa
t-SNE	t-Distributed Stochastic Neighbor Embedding	Thuật toán nhúng ngẫu nhiên các hàng xóm theo phân phối t
TF-IDF	Term Frequency-Inverse Document Frequency	Tần suất từ - Tần suất nghịch đảo văn bản
UMAP	Uniform Manifold Approximation and Projection	Phương pháp xấp xỉ và dự báo đồng nhất
URL	Uniform Resource Locator	Hệ thống định vị tài nguyên thống nhất

TÓM TẮT

Trong bối cảnh của xã hội số hiện đại, dữ liệu người dùng đã trở thành một tài sản không thể thiếu đối với các doanh nghiệp. Trên nền tảng mạng xã hội lớn nhất thế giới Facebook, hàng tỷ bài đăng được tạo ra hàng ngày, mang theo những thông tin chi tiết về sở thích, hành vi và nhu cầu của người tiêu dùng. Tuy nhiên, việc khai thác và sử dụng hiệu quả nguồn tài nguyên dữ liệu khổng lồ này vẫn là một thách thức đáng kể, đòi hỏi sự ứng dụng của các công cụ phân tích tiên tiến. Nghiên cứu này trình bày một phương pháp mới, kết hợp các kỹ thuật học máy như vector nhúng từ vựng và thuật toán phân cụm để tự động phân tích và phân loại người dùng Facebook ,kèm theo kĩ thuật trích xuất từ khóa để mô hình hóa chủ đề người dùng dựa trên nội dung bài đăng của họ. Toàn bộ quy trình, từ thu thập dữ liệu đến xử lý và phân tích, được tự động hóa nhằm đảm bảo tính khả thi và tái khả dụng. Phương pháp đề xuất không chỉ xác định chính xác sở thích và mối quan tâm của người dùng mà còn sắp xếp họ thành các nhóm riêng biệt, đặc trưng. Điều này cho phép các doanh nghiệp xác định và tập trung vào các phân khúc khách hàng mục tiêu một cách hiệu quả hơn. Kết quả nghiên cứu cung cấp một công cụ hữu ích, giúp doanh nghiệp hiểu rõ hơn về cơ cấu, nhu cầu và xu hướng của khách hàng, từ đó xây dựng và tối ưu hóa chiến lược tiếp thị phù hợp cũng như phát triển thương hiệu sản phẩm. Hơn nữa, nghiên cứu này cũng góp phần thúc đẩy sự phát triển của lĩnh vực phân tích dữ liệu và tiếp thị kỹ thuật số bằng cách đề xuất một phương pháp mới trong lĩnh vực xử lí ngôn ngữ tự nhiên, mở ra nhiều cơ hội cho các ứng dụng và đổi mới tiếp theo trong kỷ nguyên số hóa. Trong suốt quá trình nghiên cứu, chúng tôi tuân thủ các nguyên tắc khoa học như tính khách quan, thận trọng trong đánh giá và kiểm chứng kết quả, cũng như đảm bảo tính toàn vẹn và tuân thủ đạo đức trong việc xử lý dữ liệu cá nhân người dùng.

Keywords : Phân cụm người dùng, hồ sơ người dùng, phân tích văn bản, xử lí ngôn ngữ tự nhiên, trích xuất chủ đề văn bản , học máy không giám sát.

CHƯƠNG 1: GIỚI THIỆU NGHIÊN CỨU

1.1. Lý do chọn đề tài

Gần đây, số lượng tài liệu văn bản trên Internet đã tăng lên đáng kể và nhanh chóng. Sự phát triển nhanh chóng của thiết bị di động và công nghệ Internet đã khuyến khích người dùng tìm kiếm thông tin, liên lạc với bạn bè và chia sẻ ý kiến, sự quan tâm của họ trên các phương tiện truyền thông xã hội như Twitter, Instagram, Facebook. Các văn bản được tạo ra hàng ngày trên mạng xã hội là dữ liệu khổng lồ và không có cấu trúc [1]. Các văn bản ngắn thường thiếu ngữ cảnh, điều này làm cho việc tìm kiếm thông tin trong chúng trở nên khó khăn. Hơn nữa, tính đa dạng và hỗn loạn của chúng khiến việc phân tích trở nên phức tạp hơn, với sự hiện diện thường xuyên của tiếng ồn, ngôn ngữ địa phương, biểu tượng cảm xúc, lỗi chính tả, viết tắt và ngữ pháp không chính xác.

Mặc dù vậy, nếu chúng ta có thể khai thác và phân tích các văn bản này một cách thông minh, chúng có thể cung cấp thông tin quý giá về sở thích và quan tâm của người dùng. Điều này có thể giúp các doanh nghiệp hiểu sâu hơn về đối tượng mục tiêu của họ và từ đó tối ưu hóa chiến lược kinh doanh của mình để đạt được lợi ích tối đa. Trong công việc này, chúng tôi nghiên cứu vấn đề về cách gom cụm người dùng theo sở thích của họ và đưa ra đề xuất sản phẩm, hay tạo ra các cộng đồng phát triển chung cho từng nhóm dựa vào các văn bản ngắn mà họ chia sẻ trên mạng xã hội. Kết quả của công việc này có thể tiết lộ thông tin quý giá cho việc ra quyết định trong các hoạt động kinh doanh sau này của các doanh nghiệp.

Mặc dù đã có nhiều nghiên cứu về việc phân loại người dùng trên mạng xã hội, tuy nhiên, vẫn còn nhiều thách thức và câu hỏi cần được khám phá. Vì vậy, câu hỏi mà nhóm đặt ra là liệu chỉ với ID người dùng trên Facebook, chúng ta có thể hiểu được nhu cầu và sở thích của họ một cách tự động? Và liệu chúng ta có thể tập hợp các nhóm người dùng có sở thích chung thành các cộng đồng một cách tự động?

Khác với nghiên cứu của [2], chỉ dừng lại ở việc phân cụm người dùng trên nền tảng mạng xã hội hoặc chỉ thêm một bước lập mô hình chủ đề của từng cụm như nghiên cứu của [3], [4]. Trong nghiên cứu này, chúng tôi sẽ đề xuất một phương pháp kết hợp sử dụng các công nghệ như học máy và phân loại dữ liệu để tự động phân loại và gom cụm người dùng dựa vào việc phân tích chủ đề của những văn bản ngắn mà họ đã chia sẻ trên mạng xã hội. Đặc biệt hơn hết, nhóm chúng tôi sẽ sử dụng những chủ đề đó để đề xuất ra những sản phẩm phù hợp nhất cho từng nhóm cụ thể và xây dựng các cộng đồng người dùng. Nghiên cứu không chỉ mang tính lý thuyết mà còn mang lại giá trị thực tiễn cho doanh nghiệp - điều mà các nghiên cứu trước thường chưa thực hiện được hoặc chưa thực hiện một cách toàn diện và hiệu quả.

Để có kết quả phân cụm chính xác, đầu tiên, chúng tôi thử nghiệm sử dụng lần lượt các mô hình vector nhúng là E5-base, E5-small, E5-large để biểu diễn các từ trong văn bản đã được làm sạch, thành các vectơ có số chiều cố định, sau đó đánh giá hiệu suất và chọn lọc. Việc này giúp chúng tôi biểu diễn, tính toán dữ liệu văn bản một cách dễ dàng và hiệu quả. Tiếp theo, chúng tôi tiếp tục thực hiện thử nghiệm sử dụng lần lượt 3 phương pháp giảm chiều dữ liệu từ không gian đa chiều xuống không gian có số chiều thấp hơn là PCA, t-SNE, Umap, để

dễ dàng cho việc gom cụm cũng như trực quan hóa dữ liệu. Sau đó, chúng tôi gom cụm người dùng dựa trên những điểm tương đồng được phân tích từ các bài đăng của họ bằng thuật toán K-means. Cuối cùng, để hiểu rõ hơn về nội dung của từng cụm, chúng tôi sử dụng mô hình LDA [5] để xác định các chủ đề chính trong dữ liệu, LDA giúp chúng tôi phát hiện các chủ đề ẩn và phân bổ các từ vào các chủ đề này, giúp chúng tôi hiểu rõ hơn về nội dung của từng cụm. Qua quy trình này, chúng tôi có thể tổ chức và hiểu rõ hơn về dữ liệu văn bản của mình, giúp chúng tôi trích xuất thông tin quan trọng và thực hiện các phân tích chi tiết hơn về nội dung của văn bản.

Nghiên cứu này cũng nhấn mạnh vào sự tiềm năng của việc sử dụng các phương pháp máy học và xử lý ngôn ngữ tự nhiên để tự động hóa quá trình gom cụm khách hàng. Sử dụng các công nghệ này, chúng ta có thể phân loại và nhóm các người dùng dựa trên sở thích, hành vi và tương tác trên các nền tảng mạng xã hội và ứng dụng, từ đó tạo ra các chiến lược tiếp thị cũng như chiến lược quảng bá sản phẩm được cá nhân hóa hiệu quả hơn. Điều này mang lại lợi ích lớn cho doanh nghiệp bằng cách giúp họ tối ưu hóa chiến lược tiếp thị, tăng cường tương tác với khách hàng và nâng cao hiệu suất kinh doanh.

1.2. Mục tiêu nghiên cứu

Nghiên cứu này có mục tiêu tổng quát là xác định các nhóm người dùng dựa trên các chủ đề quan tâm được rút trích từ những bài đăng của họ trên nền tảng mạng xã hội. Từ đó, sử dụng thông tin và chủ đề trên mạng xã hội ứng dụng vào các mục đích kinh doanh như: đề xuất sản phẩm tương thích với chủ đề mà họ quan tâm, tiếp thị sản phẩm và nghiên cứu thị trường, xây dựng các cộng đồng có chung chủ đề quan tâm để tự nhiên hóa quá trình phát triển thương hiệu từ các trang mạng xã hội. Để thực hiện được như thế, nhóm nghiên cứu đã đề ra các mục tiêu cụ thể như sau:

- Phân tích hành vi sử dụng mạng xã hội của người dùng và xác định các nhóm cụm người dùng dựa trên các mẫu quan tâm, hoạt động, và tương tác trên các nền tảng mạng xã hội .
- Xác định các đặc điểm chung và khác biệt giữa các nhóm người dùng, bao gồm sở thích, nhu cầu, và mong muốn khi tương tác trên mạng xã hội.
- Đề xuất các chiến lược và phương pháp tiếp cận nhóm người dùng cụ thể để tối ưu hóa hiệu quả kinh doanh và tương tác trên mạng xã hội, bằng cách quảng cáo các sản phẩm và dịch vụ phù hợp với nhu cầu và quan tâm của từng nhóm người dùng, xây dựng các cộng đồng người dùng có chung chủ đề quan tâm nhằm phát triển và quảng bá thương hiệu sản phẩm.
- Đánh giá và đề xuất các phương tiện công nghệ và công cụ hỗ trợ để thu thập và phân tích dữ liệu từ mạng xã hội một cách hiệu quả, nhằm tối ưu hóa quá trình gom cụm người dùng và ứng dụng trong doanh nghiệp.

1.3. Câu hỏi nghiên cứu

Để đạt được mục tiêu nghiên cứu cụ thể, đề tài cần trả lời các câu hỏi nghiên cứu sau:

- Làm thế nào để phân tích và nhận diện các nhóm cụm người dùng trên nền tảng mạng xã hội dựa trên hành vi và tương tác của họ?

- Có những đặc điểm chung và khác biệt nào giữa các nhóm người dùng được phát hiện, bao gồm sở thích, nhu cầu và mức độ quan tâm đến các chủ đề cụ thể trên mạng xã hội?

- Có những cơ hội và thách thức gì khi áp dụng thông tin về các nhóm người dùng từ mạng xã hội vào chiến lược kinh doanh và tiếp thị?

- Có những công nghệ và công cụ nào có thể hỗ trợ việc thu thập và phân tích dữ liệu từ mạng xã hội một cách hiệu quả để gom cụm người dùng và ứng dụng trong kinh doanh?

1.4. Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:**

Đối tượng nghiên cứu của đề tài là các bài đăng của những nhân vật nổi tiếng trên nền tảng mạng xã hội. Đối tượng nghiên cứu này đại diện cho một phạm vi rộng lớn của cộng đồng mạng, từ các người yêu thích bóng đá, âm nhạc, ẩm thực cho đến các vấn đề kinh doanh, chính trị, …

- **Phạm vi nghiên cứu:**

Nghiên cứu sẽ tập trung vào việc phân tích 716,649 bài viết của 302 đối tượng nghiên cứu trên mạng xã hội Facebook. Mục tiêu là hiểu rõ hơn về sở thích, quan điểm, và tương tác của người dùng với các chủ đề mà họ quan tâm. Sử dụng kỹ thuật mô hình hóa chủ đề LDA để phát hiện các chủ đề tiềm ẩn trong bài viết. Gom cụm người dùng dựa trên các chủ đề quan trọng mà họ quan tâm, từ đó đề xuất các sản phẩm hoặc dịch vụ phù hợp và xây dựng các cộng đồng với từng nhóm người dùng có chung chủ đề.

1.5. Phương pháp nghiên cứu

Phương pháp nghiên cứu của chúng tôi được áp dụng trong đề tài là kết hợp cả 2 phương pháp định tính và định lượng. Hai phương pháp này nhằm mục đích cung cấp một cái nhìn toàn diện về các khía cạnh của đối tượng nghiên cứu, từ các chi tiết cụ thể đến các xu hướng và mối quan hệ tổng quát.

Trong quá trình nghiên cứu, chúng tôi đã sử dụng phương pháp nghiên cứu định tính để tiếp cận và hiểu sâu hơn về trải nghiệm và ý kiến của người dùng về việc chia sẻ các bài đăng trên mạng xã hội. Phương pháp này đã cho phép chúng tôi tiếp cận các thông tin không chỉ từ các bài đăng cụ thể mà còn từ ngữ cảm xúc và cảm nhận của người dùng về nội dung mà họ chia sẻ. Điều này giúp chúng tôi hiểu rõ hơn về ngữ cảnh và ý nghĩa của các bài đăng. Cụ thể tính định tính được thể hiện qua các bước làm như, xây dựng danh mục mã người dùng phục vụ cho việc thu thập dữ liệu bài đăng của họ trên nền tảng mạng xã hội, xử lí làm sạch dữ liệu văn bản để phục vụ cho việc vector số hóa văn bản .

Ngoài ra, chúng tôi cũng sử dụng phương pháp nghiên cứu định lượng để phân tích và mô hình hóa dữ liệu thu thập được từ dưới dạng các bài đăng văn bản thành các vector số học. Việc này giúp chúng tôi đo lường sự tương đồng giữa các người dùng và phân tích các mẫu và xu hướng trong dữ liệu một cách toàn diện hơn, từ đó đưa ra những kết luận và dự đoán cụ thể về hành vi của người dùng trên mạng xã hội.

Kết hợp cả hai phương pháp nghiên cứu định tính và định lượng giúp chúng tôi nhìn nhận và hiểu rõ hơn về đối tượng nghiên cứu từ nhiều góc độ khác nhau, từ chi tiết cụ thể đến các xu hướng tổng quát, từ các cảm nhận cá nhân đến các xu hướng và đặc điểm của cộng đồng người dùng. Cách tiếp cận này giúp tăng cường tính toàn diện và độ chính xác của nghiên cứu của chúng tôi.

1.6. Đóng góp của nghiên cứu

Nghiên cứu về việc gom cụm người dùng trên nền tảng mạng xã hội và ứng dụng trong kinh doanh là một bước quan trọng trong việc hiểu sâu hơn về hành vi của người dùng trực tuyến và tạo ra các gợi ý sản phẩm phù hợp, xây dựng các cộng đồng có cùng chủ đề trong sở thích cá nhân . Sự kết hợp giữa các mô hình vector nhúng , thuật toán phân cụm K-means, và mô hình hoá chủ đề LDA giúp hiểu rõ hơn về các nhóm người dùng và sở thích của họ, cho phép đề xuất các sản phẩm và dịch vụ phù hợp với từng nhóm mục tiêu, hay có thể xây dựng lên những cộng đồng trên các trang mạng xã hội để mang lại giá trị lớn cho việc quảng bá sản phẩm và chiến lược kinh doanh, giúp doanh nghiệp hiểu rõ hơn về thị trường và nhu cầu của người tiêu dùng, từ đó tối ưu hóa chiến lược tiếp thị và tăng cường tương tác với khách hàng.

1.7. Cấu trúc của nghiên cứu

Nghiên cứu được kết cấu thành 5 chương như sau:

Chương 1 – Giới thiệu nghiên cứu: Chương này giới thiệu chung về đề tài nghiên cứu, lý do chọn đề tài, mục tiêu nghiên cứu, câu hỏi nghiên cứu, đối tượng, phạm vi nghiên cứu, đóng góp và cấu trúc của nghiên cứu.

Chương 2 - Cơ sở lý thuyết: Chương này cung cấp một nền tảng lý thuyết cho nghiên cứu, giải thích các khái niệm, lý thuyết, mô hình, hoặc các kết quả nghiên cứu trước đó liên quan đến chủ đề được nghiên cứu.

Chương 3 - Phương pháp nghiên cứu: Chương này mô tả các phương pháp và quy trình được sử dụng để thu thập dữ liệu, phân tích dữ liệu và đánh giá kết quả của nghiên cứu.

Chương 4 - Kết quả nghiên cứu: Chương này trình bày các kết quả chính mà nghiên cứu đã đạt được dựa trên quá trình thu thập và phân tích dữ liệu.

Chương 5 - Kết luận và kiến nghị: Chương này tổng hợp và rút ra những kết luận cũng như những điểm hạn chế từ đề tài nghiên cứu, đề xuất các hướng tiếp cận và hướng phát triển trong tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Các khái niệm liên quan

2.1.1. Gom cụm người dùng trên mạng xã hội.

- **Khái niệm chung**

Gom cụm người dùng trên mạng xã hội là quá trình sử dụng các phương pháp phân cụm dữ liệu để nhóm các người dùng dựa trên hành vi như hành vi và sở thích thích, bình luận,

chia sẻ và tương tác với nội dung trên mạng xã hội. Mục tiêu là hiểu rõ hơn về sở thích và nhu cầu của người dùng để cung cấp dịch vụ cá nhân hóa và tối ưu hóa chiến lược tiếp thị [6].

- **Khái niệm trong phạm vi nghiên cứu**

Khái niệm gom cụm người dùng trên mạng xã hội dựa trên việc phân tích bài viết văn bản, các dòng trạng thái ngắn là quá trình tổ chức và nhóm hóa các cá nhân hoặc tài khoản trên các nền tảng mạng xã hội dựa trên nội dung của bài đăng. Thông qua phân tích văn bản, các thuật toán và công cụ máy học được sử dụng để đọc, hiểu và xử lý các bài viết trên mạng xã hội, từ đó nhận biết các đặc điểm chung, quan điểm hoặc sở thích giống nhau giữa các người dùng. Quá trình này giúp tạo ra các nhóm cụm người dùng có sở thích tương tự hoặc quan tâm đến các chủ đề cụ thể, từ đó các doanh nghiệp có thể tùy chỉnh và tối ưu hóa chiến lược tiếp thị của họ để tiếp cận đúng đối tượng mục tiêu và tăng cường tương tác trên mạng xã hội.

2.1.2. Vector embedding (Vector nhúng)

Tiến bộ đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên đã dẫn đến sự phát triển mạnh mẽ của các phương pháp biểu diễn từ ngữ. Trong đó, vector embedding (vector nhúng) đã trở thành một công cụ mạnh mẽ cho việc biểu diễn văn bản, từ ngữ trong không gian số hóa, và được sử dụng rộng rãi trong nhiều ứng dụng của lĩnh vực này.

Vector embedding là một biểu diễn số hóa của từ hoặc câu, trong đó từng phần tử của vector tương ứng với một chiều trong không gian n chiều. Mỗi từ hoặc câu sẽ được biểu diễn bằng một vector trong không gian N chiều này, trong đó mỗi chiều đại diện cho một đặc trưng nào đó của từ hoặc câu đó. Việc biểu diễn này giúp chúng ta có thể ánh xạ từ ngữ vào không gian số hóa một cách hiệu quả, từ đó tạo ra những thông tin có ý nghĩa và dễ dàng được sử dụng cho các mục đích xử lý ngôn ngữ tự nhiên.

Trong nghiên cứu này, vector embedding là một biểu diễn số học của các bài viết văn bản ngắn trong không gian nhiều chiều. Mỗi vector embedding biểu diễn một bài viết văn bản bằng một vectơ số có độ dài cố định, trong đó mỗi chiều của vectơ tương ứng với một thuộc tính hoặc đặc trưng của bài viết. Quá trình tạo ra vector embedding cho các bài viết văn bản thường được thực hiện bằng cách sử dụng các mô hình embedding trong lĩnh vực xử lý ngôn ngữ tự nhiên phổ biến như TF-IDF, Cbow, Word2Vec, FastText….Hay là những mô hình ứng dụng của phương pháp học chuyển đổi (Transfer learning) mới nhất hiện nay như Gte-large, Gte-base, Gte-small, E5-small, E5-base, E5-large…Các mô hình này được huấn luyện trên dữ liệu lớn và chất lượng cao để học cách biểu diễn các từ hoặc cụm từ dưới dạng các vectơ số. Sau đó, các bài viết văn bản có thể được biểu diễn bằng cách kết hợp các embedding của các từ hoặc cụm từ trong bài viết đó.

2.1.3. Giảm chiều dữ liệu

Giảm chiều dữ liệu trong máy học là quá trình giảm số chiều của dữ liệu ban đầu mà vẫn giữ được thông tin quan trọng nhất. Việc này thường được thực hiện để giảm thiểu chi phí tính toán và cải thiện hiệu suất của các mô hình máy học. Phương pháp giảm chiều dữ liệu bao gồm các kỹ thuật như PCA ,t-SNE và UMAP…. Mục tiêu của quá trình này là tạo ra một

biểu diễn dữ liệu có chiều thấp hơn nhưng vẫn giữ được càng nhiều thông tin càng tốt, giúp cải thiện khả năng khám phá và hiểu được dữ liệu. [7]

2.1.4. Trích xuất từ khóa (key word) và mô hình hóa chủ đề (Topic Modeling)

Trích xuất từ khóa trong dữ liệu văn bản là quá trình tự động hoặc bán tự động nhằm xác định và rút trích các từ hoặc cụm từ có ý nghĩa cao nhất từ mỗi văn bản. Mục tiêu của trích xuất từ khóa là để phục vụ quá trình mô hình hóa chủ đề cho một văn bản, nhằm để tóm tắt nội dung của văn bản một cách ngắn gọn, giúp người đọc hiểu được chủ đề và nội dung chính của văn bản mà không cần đọc toàn bộ. [8] Việc trích xuất từ khóa (keyword) và mô hình hóa chủ đề trong dữ liệu văn bản có thể được thực hiện bằng các mô hình học máy trong lĩnh vực xử lí ngôn ngữ tự nhiên như TF-IDF, LDA, TextRank, Word embedding….

2.2. Các lý thuyết liên quan

2.2.1. Các phương pháp giảm chiều dữ liệu

2.2.1.1. Phương pháp PCA

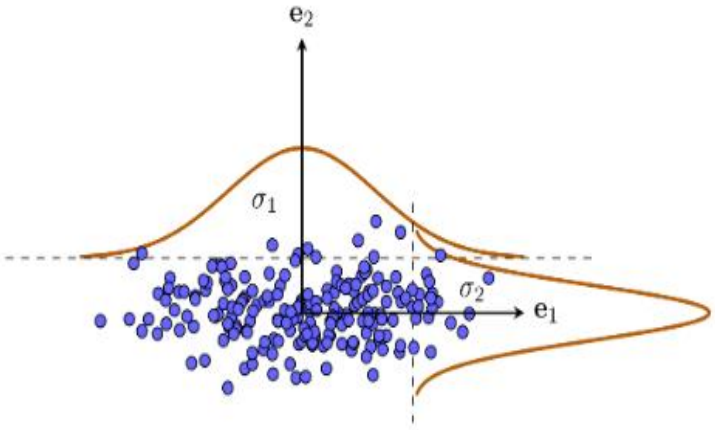
- **Định nghĩa**

Phân tích thành phần chính PCA (Principal Component Analysis) là một phương pháp thống kê cổ điển để chuyển đổi các thuộc tính của tập dữ liệu thành tập hợp mới gồm các thuộc tính không tương quan được gọi là thành phần chính (PC). PCA có thể được sử dụng để giảm kích thước của tập dữ liệu, trong khi vẫn giữ lại càng nhiều biến thể của tập dữ liệu càng tốt [9]

Giảm chiều dữ liệu, nói một cách đơn giản, là việc đi tìm một hàm số, hàm số này lấy đầu vào là một điểm dữ liệu ban đầu $x \epsilon R^D$ với D rất lớn, và tạo ra một điểm dữ liệu mới $z \epsilon R^K$ có số chiều $K < D$

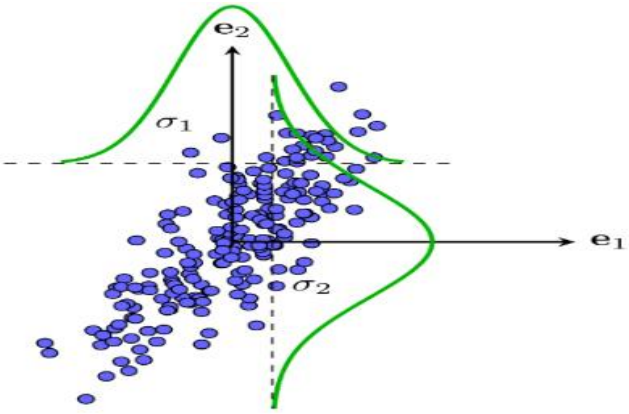
+ Yêu cầu của PCA: $x \in R^D \rightarrow z \in R^K,\ K < D$

+ Cách đơn giản nhất để giảm chiều dữ liệu từ D về $K < D$ là chỉ giữ lại K phần tử quan trọng.



Hình 2.1 Giảm chiều chưa tối ưu

Nguồn: [10]



Hình 2.2 Giảm chiều tối ưu

Nguồn: [10]

Hình 2.1: Phương sai của chiều thứ hai nhỏ hơn phương sai của chiều thứ nhất.

Hình 2.2 : Phương sai của hai chiều là bằng nhau.

Trong PCA, mục đích là tìm ra các trục (thành phần chính) mới sao cho phần lớn phương sai của dữ liệu được giữ lại, đồng thời loại bỏ những chiều ít quan trọng để giảm số chiều của không gian dữ liệu. Điều này giúp giảm chi phí tính toán và tăng hiệu suất trong các bài toán học máy.

- **Cách thức hoạt động**

Theo [10]

Bước 1: Tính vector trung bình của dữ liệu $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Bước 2: Tính vector chuẩn hóa của dữ liệu $\widehat{x_i} = x_i - \bar{x}$

Bước 3: Đặt $\hat{X} = [\widehat{x_1}, \widehat{x_2}, \dots, \widehat{x_D}]$ là ma trận dữ liệu chuẩn hóa, tính ma trận hiệp phương sai $S = \frac{1}{N}\hat{X}\hat{X}^T$

Bước 4: Tính các trị riêng, vector riêng tương ứng có $l_2\ norm$ bằng 1 của S, xếp chúng theo thứ tự giảm dần của trị riêng

Bước 5: Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận U_K

Bước 6: Chiếu dữ liệu chuẩn hóa $\hat{X}$ xuống không gian con U_K

Bước 7: Dữ liệu mới là tọa độ của các điểm dữ liệu trong không gian mới $Z = U_K^T\hat{X}$

2.2.1.2. Phương pháp t-SNE

- **Định nghĩa**

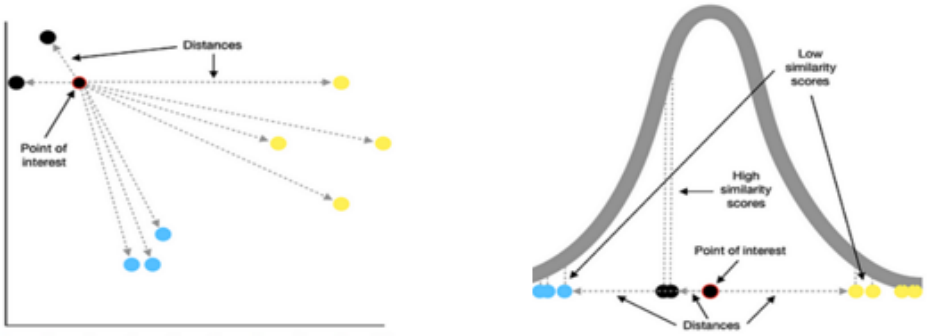
t-SNE (t-Distributed Stochastic Neighbor Embedding) là phương pháp giảm chiều dữ liệu phi tuyến tính, biểu diễn dữ liệu số chiều cao trong không gian số chiều thấp sao cho các điểm tương đồng trong không gian gốc vẫn giữ tính tương đồng trong không gian mới. Phương pháp này thường được sử dụng để khám phá mối quan hệ giữa các điểm dữ liệu chiều cao, tạo biểu đồ phân tán, biểu đồ nhúng để biểu diễn cụm hay cấu trúc dữ liệu phức tạp. t-SNE mang lại cảm giác, trực giác về cách dữ liệu được sắp xếp ở chiều cao hơn, thường được dùng

để trực quan hóa dữ liệu phức tạp xuống 2, 3 chiều, giúp hiểu mẫu, mối quan hệ cơ bản trong dữ liệu. [11]

- **Cách thức hoạt động**

Theo [12].

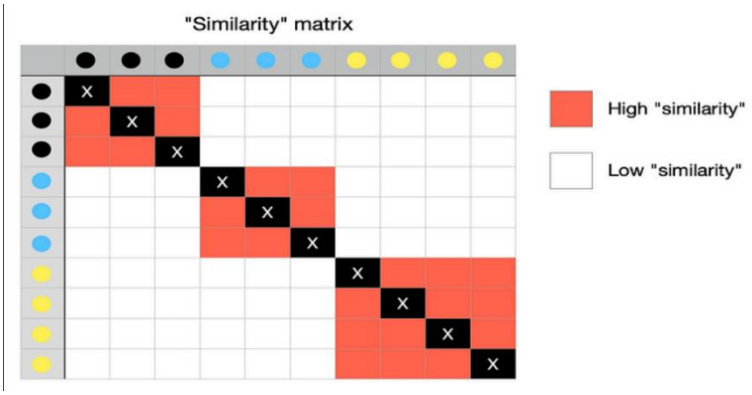
Bước 1: t-SNE bắt đầu bằng việc xác định "sự giống nhau"(similarity) của các điểm dựa trên khoảng cách giữa chúng. Các điểm ở gần được coi là "tương tự"(similar), trong khi các điểm ở xa được coi là "không giống nhau"(dissimilar). Nó đạt được điều này bằng cách đo khoảng cách giữa điểm quan tâm và các điểm khác, sau đó đặt chúng trên đường cong Bình thường. Nó thực hiện điều này cho mọi điểm, áp dụng một số tỷ lệ để tính đến sự thay đổi về mật độ của các vùng khác nhau. Ví dụ: hình minh họa bên dưới có mật độ cao hơn ở vùng có các điểm màu xanh lam và mật độ thấp hơn ở vùng có các điểm màu vàng.



Hình 2.3 Không gian đa chiều gốc và đường cong phân phối chuẩn

Nguồn: [12]

Kết quả của những phép tính này là một ma trận chứa điểm tương đồng giữa từng cặp điểm từ không gian đa chiều ban đầu.



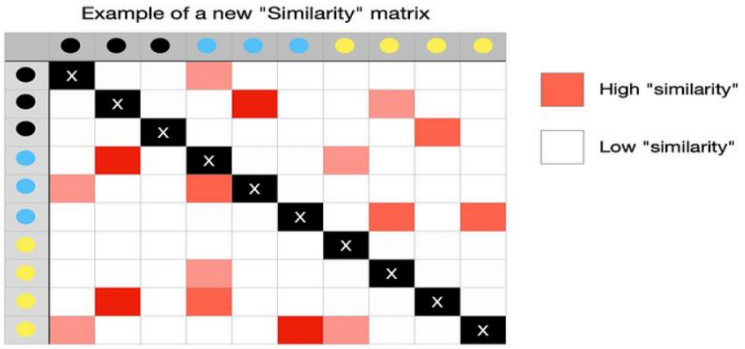
Hình 2.4 Ma trận tương đồng ban đầu

Nguồn: [12]

Bước 2: Tiếp theo, t-SNE ánh xạ ngẫu nhiên tất cả các điểm vào không gian có chiều thấp hơn và tính toán "sự tương đồng" giữa các điểm như được mô tả trong quy trình trên. Tuy

nhiên, có một điểm khác biệt là lần này thuật toán sử dụng phân phối t thay vì phân phối chuẩn.

Tuy nhiên, không có gì đáng ngạc nhiên khi ma trận "tương tự" mới khác biệt đáng kể so với ma trận ban đầu do ánh xạ ngẫu nhiên. Đây là một ví dụ về những gì nó có thể trông như thế nào



Hình 2.5 Ma trận tương đồng mới

Nguồn: [12]

Bước 3: Bây giờ mục tiêu của thuật toán là làm cho ma trận "tương tự" mới trông giống ma trận ban đầu bằng cách sử dụng phương pháp lặp. Với mỗi lần lặp lại, các điểm sẽ di chuyển về phía "hàng xóm gần nhất" của chúng từ không gian có chiều cao hơn ban đầu và cách xa những không gian ở xa. Ma trận "tương tự" mới dần dần bắt đầu trông giống ma trận ban đầu hơn. Quá trình tiếp tục cho đến khi đạt được số lần lặp tối đa hoặc không thể cải thiện thêm nữa. Theo thuật ngữ khoa học hơn, lời giải thích ở trên mô tả quá trình của một thuật toán cố gắng giảm thiểu sự phân kỳ Kullback–Leibler (phân kỳ KL) thông qua việc giảm độ dốc.

2.2.1.3. Phương pháp UMAP

- **Định nghĩa**

UMAP (Uniform Manifold Approximation and Projection) là một kỹ thuật học đa tạp mới để giảm kích thước. UMAP được xây dựng từ khung lý thuyết dựa trên hình học Riemannian và cấu trúc liên kết đại số. kết quả là một thuật toán có thể mở rộng thực tế có thể áp dụng cho dữ liệu trong thế giới thực. Thuật toán UMAP cạnh tranh với t-SNE về chất lượng hiển thị và được cho là bảo toàn được nhiều cấu trúc tổng thể hơn với hiệu suất thời gian chạy vượt trội. Hơn nữa, UMAP không có hạn chế tính toán đối với việc nhúng kích thước, khiến nó trở thành một kỹ thuật giảm kích thước cho mục đích chung cho học máy. [13]

Ý tưởng cơ bản đằng sau UMAP là xây dựng cách biểu diễn dữ liệu theo chiều thấp để bảo toàn cấu trúc cục bộ và toàn cục của không gian chiều cao. UMAP sử dụng cách tiếp cận dựa trên biểu đồ để xây dựng biểu diễn tôpô của dữ liệu, sau đó được nhúng vào không gian có chiều thấp bằng cách sử dụng độ dốc giảm dần ngẫu nhiên. [14]

- **Cách thức hoạt động**

Theo [15]

Chúng ta có thể chia UMAP thành hai bước chính:

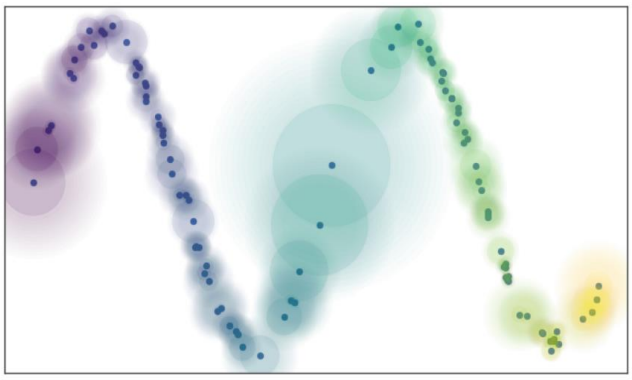
Bước 1: Tìm hiểu cấu trúc đa tạp (Learning the manifold structure)

Để ánh xạ dữ liệu tới các chiều thấp hơn, trước tiên chúng ta cần tìm hiểu xem nó trông như thế nào trong không gian chiều cao hơn.

Bước 1.1: Tìm hàng xóm gần nhất (Finding nearest neighbors): UMAP bắt đầu bằng cách tìm hàng xóm gần nhất, chúng ta có thể chỉ định số lượng hàng xóm gần nhất mà chúng ta muốn sử dụng bằng cách điều chỉnh siêu tham số n_neighbors của UMAP. Điều quan trọng là phải thử nghiệm số lượng n_neighbor vì nó kiểm soát cách UMAP cân bằng cấu trúc cục bộ và toàn cầu trong dữ liệu . Nó thực hiện điều đó bằng cách hạn chế kích thước của vùng lân cận cục bộ khi cố gắng tìm hiểu cấu trúc đa dạng. Về cơ bản, một giá trị nhỏ cho n_neighbors có nghĩa là chúng ta muốn một cách diễn giải rất cục bộ để nắm bắt chính xác từng chi tiết nhỏ của cấu trúc. Ngược lại, giá trị n_lân cận lớn có nghĩa là ước tính của chúng tôi sẽ dựa trên các vùng lớn hơn, do đó chính xác hơn trên toàn bộ đa tạp.

Bước 1.2: Xây dựng biểu đồ (Constructing neighbour graph): Tiếp theo, UMAP cần xây dựng biểu đồ bằng cách kết nối các lân cận gần nhất đã được xác định trước đó. Để hiểu quá trình này, chúng ta cần xem xét một số thành phần phụ giải thích cách biểu đồ lân cận xuất hiện.

Bước 1.2.1: Khoảng cách khác nhau (Varying distance) **:** Như đã nêu trong phân tích tên của UMAP, chúng tôi giả định sự phân bố đồng đều các điểm trên đa tạp, cho thấy rằng không gian giữa chúng đang giãn ra hoặc co lại tùy theo nơi dữ liệu có vẻ thưa thớt hơn hoặc dày đặc hơn. Về cơ bản, điều đó có nghĩa là thước đo khoảng cách không phổ biến trên toàn bộ không gian và thay vào đó, nó khác nhau giữa các vùng khác nhau. Chúng ta có thể hình dung nó bằng cách vẽ các vòng tròn/hình cầu xung quanh mỗi điểm dữ liệu. Các điểm này dường như có kích thước khác nhau do thước đo khoảng cách khác nhau (xem hình minh họa bên dưới).



Hình 2.6 Mô hình kết nối cục bộ và tập mờ mờ

Nguồn: [15]

Bước 1.2.2: Kết nối cục bộ (Local connectivity): Tiếp theo, chúng tôi muốn đảm bảo rằng cấu trúc đa dạng mà chúng tôi đang cố gắng tìm hiểu không dẫn đến nhiều điểm không được kết nối. May mắn thay, chúng ta có thể sử dụng một siêu tham số khác gọi là local_connectivity (giá trị mặc định = 1) để giải quyết vấn đề tiềm ẩn này/ Khi đặt

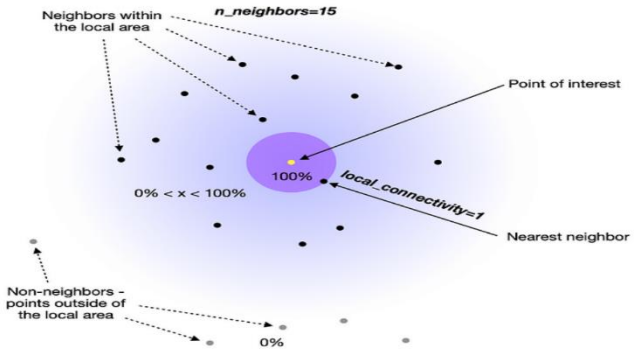
local_connectivity=1, chúng tôi cho thuật toán biết rằng mọi điểm trong không gian có nhiều chiều hơn đều được kết nối với ít nhất một điểm khác. Chúng ta có thể thấy trong hình minh họa ở trên mỗi vòng tròn liền nét chạm vào ít nhất một điểm dữ liệu như thế nào.

Bước 1.2.3: Vùng mờ (Fuzzy area): Hình minh họa ở trên cũng chứa các vòng tròn mờ mở rộng ra ngoài vùng lân cận gần nhất cho chúng ta biết rằng độ chắc chắn của sự kết nối với các điểm khác sẽ giảm đi khi chúng ta càng rời xa điểm quan tâm. Cách dễ nhất để nghĩ về nó là xem hai siêu tham số (local_connectivity và n_neighbors) dưới dạng giới hạn dưới và giới hạn trên :

+ local_connectivity (mặc định=1): chắc chắn 100% rằng mỗi điểm được kết nối với ít nhất một điểm khác (giới hạn thấp hơn đối với một số kết nối).

+ n_neighbors (mặc định=15): có 0% khả năng một điểm được kết nối trực tiếp với điểm lân cận thứ 16+ vì nó nằm ngoài khu vực cục bộ được UMAP sử dụng khi xây dựng biểu đồ.

Hàng xóm từ 2 đến 15 - có một số mức độ chắc chắn (>0% nhưng <100%) rằng một điểm được kết nối với hàng xóm thứ 2 đến thứ 15 của nó.

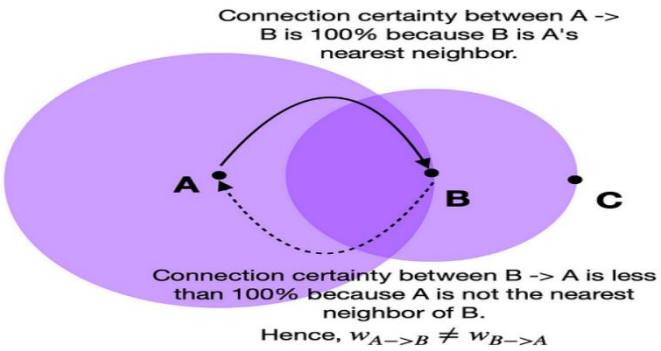


Hình 2.7 Vùng lân cận và xu hướng các điểm hàng xóm

Nguồn: [15]

Bước 1.2.4: Hợp nhất các cạnh (Merging of edges): Cuối cùng, chúng ta cần hiểu rằng độ chắc chắn của kết nối được thảo luận ở trên được thể hiện thông qua các trọng số của cạnh (w).

Vì chúng ta đã sử dụng một cách tiếp cận khoảng cách khác nhau nên chắc chắn sẽ có trường hợp trọng số của các cạnh không thẳng hàng khi nhìn từ góc độ của từng điểm. Ví dụ: trọng số cạnh của điểm $A \rightarrow B$ sẽ khác với trọng số cạnh của $B \rightarrow A$



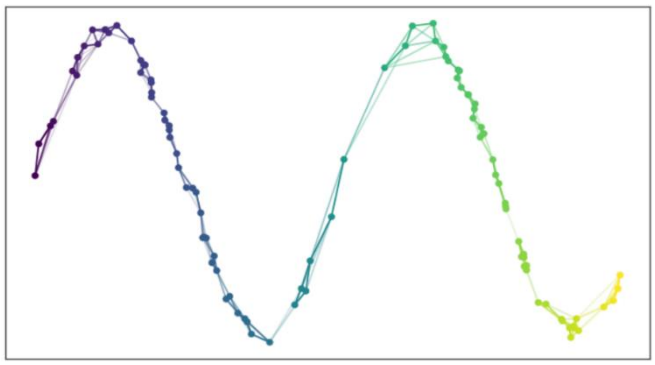
Hình 2.8 Trọng lượng cạnh không đồng ý

Nguồn: [15]

UMAP khắc phục vấn đề về trọng số các cạnh không đồng nhất mà chúng ta vừa mô tả bằng cách lấy sự kết hợp của hai cạnh.

Nếu chúng ta muốn hợp nhất hai cạnh không đồng nhất với nhau có trọng số a và b thì chúng ta cần có một cạnh duy nhất có tổng trọng số $a + b - a.b$. Cách nghĩ về điều này là các trọng số thực sự là xác suất tồn tại của một cạnh (1-đơn hình). Trọng số kết hợp khi đó là xác suất tồn tại ít nhất một trong các cạnh.

Cuối cùng, chúng ta có được một biểu đồ vùng lân cận được kết nối trông như thế này:



Hình 2.9 Đồ thị có trọng số cạnh thích hợp

Nguồn: [15]

Bước 2: Tìm biểu diễn có chiều thấp (Finding a low-dimensional representation)

Sau khi tìm hiểu đa tạp gần đúng từ không gian có chiều cao hơn, bước tiếp theo của UMAP là chiếu nó (ánh xạ nó) sang không gian có chiều thấp hơn.

Bước 2.1: Khoảng cách tối thiểu (Minimum distance): Không giống như bước đầu tiên, chúng tôi không muốn thay đổi khoảng cách trong biểu diễn không gian có chiều thấp hơn. Thay vào đó, chúng ta muốn khoảng cách trên đa tạp là khoảng cách Euclide tiêu chuẩn đối với hệ tọa độ tổng thể.

Việc chuyển từ khoảng cách khác nhau sang khoảng cách tiêu chuẩn cũng ảnh hưởng đến khoảng cách gần nhất với những người hàng xóm gần nhất. Do đó, chúng ta phải chuyển một

siêu tham số khác gọi là min_dist (mặc định=0,1) để xác định khoảng cách tối thiểu giữa các điểm được nhúng.

Về cơ bản, chúng ta có thể kiểm soát mức độ phân tán điểm tối thiểu, tránh các tình huống có nhiều điểm nằm chồng lên nhau trong quá trình nhúng chiều thấp hơn.

Bước 2.2: Giảm thiểu hàm chi phí (Minimizing the cost function): Với khoảng cách tối thiểu được chỉ định, thuật toán có thể bắt đầu tìm kiếm một biểu diễn đa tạp có chiều thấp tốt. UMAP thực hiện điều đó bằng cách giảm thiểu hàm chi phí sau, còn được gọi là Entropy chéo (CE):

Mục tiêu cuối cùng là tìm trọng số tối ưu của các cạnh trong biểu diễn chiều thấp. Các trọng số tối ưu này xuất hiện khi hàm chi phí Cross-Entropy ở trên được giảm thiểu sau quá trình giảm độ dốc ngẫu nhiên lặp đi lặp lại. Và đó là nó! Công việc của UMAP hiện đã hoàn tất và chúng ta được cung cấp một mảng chứa tọa độ của từng điểm dữ liệu trong một không gian có chiều thấp hơn được chỉ định.

2.2.2. Các mô hình Vector embedding (Vector nhúng).

2.2.2.1. Tổng quan về mô hình Vector embedding

Sự tương đồng của văn bản ngắn ngày nay trên các trang mạng xã hội đã thu hút nhiều sự chú ý hơn trong những năm gần đây và việc hiểu ngữ nghĩa chính xác giữa các tài liệu là một thách thức để hiểu sự đa dạng và mơ hồ về từ vựng [16]. Đại diện cho văn bản ngắn là rất quan trọng trong xử lí ngôn ngữ tự nhiên nhưng đầy thách thức do sự thưa thớt của nó; tính chiều hướng cao; phức tạp; khối lượng lớn và nhiều thông tin không liên quan, dư thừa và ồn ào [17]. Kết quả là, các phương pháp tính toán tương tự ngữ nghĩa truyền thống là một rào cản đáng kể vì chúng không hiệu quả trong các trường hợp khác nhau. Nhiều hệ thống truyền thống hiện tại không giải quyết được các thuật ngữ không được bao phủ bởi các từ đồng nghĩa và không thể xử lý các chữ viết tắt, từ viết tắt, tên thương hiệu và các thuật ngữ khác [18]. Ví dụ về các hệ thống truyền thống này là BOW và TF-IDF, đại diện cho văn bản dưới dạng vectơ giá trị thực để giúp tính toán tương tự ngữ nghĩa. Tuy nhiên, những chiến lược này không thể giải thích cho thực tế là các từ có ý nghĩa đa dạng và các từ khác nhau có thể được sử dụng để đại diện cho cùng một khái niệm. Ví dụ, hãy xem xét hai câu: 'A đi đá banh' và 'A yêu bóng đá'. Mặc dù hai câu này có cùng ý nghĩa là nhân vật A có niềm yêu thích về khái niệm ‘bóng đá’, nhưng chúng không sử dụng cùng một từ.

Các phương pháp này nắm bắt các tính năng từ vựng của văn bản và rất đơn giản để thực hiện. Tuy nhiên, họ bỏ qua các tính năng ngữ nghĩa và cú pháp của văn bản. Để giải quyết vấn đề này, trong bài nghiên cứu , chúng tôi đề xuất sử dụng phương pháp học chuyển đổi (transfer learning) , trong đó các mô hình pretrained được huấn luyện trên dữ liệu lớn trước đó được sử dụng như một bước khởi đầu để huấn luyện lại trên tập dữ liệu cụ thể của mình, đã chứng tỏ sức mạnh đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên. Thay vì phải bắt đầu từ đầu và huấn luyện mô hình từ tập dữ liệu nhỏ, việc sử dụng các mô hình đã được huấn luyện trước (pretrained model) giúp tiết kiệm thời gian và công sức đáng kể. Những mô hình này đã được huấn luyện trên dữ liệu lớn và đa dạng, giúp chúng hiểu biết ngôn ngữ tự nhiên

một cách toàn diện hơn. Khi được huấn luyện lại trên tập dữ liệu cụ thể, các mô hình này có khả năng học được các đặc trưng và mối quan hệ ngữ nghĩa đặc biệt của tập dữ liệu đó, mang lại hiệu suất dự đoán cao hơn và giảm thiểu công việc phải thực hiện lặp lại một cách đáng kể. Cụ thể chúng tôi tiến hành thực hiện và đánh giá trên 3 mô hình pretrained E5 là E5-small, E5-base, E5-large.

2.2.2.2. Lý thuyết các mô hình vector nhúng E5.

Theo nghiên cứu của [19]

- **Định nghĩa**

Mô hình vector nhúng E5 được phát triển bởi Google AI, một nhóm nghiên cứu hàng đầu trong lĩnh vực trí tuệ nhân tạo và được giới thiệu lần đầu tiên vào năm 2020 trong bài báo "E5: A Large-scale Language Model for Text-to-text Transfer Learning". Mô hình yêu cầu cung cấp một tập dữ liệu gồm đầu vào (văn bản) và sẽ cho ra đầu ra mong muốn (biểu diễn vector số), sau đó mô hình học cách tự động tạo ra biểu diễn số cho văn bản mới. E5 được phát minh với mục tiêu có thể sử dụng cho bất kỳ tác vụ nào cần biểu diễn văn bản bằng một vector duy nhất, hoạt động tốt trong cả hai trường hợp không cần tinh chỉnh và được tinh chỉnh. Với 3 phiên bản để tùy chỉnh cho các kích cỡ dữ liệu của người áp dụng là E5-small, E5-base, E5-large.

- **Xây dựng mô hình**

Mô hình vector nhúng E5 nhằm mục đích cung cấp các nhúng văn bản có sẵn mạnh mẽ phù hợp với mọi tác vụ yêu cầu biểu diễn vector đơn trong cả cài đặt zero-shot hoặc tinh chỉnh. Để đạt được mục tiêu này, thay vì dựa vào dữ liệu được gắn nhãn hạn chế hoặc các cặp văn bản tổng hợp chất lượng thấp, nghiên cứu ngược lại đào tạo nhúng E5 từ CCPairs, một bộ dữ liệu cặp văn bản quy mô web được quản lý có chứa tín hiệu đào tạo không đồng nhất. Nghiên cứu xây dựng bộ dữ liệu CCPairs bằng cách kết hợp các nguồn dữ liệu bán cấu trúc khác nhau như CommunityQA, Common Crawl và Scientific papers, đồng thời thực hiện lọc tích cực với bộ lọc dựa trên tính nhất quán [20] để cải thiện chất lượng dữ liệu. nghiên cứu chọn một công thức học tập tương phản đơn giản bằng cách sử dụng âm bản hàng loạt với kích thước lô lớn để đào tạo mô hình. Kết quả trên 56 bộ dữ liệu từ điểm chuẩn MTEB được giới thiệu gần đây [21] cho thấy E5base cạnh tranh với GTRxxl và Sentence-T5xxl, có nhiều hơn 40× tham số.

- **Kiểm định:**

Các mô hình E5 được đánh giá trên một số nhiệm vụ xử lí ngôn ngữ tự nhiên khác nhau, bao gồm: Truy xuất thông tin, tóm tắt văn bản, dịch máy, trả lời câu hỏi. Mô hình đã đạt được kết quả ấn tượng trên các nhiệm vụ này, cho thấy hiệu quả của chúng trong việc xử lý ngôn ngữ tự nhiên. Dưới đây là các kết quả chỉ số đánh giá trung bình của các mô hình E5 so với các mô hình khác, khi được cùng huấn luyện và thử nghiệm từ 56 bộ dữ liệu theo điểm chuẩn MTEB trải dài trên 6 loại: Phân loại (Class.), Phân cụm (Clust.), Phân loại cặp (PairClass.), Xếp hạng lại, Truy xuất (Retr.), STS và Tóm tắt (Summ.)

# of datasets →	Class. 12	Clust. 11	PairClass. 3	Rerank 4	Retr. 15	STS 10	Summ. 1	Avg 56
Unsupervised models								
Glove	57.3	27.7	70.9	43.3	21.6	61.9	28.9	42.0
BERT	61.7	30.1	56.3	43.4	10.6	54.4	29.8	38.3
SimCSE-BERT-unsup	62.5	29.0	70.3	46.5	20.3	74.3	31.2	45.5
E5-PT$_{small}$	67.0	41.7	78.2	53.1	40.8	68.8	**32.7**	54.3
E5-PT$_{base}$	67.9	43.4	79.2	53.5	42.9	69.5	31.1	55.6
E5-PT$_{large}$	69.0	**44.3**	80.3	54.4	44.2	69.9	32.6	56.6
Supervised models								
SimCSE-BERT-sup	67.3	33.4	73.7	47.5	21.8	79.1	23.3	48.7
BERT-FT$_{base}$	68.7	33.9	82.6	50.5	41.5	79.2	29.0	55.2
Contriever	66.7	41.1	82.5	53.1	41.9	76.5	30.4	56.0
GTR$_{large}$	67.1	41.6	85.3	55.4	47.4	78.2	29.5	58.3
Sentence-T5$_{large}$	72.3	41.7	85.0	54.0	36.7	81.8	29.6	57.1
E5$_{small}$	71.7	39.5	85.1	54.5	46.0	80.9	31.4	58.9
E5$_{base}$	72.6	42.1	85.1	55.7	48.7	81.0	31.0	60.4
E5$_{large}$	**73.1**	43.3	**85.9**	**56.5**	**50.0**	**82.1**	31.0	**61.4**
Larger models								
GTR$_{xxl}$	67.4	42.4	86.1	56.7	48.5	78.4	30.6	59.0
Sentence-T5$_{xxl}$	73.4	43.7	85.1	56.4	42.2	82.6	30.1	59.5

Hình 2.10 Kết quả điểm chuẩn MTEB giữa các mô hình

Nguồn: [19]

Các mô hình E5 không chỉ vượt trội đáng kể so với các mô hình hiện có có kích thước tương tự, mà còn đạt được kết quả tương đương với các mô hình lớn hơn nhiều. 2 mô hình đứng đầu trên bảng xếp hạng MTEB là 7 GTRxxl và Sentence-T5xxl có 4,8 tỷ tham số, trong khi mô hình E5-large nhỏ hơn 10 lần với 300 triệu tham số. Đối với hầu hết các loại tác vụ, hiệu suất được cải thiện sau khi điều chỉnh tinh được giám sát. Phù hợp với các công trình trước đây, điều này một lần nữa cho thấy tầm quan trọng của việc kết hợp kiến thức của con người để học các nhúng văn bản tốt hơn.

Đặc biệt trong tác vụ phân cụm (Clust.), các mô hình E5 đều thể hiện sự vượt trội về chỉ số đánh giá. Đây cũng là lí do nhóm nghiên cứu chúng tôi ưu tiên lựa chọn các mô hình E5, cụ thể là 3 mô hình E5-base, E5-small , E5-large để thực hiện đề tài nghiên cứu gom cụm người dùng dựa trên chủ đề trích xuất được từ các bài đăng của họ trên mạng xã hội, nhằm đưa ra kết quả tốt nhất cho bài nghiên cứu.

- **Ưu điểm chung:**

+ Hiệu quả: Các mô hình E5 cung cấp hiệu suất tốt với tốc độ và chi phí tính toán thấp so với các mô hình nhúng văn bản khác.

+ Đa dạng: Có nhiều phiên bản E5 với kích thước và độ chính xác khác nhau, phù hợp với nhiều nhu cầu sử dụng.

+ Khả năng mở rộng: Các mô hình E5 có thể được huấn luyện trên tập dữ liệu lớn, giúp nâng cao hiệu suất cho các tác vụ cụ thể.

+ Hỗ trợ nhiều ngôn ngữ: Các mô hình E5 hỗ trợ nhiều ngôn ngữ, giúp mở rộng khả năng ứng dụng.

+ Mã nguồn mở: E5 là mã nguồn mở, cho phép người dùng tùy chỉnh và phát triển thêm.

2.2.3. Phương pháp mô hình hóa chủ đề LDA

- **Lý thuyết**

Mô hình hóa chủ đề LDA (Latent Dirichlet Allocation) là một mô hình thống kê đại diện cho một tập hợp các tài liệu, với mỗi tài liệu được coi là một tập hợp các chủ đề, và mỗi chủ đề là một phân phối xác suất trên một tập hợp từ vựng cố định. LDA là một mô hình sinh ra văn bản, có nghĩa là nó cung cấp một quá trình ngẫu nhiên có thể sinh ra tài liệu thực tế. Mô hình này được sử dụng rộng rãi trong các lĩnh vực như khai thác dữ liệu văn bản, xử lý ngôn ngữ tự nhiên và lọc thông tin.

- **Cách thức hoạt động**

Bước 1: Khởi tạo ngẫu nhiên các tham số:

+ Khởi tạo ngẫu nhiên phân phối từ $\emptyset k$ cho mỗi chủ đề k theo phân phối Dirichlet β.

+ Khởi tạo ngẫu nhiên phân phối chủ đề θd cho mỗi tài liệu d theo phân phối Dirichlet α.

Bước 2: Gán chủ đề cho mỗi từ trong tập dữ liệu:

+ Cho mỗi từ wn trong tập dữ liệu văn bản:

+ Tính xác suất có điều kiện $P(zn = k|wn, z-n, \alpha, \beta)$ cho mỗi chủ đề k, với $z - n$ là các gán chủ đề cho các từ khác.

Gán từ wn vào chủ đề k với xác suất tương ứng.

Bước 3: Cập nhật phân phối từ $\emptyset k$ cho mỗi chủ đề k:

+ Tính tỷ lệ số lần từ w xuất hiện trong chủ đề k trên tổng số từ của chủ đề k.

Bước 4: Cập nhật phân phối chủ đề θd cho mỗi tài liệu d:

+ Tính tỷ lệ số lần chủ đề k xuất hiện trong tài liệu d trên tổng số chủ đề của tài liệu d.

Bước 5: Lặp lại bước 2 tới bước 4 cho đến khi mô hình hội tụ:

+ Thường sử dụng phương pháp Gibbs sampling để lấy mẫu chủ đề mới cho các từ.

+ Sau mỗi lần lặp, các phân phối $\emptyset k$ và θd được cập nhật.

+ Quá trình lặp lại cho đến khi không có sự thay đổi lớn về phân phối hoặc đạt số vòng lặp tối đa.

+ Kết quả cho ra cuối cùng là các danh sách chứa các từ wn có xác suất cao nhất ứng với từng chủ đề k, các từ này sẽ đại diện cho chủ đề k để người áp dụng thuật toán có thể xác định ra chủ đề k là gì.

2.3. Các nghiên cứu liên quan

2.3.1. Nghiên cứu trong nước

Nghiên cứu của [22] đã tiến hành khảo sát và trình bày một cái nhìn tổng quan về mạng xã hội, bài toán phân nhóm trong mạng xã hội và các thuật toán thông dụng được áp dụng trong việc phân nhóm. Nghiên cứu đã phân tích các ưu và nhược điểm của từng thuật toán.

Dựa trên những hiểu biết thu được từ các thuật toán đã nghiên cứu, nghiên cứu đã đề xuất một phương pháp mới mang tên INC(Incre-Comm-Extraction). Phương pháp này được phát triển dựa trên cơ sở của thuật toán CNM (Clauset-Newman-Moore), với việc sử dụng phương pháp tiếp cận đệ quy để tìm kiếm các nhóm con có ý nghĩa bổ sung trong các nhóm lớn mà thuật toán CNM tạo ra, đồng thời cải thiện chất lượng của các nhóm so với CNM.

Mạng xã hội được tạo ra trong phạm vi của dự án nghiên cứu này dựa trên hành vi tương tác của người dùng. Điều này làm cho mạng trở nên rất động, không giống như các nghiên cứu trước đó chỉ tập trung vào các mạng tĩnh với các mối quan hệ như bạn bè hoặc theo dõi. Mạng xã hội được xây dựng bằng phương pháp INC là một đồ thị vô hướng có trọng số. Trọng số của mỗi cạnh thể hiện mức độ mạnh mẽ của mối quan tâm giữa hai đỉnh trong đồ thị.

Trong phạm vi của dự án nghiên cứu này, mạng xã hội được nghiên cứu là Facebook - mạng xã hội phổ biến nhất ở Việt Nam và trên toàn thế giới. Các đỉnh trong mạng xã hội này thường là các trang Facebook, đại diện cho cá nhân, nhóm, tổ chức, công ty hoặc tập đoàn hoạt động trong nhiều lĩnh vực khác nhau. Việc khám phá các nhóm con có ý nghĩa trong mạng xã hội này đóng vai trò quan trọng trong thực tế, như trong bài toán tiếp thị sản phẩm tới các thành viên trong nhóm có sự quan tâm chung tới sản phẩm đó.

2.3.2. Nghiên cứu ngoài nước

Nghiên cứu của [23] tập trung vào tổng hợp các tài liệu xuất bản liên quan đến phân cụm văn bản ngắn STC (Short Text Clustering) và trình bày các ứng dụng của nó. Nghiên cứu cung cấp cái nhìn tổng quan về STC, mô tả chi tiết các giai đoạn của quá trình phân cụm. Các phương pháp biểu diễn văn bản ngắn dưới dạng vector số học, ưu nhược điểm và tác động khi áp dụng các phương pháp khác nhau đã được trình bày. Đồng thời, nghiên cứu giải thích các phương pháp học sâu cơ bản trong xử lý văn bản. Một số phương pháp hoạt động tốt trong nghiên cứu này nhưng kém hiệu quả trong nghiên cứu khác, dẫn đến vectơ đặc trưng thưa thớt, chiều cao ít đặc biệt để đo khoảng cách như TF-IDF, CBOW... Các nghiên cứu tiếp theo có thể giải quyết vấn đề biểu diễn văn bản ngắn và cải thiện độ chính xác phân cụm.

Nghiên cứu của [24] khảo sát tác động của chủ đề và tình cảm trong bài đăng trên fanpage Facebook đến mức độ tương tác của người dùng. Nghiên cứu trích xuất dữ liệu bài đăng không cấu trúc từ fanpage Toyota, phân loại bài đăng dựa trên nghiên cứu trước, rồi trích xuất chủ đề và tình cảm bằng ngôn ngữ R. Sau đó, kiểm tra giả thuyết bằng mô hình hồi quy nhị phân. Kết quả xác nhận hai giả thuyết: Các chủ đề khác nhau có ảnh hưởng khác nhau đến tương tác, trong đó chủ đề bán hàng và chiến dịch tiếp thị hiệu quả nhất. Mức độ tương tác tăng lên khi kích thích cảm xúc tích cực hoặc tiêu cực của người dùng.

Nghiên cứu của [25] cũng giới thiệu kỹ thuật mô hình hoá chủ đề để khai thác văn bản. Tùy mục đích, các kỹ thuật khác có thể phù hợp hơn. Mô hình hoá chủ đề phù hợp để nghiên

cứu quy nạp mẫu trong các bộ văn bản lớn, đặc biệt cho nghiên cứu thăm dò, mở rộng lý thuyết. Ngược lại, phương pháp dựa trên từ điển ít phù hợp nếu đối tượng nghiên cứu chưa xác định. Phương pháp từ điển dễ tạo ra từ điển, quy tắc phù hợp với giả thuyết nhưng tiềm năng khám phá bị giới hạn. Cuối cùng, nghiên cứu giải thích hai phương pháp phức tạp LDA và LASSO theo cách dễ hiểu, nhà nghiên cứu quan tâm nên nghiên cứu tài liệu liên quan để hiểu rõ hơn trước khi giải thích kết quả.

Tóm tắt chương 2

Chương này cung cấp nền tảng lý thuyết bằng cách giới thiệu các khái niệm quan trọng như gom cụm người dùng, vector nhúng và giảm chiều dữ liệu. Tiếp theo, trình bày lý thuyết của 3 phương pháp giảm chiều dữ liệu phổ biến: PCA, T-SNE và UMAP. Chương cũng tập trung vào ứng dụng transfer learning với mô hình đã được đào tạo sẵn trong xử lý ngôn ngữ tự nhiên và mô hình chủ đề LDA. Cuối cùng, tổng hợp và phân tích các nghiên cứu liên quan trước đây để xác định khoảng trống và cơ hội đóng góp mới.

Tóm lại, chương này đóng vai trò quan trọng trong việc thiết lập nền tảng lý thuyết vững chắc, cung cấp các khái niệm, lý thuyết và bối cảnh nghiên cứu cần thiết để hỗ trợ cho quá trình thực hiện và phân tích kết quả của nghiên cứu này.

CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

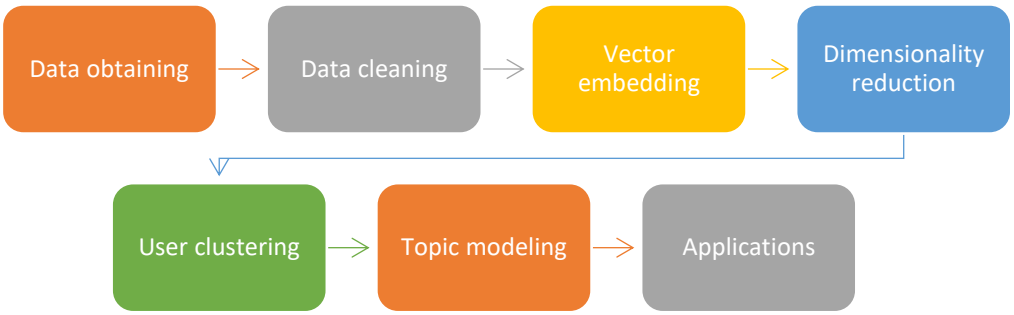
Trong đề tài này, chúng tôi tiến hành một loạt các bước phương pháp nhằm nắm bắt và hiểu sâu hơn về dữ liệu thu thập được từ các bài viết của người dùng trên mạng xã hội, từ đó có thể hiểu được vấn đề người dùng đang quan tâm nhằm đưa ra những ứng dụng thiết thực.

Bước đầu tiên, "Data obtaining", tập trung vào việc thu thập dữ liệu từ các nguồn đa dạng, đảm bảo dữ liệu thu được là đủ và đại diện cho phạm vi nghiên cứu. Tiếp theo, "Data Cleaning" là quá trình loại bỏ dữ liệu không đúng, xử lý dữ liệu còn thiếu và tiền xử lý để chuẩn bị cho các bước tiếp theo.

Sau đó, chúng tôi sử dụng "Vector Embedding" để biểu diễn dữ liệu dưới dạng vector, giúp thu thập thông tin quan trọng từ dữ liệu không cấu trúc như văn bản. "Dimensionality Reduction" là bước tiếp theo, giúp giảm số chiều của dữ liệu mà vẫn giữ lại đặc trưng quan trọng, từ đó tối ưu hóa hiệu suất phân tích.

Tiếp theo, chúng tôi sử dụng "User Clustering" để phân cụm người dùng dựa trên hành vi hoặc thuộc tính của họ, từ đó hiểu sâu hơn về cấu trúc và đặc điểm của người dùng trong dữ liệu. Sau đó là "Topic Modeling" được áp dụng để xác định các chủ đề chính trong dữ liệu, giúp trích xuất thông tin quan trọng và hiểu rõ hơn về nội dung của dữ liệu. Cuối cùng, đưa ra những đề xuất thiết thiết thực có thể áp dụng vào thực tế.

Những bước này cùng nhau tạo nên một phương pháp toàn diện để nghiên cứu và phân tích dữ liệu, giúp chúng tôi khám phá và hiểu sâu hơn về thông tin ẩn trong dữ liệu, từ đó đưa ra những kết luận và nhận định có ý nghĩa về đối tượng nghiên cứu của chúng tôi. [25]



Hình 3.1 Tổng quan quá trình nghiên cứu

3.1. Thu thập dữ liệu nghiên cứu (Data obtaining)

Trong nghiên cứu này, dữ liệu được thu thập bằng cách sử dụng kỹ thuật scraping dữ liệu bài đăng trên mạng xã hội Facebook thông qua API (Application Programming Interface) của Facebook. API cung cấp một giao diện lập trình ứng dụng cho phép truy cập và tương tác với dữ liệu của Facebook một cách an toàn và được kiểm soát.

Để có thể sử dụng API của Facebook, trước tiên chúng tôi đã đăng ký một ứng dụng trên nền tảng của Facebook. Quá trình đăng ký bao gồm cung cấp thông tin về ứng dụng, mục đích sử dụng và tuân thủ các chính sách bảo mật dữ liệu của Facebook. Sau khi ứng dụng được phê duyệt, chúng tôi nhận được một access token, đóng vai trò như một "chìa khóa" để truy cập dữ liệu thông qua API. Tiếp theo, chúng tôi sử dụng ngôn ngữ lập trình Python và thư viện Requests để gửi yêu cầu HTTP đến các điểm cuối (endpoints) của API Facebook nhằm lấy dữ liệu bài đăng. Các yêu cầu này bao gồm thông tin xác thực (access token), tham số lọc dữ liệu (như ID người dùng, ID trang, thời gian đăng bài, vv.) và các tùy chọn khác để điều chỉnh kết quả trả về.

Dữ liệu bài đăng thu được từ API Facebook được trả về dưới dạng JSON (JavaScript Object Notation), một định dạng dữ liệu phổ biến và dễ đọc cho máy tính. Chúng tôi sử dụng thư viện JSON trong Python để phân tích và xử lý dữ liệu này. Các thông tin thu thập bao gồm nội dung bài đăng, thời gian đăng, số lượng lượt thích, bình luận, chia sẻ, cũng như các siêu dữ liệu khác liên quan đến bài đăng và người đăng. Quá trình thu thập dữ liệu tuân thủ nghiêm ngặt các chính sách và quy định của Facebook về việc sử dụng API, bảo vệ quyền riêng tư người dùng và xử lý dữ liệu. Chúng tôi cũng đảm bảo rằng việc thu thập dữ liệu không gây ra bất kỳ sự cố nào đối với hoạt động của Facebook hoặc ảnh hưởng đến trải nghiệm của người dùng.

3.2. Làm sạch dữ liệu (Data cleaning)

Quá trình data cleaning là bước quan trọng tiếp theo mà chúng tôi thực hiện để chuẩn bị dữ liệu cho việc phân tích. Data cleaning bao gồm loạt các bước như loại bỏ dữ liệu trùng lặp, xử lý các giá trị bị thiếu, chuẩn hóa định dạng dữ liệu, kiểm tra và sửa lỗi các giá trị ngoại lai, và làm sạch dữ liệu từ các ký tự không mong muốn hoặc ký tự đặc biệt. Mục tiêu của quá

Hình 3.2 Quá trình làm sạch dữ liệu

trình data cleaning là tạo ra một tập dữ liệu chất lượng cao và đáng tin cậy, giúp cho quá trình phân tích và phát triển mô hình sau này trở nên chính xác và có ý nghĩa hơn.

3.2.1. Tách từ và chuẩn hóa (Tokenization và Normalization)

Mã thông báo (tokenization) là quá trình chuyển đổi văn bản thành các từ hoặc cụm từ riêng biệt, nhằm phân tách và phân loại văn bản ngôn ngữ tự nhiên thành các thành phần quan trọng gọi là mã thông báo.

Chuẩn hóa dữ liệu (Normalization) là một bước quan trọng khác trong quá trình tiền xử lý dữ liệu, nhằm loại bỏ các thành phần không cần thiết và nhiễu trong dữ liệu, chẳng hạn như số, ký hiệu, thẻ mã và ký tự đặc biệt. Khi trích xuất các kết quả, ví dụ như đoạn mã theo ngữ cảnh được cung cấp làm đầu vào, có thể bao gồm tên tệp, URL và các ký tự đặc biệt phân định các phần trong toàn bộ tài liệu, như các ký tự dấu chấm lửng (@, %, &, v.v.) và các ký hiệu khác không rõ ý nghĩa. Việc lọc nhiễu là một nhiệm vụ quan trọng để đảm bảo tính chính xác và hiệu suất của việc phân loại kết quả tìm kiếm.

Bảng 3.1 Ví dụ minh họa kết quả tách từ và chuẩn hóa dữ liệu

Input	Out put
Hello, I am a student !!!	Hello_I_am_a_student

Xóa số (2, 1…).
Xóa dấu chấm câu ('!', ', -, ”, :, ?, [], \, …).
Xóa các ký tự đặc biệt (~, @, #, $, %, &, =, +).
Xóa các ký hiệu (ví dụ: 😀).

3.2.2. Loại bỏ từ dừng (Stop words removing)

Các từ dừng (stop words) được sử dụng trong ngôn ngữ như một phần của ngữ pháp khi không có ngữ cảnh, thay vì chỉ định chức năng hoặc ý nghĩa ngữ nghĩa. Tuy nhiên, các từ dừng được coi là ít hữu ích hơn trong văn bản so với các thuật ngữ khác và có thể ảnh hưởng trực tiếp đến ý nghĩa của văn bản. Trong hầu hết các trường hợp, tài liệu bao gồm nhiều từ không cần thiết. Người viết thường sử dụng các từ dừng để cải thiện cấu trúc ngôn ngữ của văn bản. Ví dụ về các từ dừng bao gồm các từ chỉ định như 'this', 'that' và 'those', cũng như các mạo từ như 'the', 'a' và 'an'. Bằng cách loại bỏ các từ dừng thường xuyên này khỏi tài liệu văn bản, số lượng từ mà mỗi cụm từ tìm kiếm phải khớp sẽ giảm, làm tăng đáng kể thời gian cần thiết để truy xuất kết quả mà không ảnh hưởng đến độ chính xác.

Với nhận thức về điều này, việc loại bỏ các từ không cần thiết sẽ giúp cải thiện khả năng truyền tải ý nghĩa của nội dung văn bản hoặc tài liệu và giúp chúng ta hiểu nội dung đó dễ dàng hơn bằng cách sử dụng các phương pháp học máy. Hình dưới hiển thị một mẫu tài liệu văn bản sau khi các từ dừng đã được loại bỏ.

Bảng 3.2 Ví dụ minh hoạ kết quả loại bỏ từ dừng

Input	Output
Hello, I am a student at HUB university	Student HUB university

3.2.3. Loại bỏ đường dẫn (Link, URL removing)

Có một số lý do nên loại bỏ các đường link trong tệp dữ liệu :

- Bảo mật thông tin nhạy cảm: Trong một số trường hợp, tệp dữ liệu có thể chứa các đường link đến tài liệu hoặc trang web chứa thông tin nhạy cảm, chẳng hạn như thông tin cá nhân, mật khẩu, hay dữ liệu kinh doanh nhạy cảm. Loại bỏ các đường link trong tệp dữ liệu giúp tránh tiết lộ thông tin quan trọng và bảo vệ quyền riêng tư.
- Phân tích ngôn ngữ tự nhiên : Trong các tác vụ phân tích ngôn ngữ tự nhiên, ví dụ như phân loại văn bản, phân tích ý kiến, hay tạo ra mô hình dự đoán, các đường link thường không mang giá trị ngữ nghĩa và chỉ gây nhiễu. Loại bỏ các đường link giúp tập trung vào nội dung chính của văn bản và cải thiện chất lượng và hiệu suất của mô hình xử lí ngôn ngữ tự nhiên.
- Tối ưu hóa kích thước dữ liệu: Trong một số trường hợp, tệp dữ liệu có thể chứa nhiều đường link, và việc giữ lại chúng trong tệp có thể làm tăng kích thước của dữ liệu. Khi lưu trữ hoặc truyền tải dữ liệu, việc loại bỏ các đường link có thể giúp giảm kích thước tệp và tiết kiệm không gian lưu trữ hoặc băng thông mạng.

Bảng 3.3 Ví dụ minh họa kết quả loại bỏ đường dẫn

Input	Output
This video is great : Htttp://www.youtube.com/abcd	This video is great :

3.2.4. Chuyển từ về dạng gốc (Word stemming)

Phương pháp stemming trong xử lý ngôn ngữ tự nhiên là một quá trình nhằm đưa các từ về dạng gốc hay hình thức cơ bản của chúng, gọi là "stem" (gốc từ). Mục đích chính của stemming là rút gọn từ vựng và xử lý các biến thể từ một cách thống nhất, từ đó cải thiện hiệu suất của các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, phân tích ý kiến, truy xuất thông tin và tìm kiếm.

Cách thức hoạt động của phương pháp stemming thường dựa trên một tập hợp các quy tắc ngữ pháp đơn giản và heuristics (phương pháp ước lượng), chúng xem xét các đặc điểm của từ để loại bỏ các đuôi từ và chuyển đổi từ về dạng gốc tương ứng. Ví dụ, trong tiếng Anh,

phương pháp stemming có thể loại bỏ các đuôi như -s, -ing, -ed, -er, -tion….. từ các từ về dạng gốc của chúng. [23]

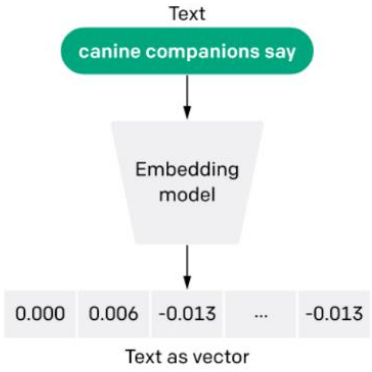
Bảng 3.4 Ví dụ minh họa kết quả chuyển từ về dạng gốc

Input	Output
Running	Run
Happiness	Happy
Collection	collect

3.3. Vector nhúng (Vector embedding)

Khi làm việc với dữ liệu văn bản, việc áp dụng các mô hình máy học trực tiếp có thể gặp phải những hạn chế đáng kể. Mặc dù các mô hình này có thể hiệu quả trong một số trường hợp, nhưng chúng thường không thể hiện được sự liên quan ngữ nghĩa giữa các từ hoặc câu một cách chính xác. Điều này làm giảm khả năng của mô hình trong việc hiểu và xử lý ngôn ngữ tự nhiên một cách chính xác.

Do đó, bước quan trọng tiếp theo là chuyển đổi danh sách các bài viết thành các Vector embedding. Điều này có nghĩa là chúng ta cần mã hóa thông tin ngữ nghĩa và cú pháp của từng từ hoặc câu trong một không gian vector. Vector embedding là biểu diễn số học của từng đối tượng văn bản trong không gian vectơ đa chiều, trong đó mỗi chiều có thể đại diện cho một thuộc tính cụ thể của từ hoặc câu.



Hình 3.3 Quy trình nhúng văn bản thành vector

Nguồn: [26]

Sử dụng vector nhúng giúp mô hình máy học hiểu được ý nghĩa và mối quan hệ giữa các từ và câu dựa trên ngữ cảnh và sự liên kết. Thay vì chỉ sử dụng biểu diễn đơn giản như chỉ số hoặc one-hot vector, vector nhúng mang thông tin phong phú về từ ngữ và mối quan hệ giữa các từ. Điều này giúp mô hình hiểu các khái niệm, mối quan hệ và ngữ nghĩa trong ngôn ngữ tự nhiên một cách sâu sắc hơn.

Trong nghiên cứu này, chúng tôi đã áp dụng phương pháp học chuyển đổi Transfer learning kế thừa và sử dụng lại các mô hình học máy trong lĩnh vực xử lí ngôn ngữ tự nhiên, cụ thể là 3 mô hình 'E5-base, ‘E5-small’, ‘E5-large’ đã được đề cập trong phần cơ sở lý thuyết.

3.4. Giảm chiều dữ liệu (Dimensionality reduction)

Sau khi thực hiện Vector embedding, chúng tôi đã thu được dữ liệu có số chiều rất lớn, điều này gây khó khăn cho việc thực hiện gom cụm và trực quan hóa. Để giải quyết vấn đề này, chúng tôi sẽ áp dụng các phương pháp giảm chiều dữ liệu, được gọi là Dimensionality reduction.

Giảm chiều dữ liệu là quá trình giảm số chiều của tập dữ liệu ban đầu thành một không gian dữ liệu mới có số chiều thấp hơn. Khi làm việc với dữ liệu có số chiều cao, việc giảm chiều dữ liệu có thể giúp giảm độ phức tạp tính toán, giảm không gian lưu trữ, loại bỏ thông tin không cần thiết và cải thiện hiệu suất và khả năng hiểu dữ liệu.

Ví dụ, nếu có một tập dữ liệu gồm các điểm trong không gian ba chiều (x, y, z), giảm chiều dữ liệu có thể giúp chuyển đổi tập dữ liệu này thành một không gian dữ liệu mới chỉ gồm hai chiều (x, y). Trong quá trình này, thông tin không gian z bị loại bỏ, nhưng vẫn giữ lại một phần lớn thông tin quan trọng của dữ liệu ban đầu.

Chúng tôi thực hiện giảm chiều dữ liệu bằng 3 phương pháp nổi bật nhất hiện nay là PCA, t-SNE, Umap, và thực hiện so sánh đánh giá để chọn ra phương pháp phù hợp nhất cho bài nghiên cứu này.

3.5. Gom cụm người dùng (Users clustering)

Phân cụm người dùng (Users clustering) là một kỹ thuật quan trọng trong phân tích dữ liệu nhằm nhóm các người dùng vào các nhóm có đặc trưng hoặc hành vi tương tự nhau.

Trong việc xử lí dữ liệu văn bản từ các bài viết của người dùng trên mạng xã hội Facebook, chúng ta thường đối mặt với một số thách thức, đặc biệt là khi không có sẵn nhãn để hướng dẫn quá trình huấn luyện mô hình (unsupervised data). Trong tình huống như vậy, việc sử dụng các phương pháp phân cụm không giám sát để nhóm người dùng là một lựa chọn phù hợp. Chúng tôi lựa chọn áp dụng phương pháp phân cụm không giám sát K-means để thực hiện phân cụm người dùng.

Sau khi phân cụm xong, có thể tiến hành phân tích và khám phá các đặc điểm của từng nhóm người dùng. Điều này giúp hiểu rõ hơn về cấu trúc và đặc điểm của cộng đồng người dùng trên mạng xã hội Facebook, từ đó có thể đưa ra các chiến lược hoặc quyết định phù hợp cho việc tương tác và phục vụ người dùng một cách hiệu quả hơn.

3.6. Mô hình hóa chủ đề (Topic modeling)

Trong phần User Clustering, chúng tôi đã sử dụng phương pháp K-means để nhóm các người dùng dựa trên các bài viết trên mạng xã hội Facebook. Bằng cách này, chúng tôi đã tạo ra các nhóm người dùng có đặc điểm và sở thích tương đồng. Tiếp theo, chúng tôi tiến hành phân tích bài đăng của những người dùng trong cùng 1 cụm để xác định chủ đề chính mà mỗi nhóm người dùng quan tâm bằng cách sử dụng phương pháp mô hình hóa chủ đề LDA một kỹ thuật máy học không giám sát giúp phân tích và xác định các chủ đề trong dữ liệu văn bản. Bằng cách này, chúng tôi có thể hiểu rõ hơn về nội dung và sở thích của từng nhóm người

dùng, từ đó điều chỉnh chiến lược tương tác và cung cấp nội dung một cách phù hợp và hiệu quả hơn, dựa trên đặc điểm cụ thể của từng nhóm người dùng.

3.7. Ứng dụng nghiên cứu (Aplications)

Kết quả của nghiên cứu này có nhiều ứng dụng tiềm năng trong việc phát triển cộng đồng người dùng trực tuyến và hệ thống đề xuất sản phẩm thông minh. Thông qua kỹ thuật gom cụm người dùng và mô hình hóa chủ đề, chúng tôi có thể đạt được hai mục tiêu chính sau:

+ Xây dựng cộng đồng người dùng có cùng sở thích và chủ đề cá nhân bằng cách gom cụm người dùng dựa trên dữ liệu bài đăng và chủ đề được mô hình hóa, chúng ta có thể nhóm những người dùng có sở thích, quan tâm tương đồng vào cùng một cộng đồng trực tuyến.

+ Hệ thống đề xuất sản phẩm, dịch vụ phù hợp cho từng cá nhân thông qua việc hiểu rõ sở thích, nhu cầu của từng nhóm người dùng, các doanh nghiệp có thể triển khai các hệ thống đề xuất sản phẩm, dịch vụ chính xác và phù hợp.

Tóm tắt chương 3

Chương 3 trình bày chi tiết phương pháp luận nghiên cứu, bao gồm thu thập dữ liệu bằng kỹ thuật scraping từ mạng xã hội Facebook thông qua API. Tiếp theo là làm sạch dữ liệu bằng cách loại bỏ dữ liệu nhiễu, trùng lặp và các bước tiền xử lý như xóa ký tự đặc biệt, chuẩn hóa văn bản. Sau đó, áp dụng vector nhúng văn bản bằng 3 mô hình pretrained E5. Kế đến, sử dụng 3 phương pháp giảm chiều PCA, t-SNE, Umap để chuyển dữ liệu xuống 2 chiều nhằm trực quan hóa. Tiếp theo, mô tả phương pháp gom cụm người dùng K-means và mô hình hóa chủ đề của người dùng bằng kỹ thuật LDA để phát hiện chủ đề tiềm ẩn từ dữ liệu văn bản.

Tóm lại, Chương 3 cung cấp một cái nhìn toàn diện về phương pháp luận được áp dụng trong nghiên cứu, đảm bảo tính minh bạch và khoa học. Mỗi bước được trình bày rõ ràng, logic và có sự hỗ trợ từ các lý thuyết, công cụ và kỹ thuật phù hợp. Chương này cho thấy nghiên cứu được thực hiện một cách bài bản và khoa học, đảm bảo tính xác thực của kết quả thu được.

CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU

4.1. Khai báo các thư viện cần thiết

Chúng tôi thực hiện nghiên cứu này bằng công cụ Jupyter Notebook trên môi trường Python. Trước hết, chúng tôi thực hiện khai báo các thư viện phục vụ cho nghiên cứu này bao gồm:

```
import json  # hỗ trợ làm việc với các tập tin JSON.
import pandas as pd  # hỗ trợ thực hiện các thao tác liên quan đến dữ liệu dạng bảng.
import numpy as np  # hỗ trợ thực hiện các thao tác liên quan đến mảng và ma trận.
import os  # hỗ trợ thực hiện các thao tác liên quan đến hệ điều hành.
import glob  # hỗ trợ thực hiện các thao tác liên quan đến quản lý đường dẫn tập tin.
```

```
from sentence_transformers import SentenceTransformer  # hỗ trợ cung cấp các mô hình vector embedding

import nltk  # hỗ trợ các tác vụ liên quan đến xử lý ngôn ngữ tự nhiên.
from nltk.corpus import stopwords  # hỗ trợ loại bỏ stopword.
import re  # hỗ trợ các phép biến đổi trên các chuỗi văn bản.
import string  # hỗ trợ các thao tác trên các chuỗi ký tự.
import emoji  # xử lý các biểu tượng cảm xúc trong văn bản.
from nltk.tokenize import word_tokenize, sent_tokenize, PunktSentenceTokenizer  # hỗ trợ tách các câu từ văn bản.
from googletrans import Translator  # hỗ trợ dịch văn bản giữa các ngôn ngữ.
from nltk.tag import pos_tag  # gán phần loại từ cho từng từ trong văn bản.

from umap import UMAP  # hỗ trợ phương pháp giảm chiều dữ liệu Umap
from sklearn.manifold import TSNE  # hỗ trợ phương pháp giảm chiều dữ liệu t-SNE.
from sklearn.decomposition import PCA  hỗ trợ phương pháp giảm chiều dữ liệu (PCA).

from sklearn.cluster import K-means  # hỗ trợ phương pháp phân cụm dữ liệu.
from sklearn.metrics import silhouette_score, davies_bouldin_score  # hỗ trợ đánh giá chỉ số chất lượng của phân cụm.

import matplotlib.pyplot as plt  # hỗ trợ trực quan hóa dữ liệu.
import seaborn as sns  # hỗ trợ trực quan hóa dữ liệu.

from sklearn.feature_extraction.text import TfidfVectorizer  # trích xuất đặc trưng từ văn bản sử dụng TF-IDF.
from sklearn.decomposition import LatentDirichletAllocation  # mô hình hóa chủ đề trong văn bản.
```

4.2. Kết quả gom cụm

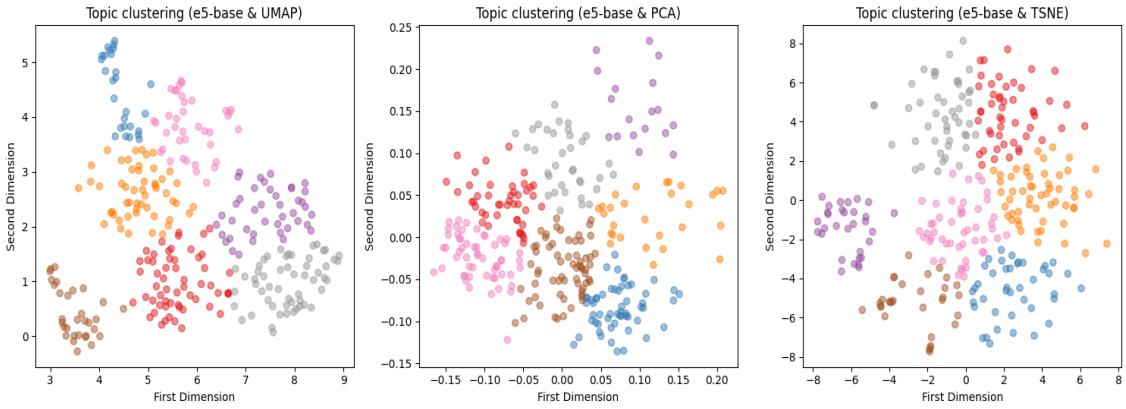
Trong quá trình nghiên cứu, chúng tôi đã tiến hành một loạt các phương pháp nhúng văn bản như 'E5-base', 'E5-small' và 'E5-large'. Mục đích của việc này là tìm hiểu xem phương pháp nào sẽ mang lại kết quả tốt nhất cho việc gom cụm dữ liệu mà chúng tôi đã thu thập được. Sau khi thực hiện nhúng văn bản, chúng tôi tiếp tục áp dụng các phương pháp giảm chiều dữ liệu như PCA, Umap và t-SNE để thu được một biểu diễn dữ liệu có số chiều thấp hơn. Mục tiêu của việc giảm chiều dữ liệu là giữ lại những đặc trưng quan trọng và đồng thời loại bỏ nhiễu và thông tin không cần thiết. Bằng cách kết hợp các phương pháp nhúng với phương pháp giảm chiều dữ liệu khác nhau, chúng tôi hy vọng có thể tìm ra cách tiếp cận tốt nhất để gom cụm và hiểu sâu hơn về dữ liệu mà chúng tôi đang nghiên cứu.

- Để đánh giá hiệu quả gom cụm, chúng tôi sử dụng các chỉ số:

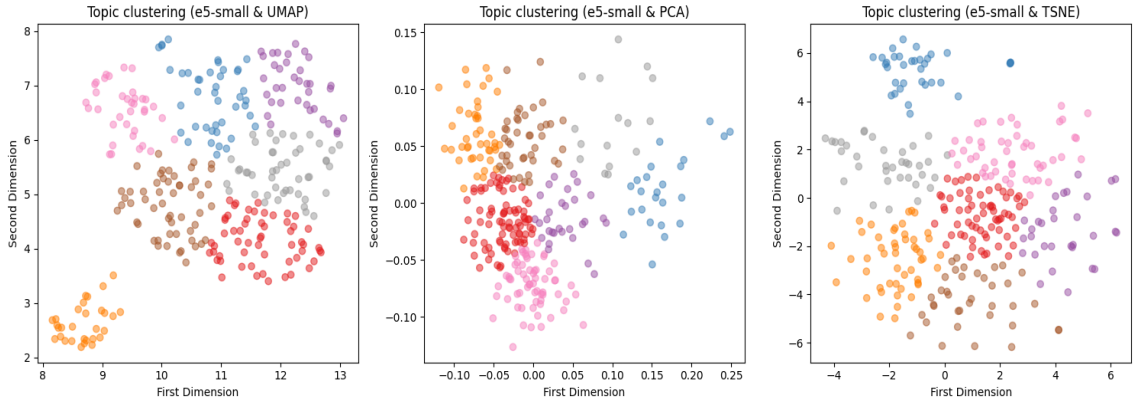
- **Davies-Bouldin score:** Đo lường độ hợp lý của cụm trong phân cụm, dựa trên sự gần gũi của dữ liệu đến trung tâm của cụm và khoảng cách giữa các trung tâm cụm. Mục tiêu là tối thiểu hóa chỉ số này.
- **Silhouette score:** là thước đo mức độ giống nhau của một đối tượng với cụm của chính nó (sự gắn kết) so với các cụm khác (sự tách biệt). Silhouette score nằm trong khoảng từ −1 đến +1, mục tiêu là tối đa hóa giá trị để đạt hiệu quả tốt nhất.
- **Calinski-Harabasz index:** Đo lường sự tách biệt giữa các cụm khác nhau trong việc gom cụm. Nó tính toán tỷ lệ giữa phương sai giữa các cụm và phương sai trong cùng một cụm, mục tiêu tối đa hóa giá trị để đạt hiệu quả tốt nhất.

- Việc lựa chọn số lượng cụm trong ứng dụng thuật toán phân cụm là 1 công đoạn tối quan trọng, số lượng cụm có thể được xác định dựa trên những phương pháp định lượng như phương pháp Elbow, Silhouette, Gap Statistic... Nhưng trong bài nghiên cứu mang tính áp dụng thực tế cho doanh nghiệp cao, chúng tôi không xác định số lượng cụm bằng những phương pháp định lượng, mà chúng tôi đề cao các phương pháp định tính hơn, bằng việc xem xét các yếu tố kinh doanh và thực tiễn như quy mô và chiến lược kinh doanh của doanh nghiệp, nguồn lực sẵn có để phục vụ các nhóm khách hàng khác nhau, khả năng phân biệt và tiếp cận các nhóm khách hàng riêng biệt và mục đích sử dụng kết quả phân cụm (marketing, phân khúc thị trường, dịch vụ chăm sóc khách hàng, ….). Ở trong bài nghiên cứu này, chúng tôi giả định doanh nghiệp đang cần tạo ra 7 nhóm khách hàng riêng biệt để áp dụng các chiến lược marketing và dịch vụ phù hợp nhằm đáp ứng nhu cầu cụ thể của từng nhóm. Với 7 cụm đã được xác định, các doanh nghiệp có thể thiết kế các chương trình quảng cáo, khuyến mãi, sản phẩm dịch vụ riêng biệt cho từng phân khúc để tối ưu hiệu quả kinh doanh. Ngoài ra, kết quả nghiên cứu này cũng có thể được ứng dụng trong việc cá nhân hóa trải nghiệm người dùng, đề xuất sản phẩm phù hợp, và nâng cao chất lượng dịch vụ khách hàng. Với 7 nhóm người dùng rõ ràng, các doanh nghiệp sẽ dễ dàng xác định nhu cầu của từng phân khúc và điều chỉnh hoạt động kinh doanh cho phù hợp.

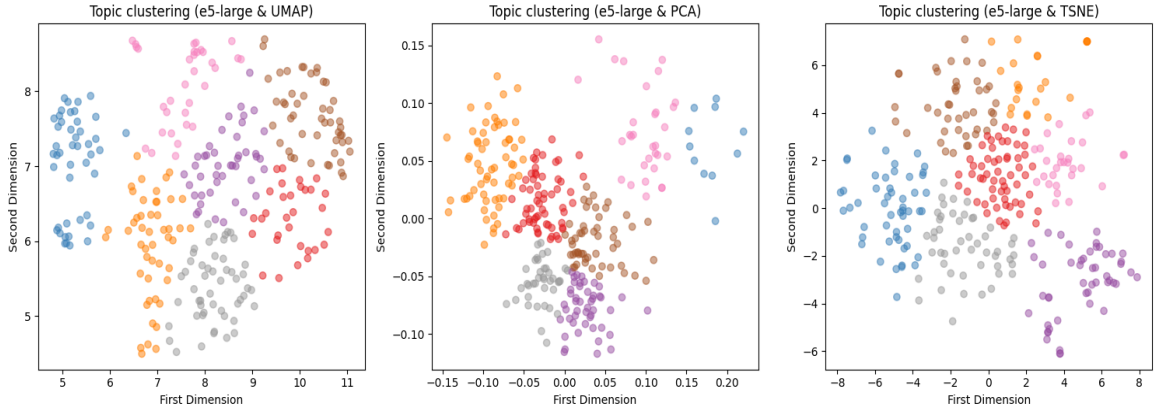
- Trực quan kết quả :



Hình 4.1 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-base và 3 phương pháp giảm chiều dữ liệu)



Hình 4.2 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-small và 3 phương pháp giảm chiều dữ liệu)



Hình 4.3 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-large và 3 phương pháp giảm chiều dữ liệu)

Bảng 4.1 Bảng kết quả đánh giá các chỉ số gom cụm

E5-base

Method	Davies-Bouldin Index	Silhouette Score	Calinski-Harabaz Index
PCA	0.8683	0.3584	266.5079
TSNE	0.8429	0.3783	298.4545
UMAP	0.8008	0.4305	392.6561

E5-small

Method	Davies-Bouldin Index	Silhouette Score	Calinski-Harabaz Index
PCA	0.8927	0.3536	275.8334
TSNE	0.8140	0.3780	264.0097
UMAP	0.7931	0.4131	384.8666

E5-large

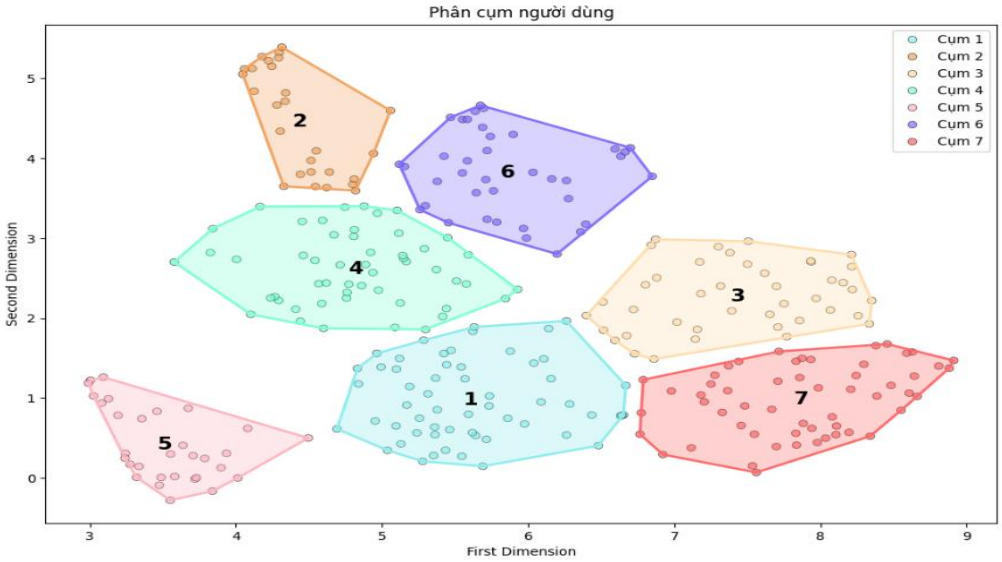
Method	Davies-Bouldin Index	Silhouette Score	Calinski-Harabaz Index
PCA	0.8951	0.3622	297.4207
TSNE	0.8039	0.3736	269.7354
UMAP	0.8566	0.3997	344.5047

Dựa vào thông số đánh giá, ta có thể nhận thấy rằng việc tiến hành vector nhúng trên văn bản bằng 3 mô hình E5 khi có áp dụng phương pháp giảm chiều dữ liệu Umap đều đạt chỉ số đánh giá tốt vượt trội. Trong đó, phương pháp vector nhúng văn bản 'E5-base' kèm áp dụng phương pháp giảm chiều UMAP có kết quả gom cụm tốt nhất..

Điều này đã phản ánh việc kết hợp giữa phương pháp nhúng văn bản 'E5-base' cùng phương pháp giảm chiều dữ liệu UMAP vào thuật toán gom cụm K-means, đã đem lại hiệu quả gom cụm cao nhất.

4.3. Cụm người dùng và chủ đề

Sau khi thực hiện đánh giá các mô hình vector nhúng và các phương pháp giảm chiều dữ liệu, thành công lựa chọn được mô hình E5-base kết hợp với phương pháp giảm chiều dữ liệu Umap cho kết quả phân cụm người dùng tốt nhất (Mỗi người dùng được phân vào 1 trong 7 cụm thích hợp nhất). Kết quả phân cụm như hình sau:



Hình 4.4 Trực quan các cụm người dùng (Thực hiện bởi mô hình E5-base và phương pháp giảm chiều dữ liệu Umap)

Đưa ra chủ đề cho từng cụm người dùng bằng việc trích xuất các từ ngữ có trọng số cao nhất tương ứng với mỗi cụm trong quá trình mô hình hóa chủ đề bằng phương pháp LDA. Mỗi người dùng thường quan tâm đến rất nhiều chủ đề khác nhau, ở đây chúng tôi xác định 2 tới 3 nội dung chủ yếu mà mỗi cụm người quan tâm.

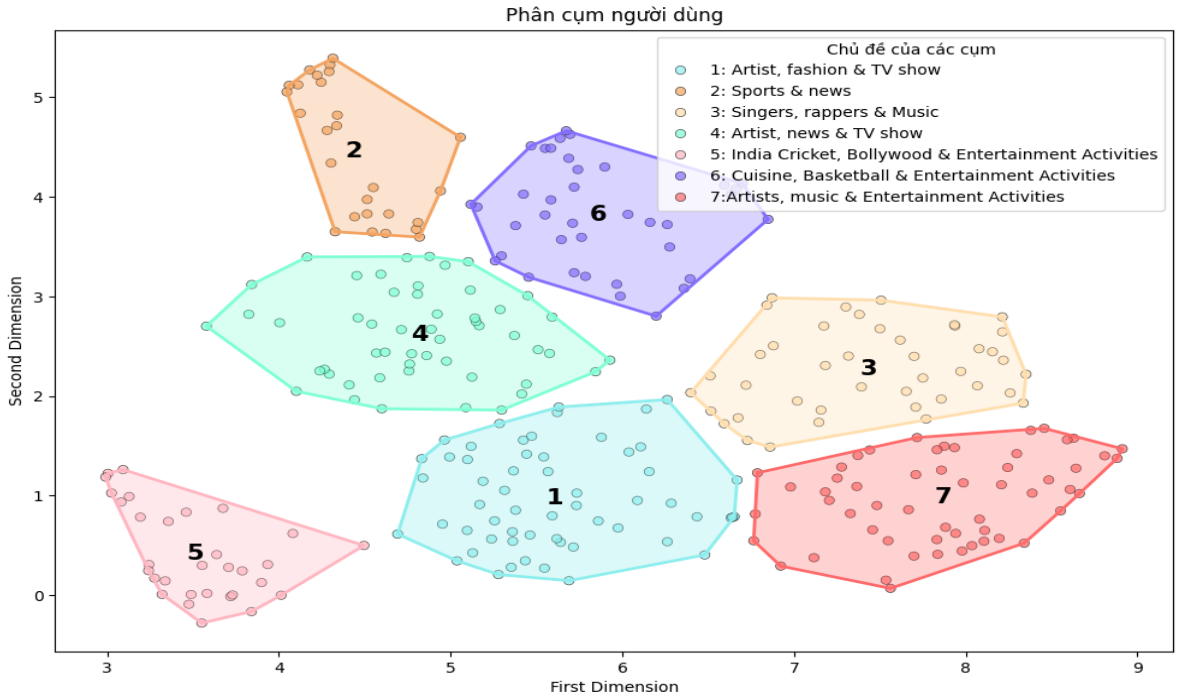
Bảng 4.2 Mô hình chủ đề cho từng cụm người dùng

Cụm	Từ ngữ được trích xuất	Chủ đề
1	**Nội dung 1:** ['tom hanks', 'hanks', 'lindsay lohan', 'tom', 'lindsay', 'hanks', 'asg', 'lindsay lohan', 'hanks', 'admin', 'lohan', 'rita wilson', 'lindsay', 'tom hanks fan', 'intern', 'best actor', 'tom hanks fan site', 'www lindsay lohan us', 'actor'] **Nội dung 2:** ['thank', 'love', 'kardashian', 'video', 'jenner', 'kylie', 'new', 'night', 'show', 'kim', 'music', 'life', 'kardashian jenner news update', 'friend', 'world', 'birthday', 'girl', 'everyone', 'fifth', 'family']	Artist, fashion & TV show
2	**Nội dung 1:** ['fantastico combeep', 'ronaldo', 'neymar', 'world', 'cristiano', 'team', 'goal', 'nadal', 'player', 'game', 'thank', 'madrid', 'real madrid', 'deviltim', 'roger', 'people', 'qué', 'federer', 'man utd', 'messi'] **Nội dung 2:** ['trump', 'biden', 'people', 'obama', 'ufc', 'rudo', 'president', 'anderson', 'bill', 'world', 'nye', 'science', 'family', 'life', 'live', 'cooper', 'donald', 'election', 'rt', 'america']	Sports & news
3	**Nội dung 1:** ['music', 'video', 'justin', 'album', 'eminem', 'aaliyah', 'kendrick', 'song', 'fifth', 'normani', 'love', 'metallica', 'god did', 'thank', 'harmony', 'dinah', 'kbs', 'jauregui', 'bieber', 'love fl'] **Nội dung 2:** ['video', 'music', 'song', 'thank', 'taylor', 'people', 'zayn', 'guitar', 'love', 'cole', 'wayne', 'new', 'night', 'harry', 'iggy', 'show', 'life', '#nodaysoff', 'tour', 'world']	Singers, rappers & Music
4	**Nội dung 1:** ['hollywood actresses', 'timeless beautiful goddess', 'diane lane', 'eva mendes', 'jessica chastain', 'jessica biel', 'jennifer garner', 'charlize theron', 'brie larson', 'jennifer aniston', podcast',	Artist, news & TV show

	'silverman', 'conan', 'hulu', 'sarah', 'fantastico beep'] **Nội dung 2:** ['fantastico beep', 'thank', 'show jessica alba', 'ilya merica', 'alexandra daddario', 'jennifer lawrence', 'sarah silverman ', 'world', 'people', 'life', 'night', 'love', 'read', 'god', 'family', 'blast', 'way', 'live', 'thing', 'birthday', 'friend', 'new', 'video', 'jimmy']	
5	**Nội dung 1:** ['india', 'bollywood', 'love', 'birthday', 'cricket', 'deepika padukone', 'dhoni', 'thank', 'team', 'video', 'world', 'life', 'film', 'khan', 'virat', 'kohli', 'morning', 'playbold', 'kapoor', 'sharma'] **Nội dung 2:** ['deepika padukone the dreamy girl', 'deepika padukone', 'padukone', 'deepika', 'playbold', 'asli akhada', 'fantasy akhada', 'gain follow', 'normani', 'bhogle', 'salman khan superfan', 'fifth', 'salman khan fan', 'tamannaah', 'akhada', 'cricket played loud', 'topstar celebrity', 'bollywood star', 'gain like', 'bollywood']	India Cricket, Bollywood & Entertainment Activities
6	**Nội dung 1:** ['ramsay', 'gordon', 'https', 'twitter com', 'recip', 'kitchen', 'index html', 'snowsmanenalib', 'abacom', 'gordon ramsay', 'chef', 'attn', 'cook', 'masterchef', 'wendi', 'gx', 'dish', 'sticker', 'gino', 'burger'] **Nội dung 2:** ['fantastico com beep', 'laker', 'nba', 'lebron', 'wwe', 'game', 'thank', 'team', 'season', 'pts', 'steven harvey', 'birthday', 'read', 'night', 'world', 'ticket', 'show', 'player', 'blast', 'people']	Cuisine, Basketball & Entertainment Activities
7	**Nội dung 1:** ['yeng', 'constantino', 'ebru', 'codi', 'shreya', 'ghoshal', 'martin garrix fan', 'watermark', ‘The Media Nanny‘, 'michael ross george', 'martin garrix', 'carrie underwood', 'carrie', 'freeview', 'shreya ghoshal', 'metazoo', 'nicki minaj', 'mariah', 'simpson', 'carey']	Artists, music & Entertainment Activities

	Nội dung 2: ['music', 'thank', 'show', 'song', 'video', 'album', 'love', 'ticket', 'tour', 'night', 'world', 'fan', 'new', 'adam', 'life', 'live', 'photo', 'people', 'levin', 'friend']	

Sau cùng, kết quả thu được như hình sau:



Hình 4.5 Trực quan các cụm người dùng kèm chủ đề tương ứng (Thực hiện bởi mô hình E5-base và phương pháp giảm chiều dữ liệu Umap)

4.4. Ứng dụng

4.4.1. Xây dựng cộng đồng người dùng

Bảng 4.3 Chi tiết thông tin người dùng trong các nhóm chủ đề

Chủ đề	Tên thật/Tên fanpage/Biệt danh của người dùng	Số lượng
Artist, fashion & TV show	'143redangel', 'AustinMahone', 'CherLloyd', 'DollyParton', 'EvaLongoria', 'HilaryDuff', 'IAMJHUD', 'IAMQUEENLATIFAH', 'JLo', 'JessicaSimpson', 'KELLYROWLAND', 'KendallJenner', 'KeriHilson', 'KimKardashian', 'KrisJenner', 'KylieJenner', 'LennyKravitz', 'Ludacris', 'LukE5SOS', 'MileyCyrus', 'MirandaCosgrove', 'MirandaKerr', 'ParisHilton', 'RitaOra', 'ScottDisick',	57

	'VanessaHudgens', 'VictoriaJustice', 'ZooeyDeschanel', 'aliciakeys', 'annecurtissmith', 'bellathorne', 'britneyspears', 'chelseahandler', 'cher', 'ciara', 'dylanobrien', 'itsgabrielleu', 'joejonas', 'justdemi', 'katyperry', 'kerrywashington', 'khloekardashian', 'kourtneykardash', 'ladygaga', 'lindsaylohan', 'lucyhale', 'maryjblige', 'mindykaling', 'mirandalambert', 'ninadobrev', 'robkardashian', 'russwest44', 'selenagomez', 'serenawilliams', 'snooki', 'tomhanks', 'victoriabeckham'	
Sports & news	'10Ronaldinho', 'BillGates', 'BillNye', 'BillSimmons', 'Cristiano', 'DjokerNole', 'GarethBale11', 'IAmSteveHarvey', 'JohnCleese', 'KAKA', 'LewisHamilton', 'MesutOzil1088', 'RafaelNadal', 'Schwarzenegger', 'TimTebow', 'WayneRooney', 'andersoncooper', 'andresiniesta8', 'barackobama', 'billmaher', 'danawhite', 'hazardeden10', 'maddow', 'neymarjr', 'rioferdy5', 'rogerfederer', 'rustyrockets'	27
Singers, rappers & Music	'BigSean', 'CalvinHarris', 'Drake', 'Eminem', 'Fergie', 'HARDWELL', 'Harry', 'IGGYAZALEA', 'JColeNC', 'LilTunechi', 'MeekMill', 'Metallica', 'Michael5SOS', 'MissyElliott', 'NICKIMINAJ', 'NickCannon', 'PerezHilton', 'Pharrell', 'Pink', 'RobertDowneyJr', 'Skrillex', 'Tip', 'Tyga', 'Usher', 'Wale', 'Zendaya', 'bridgitmendler', 'chrisbrown', 'chrisrock', 'djkhaled', 'jason', 'justinbieber', 'justintimberlake', 'kanyewest', 'kendricklamar', 'official', 'paramore', 'taylorswift', 'troyesivan', 'wizkhalifa', 'xtina', 'zaynmalik'.	42
Artist, news & TV show	'ActuallyNPH', 'Beyonce', 'Bourdain', 'Caradelevingne', 'Caspar', 'ConanOBrien', 'Fearnecotton', 'GaryLineker', 'IanMcKellen', 'JeremyClarkson', 'JerrySeinfeld', 'JimmyFallon', 'JoelOsteen', 'JoeyGraceffa', 'KevinSpacey', 'LeoDiCaprio', 'Lord', 'Oprah', 'RealHughJackman', 'RyanSeacrest', 'SamuelLJackson', 'SarahKSilverman', 'Schofe', 'Sethrogen', 'StephenAtHome', 'SteveMartinToGo', 'TomCruise', 'ZacEfron', 'antanddec', 'aplusk', 'camerondallas', 'channingtatum', 'charliesheen', 'hitRECordJoe', 'hollywills', 'jackwhitehall', 'jessicaalba', 'jimmycarr', 'jk', 'johngreen', 'neiltyson', 'piersmorgan', 'prattprattpratt', 'priyankachopra', 'richardbranson', 'rickygervais', 'simonpegg', 'twhiddleston', 'tyleroakley', 'tylerperry', 'wossy', 'yokoono'	52
India Cricket, Bollywood & Entertainment Activities	'ABdeVilliers17', 'AnushkaSharma', 'BeingSalmanKhan', 'ImRaina', 'ImRo45', 'MirzaSania', 'ParineetiChopra', 'Riteishd', 'ShraddhaKapoor', 'SrBachchan', 'TheFarahKhan', 'YUVSTRONG12', 'akshaykumar', 'aliaa08', 'arrahman', 'bhogleharsha', 'bipsluvurself', 'deepikapadukone', 'henrygayle', 'iHrithik', 'iamsrk', 'ilovegeorgina', 'imVkohli', 'msdhoni', 'rampalarjun', 'realpreityzinta', 'sachin', 'sardesairajdeep', 'shrutihaasan', 'sonamakapoor', 'virendersehwag'	31

Cuisine, Basketball & Entertainment Activities	'AnnaKendrick47', 'CP3', 'DwightHoward', 'DwyaneWade', 'FloydMayweather', 'GordonRamsay', 'JHarden13', 'JohnCena', 'JonahHill', 'KDTrey5', 'KevinHart4real', 'KingJames', 'KyrieIrving', 'MagicJohnson', 'MichelleObama', 'Nashgrier', 'RandyOrton', 'SHAQ', 'ShawnMichaels', 'SnoopDogg', 'StephenCurry30', 'TheRock', 'TigerWoods', 'TripleH', 'carmeloanthony', 'danieltosh', 'iamjamiefoxx', 'iansomerhalder', 'icecube', 'jennettemccurdy', 'jimmykimmel', 'kobebryant', 'mcuban', 'paulpierce34', 'paulwesley', 'rainnwilson', 'tonyhawk', 'usainbolt'	38
Artists, music & Entertainment Activities	'50cent', '5SOS', 'Adele', 'Ashton5SOS', 'AvrilLavigne', 'BradPaisley', 'BrunoMars', 'Calum5SOS', 'CodySimpson', 'FifthHarmony', 'GaryBarlow', 'GreenDay', 'Imaginedragons', 'JanetJackson', 'JessieJ', 'KeithUrban', 'LittleMix', 'Louis', 'MacMiller', 'MariahCarey', 'MartinGarrix', 'MsLeaSalonga', 'NeYoCompound', 'NiallOfficial', 'NicoleScherzy', 'OzzyOsbourne', 'PaulMcCartney', 'ShawnMendes', 'SimonCowell', 'TheVampsband', 'Trevornoah', 'YengPLUGGEDin', 'Zedd', 'adamlevine', 'blakeshelton', 'carlyraejepsen', 'carrieunderwood', 'coldplay', 'davidguetta', 'ddlovato', 'edsheeran', 'elliegoulding', 'johnlegend', 'kevinjonas', 'linkinpark', 'maroon5', 'nickjonas', 'onedirection', 'pitbull', 'shakira', 'shreyaghoshal', 'steveaoki', 'thebeatles', 'thekillers', 'theweeknd'	55

Từ kết quả gom cụm, có thể xây dựng các cộng đồng người dùng dựa trên các chủ đề hoặc quan tâm chung của họ. Kết quả gom cụm giúp phân loại người dùng vào các nhóm dựa trên sự tương đồng trong hành vi, sở thích hoặc dữ liệu khác. Việc xây dựng các cộng đồng dựa trên gom cụm có thể mang lại nhiều lợi ích:

+ Tạo nền tảng giao lưu và chia sẻ thông tin
+ Tăng cơ hội kết nối và hợp tác
+ Tạo ra một môi trường hỗ trợ
+ Thúc đẩy sự phát triển và chia sẻ kiến thức
+ Tạo ra cơ hội kinh doanh và tiếp thị

4.4.2. Đề xuất sản phẩm cho người dùng

Cuối cùng là mục tiêu nghiên cứu ban đầu của chúng tôi, là có thể đề xuất được sản phẩm thích hợp cho các nhóm người dùng thông qua việc phân tích hành vi, sở thích của họ thông qua các bài đăng trên mạng xã hội Facebook. Xác định chủ đề mà người dùng quan tâm là một bước quan trọng để tối ưu hóa chiến dịch quảng cáo và tăng hiệu quả bán hàng. Qua việc tiếp cận thị trường một cách cẩn thận, tạo ra nội dung chất lượng và sử dụng công cụ quảng cáo đích đến, chúng tôi có thể truyền đạt thông điệp một cách chính xác và hiệu quả tới từng nhóm người dùng cụ thể. Chiến lược này không chỉ giúp giảm chi phí quảng cáo mà còn tạo ra một môi trường thuận lợi để tăng cơ hội mua sản phẩm và dịch vụ từ phía khách hàng.

Chúng tôi thực hiện tìm 1 số những loại sản phẩm tiêu biểu mô phỏng cho mỗi chủ đề , phục vụ cho việc đề xuất cho từng nhóm người dùng :

Bảng 4.4 Minh họa sản phẩm được đề xuất tới các cụm chủ đề người dùng

Chủ đề	Nhóm sản phẩm	Sản phẩm chi tiết
1. Artist, fashion & TV show	Clothes products	○ Sleek Stretch Cutout Skirt ○ Sueded Stretch Twist Maxi Dress
	Makeup products	○ Lipsticks ○ Bronzers ○ Mascaras
	CD, DVD or Blu-ray disc	○ Lindsay Lohan Blu-ray ○ Rita wilson DVD
	Books related to famous artists	○ Tom Hanks Books
2. Sports & news	Sport product	○ Football player posters ○ Tennis racket ○ Sport shoes
	News service	○ Online newspaper service ○ Magazine ○ News aggregation service
	Economic and political magazines	○ Economic Magazines ○ Political Magazines ○ Donald trump book
3. Singers, rappers & Music	Album or CD, DVD	○ Justin bieber album ○ Taylor swift CD ○ Eminem DVD
	Music souvenir	○ Poster of singer, rapper ○ Concert banner, flag
	Audio equipment	○ Headphone ○ Mp3 player ○ Microphone
4. Artist, News & TV show	Books and magazines related to artists	○ Eva Mendes Biography ○ Books related to Jessica ○ Paperback related to Sarah Silverman

	Products related to artist	○ Charlize Theron Dior perfume ○ Brie Larson T-shirt ○ Haircare by Jennifer Aniston
	Movies and TV Shows	○ Jennifer Lawrence's Movies ○ Conan O'Brien's Podcasts
5. India Cricket, Bollywood & Entertainment Activities	Books and Magazines About Bollywood	○ Bollywood Magazines from India ○ Bollywood Celebrities Special Monthly Magazines
	Sports products related to cricket	○ Cricket Bats ○ Cricket Balls ○ Cricket Protective Gear
	Bollywood Movies and TV Shows	○ Salman Khan Movies ○ Deepika Padukone movies ○ Salman Khan: Movies, TV, and Bio
6. Cuisine, Basketball & Entertainment Activities	Food recipe books	○ Best Cookbooks: Food, Wine, and Baking Books ○ Cookbooks, Food & Wine: Books ○ Free Recipe Books
	Products related to Basketball	○ Basketball Shoes ○ Lebron james poster ○ Basketball match ticket
	Products related to outdoor activities	○ Camping tent, outdoor stove ○ Fishing tools
7. Artist, Music & Entertainment Activities	Tickets to live performances or tours	○ Nicki Minaj Tickets ○ Carrie Underwood Tour 2024
	DVD or Blu-ray disc	○ Mariah Carey Blu Ray ○ The Martin Garrix DVD
	Musical accessories	○ Instrument Accessories ○ Music Accessories Store

4.5. So sánh với các nghiên cứu trước

So với nghiên cứu trước trong lĩnh vực phân tích cảm xúc, sở thích, và chủ đề cá nhân trên mạng xã hội, cụ thể là nghiên cứu của [27], nghiên cứu của chúng tôi đề xuất một phương pháp tiên tiến hơn. Trong khi nghiên cứu của [27] tập trung vào việc thống kê số lượng, phân bổ các lượt bày tỏ cảm xúc ở các bài đăng để hiểu sở thích cá nhân, cảm xúc của người dùng đối với một chủ đề nào đó, thì nghiên cứu của chúng tôi đi sâu vào việc phân tích nội dung

văn bản của người dùng để hiểu biết sâu sắc hơn về sở thích, chủ đề, và tư duy của họ. Phương pháp của chúng tôi tập trung vào việc phân tích và hiểu biết ngữ cảnh xung quanh các bài đăng trên mạng xã hội. Thay vì chỉ đếm lượt bày tỏ cảm xúc một cách cơ bản, chúng tôi sử dụng các công nghệ và phương pháp tiên tiến để phân tích nội dung văn bản. Qua đó, chúng tôi nhằm nhận biết và hiểu rõ hơn về cảm xúc, ý kiến, và quan điểm của người dùng, từ đó mang lại cái nhìn toàn diện và sâu sắc về hành vi và tư duy của họ trên mạng xã hội.

So với các phương pháp truyền thống như TF-IDF được sử dụng trong nghiên cứu của [28], nghiên cứu của chúng tôi đề xuất một cách tiếp cận mới sử dụng các mô hình vector embedding pretrained tiên tiến trong lĩnh vực transfer learning để vector hóa văn bản. Trong khi TF-IDF gặp phải những hạn chế đáng kể như không xác định được ngữ nghĩa của từ, không thể kiểm tra sự xuất hiện đồng thời của các từ, và không hiệu quả khi văn bản cần phân loại không đồng nhất, phương pháp của chúng tôi vượt qua những giới hạn này và mang lại nhiều lợi ích đáng kể. Việc sử dụng các mô hình pretrained đã được huấn luyện trên dữ liệu lớn nên có khả năng hiểu biết ngôn ngữ phong phú và biểu diễn ngữ cảnh một cách hiệu quả. Điều này giúp tăng cường khả năng hiểu biết về ý nghĩa của từng từ và cụm từ trong văn bản. Ngoài ra, sử dụng transfer learning cũng giúp chúng tôi tận dụng được kiến thức đã học từ các tác vụ liên quan trước đó, giúp cải thiện hiệu suất và độ chính xác của mô hình. Mặt khác, trong khi TF-IDF chỉ tập trung vào cấp độ từ vựng, phương pháp của chúng tôi có thể hiểu được ngữ nghĩa của văn bản và đồng thời xem xét mối quan hệ giữa các từ.

Tóm tắt chương 4

Ở chương này, chúng tôi thực hiện đánh giá kết quả gom cụm từ kết quả của các phương pháp nhúng khác nhau kết hợp cùng các phương pháp giảm chiều dữ liệu. Từ đó đưa ra kết luận sự kết hợp giữa phương pháp nhúng văn bản ‘E5-base’ cùng với phương pháp giảm chiều Umap đem lại kết quả gom cụm tốt nhất. Tiếp theo, từ kết quả gom cụm chúng tôi đưa ra các ứng dụng thiết thực: xây dựng các cộng đồng người dùng và đề xuất các sản phẩm dựa trên nội dung mà họ quan tâm.

CHƯƠNG 5: KẾT LUẬN VÀ KIẾN NGHỊ

5.1. Kết luận

Trong nghiên cứu hiện tại, chúng tôi đã đề xuất và phát triển một phương pháp tiên tiến nhằm tự động hóa toàn bộ quy trình phân tích và gom cụm người dùng trên nền tảng mạng xã hội Facebook dựa trên nội dung bài đăng của họ. Bằng việc kết hợp một cách hiệu quả các kỹ thuật học máy hiện đại như mô hình vector nhúng từ vựng và các thuật toán phân cụm tiên tiến, phương pháp được trình bày đã thể hiện khả năng xử lý dữ liệu quy mô lớn, bao gồm việc thu thập, tiền xử lý, trích xuất đặc trưng và cuối cùng là phân loại người dùng.

Kết quả đạt được từ nghiên cứu này cho thấy phương pháp đề xuất đã thành công trong việc phân chia tập người dùng Facebook thành các cụm riêng biệt, trong đó mỗi cụm đại diện cho một tập hợp sở thích và mối quan tâm đặc trưng. Nhờ khả năng xác định chính xác các chủ đề then chốt liên quan đến từng nhóm người dùng, nghiên cứu đã đóng góp một công cụ

phân tích dữ liệu đắc lực, giúp các tổ chức doanh nghiệp nâng cao mức độ hiểu biết và nhận thức về phân khúc khách hàng tiềm năng của mình.

Với khả năng phân tích sâu rộng này, các nhà kinh doanh và nhà tiếp thị có thể thiết kế và triển khai các chiến lược tiếp cận, quảng bá sản phẩm hay dịch vụ một cách hiệu quả hơn, phù hợp với nhu cầu và sở thích cụ thể của từng phân khúc khách hàng. Trong phạm vi nghiên cứu, chúng tôi đã có thể xây dựng các cộng đồng người dùng có cùng chủ đề sở thích cá nhân, đề xuất và trình bày một số ví dụ sản phẩm mẫu trong các lĩnh vực chủ đề được xác định từ dữ liệu thực tế, nhằm minh chứng tiềm năng ứng dụng thực tiễn cao của phương pháp được đề xuất.

Tóm lại, nghiên cứu hiện tại đã đạt được mục tiêu ban đầu đề ra là thực hiện phân cụm người dùng Facebook một cách hiệu quả dựa trên phân tích nội dung bài đăng của họ. Thành quả chính là việc xác định rõ ràng các chủ đề quan tâm nổi bật của từng nhóm người dùng, từ đó hỗ trợ việc đưa ra các đề xuất sản phẩm hay dịch vụ thiết thực, hay hỗ trợ xây dựng cộng đồng người dùng trên các nền tảng mạng xã hội. Những kết quả thu được không chỉ mở ra nhiều triển vọng ứng dụng thực tế trong lĩnh vực tiếp thị, phân khúc khách hàng và nâng cao hiệu quả kinh doanh, mà còn đóng góp vào sự phát triển của lĩnh vực phân tích dữ liệu quy mô lớn và học máy ứng dụng nói chung.

5.2. Hạn chế của đề tài

Mặc dù nghiên cứu hiện tại đã đạt được những kết quả đáng khích lệ, song vẫn tồn tại một số hạn chế nhất định cần được xem xét và khắc phục trong các công trình nghiên cứu tiếp theo:

Thứ nhất, phương pháp phân tích chỉ dựa trên dữ liệu nội dung bài đăng của người dùng trên Facebook, chưa tích hợp các nguồn thông tin bổ sung khác như đặc điểm nhân khẩu học, vị trí địa lý, lịch sử tương tác hay mối quan hệ xã hội. Việc kết hợp các dạng dữ liệu đa chiều này sẽ giúp tạo ra bức tranh toàn cảnh, chi tiết và chính xác hơn về hồ sơ người dùng, từ đó nâng cao hiệu quả phân loại.

Hạn chế thứ hai, liên quan đến khả năng xử lý ngôn ngữ tự nhiên hiện tại của phương pháp. Mặc dù đã sử dụng các kỹ thuật tiên tiến, song vẫn có thể gặp khó khăn trong việc phân tích các dạng ngôn ngữ phức tạp như ẩn dụ, đa nghĩa hay kiểu ngôn ngữ đặc thù của riêng cộng đồng người dùng. Việc kết hợp các mô hình xử lý ngôn ngữ hiện đại, tiên tiến hơn sẽ góp phần nâng cao hiệu suất và tính chính xác của hệ thống.

Một hạn chế khác thứ ba, là phương pháp gom cụm chưa xem xét sự tương tác và mối liên hệ giữa các chủ đề. Điều này có thể dẫn đến việc bỏ qua những mối liên kết quan trọng giữa các nhóm người dùng khác nhau. Để khắc phục, các nghiên cứu tiếp theo cần phát triển các mô hình phân cụm nâng cao với khả năng xem xét mối quan hệ chủ đề, từ đó phát hiện các nhóm người dùng phức tạp, đan xen nhiều chủ đề liên quan.

Cuối cùng, mức độ hiệu quả và khả năng mở rộng của phương pháp có thể bị ảnh hưởng bởi kích thước và chất lượng của tập dữ liệu đầu vào. Những cải tiến về hiệu năng xử lý dữ liệu lớn, khả năng song song hóa và lưu trữ phân tán sẽ giúp nâng cao tính khả thi và khả

năng mở rộng của hệ thống trong tương lai. Tuy nhiên, những hạn chế này không làm giảm đi các đóng góp và ý nghĩa của nghiên cứu hiện tại. Đồng thời, chúng cũng mở ra nhiều hướng phát triển tiềm năng nhằm nâng cao chất lượng và hiệu năng của phương pháp trong tương lai.

5.3. Hướng phát triển đề tài

Dựa trên những kết quả đạt được và hạn chế được nhận diện từ nghiên cứu hiện tại, nhiều hướng phát triển tiềm năng đã được khởi tạo, mở ra các triển vọng nghiên cứu mới trong lĩnh vực phân tích dữ liệu người dùng quy mô lớn và ứng dụng học máy trong tiếp thị.

Một trong những hướng đi đó là tích hợp các nguồn dữ liệu phong phú hơn, bao gồm thông tin nhân khẩu học, vị trí địa lý, lịch sử tương tác và mối liên hệ xã hội. Việc kết hợp các dạng dữ liệu đa chiều này sẽ tạo ra bức tranh toàn cảnh về người dùng, giúp các mô hình có thể nắm bắt đầy đủ hơn các khía cạnh về nhu cầu, hành vi và đặc điểm của từng cá nhân. Từ đó, khả năng phân loại và dự đoán sẽ được nâng cao đáng kể.

Một hướng phát triển quan trọng khác là nghiên cứu và ứng dụng các kỹ thuật tiên phong trong lĩnh vực xử lý ngôn ngữ tự nhiên như các mô hình ngôn ngữ lớn (large language models). Nhờ khả năng hiểu được ý nghĩa ngữ cảnh phức tạp và nắm bắt các quan hệ ngôn ngữ tinh tế, những công cụ xử lý ngôn ngữ tự nhiên hiện đại này sẽ nâng cao đáng kể tính chính xác và hiệu suất của hệ thống phân tích nội dung bài đăng người dùng.

Trong quá trình phân cụm, các nghiên cứu tiếp theo có thể tập trung vào phát triển các mô hình gom cụm nâng cao, có khả năng xem xét sự tương quan giữa các chủ đề. Điều này sẽ cho phép phát hiện ra những mối liên hệ phức tạp, đan xen giữa các nhóm người dùng, mở ra cơ hội nhận diện các phân khúc khách hàng tiềm năng mới.

Để đảm bảo khả năng mở rộng và ứng dụng thực tế của phương pháp, việc phát triển các giải pháp xử lý song song, phân tán và lưu trữ dữ liệu quy mô lớn cũng là một yêu cầu cấp thiết. Những cải tiến về hiệu năng này sẽ đảm bảo khả năng xử lý dữ liệu khổng lồ từ các nền tảng mạng xã hội khác nhau và mở rộng phạm vi áp dụng của phương pháp trong tương lai.

Cuối cùng, xây dựng giao diện người dùng (GUI) trực quan, tương tác là một nhiệm vụ quan trọng nhằm trình bày và khai thác hiệu quả các kết quả phân tích. Với một giao diện thân thiện và dễ sử dụng, các nhà kinh doanh, nhà tiếp thị có thể dễ dàng khai thác, trực quan hóa và đưa ra các quyết định kinh doanh dựa trên kết quả phân tích phân khúc khách hàng.

Trong quá trình phát triển các hướng nghiên cứu mới, đảm bảo tuân thủ các quy định về bảo vệ quyền riêng tư và xử lý dữ liệu cá nhân người dùng sẽ luôn là ưu tiên hàng đầu. Việc xây dựng môi trường nghiên cứu minh bạch, đạo đức và có trách nhiệm là một yêu cầu cốt lõi để đảm bảo sự phát triển bền vững của lĩnh vực. Với những hướng phát triển đầy triển vọng này, các nghiên cứu tiếp theo sẽ góp phần quan trọng vào việc nâng cao năng lực phân tích dữ liệu và ứng dụng học máy trong lĩnh vực tiếp thị và kinh doanh. Những tiến bộ trong tương lai sẽ giúp mở rộng phạm vi áp dụng, cải thiện hiệu quả và nâng cao tính thực tiễn của các phương pháp phân tích dữ liệu khách hàng, từ đó nâng cao lợi thế cạnh tranh của các doanh nghiệp trong kỷ nguyên thời đại số.

TÀI LIỆU THAM KHẢO

[1] S. Yang, G. Huang, B. Ofoghi and J. Yearwood, "Short text similarity measurement using context-aware weighted biterms," *Concurrency and Computation: Practice and Experience,* vol. 34(8), p. e5765, 2022.

[2] C. Virmani, A. Pillai and D. Juneja, "Clustering in Aggregated User Profiles across Multiple Social Networks," *International Journal of Electrical & Computer Engineering,* vol. 7(6), pp. 2088-8708, 2017.

[3] V. Gurusamy, S. Kannan and J. R. Prabhu, "Mining the attitude of social network users using k-means clustering," *International Journal,*, vol. 7(5), 2017.

[4] Z. Qiu and H. Shen, "User clustering in a dynamic social network topic model for short text streams," *Information Sciences,* vol. 414, pp. 102-116, 2017.

[5] K. D. Joshi and P. S. Nalwade, "Modified K-Means for Better Initial Cluster Centres," *International Journal of Computer Science and Mobile Computing,* vol. 2(7), pp. 219-223, 2013.

[6] H. Jiawei, M. Kamber and I. Pei, "Data mining: concepts and techniques," *Morgan kaufmann,* 2006.

[7] T. Hastie, R. Tibshirani, J. H. Friedman and J. H. Friedman, "The elements of statistical learning: data mining, inference, and prediction," *Springer,* vol. 2, pp. 1-758, 2009.

[8] S. Rose, D. Engel, N. Cramer and W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory,* pp. 1-20, 2010.

[9] T. Howley, M. G. Madden, M. L. O'Connell and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high dimensional spectral data," *In International Conference on Innovative Techniques and Applications of Artificial Intelligence,* pp. 209-222, 12 2005.

[10] M. L. c. bản, "Bài 27: Principal Component Analysis (phần 1/2)," 15 06 2017. [Online]. Available: https://machinelearningcoban.com/2017/06/15/pca/.

[11] Datacamp, "Introduction to t-SNE," 3 2023. [Online]. Available: https://www.datacamp.com/tutorial/introduction-t-sne.

[12] S. Dobilas, "t-SNE Machine Learning Algorithm — A Great Tool for Dimensionality Reduction in Python," 26 9 2021. [Online]. Available: https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e.

[13] L. McInnes, J. Healy and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426,* 2018.

[14] h. vijay, "Dimensionality Reduction : PCA, tSNE, UMAP," 18 5 2023. [Online]. Available: https://aurigait.com/blog/blog-easy-explanation-of-dimensionality-reduction-and-techniques/.

[15] S. Dobilas, "UMAP降维算法原理详解和应用示例," 13 11 2021. [Online]. Available: https://zhuanlan.zhihu.com/p/432805218.

[16] T. N. T. Zakaria, M. J. Ab Aziz, M. R. Mokhtar and S. Darus, "Semantic similarity measurement for Malay words using WordNet Bahasa and Wikipedia Bahasa Melayu: Issues and proposed solutions," *International Journal of Software Engineering and Computer Systems,* vol. 6(1), pp. 25-40, 2020.

[17] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 233-242, 8 2014.

[18] A. Sabah, S. Tiun, N. S. Sani, M. Ayob and A. Y. Taha, "Enhancing web search result clustering model based on multiview multirepresentation consensus cluster ensemble (mmcc) approach," *PloS one,* vol. 16(1), p. e0245264, 2021.

[19] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533,* 2022.

[20] Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu and M. W. Chang, "Promptagator: Few-shot dense retrieval from 8 examples," *arXiv preprint arXiv:2209.11755*, 2022.

[21] N. Muennighoff, N. Tazi, L. Magne and N. Reimers, "MTEB: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022.

[22] A. N. Lê, "Phân nhóm người dùng dựa vào hành vi tương tác trong mạng xã hội," *Trường Đại học Bách khoa - Đại học Đà Nẵng*, 2018.

[23] M. H. Ahmed, S. Tiun, N. Omar and N. S. Sani, "Short Text Clustering Algorithms, Application and Challenges: A Survey," *Applied Sciences*, vol. 13(1), p. 342, 2022.

[24] H. Chun, B. H. Leem and H. & Suh, "Using text analytics to measure an effect of topics and sentiments on social-media engagement: Focusing on Facebook fan page of Toyota," *International Journal of Engineering Business Management*, vol. 13, 2021.

[25] S. Debortoli, O. Müller, I. Junglas and J. Vom Brocke, "Text mining for information systems researchers: An annotated topic modeling tutorial," *ommunications of the Association for Information Systems (CAIS)*, vol. 39(1), p. 7, 2016.

[26] J. Thomaz, "What is a vector embedding?," 18 9 2023. [Online]. Available: https://dev.to/josethz00/what-is-a-vector-embedding-3335.

[27] F. T. Giuntini, L. P. Ruiz, L. D. F. Kirchner, D. A. Passarelli, M. D. J. D. Dos Reis, A. T. Campbell and J. Ueyama, "How do I feel? Identifying emotional expressions on Facebook reactions using clustering mechanism," *IEEE Access*, vol. 7, pp. 53909-53921, 2019.

[28] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181(1), pp. 25-29, 2018.

[29] B. Arsić, M. Bašić, P. Spalević, M. Ilić and M. Veinović, "Facebook profiles clustering," *In Proceedings of 6th International Conference on Information Society and Technology*, vol. ICIST 2016, pp. 154-158, 2016.

[30] K. K. Bain, I. Firli and S. Tri, "A Genetic Algorithm For Optimized Initial Centers K-Means Clustering In SMEs," *Journal of Theoretical and Applied Information Technology* , vol. 90(1), p. 23, 2016.

[31] P. Hu, W. Liu, W. Jiang and Z. Yang, "Latent topic model based on Gaussian-LDA for audio retrieval," *In Pattern Recognition: Chinese*

Conference, CCPR 2012, Beijing, China, September 24-26, 2012. Proceedings., vol. 321 of CCIS, pp. 5 (pp. 556-563), 2012.

[32] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science,* vol. 344(6191), p. 1492–1496, 2014.

[33] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie and M. & Zhang, "Towards General Text Embeddings with Multi-stage Contrastive Learning," *arXiv preprint arXiv:2308.03281,* 2023.

[34] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9(11), 2008.

[35] L. McInnes, "How UMAP Works," 2018. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html.

[36] M. H. Ahmed and S. Tiun, "K-means based algorithm for islamic document clustering," *In Proceedings of International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2013),* pp. 2-9, 7 2013.

[37] A. S. Abdulameer, S. Tiun, N. S. Sani, M. Ayob and A. Y. Taha, "Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering," *Journal of King Saud University-Computer and Information Sciences,* vol. 34(3), pp. 840-850, 2022.

[38] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study," *In Conference on Data Mining| DMIN,* vol. 6, p. 79, 2006.

[39] S. Dobilas, "UMAP Dimensionality Reduction — An Incredibly Robust Machine Learning Algorithm," 25 10 2021. [Online]. Available: https://towardsdatascience.com/umap-dimensionality-reduction-an-incredibly-robust-machine-learning-algorithm-b5acb01de568.

[40] A. Vysala and D. J. Gomes, "Evaluating and validating cluster results," *arXiv preprint arXiv:2007.08034,* 2020.

[41] M. Mughnyanti, S. Efendi and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *In IOP Conference Series: Materials Science and Engineering,* vol. 725(1), p. 012128, 2020.

[42] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987.

[43] R. B. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communication in statistics,* 1974.

[44] S. N. Kim, O. Medelyan, M. Y. Kan, T. Baldwin and L. P. Pingar, "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific," *Proc. 5th Int. Workshop Semantic Eval,* pp. 21-26, 2010.

[45] F. Boudin, "PKE: an open source python-based keyphrase extraction toolkit," *In Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations*, pp. 69-73, 12 2016.

[46] M. T. Luong, T. D. Nguyen and M. Y. Kan, "Logical structure recovery in scholarly articles with rich document features," *In Multimedia Storage and Retrieval Innovations for Digital Library Systems,* pp. 270-292, 2012.

[47] S. R. El-Beltagy and A. Rafea, "Kp-miner: Participation in semeval-2," *In Proceedings of the 5th international workshop on semantic evaluation,* pp. 190-193, 7 2010.

[48] X. Wan and J. Xiao, "CollabRank: towards a collaborative approach to single-document keyphrase extraction," *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* pp. 969-976, 8 2008.

[49] A. Bougouin, F. Boudin and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction," *In International joint conference on natural language processing (IJCNLP),* pp. 543-551, 10 2013.

[50] L. Sterckx, T. Demeester, J. Deleu and C. Develder, "Topical word importance for fast keyphrase extraction," *In Proceedings of the 24th International Conference on World Wide Web*, pp. 121-122, 5 2015.

[51] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," *In Proceedings of the 55th annual meeting of the association for computational linguistics*, vol. 1: long papers, pp. 1105-1115, 7 2017.

[52] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," *arXiv preprint arXiv:1803.08721,* 2018.

[53] I. H. Witten and G. W. F. E. G. C. &. N.-M. Paynter, "KEA: Practical automatic keyphrase extraction," *In Proceedings of the fourth ACM conference on Digital libraries*, pp. 254-255, 8 1999.

[54] T. D. Nguyen and M. T. Luong, "WINGNUS: Keyphrase extraction utilizing document logical structure," *In Proceedings of the 5th international workshop on semantic evaluation*, pp. 166-169, 7 2010.

[55] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia tools and applications*, vol. 78, pp. 15169-15211, 2019.

[56] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert and S. Adam, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology," *In Computational methods for communication science*, pp. 13-38, 2021.

[57] Y. Song, S. Pan, S. Liu, M. X. Zhou and W. Qian, "Topic and keyword re-ranking for LDA-based topic modeling," *In Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1757-1760, 9 2009.

[58] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation. Journal of machine Learning research," pp. 993-1022, 3 1 2003.

[59] D. M. Blei and J. D. Lafferty, "Topic models," *In Text mining*, pp. 101-124, 2009.

[60] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55(4), pp. 77-84, 2012.

[61] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[62] V. Duc-Vinh, K. Jessada and H. Van-Nam, "An integrated framework of learning and evidential reasoning for user profiling using short texts," *Information Fusion*, vol. 70, pp. 27-42, 2021.

[63] D. -V. Vo, T. -T. Tran, K. Shirai and V. -N. Huynh, "Deep Generative Networks Coupled With Evidential Reasoning for Dynamic User Preferences Using Short Texts," *in IEEE Transactions on Knowledge and Data Engineering*, vol. 35(7), pp. 6811-6826, 1 7 2023.