BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC
HỌC PHẦN: MẠNG XÃ HỘI

TÊN ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU CÁC BÀI BÁO TRÊN MẠNG XÃ HỘI VNEXPRESS

Ngành: **KHOA HỌC DỮ LIỆU**

Chuyên ngành: **KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: **LÊ NHẬT TÙNG**

Lớp: **21DKHA1**

Sinh viên thực hiện:	MSSV:
Nguyễn Bá Tuấn Anh	2186400123
Trần Minh Chiến	2186400225
Nguyễn Hồng Nhất Linh	2186400278

TP. Hồ Chí Minh, 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

HCM, Ngày … tháng … năm 2024

Giáo viên hướng dẫn

(Ký tên, đóng dấu)

MỤC LỤC

DANH MỤC HÌNH

LỜI CẢM ƠN

Để hoàn thành tốt đồ án này, chúng tôi xin chân thành gửi lời cảm ơn sâu sắc đến thầy Lê Nhật Tùng người đã luôn đồng hành, hướng dẫn và chia sẻ những kiến thức quý báu trong suốt quá trình học tập và làm báo cáo đồ án. Sự tận tâm và nhiệt huyết của thầy đã giúp chúng tôi vượt qua nhiều khó khăn và thách thức, đồng thời truyền cảm hứng cho chúng tôi trong việc nghiên cứu và phát triển dự án này.

Bên cạnh đó, chúng tôi cũng muốn bày tỏ lòng biết ơn đối với những người bạn đồng môn đã hỗ trợ, động viên và cùng nhau làm việc để hoàn thành đồ án. Những cuộc thảo luận sôi nổi và sự chia sẻ ý tưởng từ các bạn đã góp phần làm phong phú thêm nội dung của báo cáo. Chúng tôi rất trân trọng sự gắn bó và tình bạn trong suốt quá trình này.

Ngoài ra, chúng tôi cũng xin gửi lời cảm ơn đến quý thầy cô và các chuyên gia đã dành thời gian để góp ý và cho chúng tôi những chỉ dẫn quý giá. Những phản hồi của các thầy cô là nguồn động lực lớn giúp chúng tôi hoàn thiện hơn trong từng bước đi.

Chúng tôi nhận thức được rằng, do thời gian hạn hẹp và những thiếu sót trong quá trình thực hiện, đồ án vẫn còn nhiều điểm cần cải thiện. Chúng tôi rất mong nhận được sự thông cảm và những ý kiến đóng góp chân thành từ thầy và các bạn để có thể hoàn thiện hơn trong các dự án sau này.

Một lần nữa, xin chân thành cảm ơn tất cả mọi người đã hỗ trợ và đồng hành cùng chúng tôi.

Chúng tôi xin chân thành cảm ơn!

Sinh viên thực hiện

Nguyễn Hồng Nhất Linh

Trần Minh Chiến

Nguyễn Bá Tuấn Anh

LỜI MỞ ĐẦU

Trong thời đại kỹ thuật số hiện nay, mạng xã hội đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của con người. Đây không chỉ là nơi giao tiếp, chia sẻ thông tin mà còn là nền tảng quan trọng để các cơ quan báo chí tiếp cận độc giả. Trong số các tờ báo lớn tại Việt Nam, VnExpress nổi bật với vai trò tiên phong trong việc khai thác sức mạnh của mạng xã hội để lan tỏa thông tin một cách nhanh chóng và hiệu quả.

Tuy nhiên, việc các bài báo trên mạng xã hội thu hút sự chú ý như thế nào, chủ đề nào được quan tâm nhiều nhất, hay yếu tố nào quyết định mức độ tương tác của người đọc vẫn là những câu hỏi đáng được nghiên cứu. Những câu hỏi này không chỉ giúp hiểu rõ hơn về hành vi của người dùng trên mạng xã hội mà còn cung cấp cơ sở để cải thiện chiến lược truyền thông cho các cơ quan báo chí.

Với đề tài *"Phân tích dữ liệu các bài báo trên trang mạng xã hội VnExpress"*, bài báo cáo này sẽ tập trung nghiên cứu các khía cạnh như xu hướng chủ đề và mức độ tương tác. Qua đó, bài viết không chỉ giúp làm sáng tỏ những đặc điểm quan trọng trong việc truyền thông báo chí trên mạng xã hội mà còn góp phần đề xuất các giải pháp tối ưu hóa nội dung nhằm đáp ứng tốt hơn nhu cầu của độc giả trong kỷ nguyên số.

CHƯƠNG I: TỔNG QUAN

1. Sơ lượt về tài

Mạng xã hội ngày nay không chỉ là nơi kết nối mọi người mà còn là kênh thông tin quan trọng đối với báo chí. VnExpress, một trong những tờ báo điện tử hàng đầu Việt Nam, đã và đang tận dụng mạnh mẽ các nền tảng mạng xã hội để tăng cường mức độ tiếp cận và tương tác với độc giả. Đề tài *"Phân tích dữ liệu các bài báo trên trang mạng xã hội VnExpress"* hướng tới việc khám phá những đặc điểm nổi bật trong hoạt động truyền thông báo chí trên mạng xã hội, giúp làm rõ cách mà nội dung báo chí được lan tỏa và đón nhận.

1.2 Tính cấp thiết của đề tài

Sự bùng nổ của mạng xã hội đã thay đổi cách thức mà con người tiếp cận và tiêu thụ thông tin. Trong bối cảnh cạnh tranh khốc liệt giữa các tờ báo, việc hiểu rõ xu hướng tương tác, hành vi độc giả, và các yếu tố ảnh hưởng đến sự phổ biến của các bài viết là điều cần thiết. Đặc biệt, các nghiên cứu chuyên sâu về dữ liệu tương tác trên mạng xã hội hiện tại vẫn còn hạn chế, nhất là trong lĩnh vực báo chí. Điều này tạo nên tính cấp thiết cho việc phân tích dữ liệu từ các bài báo của VnExpress nhằm cung cấp những hiểu biết mới và hữu ích cho lĩnh vực truyền thông hiện đại.

1.3 Ý nghĩa khoa học và thực tiễn

- **Ý nghĩa khoa học:** Nghiên cứu góp phần vào việc xây dựng cơ sở dữ liệu về hoạt động báo chí trên mạng xã hội tại Việt Nam, đồng thời bổ sung những lý luận về hành vi người dùng và cách thức lan tỏa nội dung trên không gian số.

- **Ý nghĩa thực tiễn**: Kết quả nghiên cứu có thể hỗ trợ các cơ quan báo chí trong việc tối ưu hóa nội dung, cải thiện chiến lược tiếp cận độc giả, và nâng cao hiệu quả truyền thông. Ngoài ra, đề tài còn cung cấp các gợi ý hữu ích cho những nhà quản lý mạng xã hội, giúp họ hiểu rõ hơn về nhu cầu và thị hiếu của người dùng.

1.4 Phương pháp nghiên cứu

- **Phương pháp thu thập dữ liệu**: Sử dụng các công cụ thu thập dữ liệu để thu thập thông tin về bài báo:

    1. **URL**: Đường dẫn (liên kết) của bài báo.
    2. **Date**: Ngày đăng bài.
    3. **Category**: Chuyên mục (danh mục) của bài báo (ví dụ: Kinh tế, Xã hội, Giải trí).
    4. **Title**: Tiêu đề bài báo.
    5. **Comment Count**: Số lượng bình luận của bài báo.

- **Phương pháp phân tích**: Áp dụng các kỹ thuật phân tích dữ liệu để khám phá xu hướng và các mối liên hệ trong dữ liệu.
- **Phương pháp trực quan hóa**: Sử dụng biểu đồ để thể hiện xu hướng và so sánh các chỉ số quan trọng.

1.5 Cấu trúc đồ án

Đồ án được cấu trúc thành các chương như sau:

- **Chương 1: Giới thiệu**: Trình bày tổng quan về đề tài nghiên cứu, tính cấp thiết, ý nghĩa khoa học và thực tiễn, phương pháp nghiên cứu, đối tượng sử dụng và cấu trúc đồ án.
- **Chương 2: Cơ sở lý thuyết**: Giới thiệu về các mô hình học máy, trí tuệ nhân tạo được sử dụng trong nghiên cứu và bộ dữ liệu.
- **Chương 3: Kết quả và thực nghiệm**
- **Chương 4: Kết luận**

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2. Tổng quan về dữ liệu

Bộ dữ liệu được xây dựng bằng cách cào dữ liệu từ trang báo điện tử VnExpress, bao gồm 5 nhãn thông tin: URL**,** Date**,** Category**,** Title**, và** Comment Count. Tổng cộng, bộ dữ liệu chứa 822 hàng dữ liệu, mỗi hàng đại diện cho một bài báo khác nhau. Các nhãn mang ý nghĩa cụ thể:

- **URL**: Đường dẫn liên kết đến bài báo.
- **Date**: Ngày và thời gian đăng bài.
- **Category**: Chuyên mục của bài báo (ví dụ: Sức khỏe, Khoa học, Thời sự, Giáo dục).
- **Title**: Tiêu đề của bài báo, giúp nhận diện nhanh nội dung.
- **Comment Count**: Số lượng bình luận cho mỗi bài báo, biểu thị mức độ tương tác từ độc giả.

Bộ dữ liệu cung cấp cái nhìn tổng quan về nội dung và mức độ quan tâm của độc giả đối với các bài viết trên VnExpress, tạo nền tảng cho việc phân tích chuyên sâu về xu hướng chủ đề, hành vi người dùng và mức độ tương tác trên nền tảng mạng xã hội.

2.1. Đề tài được viết bằng ngôn ngữ gì?

Đề tài sử dụng ngôn ngữ python để phân tích và cào dữ liệu.

2.2 Ngôn ngữ Python là gì?

Python được sáng tạo bởi Guido van Rossum, một nhà nghiên cứu người Hà Lan. [1] Phiên bản chính của Python là Python 2 và Python 3, với Python 3 được khuyến khích sử dụng sau khi hỗ trợ cho Python 2 chấm dứt vào tháng 1 năm 2020.

Ngoài ra python còn có những lợi ích trong việc nghiên cứu:

[1] Phiên bản đầu tiên (Python 0.9.0) được ra mắt vào tháng 2 năm 1991. Tên "Python" xuất phát từ đam mê của Guido với một chương trình truyền hình hài nổi

tiếng. Python là một ngôn ngữ lập trình thông dịch (interpreted), có cú pháp đơn giản, dễ đọc, và có hệ sinh thái phong phú với hàng nghìn thư viện và framework hỗ trợ. Phiên bản đầu tiên (Python 0.9.0) ra mắt vào tháng 2 năm 1991. Tên "Python" xuất phát từ đam mê của Guido với một chương trình truyền hình hài nổi tiếng. Python là một ngôn ngữ lập trình thông dịch (interpreted), có cú pháp đơn giản, dễ đọc, và có hệ sinh thái phong phú với hàng nghìn thư viện và framework hỗ trợ.

+ Cú pháp đơn giản và dễ hiểu: Giúp lập trình viên dễ dàng viết và duy trì mã nguồn.
+ Cộng đồng mạnh mẽ: Hỗ trợ tốt, tài liệu phong phú và nhiều dự án mã nguồn mở, giúp lập trình viên giải quyết vấn đề nhanh chóng.
+ Tính linh hoạt và hiệu suất cao: Python được sử dụng rộng rãi trong nhiều lĩnh vực từ phát triển web, khoa học dữ liệu, đến trí tuệ nhân tạo và machine learning, giúp nâng cao hiệu quả công việc.

3. Độ đo cơ bản của mạng (Basic Network Metrics)

3.1. Average Degree

Ý nghĩa: [2] Average degree, hay bậc trung bình, là một chỉ số quan trọng trong lý thuyết đồ thị, thể hiện mức độ kết nối trung bình của các đỉnh trong một đồ thị. Nó cho thấy một cái nhìn tổng quan về độ dày đặc của các cạnh, tức là trung bình một đỉnh sẽ kết nối với bao nhiêu đỉnh khác. Đồ thị có average degree cao thường có nhiều cạnh hơn so với số đỉnh, cho thấy các đỉnh có xu hướng kết nối với nhiều đỉnh khác và ngược lại. [2]

Công thức:

$$(\text{avg\_deg}) = (2 * |E|) / |V|$$

$|E|$ là số lượng cạnh của đồ thị

$|V|$ là số lượng đỉnh của đồ thị

Phạm vi: Average degree là một số thực không âm và có thể có giá trị là số nguyên hoặc thập phân. Với đồ thị vô hướng, average degree nằm trong khoảng từ 0

(đồ thị rỗng) đến |V| - 1 (đồ thị đầy đủ). Trong khi đó, với đồ thị có hướng, average degree có thể vượt quá giới hạn |V| - 1 vì mỗi cạnh có hướng riêng biệt.

3.2. Network Diameter

Ý nghĩa: [3] Network diameter, hay đường kính mạng, cho biết số bước đi tối đa cần thiết để đi từ một đỉnh bất kỳ đến một đỉnh bất kỳ khác trong đồ thị. Một đường kính mạng nhỏ thường cho thấy mạng kết nối tốt hơn và có hiệu quả cao hơn.

Công thức:

$$diam(G) = \max\{d(u, v) \mid u, v \in V\}$$

d(u, v): Khoảng cách ngắn nhất giữa đỉnh *u* và đỉnh *v*.

diam(G): Đường kính của đồ thị *G*

V là tập hợp các đỉnh của đồ thị *G*.

max là hàm tìm giá trị lớn nhất.

Phạm vi: [3] Network diameter, hay đường kính mạng, là một số nguyên không âm, thể hiện khoảng cách lớn nhất giữa bất kỳ cặp đỉnh nào trong một đồ thị. Giá trị này có ý nghĩa quan trọng trong việc đánh giá tính kết nối và hiệu quả của mạng. Phạm vi của network diameter phụ thuộc vào cấu trúc và tính liên thông của đồ thị. Phạm vi từ 1 (đồ thị đầy đủ) đến |V| - 1 (đồ thị đường thẳng) trong đồ thị liên thông, và có thể là vô cùng (∞) trong đồ thị không liên thông. Giá trị thực tế của đường kính mạng sẽ phụ thuộc vào cấu trúc cụ thể của từng đồ thị.

3.3. Graph Density

Ý nghĩa: [4] Graph density, hay mật độ đồ thị, là một chỉ số đo lường mức độ "đặc" của một đồ thị. Nó cho biết tỷ lệ giữa số lượng cạnh thực tế trong đồ thị so với số lượng cạnh tối đa mà đồ thị đó có thể chứa. Một đồ thị có mật độ cao có nghĩa là các đỉnh kết nối với nhau nhiều, trong khi một đồ thị có mật độ thấp thì các đỉnh ít kết nối hơn. Graph density giúp chúng ta so sánh mức độ kết nối giữa các đồ thị khác nhau.

Công thức:

Với đồ thị vô hướng:

$$\text{Density} = (2 * |E|) / (|V| * (|V| - 1)$$

Công thức này xuất phát từ việc số cạnh tối đa trong đồ thị vô hướng là |V| * (|V| - 1) / 2, và chúng ta nhân 2 ở tử số để loại bỏ phép chia 2 ở mẫu số.

Với đồ thị có hướng:

$$\text{Density} = |E| / (|V| * (|V| - 1))$$

Trong trường hợp này, số cạnh tối đa có thể có là |V| * (|V| - 1), vì mỗi cạnh có một hướng xác định.

Phạm vi: Graph density là một số thực nằm trong khoảng từ 0 đến 1. Giá trị 0 tương ứng với một đồ thị rỗng (không có cạnh nào), và giá trị 1 tương ứng với một đồ thị đầy đủ (mọi đỉnh đều kết nối với mọi đỉnh khác). Giá trị density càng gần 1, đồ thị càng "đặc", và ngược lại, càng gần 0, đồ thị càng "thưa".

3.4. Connected Components

Ý nghĩa: [4] Trong lý thuyết đồ thị, một connected component (thành phần liên thông) của một đồ thị vô hướng là một tập hợp các đỉnh mà trong đó có một đường đi giữa bất kỳ hai đỉnh nào trong tập đó. Nói cách khác, các đỉnh trong cùng một thành phần liên thông có thể "đến được" nhau thông qua các cạnh của đồ thị. Thành phần liên thông giúp chúng ta hiểu rõ cấu trúc kết nối của đồ thị, phân tách nó thành các "khối" độc lập.

Phạm vi: Trong đồ thị có hướng, khái niệm "thành phần liên thông" có hai biến thể: weakly connected components (thành phần liên thông yếu), nơi các đỉnh liên thông nếu bỏ qua hướng cạnh, và strongly connected components (thành phần liên thông mạnh), nơi có một đường đi có hướng giữa mọi cặp đỉnh

3.5. Average Path Length

Ý nghĩa: [4] Average Path Length (APL), hay độ dài đường đi trung bình, là một chỉ số quan trọng trong lý thuyết đồ thị, đo lường độ dài trung bình của đường đi ngắn nhất giữa tất cả các cặp đỉnh trong một đồ thị. Nó cho biết trung bình cần bao nhiêu bước để đi từ một đỉnh bất kỳ đến một đỉnh bất kỳ khác. APL càng nhỏ, các đỉnh càng "gần gũi" nhau, cho thấy tính kết nối tốt của mạng.

Công thức:

$$APL = (1 / (n * (n - 1))) * \sum \sum d(u, v)$$

n là số lượng đỉnh

d(u, v) là khoảng cách ngắn nhất giữa đỉnh u và v

Phạm vi: [4] APL là một số thực không âm. Trong một đồ thị liên thông, APL có giá trị tối thiểu là 1 (khi đồ thị là đồ thị đầy đủ) và có thể xấp xỉ n/3 trong trường hợp đồ thị là đường thẳng. Với đồ thị không liên thông, APL không được xác định rõ vì không có đường đi giữa một số cặp đỉnh; trong trường hợp này, ta có thể tính APL của từng thành phần liên thông riêng lẻ.

3.6. Average Clustering Coefficent

Ý nghĩa: [4] Average Clustering Coefficient (ACC), hay hệ số cụm trung bình, là một chỉ số quan trọng trong lý thuyết đồ thị, đặc biệt khi phân tích mạng xã hội và các mạng phức tạp khác. Nó đo lường mức độ mà các đỉnh trong đồ thị có xu hướng hình thành các "cụm" hay "tam giác". ACC cho biết, trung bình, các láng giềng của một đỉnh có xu hướng kết nối với nhau như thế nào, hay nói cách khác, nó đo lường mật độ của các đỉnh trong mạng.

Công thức:

$$ACC = (1 / |V|) * \sum C(v)$$

|V| là số lượng đỉnh

C(v) là Clustering Coefficient của đỉnh v

Phạm vi: ACC là một số thực nằm trong khoảng từ 0 đến 1. Một đồ thị với ACC bằng 0 cho thấy các láng giềng của các đỉnh hầu như không kết nối với nhau, còn một đồ thị có ACC bằng 1 cho thấy tất cả các láng giềng của mỗi đỉnh đều kết nối với nhau (tạo thành các clique). ACC càng gần 1, đồ thị càng có tính "cụm" mạnh, cho thấy các đỉnh có xu hướng hình thành các nhóm chặt chẽ.

4. Độ đo tính trung tâm (Centrality Metrics)

4.1. Degree Centrality (In-degree và Out-degree với đồ thị có hướng)

Ý nghĩa: Degree Centrality là một thước đo đơn giản nhưng mạnh mẽ, dùng để xác định tầm quan trọng của một đỉnh trong mạng dựa trên số lượng kết nối trực tiếp mà nó có. Trong đồ thị vô hướng, Degree Centrality của một đỉnh đơn giản là số lượng cạnh mà đỉnh đó kết nối với các đỉnh khác. Nó phản ánh mức độ kết nối của một đỉnh trong mạng. Một đỉnh có nhiều kết nối hơn sẽ có Degree Centrality cao hơn, cho thấy nó có vai trò trung tâm hơn trong mạng.

Công thức:

$$C_D(v) = \frac{\deg(v)}{n-1}$$

n: là số đỉnh của đồ thị

deg(*v*): tổng số liên kết trực tiếp đến đỉnh v (bậc của đỉnh)

Phạm vi: Degree Centrality (và In-degree, Out-degree Centrality) là một số nguyên không âm. Giá trị tối thiểu là 0 (đỉnh cô lập), và giá trị tối đa là |V| - 1 (trong đó |V| là số lượng đỉnh) nếu không có self-loop. Các giá trị có thể được chuẩn hóa bằng cách chia cho |V| - 1 để có giá trị trong khoảng [0,1].

4.2. Betweenness Centrality

Ý nghĩa: [5] Betweenness Centrality (BC) là một thước đo đánh giá tầm quan trọng của một đỉnh trong mạng dựa trên số lần nó xuất hiện trên các đường đi ngắn nhất giữa các cặp đỉnh khác trong mạng. Nói cách khác, nó đo lường mức độ mà một đỉnh nằm "giữa" các đỉnh khác, đóng vai trò là cầu nối hoặc trung gian trong luồng thông tin, tài nguyên, hoặc ảnh hưởng trong mạng.

Công thức:

$$BC(v) = \sum (\sigma(s, t|v) / \sigma(s, t))$$

v là đỉnh mà chúng ta đang tính betweenness centrality.

s và t là hai đỉnh khác nhau trong mạng, không bao gồm v

σ(s, t): Tổng số đường đi ngắn nhất giữa đỉnh s và đỉnh t.

σ(s, t|v): Số đường đi ngắn nhất giữa đỉnh s và đỉnh t mà đi qua đỉnh v.

∑ : Tổng của tất cả các cặp đỉnh s, t khác nhau.

Phạm vi: [5] Betweenness Centrality là một số thực không âm. Giá trị tối thiểu là 0 (đỉnh không nằm trên bất kỳ đường đi ngắn nhất nào giữa các cặp đỉnh khác). Giá trị tối đa phụ thuộc vào cấu trúc đồ thị, nhưng thường xảy ra ở các đỉnh nằm ở trung tâm của các mạng có cấu trúc "cây" hoặc "sao". Giá trị BC có thể được chuẩn hóa bằng cách chia cho số cặp đỉnh có thể có để có giá trị trong khoảng [0, 1], giúp so sánh giữa các đồ thị khác nhau.

4.3. Closeness Centrality

Ý nghĩa: [5] Closeness Centrality là một số thực không âm. Giá trị tối thiểu là 0 (đỉnh không thể kết nối đến tất cả các đỉnh khác hoặc đồ thị không liên thông). Giá trị tối đa thường xảy ra ở các đỉnh trung tâm có thể tiếp cận tất cả các đỉnh khác một cách nhanh chóng. CC không có ý nghĩa trong đồ thị không liên thông.

Công thức:

$$CC(v) = (|V| - 1) / \sum d(v, u)$$

trong đó d(v, u) là khoảng cách ngắn nhất giữa đỉnh v và đỉnh u, và tổng ∑ được tính trên tất cả các đỉnh u khác v. Chúng ta cũng có thể chuẩn hóa công thức thành $CC(v) = (|V| - 1) / \sum d(v, u)$ để có giá trị dễ so sánh hơn.

Phạm vi: Closeness Centrality là một số thực không âm. Giá trị tối thiểu là 0 (đỉnh không thể kết nối đến tất cả các đỉnh khác hoặc đồ thị không liên thông). Giá trị tối đa thường xảy ra ở các đỉnh trung tâm có thể tiếp cận tất cả các đỉnh khác một cách nhanh chóng. CC không có ý nghĩa trong đồ thị không liên thông.

4.4. Eigenvector Centrality

Ý nghĩa: [5] Eigenvector Centrality (EC) là một thước đo đánh giá tầm quan trọng của một đỉnh trong mạng dựa trên ý tưởng rằng một đỉnh quan trọng là đỉnh được kết nối với các đỉnh quan trọng khác. Thay vì chỉ đếm số lượng kết nối trực tiếp

(như Degree Centrality), EC xem xét chất lượng của các kết nối, tức là tầm quan trọng của các đỉnh láng giềng. Một đỉnh có EC cao không chỉ có nhiều kết nối, mà còn có nhiều kết nối với các đỉnh có tầm ảnh hưởng lớn khác.

Phạm vi: Eigenvector centrality là một số thực không âm. Giá trị tối thiểu là 0, thường xảy ra ở các đỉnh không có kết nối với các đỉnh quan trọng khác. Giá trị tối đa phụ thuộc vào cấu trúc đồ thị và không có giới hạn cụ thể. EC chỉ có ý nghĩa trong các đồ thị liên thông hoặc trong các thành phần liên thông riêng lẻ.

4.5. PageRank

Công thức:

$$PR(i) = (1-d) + d * \sum (PR(j) / L(j))$$

PR(i): PageRank của trang web *i*.

d: Dampening factor (thường là 0.85), biểu thị xác suất người dùng tiếp tục truy cập trang web thông qua một liên kết, thay vì truy cập ngẫu nhiên.

j: Các trang web mà liên kết đến trang web *i*.

PR(j): PageRank của trang web *j*.

L(j): Số lượng liên kết ra khỏi trang web *j*.

∑: Tổng của tất cả các trang web *j* liên kết đến *i*.

Phạm vi: [6] Giá trị PageRank là một số thực không âm, giá trị tối thiểu có thể là 0, và giá trị tối đa phụ thuộc vào cấu trúc của mạng liên kết. PageRank có thể bị ảnh hưởng bởi các thủ thuật SEO, chẳng hạn như "trang trại liên kết", và không phải là yếu tố duy nhất quyết định thứ hạng trong kết quả tìm kiếm.

4.6. HITS (Hub and Authority)

Ý nghĩa: [6] Thuật toán Hubs and Authorities, thường được gọi là HITS, là một phương pháp phân tích mạng liên kết để xác định hai loại trang web quan trọng: "hubs" (trung tâm) và "authorities" (thẩm quyền). Hubs là các trang web đóng vai trò như danh mục hoặc bộ sưu tập liên kết đến nhiều trang web authority về một chủ đề

cụ thể. Authorities là các trang web chứa thông tin chất lượng và được nhiều hub trỏ đến. HITS giúp chúng ta đánh giá không chỉ chất lượng nội dung mà còn vai trò của trang web trong mạng lưới liên kết.

Công thức:

$$authority(i) = \sum hub(j)$$

j là các trang web liên kết đến trang web i.

$$hub(i) = \sum authority(k)$$

k là các trang web mà trang web i liên kết đến

Phạm vi: Các giá trị hub và authority là các số thực không âm, thường được chuẩn hóa để có tổng bình phương bằng 1 hoặc một giá trị khác. Giá trị tối thiểu là 0, cho các trang không có liên kết hoặc không được trỏ đến bởi các trang khác. Phạm vi giá trị phụ thuộc vào cấu trúc của mạng liên kết.

4.7. Eccentricity

Ý nghĩa: [6] Eccentricity (độ lệch tâm) của một đỉnh trong đồ thị là khoảng cách lớn nhất giữa đỉnh đó đến bất kỳ đỉnh nào khác trong đồ thị. Nói cách khác, nó đo lường khoảng cách xa nhất mà một đỉnh phải đi để đến được một đỉnh nào đó trong đồ thị. Eccentricity thể hiện mức độ "ngoại vi" hoặc "trung tâm" của một đỉnh trong mạng.

Công thức:

$$eccentricity(v) = \max \{d(v, u)\}$$

trong đó d(v, u) là khoảng cách ngắn nhất giữa đỉnh v và đỉnh u,
và max là giá trị lớn nhất của khoảng cách

Phạm vi: Eccentricity là một số nguyên không âm. Giá trị tối thiểu là 0, thường chỉ xảy ra trong đồ thị có một đỉnh. Không có giới hạn tối đa cụ thể cho eccentricity, giá trị này phụ thuộc vào kích thước và cấu trúc của đồ thị. Eccentricity có thể không được xác định trong đồ thị không liên thông, vì có thể có những đỉnh không thể đến được từ đỉnh đang xét.

5. Gephi

Gephi là một phần mềm mã nguồn mở được thiết kế để phân tích và trực quan hóa các mạng lưới phức tạp (complex networks). [7] Đây là một công cụ mạnh mẽ giúp các nhà nghiên cứu, nhà khoa học dữ liệu và người làm truyền thông phân tích các mối quan hệ và cấu trúc của dữ liệu mạng. Với giao diện trực quan và khả năng xử lý linh hoạt, Gephi hỗ trợ người dùng khám phá các đặc điểm ẩn trong mạng lưới, từ việc phân tích cấu trúc đồ thị, xác định các cộng đồng, đến việc tính toán các chỉ số trung tâm (centrality) như PageRank, Betweenness, và Closeness.

[7] Một điểm nổi bật của Gephi là khả năng trực quan hóa dữ liệu mạng lưới theo thời gian thực, giúp người dùng dễ dàng tương tác với đồ thị. Các thuật toán sắp xếp (layout) như ForceAtlas, Fruchterman-Reingold và Yifan Hu cho phép hiển thị dữ liệu dưới dạng đồ họa sinh động, làm rõ các mối quan hệ phức tạp trong mạng lưới. Gephi hỗ trợ nhiều định dạng dữ liệu phổ biến, bao gồm CSV, GEXF, GraphML, và Pajek, giúp người dùng dễ dàng nhập và xuất dữ liệu để tích hợp với các công cụ khác.

Phần mềm này được ứng dụng rộng rãi trong nhiều lĩnh vực, từ khoa học xã hội, phân tích mạng xã hội, đến sinh học, tài chính, và nghiên cứu truyền thông. Trong khoa học xã hội, Gephi giúp phân tích các mối quan hệ giữa các cá nhân, tổ chức, hoặc các bài đăng trên mạng xã hội, từ đó xác định các yếu tố ảnh hưởng và xu hướng quan trọng. Trong sinh học, nó được sử dụng để nghiên cứu mạng gen và protein.

Với tính năng phong phú và cộng đồng người dùng đông đảo, Gephi đã trở thành một trong những công cụ hàng đầu trong lĩnh vực phân tích mạng. [7] Ngoài việc hỗ trợ các nhà nghiên cứu chuyên sâu, Gephi còn thân thiện với người mới bắt đầu, giúp họ dễ dàng khám phá và tìm hiểu về mạng lưới thông qua các biểu đồ và công cụ trực quan mạnh mẽ.

CHƯƠNG III: KẾT QUẢ VÀ THỰC NGHIỆM

3.1. Data Collection:

Trong nghiên cứu này, dữ liệu được thu thập từ chuyên mục Góc nhìn trên VnExpress. Quá trình thu thập dữ liệu sử dụng công cụ Selenium, BeautyfulSoup và ngôn ngữ lập trình Python. Web scraping là một phương pháp khai thác dữ liệu (data mining) nhằm trích xuất thông tin từ các trang web không cấu trúc và chuyển đổi chúng thành định dạng có cấu trúc để phục vụ cho phân tích.

- **URL**: Đường dẫn liên kết đến bài báo.
- **Date**: Ngày và thời gian đăng bài.
- **Category**: Chuyên mục của bài báo (ví dụ: Sức khỏe, Khoa học, Thời sự, Giáo dục).
- **Title**: Tiêu đề của bài báo, giúp nhận diện nhanh nội dung.
- **Comment Count**: Số lượng bình luận cho mỗi bài báo, biểu thị mức độ tương tác từ độc giả.

URL	DATE	CATEGORY	TITLE	COMMENT COUNT
https://vnexpress.net/buong-trung-da-nang-4428...	17/02/2022	Các bệnh	Buồng trứng đa nang	0
https://vnexpress.net/duong-nao-tu-nha-trang-d...	22/12/2024	Du lịch	Đường nào từ Nha Trang đi Đà Lạt qua đèo Ngoạn...	1
https://vnexpress.net/xet-xu-vu-ba-nguyen-thi-...	07/07/2024	Pháp luật	Xét xử vụ bà Nguyễn Thị Thanh Nhàn mua chuộc c...	3

3.2. Data Preprocessing

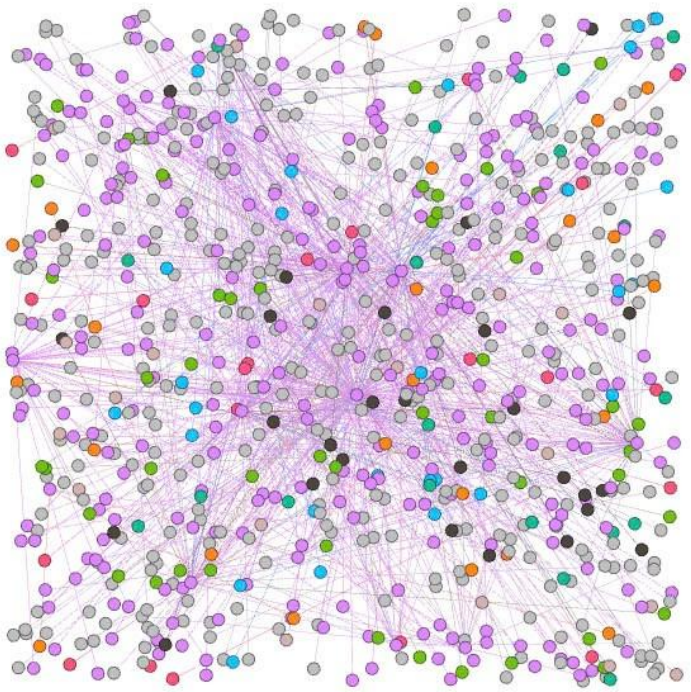
Dữ liệu mạng xã hội được thu thập gồm các bài báo với các cột: URL, Date, Category, Title, và Comment Count. Trong quá trình xử lý dữ liệu, đầu tiên tôi tiến hành kiểm tra và làm sạch dữ liệu, loại bỏ các giá trị bị thiếu hoặc không hợp lệ. Sau đó, dữ liệu được chuẩn hóa, bao gồm việc chuyển đổi các giá trị ngày tháng sang định dạng chuẩn và xử lý các dữ liệu liên quan đến thể loại (Category) để đảm bảo tính nhất quán. Các bài báo cũng được phân loại theo thể loại và tần suất bình luận (Comment Count) được sử dụng để phân tích mức độ tương tác của bài báo. Các thao tác này giúp tối ưu hóa dữ liệu, phục vụ cho việc phân tích sâu hơn, từ việc xác định xu hướng trong các thể loại bài báo đến việc đánh giá mức độ phổ biến của từng bài qua số lượng bình luận.

	URL	Date	Category	Title	Comment Count
0	https://vnexpress.net/benh-soi-tro-lai-tp-hcm-...	Thứ hai, 27/5/2024, 20:15 (GMT+7)	Sức khỏe	Bệnh sởi trở lại TP HCM sau hơn một năm vắng bóng	3
1	https://vnexpress.net/giai-cuu-gau-nuoi-nhot-o...	Thứ năm, 8/8/2024, 14:26 (GMT+7)	Khoa học	Giải cứu gấu nuôi nhốt ở Vĩnh Phúc	1
2	https://vnexpress.net/nghe-si-nhieu-nuoc-khuay...	Chủ nhật, 9/6/2024, 00:09 (GMT+7)	Thời sự	Nghệ sĩ nhiều nước khuấy động đường phố Huế	17
3	https://vnexpress.net/phu-huynh-to-truong-khon...	Thứ tư, 15/5/2024, 20:43 (GMT+7)	Giáo dục	Phụ huynh tố trường không cho con thi lớp 10	61
4	https://vnexpress.net/dai-hoc-can-tho-to-chuc-...	Thứ ba, 20/2/2024, 18:53 (GMT+7)	Giáo dục	Đại học Cần Thơ tổ chức thi đánh giá năng lực ...	NaN

Trước

	URL	Date	Category	Title	Comment Count
0	https://vnexpress.net/benh-soi-tro-lai-tp-hcm-...	27/05/2024	Sức khỏe	Bệnh sởi trở lại TP HCM sau hơn một năm vắng bóng	3
1	https://vnexpress.net/giai-cuu-gau-nuoi-nhot-o...	08/08/2024	Khoa học	Giải cứu gấu nuôi nhốt ở Vĩnh Phúc	1
2	https://vnexpress.net/nghe-si-nhieu-nuoc-khuay...	09/06/2024	Thời sự	Nghệ sĩ nhiều nước khuấy động đường phố Huế	17
3	https://vnexpress.net/phu-huynh-to-truong-khon...	15/05/2024	Giáo dục	Phụ huynh tố trường không cho con thi lớp 10	61
5	https://vnexpress.net/nhung-cong-trinh-thay-do...	29/10/2024	Thời sự	Những công trình thay đổi diện mạo đô thị Huế	8

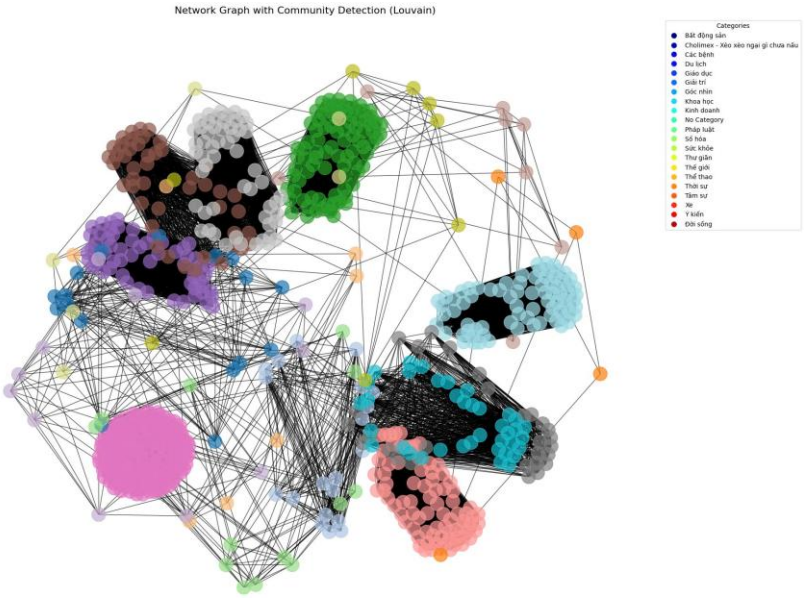
Sau



Hình 1. Mạng xã hội

Sau khi sử lý dữ liệu sẽ đưa vào Gephi để trực quan biểu đổ vơi các mối liên hệ giữa các nút và các cạnh với nhau.

3.3. Louvain

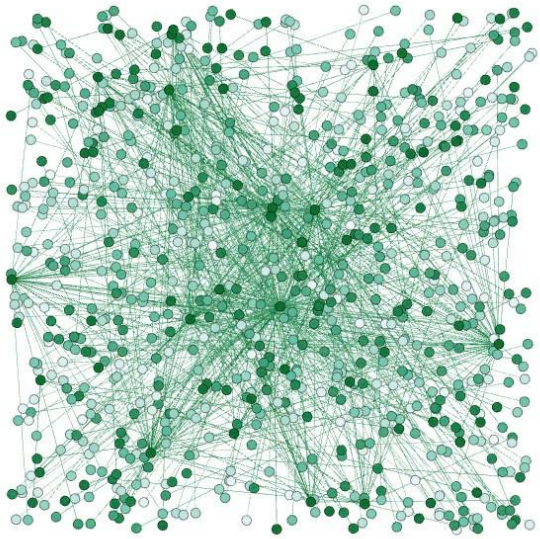


Hình 2. Biểu đồ Louvain

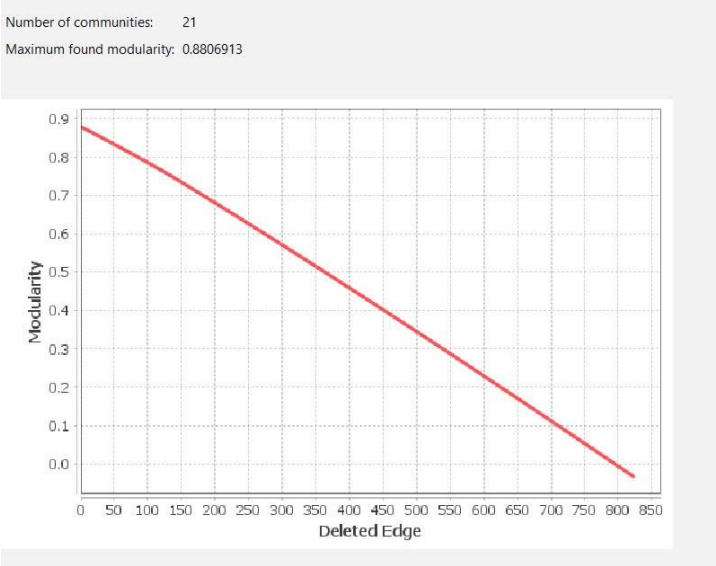
Modularity: 0.6499

Chỉ số modularity của 0.6499 là khá tốt, cho thấy rằng đồ thị đã được phân tách thành các cộng đồng tương đối mạnh mẽ. Một giá trị modularity cao (gần 1) cho thấy các nút trong cùng một cộng đồng có nhiều liên kết với nhau, trong khi liên kết giữa các cộng đồng khác ít hơn.

3.4. Girvan_Newman



Hình 3. Phân cụm Girvan_Newman



Hình 4. Kết quả phân cụm

Biểu đồ cho thấy giá trị modularity giảm dần khi số lượng cạnh bị xóa tăng lên. Điều này phản ánh rằng việc loại bỏ các cạnh có betweenness cao đã làm giảm dần sự kết nối trong mạng, từ đó chia mạng thành nhiều cộng đồng nhỏ hơn.

Ban đầu, giá trị modularity cao nhất đạt được là 0.8807, sau đó giảm dần khi nhiều cạnh bị xóa.

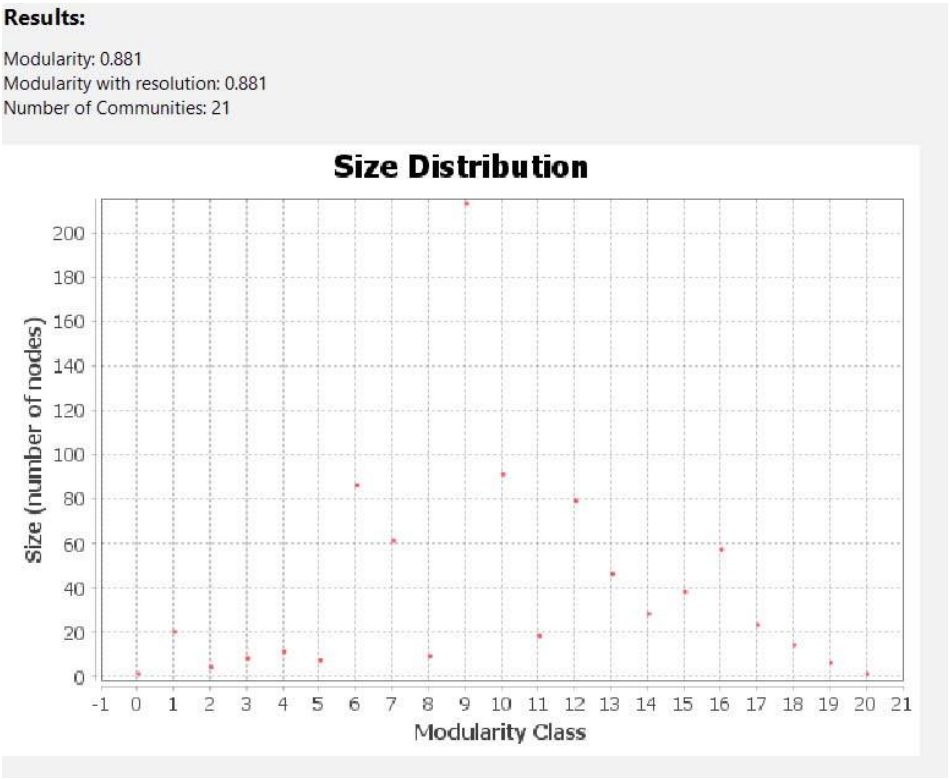
Kết thúc phân cụm:

Số lượng cộng đồng cuối cùng được tạo ra là 21 cộng đồng, tương ứng với giai đoạn khi giá trị modularity đạt mức tối ưu trước khi giảm mạnh.

Đặc điểm giảm modularity:

Ban đầu, khi chỉ một số ít cạnh bị xóa, giá trị modularity không giảm quá nhanh, chứng tỏ mạng vẫn còn duy trì được cấu trúc cụm rõ ràng.

Tuy nhiên, khi số lượng cạnh bị xóa tăng lên (sau khoảng 600 cạnh), modularity giảm mạnh hơn. Điều này có thể được giải thích là do việc loại bỏ quá nhiều cạnh đã làm phá vỡ các kết nối chính yếu, dẫn đến việc chia tách mạnh mẽ hơn các nút.



Hình 5. Kết quả phân cụm

Thuật toán này sử dụng phương pháp xóa dần các cạnh có giá trị betweenness cao nhất (những cạnh kết nối giữa các cộng đồng) để dần dần chia nhỏ mạng thành các cộng đồng. Kết quả cuối cùng sẽ xác định các cộng đồng dựa trên sự tối ưu hóa chỉ số modularity.

Kết quả từ Girvan-Newman:

Độ modularity đạt được là 0.881, cho thấy thuật toán đã phân chia mạng thành các cộng đồng một cách hiệu quả. Giá trị cao này chứng minh rằng các cạnh có betweenness cao đã được loại bỏ hợp lý, làm lộ rõ ràng cấu trúc cộng đồng.

Số lượng cộng đồng: Thuật toán đã chia mạng thành 21 cộng đồng, điều này phù hợp với tính chất của thuật toán Girvan-Newman, vốn có xu hướng phân chia mạng thành nhiều nhóm nhỏ khi tiếp tục loại bỏ các cạnh quan trọng.

Phân bố kích thước cộng đồng:

Biểu đồ phân phối kích thước cho thấy có một vài cộng đồng rất lớn (hơn 200 nút) và nhiều cộng đồng nhỏ (dưới 20 nút). Điều này phản ánh rằng trong mạng xã hội, một số nhóm có tương tác nội bộ rất mạnh (như nhóm bài viết phổ biến hoặc chủ đề chính), trong khi các nhóm nhỏ hơn có thể đại diện cho các bài viết ít phổ biến hơn hoặc chủ đề đặc thù.

Nhận xét về thuật toán:

Ưu điểm: Girvan-Newman là một thuật toán mạnh mẽ để phát hiện cấu trúc cộng đồng rõ ràng, đặc biệt trong trường hợp mạng có cấu trúc phân cụm tốt (như biểu hiện qua giá trị modularity cao trong kết quả này).

Hạn chế: Khi số lượng nút hoặc cạnh trong mạng rất lớn, thuật toán có thể trở nên tốn kém về thời gian tính toán, do việc tính toán giá trị betweenness cho mỗi cạnh là rất phức tạp.

Metric	Value
Average Degree	1.950
Graph Density	0,002

Average Degree (1.950):

Chỉ số này cho biết trung bình mỗi đỉnh (node) trong đồ thị có 1.95 cạnh (edge) kết nối với các đỉnh khác.

Giá trị thấp cho thấy rằng mạng lưới có mức độ kết nối tương đối thấp, các node thường chỉ có một vài mối liên kết trực tiếp. Điều này có thể gợi ý rằng mạng lưới có cấu trúc phân tán, không tập trung hoặc có thể có nhiều node cô lập hoặc gần như cô lập.

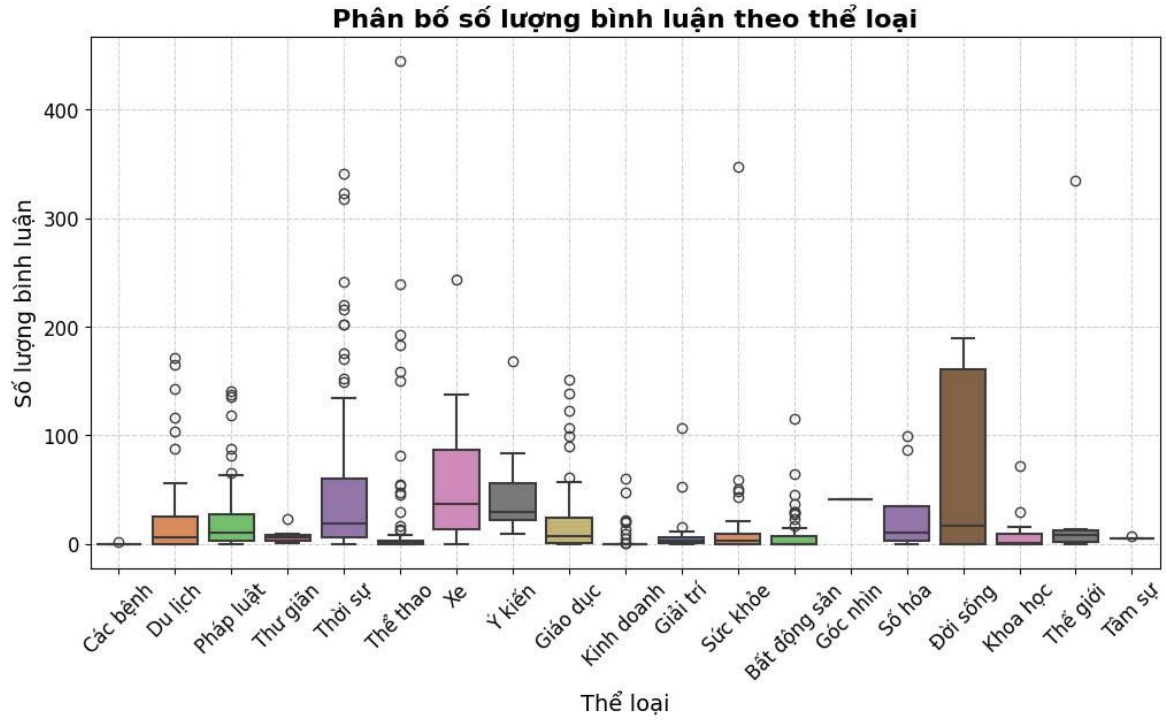
Graph Density (0.002):

Mật độ đồ thị là tỷ lệ giữa số cạnh thực tế và số cạnh tối đa có thể có trong đồ thị.

Với giá trị rất nhỏ (**0.002**), điều này chỉ ra rằng mạng lưới là đồ thị thưa (sparse graph), trong đó chỉ một phần rất nhỏ các cặp node được kết nối.

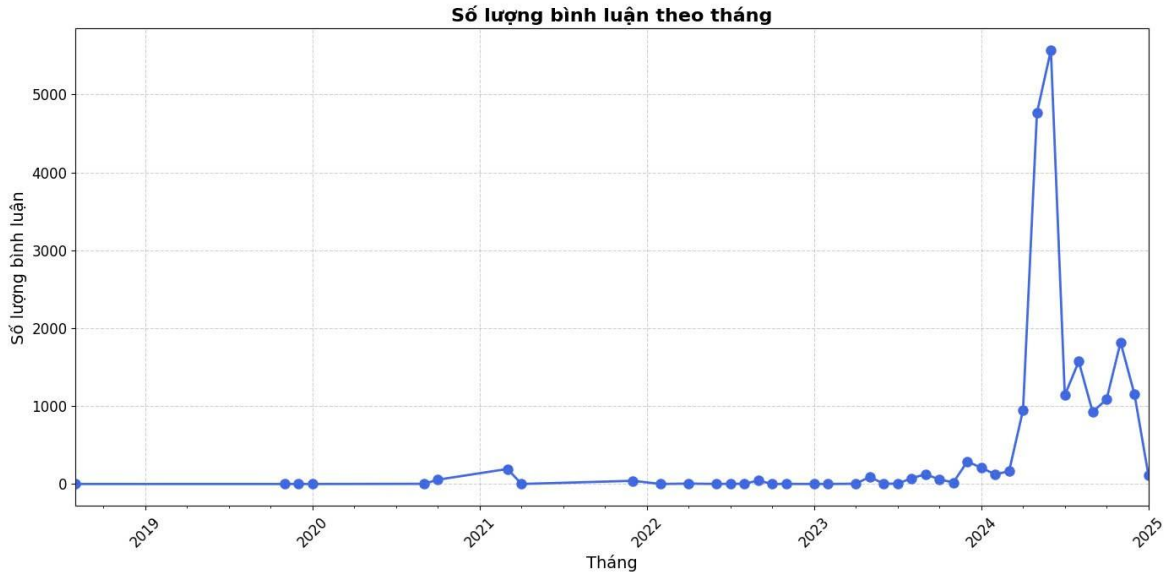
Điều này thường xảy ra trong các mạng lưới lớn, chẳng hạn như mạng xã hội hoặc mạng giao thông, nơi không phải tất cả các node đều liên kết với nhau.

3.5. Phân tích



Hình 6. Biểu đồ box plot phân bố số lượng theo thể loại

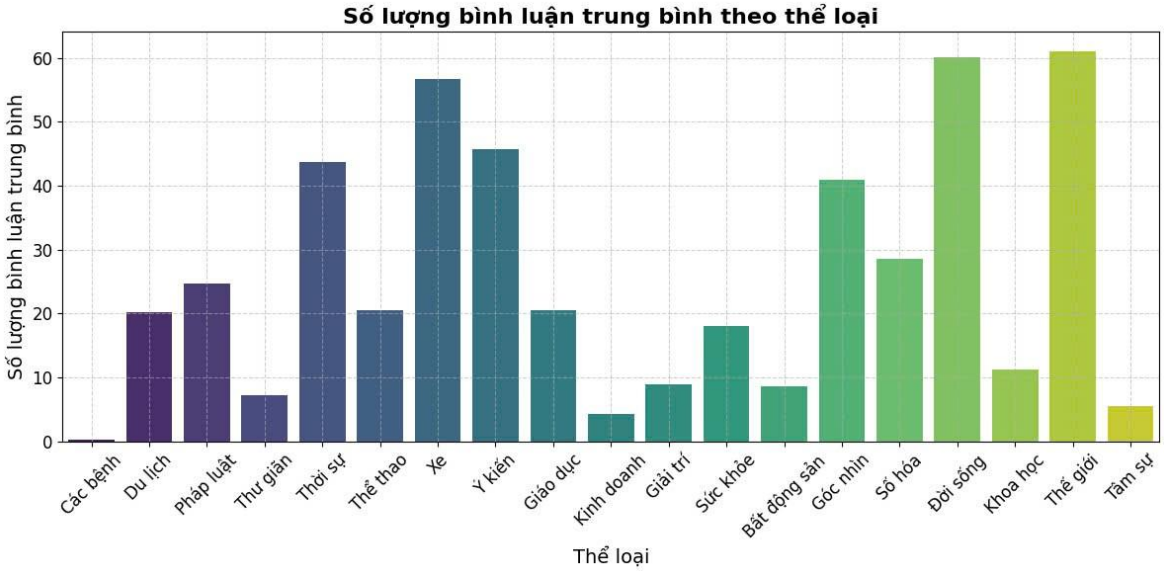
Biểu đồ box plot cho thấy phân bố số lượng bình luận cho mỗi thể loại, với các đường trung vị, tứ phân vị và các giá trị ngoại lệ được chỉ ra. Thể loại **"Đời sống"** có trung vị số lượng bình luận cao nhất, trong khi các thể loại như **"Bất động sản"** và **"Góc nhìn"** có trung vị thấp hơn. Biểu đồ này cung cấp những cái nhìn sâu sắc về những thể loại nào nhận được nhiều sự tương tác hơn về mặt bình luận.



Hình 7. Biểu đồ đường số lượng bình luận theo tháng

Biểu đồ hiển thị số lượng bình luận từ năm 2019 đến năm 2023. Trục x biểu diễn các tháng trong khoảng thời gian này, còn trục y biểu diễn số lượng bình luận, dao động từ 0 đến 5000.

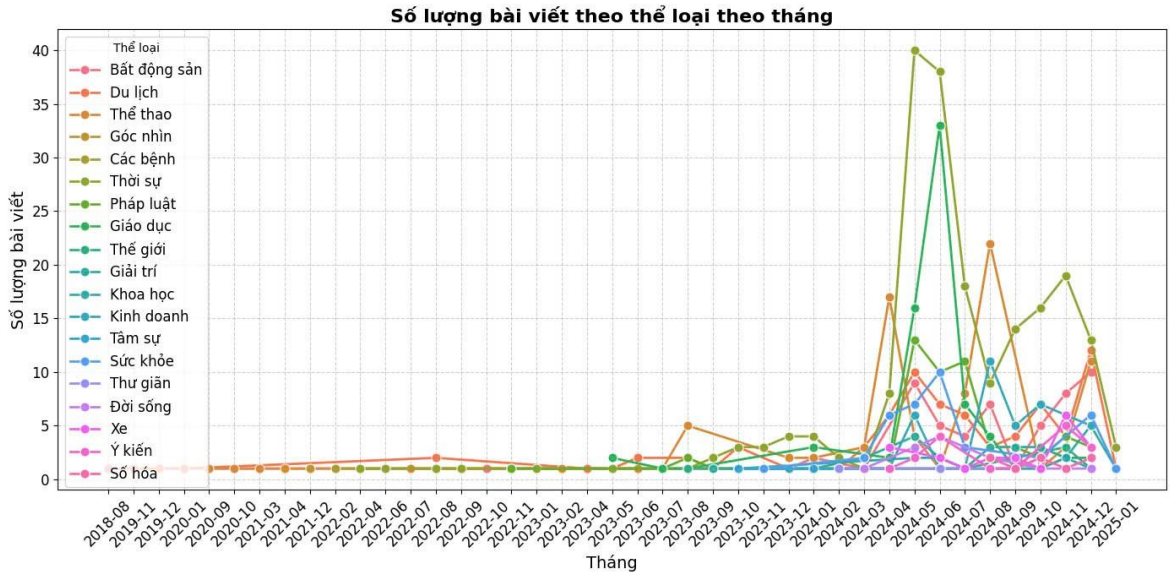
Một điểm đáng chú ý là số lượng bình luận bắt đầu tăng mạnh vào cuối năm 2023, đạt đỉnh hơn 5000 bình luận, rồi dao động và giảm mạnh về cuối năm 2023. Biểu đồ này cho thấy một xu hướng đáng kể trong số lượng bình luận trong khoảng thời gian nhất định, với sự tăng vọt hoạt động vào cuối năm 2023.



Hình 8. Biểu đồ số lượng bình luận trung bình theo thể loại

Biểu đồ cho thấy các thể loại **"Xe"** và **"Tâm sự"** có số lượng bình luận trung bình cao nhất, trong khi các thể loại như **"Pháp luật"** và **"Bất động sản"** có số lượng bình luận trung bình thấp hơn. Điều này cho thấy rằng các chủ đề về xe và tâm sự thu hút nhiều bình luận hơn so với các chủ đề khác.

Biểu đồ này cung cấp cái nhìn tổng quan về mức độ tương tác của các thể loại khác nhau thông qua bình luận, giúp nhận biết được những chủ đề nào thu hút được nhiều sự quan tâm và phản hồi từ người dùng. Bạn có thể sử dụng thông tin này để điều chỉnh nội dung hoặc chiến lược tương tác của mình.



Hình 9. Số lượng bài viết theo thể loại theo tháng

Biểu đồ hiển thị số lượng bài viết thuộc các thể loại khác nhau từ tháng 1 năm 2019 đến tháng 1 năm 2021. Trục x biểu diễn các tháng trong giai đoạn này, còn trục y biểu diễn số lượng bài viết, dao động từ 0 đến 40.

Biểu đồ cho thấy các đường màu sắc khác nhau, đại diện cho các thể loại bài viết như **"Bất động sản"** (Real Estate), **"Du lịch"** (Travel), **"Thể thao"** (Sports), và các thể loại khác. Một điểm đáng chú ý là số lượng bài viết tăng đáng kể từ khoảng tháng 3 năm 2020, đạt đỉnh vào khoảng tháng 4 năm 2020, và sau đó dao động cho đến tháng 1 năm 2021.

Xu hướng tăng mạnh vào đầu năm 2020 có thể liên quan đến các sự kiện toàn cầu hoặc khu vực trong thời gian đó, dẫn đến sự gia tăng lượng bài viết thuộc các thể loại khác nhau.

3.6. Link Prediction

```
=== Thông tin về tập dữ liệu ===
Tổng số cạnh ban đầu: 2404
Số cạnh train: 1923
Số cạnh test (cạnh sẽ xuất hiện): 481
Số cặp node test không có cạnh: 481

=== Đánh giá các phương pháp ===
           Phương pháp Accuracy Precision Recall F1-score
0     Common Neighbors    0.715     0.995  0.432    0.603
1  Jaccard Coefficient    0.784     0.979  0.580    0.728
2          Adamic/Adar    0.778     0.978  0.568    0.718
3    Cosine Similarity    0.798     0.997  0.599    0.748
```

Cosine Similarity đạt Accuracy cao nhất (0.798) và Precision gần như tuyệt đối (0.997). Điều này có nghĩa là phương pháp này dự đoán khá chính xác liệu có liên kết giữa hai bài báo hay không, và khi nó dự đoán có liên kết thì khả năng cao là đúng. Tuy nhiên, Recall của nó là 0.748, cho thấy nó bỏ sót một số liên kết thực sự tồn tại.

Jaccard Coefficient và Adamic/Adar có Accuracy tương đương (0.784 và 0.778), với Precision khá cao (0.979 và 0.978) và Recall ở mức trung bình (0.580 và 0.568).

Common Neighbors có Precision rất cao (0.995) nhưng Recall khá thấp (0.432), dẫn đến F1-score thấp hơn (0.603). Điều này cho thấy phương pháp này rất cẩn trọng trong việc dự đoán liên kết, chỉ dự đoán khi chắc chắn, nhưng bỏ sót nhiều liên kết thực sự.

Nhận xét chung:

Cosine Similarity tỏ ra hiệu quả nhất trên tập dữ liệu này, cân bằng tốt giữa Accuracy, Precision và Recall. Việc Precision gần như tuyệt đối cho thấy độ tin cậy cao của các dự đoán.

Các phương pháp dựa trên láng giềng (Common Neighbors, Jaccard, Adamic/Adar) cũng cho kết quả tốt, đặc biệt là về Precision.

Sự khác biệt về hiệu suất giữa các phương pháp cho thấy tầm quan trọng của việc lựa chọn phương pháp phù hợp với đặc điểm của dữ liệu.

CHƯƠNG 4: KẾT LUẬN

1. Tổng quan về các chỉ số

Trong quá trình phân tích dữ liệu các bài báo trên mạng xã hội, các chỉ số graph analysis đã cung cấp cái nhìn sâu sắc về cấu trúc và mức độ tương tác trong mạng lưới. Độ trung tâm (Centrality): Cho thấy các bài báo có mức độ kết nối cao, phản ánh sự chú ý và tương tác lớn từ người dùng. Clustering Coefficient: Chỉ ra các nhóm bài báo có liên quan chặt chẽ, gợi ý xu hướng nội dung hoặc các cộng đồng người dùng. Modularity: Với giá trị đạt 0.6499, cho thấy các cộng đồng trong mạng lưới được phân tách rõ ràng, phản ánh sự khác biệt về chủ đề hoặc đối tượng quan tâm. Độ trung bình của số lượng bình luận: Là 95.006, phản ánh sự phân bố không đồng đều về mức độ tương tác giữa các bài viết. Những bài báo có số lượng bình luận cao nhất đều tập trung vào các chủ đề thời sự nóng hoặc các vấn đề được công chúng quan tâm rộng rãi.

2. Kết quả từ phân tích Link Prediction

Trong phần phân tích dự đoán liên kết (Link Prediction), các thuật toán được so sánh và đánh giá dựa trên các chỉ số như Accuracy, Precision, Recall và F1-score. Cosine Similarity: Đạt Accuracy cao nhất (0.798) và Precision gần như tuyệt đối (0.997). Điều này chứng tỏ phương pháp này rất hiệu quả trong việc xác định các liên kết thực sự tồn tại. Tuy nhiên, Recall là 0.748, cho thấy vẫn còn bỏ sót một số liên kết. Jaccard Coefficient và Adamic/Adar: Có Accuracy tương đương (0.784 và 0.778), với Precision cao (0.979 và 0.978) nhưng Recall ở mức trung bình (0.580 và 0.568). Common Neighbors: Đạt Precision rất cao (0.995) nhưng Recall thấp (0.432), dẫn đến F1-score thấp hơn (0.603). Phương pháp này cẩn trọng hơn trong dự đoán nhưng bỏ sót nhiều liên kết thực sự.

Cosine Similarity tỏ ra hiệu quả nhất trên tập dữ liệu này, cân bằng tốt giữa các chỉ số Accuracy, Precision và Recall. Các phương pháp dựa trên láng giềng như Common Neighbors, Jaccard, Adamic/Adar cũng đạt hiệu quả khá, đặc biệt là về

Precision. Sự khác biệt về hiệu suất giữa các phương pháp cho thấy tầm quan trọng của việc lựa chọn thuật toán phù hợp với đặc điểm dữ liệu.

3. Ý nghĩa của đề tài

Đề tài này mang lại nhiều ý nghĩa quan trọng trong lĩnh vực phân tích mạng xã hội và truyền thông. Giúp hiểu rõ hơn về tương tác của người dùng. Giúp xác định các yếu tố ảnh hưởng đến sự chú ý và tương tác, từ đó tối ưu hóa nội dung để đáp ứng nhu cầu người dùng. Hỗ trợ xây dựng chiến lược nội dung. Cung cấp thông tin để các tổ chức truyền thông, báo chí cải thiện chiến lược xuất bản và tăng cường hiệu quả tiếp cận.

Các thông tin về xu hướng và hành vi tương tác có thể được sử dụng để thiết kế các chiến dịch quảng cáo và truyền thông phù hợp hơn. Đóng góp vào nghiên cứu khoa học: Góp phần phát triển các phương pháp và công cụ mới trong phân tích dữ liệu mạng xã hội, đặc biệt là trong lĩnh vực dự đoán liên kết và phân tích cộng đồng.

4. Hướng giải quyết

Phân tích nội dung sâu hơn: Xác định các yếu tố như tiêu đề, từ khóa, hoặc thời điểm đăng tải ảnh hưởng đến mức độ tương tác.

Tối ưu hóa chiến lược đăng tải: Dựa trên các bài báo có hiệu suất cao, điều chỉnh nội dung hoặc thời gian đăng bài để tăng lượng tiếp cận.

Khai thác thêm dữ liệu: Tích hợp dữ liệu từ các nền tảng khác hoặc phân tích phản hồi người dùng để hiểu rõ hơn về hành vi.

Ứng dụng machine learning: Dự đoán xu hướng bài viết để đề xuất nội dung tiềm năng, tăng cường hiệu quả truyền thông.

Cải thiện Recall: Với các phương pháp có Recall thấp như Common Neighbors, cần tối ưu hóa để giảm số lượng liên kết bị bỏ sót.

Kết luận, việc phân tích mạng lưới các bài báo đã mang lại nhiều thông tin giá trị, giúp nhận diện các xu hướng nội dung cũng như tối ưu hóa chiến lược truyền thông trong tương lai.

Tài Liệu Tham Khảo

[1] Guido van Rossum. “Python Programming Language and Its Development”. In: Python.org (Accessed: 2025-01-07).

[2] Shridhar C. Patekar, S. B. Rao. “On the Average Degree Eigenvalues and Average Degree Energy of Graphs”. In: *International Journal of Pure and Applied Mathematics*, vol. 118, no. 1 (2018), pp. 149-161. issn: 1311-8080. Available at: (Accessed: 2025-01-07).

[3] Trần Lanh. “Luồng cực đại”. Academia.edu. (Accessed: 2025-01-07).

[4] Alexander Strang, Oliver Haynes, Nathan D. Cahill, Darren A. Narayan. “Relationships Between Characteristic Path Length, Efficiency, Clustering Coefficients, and Graph Density”. arXiv. (Accessed: 2025-01-07).

[5] Opsahl, T., Agneessens, F., & Skvoretz, J. “Node centrality in weighted networks: Generalizing degree and shortest paths”. Social Networks, 32(3), 245-251. (Accessed: 2025-01-07).

[6] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. “Complex networks: Structure and dynamics”. Physics Reports, 424(4), 175-308. (Accessed: 2025-01-07).

[7] Mathieu Bastian, Eduardo Ramos Ibañez, Mathieu Jacomy, Cezary Bartosiak, Sébastien Heymann, Julian Bilcke, Patrick McSweeney, André Panisson, Jérémy Subtil, Helder Suzuki, Martin Skurla, Antonio Patriarca. “Gephi: An Open Source Software for Exploring and Manipulating Networks”. AAAI Publications, Third International AAAI Conference on Weblogs and Social Media. (Accessed: 2025-01-07).

HẾT