**TRƯỜNG ĐẠI HỌC PHENIKAA**

**KHOA CÔNG NGHỆ THÔNG TIN**

⸎⸎⸎⸎⸎

**Logo, company name

Description automatically generated**

**BÁO CÁO ĐỒ ÁN LIÊN NGHÀNH**

**ĐỀ TÀI : SỰ TƯƠNG QUAN GIỮA HASHTAG VÀ NỘI DUNG BÀI VIẾT TRÊN MẠNG XÃ HỘI**

**Thành viên : Nguyễn Vũ Duy Hưng, Nguyễn Đăng Đức Hải**

**Giáo viên phụ trách : Trịnh Thành**

**10 - 11 – 2023**

**MỤC LỤC**

[**I. GIỚI THIỆU ĐỀ TÀI 1**](#_Toc150947323)

[**1.1. Tổng quan về mạng xã hội Instagram 1**](#_Toc150947324)

[**1.2. Đặt vấn đề 2**](#_Toc150947325)

[**1.3. Mục tiêu và nội dung của đề tài 3**](#_Toc150947326)

[**II. ĐÁNH GIÁ TÌNH HÌNH 4**](#_Toc150947327)

[**2.1.Mạng xã hội 4**](#_Toc150947328)

[**2.2. Tác động của đại dịch Covid-19 4**](#_Toc150947329)

[**III. TẠO ĐẶC TRƯNG 5**](#_Toc150947330)

[**3.1. Giá trị tương đồng lớn nhất giữa caption với các từ phổ biến liên quan đến Covid 19. 5**](#_Toc150947331)

[**3.2. Giá trị tương đồng trung bình giữa caption với các từ phổ biến liên quan đến Covid 19 6**](#_Toc150947332)

[**3.3. Phần trăm các giá trị tương đồng của các từ trong caption với các từ phổ biến liên quan đến Covid 19 chiếm trên 20% trên tổng độ dài caption 7**](#_Toc150947333)

[**3.4. Giá trị tương đồng trung bình lớn nhất giữa caption với ba từ Coronavirus, Virus, Pandemic 8**](#_Toc150947334)

[**3.5. Giá trị tương đồng lớn nhất giữa hastags với các từ phổ biến liên quan đến Covid 19 10**](#_Toc150947335)

[**3.6. Giá trị tương đồng trung bình giữa hashtag với các từ phổ biến liên quan đến Covid 19 11**](#_Toc150947336)

[**3.7. Phần trăm các giá trị tương đồng của các từ trong hashtags với các từ phổ biến liên quan đến Covid 19 chiếm trên 20% trên tổng độ dài hashtags 12**](#_Toc150947337)

[**3.8. Giá trị tương đồng trung bình lớn nhất giữa hashtags với ba từ Coronavirus, Virus, Pandemic 13**](#_Toc150947338)

[**IV. THỰC NGHIỆM 16**](#_Toc150947339)

[**4.1.Tập dữ liệu và cách thu thập dữ liệu 16**](#_Toc150947340)

[4.1.1.Tập dữ liệu 16](#_Toc150947341)

[4.1.2.Cách thu thập dữ liệu 16](#_Toc150947342)

[**4.2. Tiền Xử lý dữ liệu 17**](#_Toc150947343)

[4.2.1.Loại bỏ những dữ liệu bị thiếu (Missing data) 17](#_Toc150947344)

[4.2.2.Loại bỏ những ký tự đặc biệt và những từ không phải là tiếng anh 18](#_Toc150947345)

[4.2.3.Loại bỏ Stop words 18](#_Toc150947346)

[4.2.4.Tách từ ghép 18](#_Toc150947347)

[**V. KẾT QUẢ 18**](#_Toc150947348)

[**5.1. Biểu đồ kết quả 18**](#_Toc150947349)

[**5.2. Nhận xét 45**](#_Toc150947350)

[**VI. TÀI LIỆU THAM KHẢO 47**](#_Toc150947351)

**DANH SÁCH HÌNH ẢNH**

*Hình 1.Logo Instagram*

*Hình 2. Cấu trúc một bài viết trên Instagram*

*Hình 3. Công thức Cosine*

*Hình 4. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Caption và từ Coronavirus*

*Hình 5. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa caption và từ Coronavirus*

*Hình 6. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa caption và từ Coronavirus*

*Hình 7 .Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Caption và từ Coronavirus lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 8. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Caption và từ Virus*

*Hình 9. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa caption và từ Virus*

*Hình 10. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa caption và từ Virus*

*Hình 11 .Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Caption và từ Virus lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 12. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Caption và từ Pandemic*

*Hình 13. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa caption và từ Pandemic*

*Hình 14. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa caption và từ Pandemic*

*Hình 15 .Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Caption và từ Pandemic lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 16. Biểu đồ Histogram thể hiện tần suất xuất hiện của các giá trị tương đồng trung bình lớn nhất giữa caption và ba từ CoronaVirus, Virus, Pandemic*

*Hình 17. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Hashtags và từ Coronavirus*

*Hình 18. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa Hashtags và từ Coronavirus*

*Hình 19. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa Hashtags và từ Coronavirus*

*Hình 20.Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Hashtags và từ Coronavirus lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 21. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Hashtags và từ Virus*

*Hình 22. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa Hashtags và từ Virus*

*Hình 23. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa Hashtags và từ Virus*

*Hình 23 .Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Hashtags và từ Virus lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 25. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Hashtags và từ Pandemic*

*Hình 26. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình giữa Hashtags và từ Pandemic*

*Hình 27. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa Hashtags và từ Pandemic*

*Hình 28 .Biểu đồ Pie thể hiện phần trăm các giá trị tương đồng của các từ trong Hashtags và từ Pandemic lớn hơn 0.2 chiếm trên 20% trên tổng độ dài của từng Caption*

*Hình 29. Biểu đồ Histogram thể hiện của các giá trị tương đồng trung bình lớn nhất giữa Hashtags và ba từ CoronaVirus, Virus, Pandemic*

# **I. GIỚI THIỆU ĐỀ TÀI**

**1.1. Tổng quan về mạng xã hội Instagram**

Instagram là một mạng xã hội chia sẻ hình ảnh và video, được phát triển bởi Kevin Systrom và Mike Krieger, và ra mắt lần đầu tiên vào tháng 10 năm 2010. Sau đó, Instagram đã nhanh chóng trở thành một trong những ứng dụng phổ biến nhất trên thế giới, thu hút hàng tỷ người dùng trên toàn cầu.



*Hình 1. Logo Instagram*

Các tính năng quan trọng của Instagram:

* Chia sẻ hình ảnh và video: Người dùng có thể chia sẻ hình ảnh và video với người theo dõi của họ hoặc công khai cho mọi người. Nền tảng này tập trung chủ yếu vào nội dung hình ảnh và trải nghiệm thị giác.
* Stories: Instagram Stories cho phép người dùng chia sẻ các bức ảnh và video tạm thời trong vòng 24 giờ. Điều này tạo ra một cách tương tác nhanh và thú vị với người theo dõi.
* IGTV (Instagram TV): IGTV là một tính năng của Instagram cho phép người dùng tải lên và xem video dài hạn, tập trung vào nội dung video chất lượng cao.
* Direct Messaging: Instagram cung cấp dịch vụ tin nhắn trực tiếp giúp người dùng gửi tin nhắn, hình ảnh và video riêng tư cho nhau.
* Explore Page: Trang Khám phá (Explore) giúp người dùng khám phá nội dung mới dựa trên sở thích và hành vi trước đó.



*Hình 2. Cấu trúc một bài viết trên Instagram*

Instagram đã trải qua nhiều cập nhật và thay đổi từ khi ra mắt, mang đến cho người dùng nhiều tính năng mới và cải thiện trải nghiệm sử dụng.

**1.2. Đặt vấn đề**

Việc dùng mạng xã hội ngày nay đã vô cùng phổ biến, tuy nhiên để đạt hiệu suất tìm kiếm cao và tăng độ tương tác đã khiến nhiều người dùng gắn nhiều thẻ (Hashtag) không liên quan trực tiếp đến chủ đề vào bài viết của mình trên instagram.

Báo cáo của chúng tôi sẽ kiểm chứng mức độ tương quan giữa các nội dung bài viết (Caption) và thẻ (Hashtag) trên mạng xã hội Instagram thông qua chủ đề Covid-19.

*\* Từ giờ chúng tôi sẽ dùng caption thay cho nội dung bài viết và hashtag cho thẻ để thuận tiện.*

**1.3. Mục tiêu và nội dung của đề tài**

Mục tiêu của đề tài: sử dụng các công cụ phân tích ngôn ngữ tự nhiên để đánh giá mức độ tương đồng giữa caption và các hashtags được gắn vào bài viết của mạng xã hội Instagram liên quan đến chủ đề đại dịch Covid-19.

Nội dung:

* Thu thập dữ liệu là các bài viết liên quan đến chủ đề Covid-19 từ năm 2020 đến năm 2023 trên Instagram
* Tiền xử lý dữ liệu: làm sạch, loại bỏ các bài viết không phải tiếng Anh,…
* Tính độ tương đồng giữa caption và hashtags để thấy độ mức độ liên quan
* Đánh giá mức độ tương quan giữa caption và hashtags của bài viết
* Kết luận và kiểm chứng mức độ tương đồng giữa caption và hashtags

# **II. ĐÁNH GIÁ TÌNH HÌNH**

## **2.1.Mạng xã hội**

Mạng xã hội là một hệ thống các mối quan hệ xã hội giữa các cá nhân, cộng đồng hoặc tổ chức. Đây là một sự kết nối giữa các thực thể thông qua các liên kết, mối quan hệ, hoặc giao tiếp, thường thông qua các nền tảng trực tuyến như Facebook, Twitter, LinkedIn, hoặc các diễn đàn trực tuyến. Mạng xã hội có thể bao gồm mối quan hệ bạn bè, đồng nghiệp, gia đình, hoặc những người có sở thích chung. Các mạng xã hội hiện nay phát triển trên nhiều nền tảng và ứng dụng khác nhau, nhưng có một số nền tảng phổ biến và quan trọng.

## **2.2. Tác động của đại dịch Covid-19**

Covid-19, một đại dịch toàn cầu do virus SARS-CoV-2 gây ra, đã có tác động sâu sắc đến mọi khía cạnh của cuộc sống xã hội. Tác động của Covid-19 có thể được tóm tắt như sau:

Thay đổi trong tương tác xã hội: Đại dịch đã thúc đẩy sự cách ly xã hội, khiến cho việc gặp gỡ và giao tiếp trực tiếp bị hạn chế. Người dân đã phải thích nghi với cuộc sống trực tuyến và mạng xã hội trở thành nền tảng chính để duy trì mối quan hệ và giao tiếp với người khác.

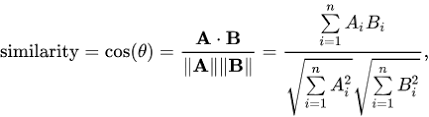
Sự lan truyền thông tin và thông điệp về Covid-19: Mạng xã hội đã đóng một vai trò quan trọng trong việc lan truyền thông tin liên quan đến đại dịch, từ thông tin y tế chính thống đến thông tin sai lệch hoặc thông tin sai trái. Việc kiểm soát thông tin và đảm bảo tính chính xác của nó trở thành một thách thức lớn.

Trong bối cảnh hiện nay, đại dịch Covid-19 đã thay đổi cách con người tương tác và hoạt động trong cuộc sống hàng ngày. Mạng xã hội, là một phần quan trọng của cuộc sống trực tuyến, đã trở thành nền tảng chính để tạo và duy trì mối quan hệ xã hội, chia sẻ thông tin, và tương tác.

# **III. TẠO ĐẶC TRƯNG**

Sử dụng mô hình GoogleNews-vectors-negative-300 là một mô hình Word2vec để Embedding các từ. Chúng tôi sử dụng thư viện Gensim [1] để tải mô hình GoogleNews-vectors-negative-300.

Sử dụng Cosine\_Similarity từ thư viện sklearn.metrics.pairwise [2] để tính toán sự tương đồng giữa hai vector. Giá trị tương đồng giữa hai vector nằm khoảng từ [-1,1], càng gần 1 thì ý nghĩa cũng như khoảng cách của chúng trong không gian vector càng gần nhau. Ngược lại càng gần -1 thì hai vector càng không tương đồng với nhau.



*Hình 3. Công thức Cosine*

## **3.1. Giá trị tương đồng lớn nhất giữa caption với các từ phổ biến liên quan đến Covid 19.**

Ví dụ: Có 1 danh sách các từ trong caption = [“virus”, “graduation”, “flu”, “happy”, “pandemic”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”.

Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Embedding\_coronavirus = Embedding[“coronavirus”]

Embedding\_virus = Embedding[“virus”]

Embedding\_pandemic = Embedding[“pandemic”]

Embedding\_virus = Embedding[“virus’”]

Embedding\_graduation = Embedding[“graduation’”]

Embedding\_flu = Embedding[“flu”]

Embedding\_happy = Embedding[“happy”]

Embedding\_pandemic = Embedding[“pandemic”]

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_graduation]= - 0.018.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_flu]= 0.469.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_happy]= 0.106.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_pandemic]= 0.373.

Ta sẽ chọn giá trị tương đồng lớn nhất đó là 0.574 của hai từ coronavirus và virus.

Tiếp tục với các từ khác như “virus”, “isolation”, “pandemic”, “vaccine”," “lockdown” và chọn ra giá trị tương đồng lớn nhất.

## **3.2. Giá trị tương đồng trung bình giữa caption với các từ phổ biến liên quan đến Covid 19**

Ví dụ: Có 1 danh sách các từ trong caption = [“virus”, “graduation”, “flu”, “happy”, “pandemic”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”.

Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_graduation]= - 0.018.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_flu]= 0.469.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_happy]= 0.106.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_pandemic]= 0.373.

Tiếp theo tôi sẽ tính giá trị tương đồng trung bình của năm cặp từ trên.

Average\_Similarity = (0.574 - 0.018 + 0.469 + 0.106 + 0.373) / 5 = 0.3.

Tiếp tục với các từ khác như “virus”, “pandemic” và chọn ra giá trị tương đồng trung bình.

## **3.3. Phần trăm các giá trị tương đồng của các từ trong caption với các từ phổ biến liên quan đến Covid 19 chiếm trên 20% trên tổng độ dài caption**

Ở đây tôi tính toán tỷ lệ phần trăm giá trị tương đồng giữa caption và các từ phổ biến và chỉ lấy những giá trị tương đồng lớn hơn mức ngưỡng mà tôi đặt ra là 0.2.

Chúng tôi sẽ giải thích tại sao chúng tôi tôi lấy ngưỡng là 0.2.

Ví dụ: Ở đây chúng tôi có các từ về 2 chủ đề khác nhau, về dịch bệnh và sự vật, sự việc trong đời sống.

Đầu tiên là chủ đề dịch bệnh gồm các cặp từ như sau:

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_epidemic]= 0.284

Cosine\_Similarity[Embedding\_headache, Embedding\_vius]= 0.291

Cosine\_Similarity[Embedding\_nose, Embedding\_flu]= 0.2344

Cosine\_Similarity[Embedding\_hospital, Embedding\_patient]= 0.6441

Cosine\_Similarity[Embedding\_vaccine, Embedding\_pen]= 0.145

Ta có thể thấy giữa hai từ “nose” và từ “flu” có giá trị tương quan là 0.23 và chúng có liên quan tới nhau, bị cúm thì có thể ảnh hưởng đến mũi, tiếp đến là từ “hospital” và từ “patient” có giá trị tương quan là 0.64 cũng liên quan nhiều đến nhau, bệnh nhân sẽ nằm ở bệnh viện. Có giá trị là 0.145 là giá trị tương quan của hai từ “vaccine” và “pen” thì rõ ràng vaccine không liên quan gì đến bút cả nên độ tương quan rất thấp

Tiếp theo là chủ đề sự việc, sự vật trong đời sống gồm các cặp từ:

Cosine\_Similarity[Embedding\_pencil, Embedding\_eraser]= 0.5699

Cosine\_Similarity[Embedding\_erase, Embedding\_note]= 0.273

Cosine\_Similarity[Embedding\_hat, Embedding\_sneaker]= 0.2653

Cosine\_Similarity[Embedding\_desk, Embedding\_monitor]= 0.2964

Cosine\_Similarity[Embedding\_plane, Embedding\_chicken]= 0.107

Ta có thể thấy giữa hai từ “hat” và từ “sneaker” có giá trị tương quan là 0.265, chúng đều là vật dụng để mang trên người và có sự tương quan đến nhau. Giữa hai từ “pencil” và “eraser” có giá trị tương quan là 0.56, bút chì thì phải đi với tẩy nên có thể thấy giá trị tương quan rất cao. Ta có cặp từ “plane” và “chicken” có giá trị tương đồng là 0.107 chúng có điểm chung là có thể bay nhưng “plane” là đồ vật còn “checken” là gia cầm nên không có sự tương quan mạnh mẽ tới nhau.

Ta dễ dàng nhận ra rằng các giá trị tương đồng từ khoảng 0.2 trở lên đã có sự tương quan tới nhau và dưới giá trị tương đồng là 0.2 thì không còn sự tương quan tới nhau nữa. Từ kết luận này nên chúng tôi đã chọn ngưỡng là 0.2.

Thay vì tính trung bình ở đây chúng tôi sẽ tính phầm trăm của các từ. Thông qua việc tính phần trăm của các từ có giá trị tương đồng lớn hơn 0.2 ta sẽ biết được chúng chiêm bao nhiêu phần trăm trên tổng số từ của caption.

Ví dụ: Có 1 danh sách các từ trong caption = [“virus”, “graduation”, “flu”, “happy”, “pandemic”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”. Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_graduation]= - 0.018.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_flu]= 0.469.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_happy]= 0.106.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_pandemic]= 0.373.

Chúng tôi lấy mốc là 0.2. Thì có ba giá trị tương đồng lớn hơn 0.3 là 0.574, 0.469, 0.373 chúng chiếm 60% và các giá trị dưới 0.2 chiếm 40%.

Tiếp tục với các từ khác như “virus”, “pandemic”.

## **3.4. Giá trị tương đồng trung bình lớn nhất giữa caption với ba từ Coronavirus, Virus, Pandemic**

Ở đây chúng tôi sẽ lấy từng từ của caption với tính giá trị tương đồng với 3 từ là “Coronavirus”, “Virus”, “Pandemic” và lấy giá trị lớn nhất. Tiếp tục như vậy với từ thứ 2 trong caption và cũng lấy giá trị lớn nhất. Sau khi tính hết tất cả các từ trong caption thì ta sẽ tính trung bình các giá trị lớn nhất đó lại.

Ví dụ: Có 1 danh sách các từ trong caption = [“virus”, “graduation”, “flu”, “happy”, “pandemic”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”. Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574

Cosine\_Similarity[Embedding\_virus, Embedding\_virus]= 1.0

Cosine\_Similarity[Embedding\_pandemic, Embedding\_virus]= 0.513

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_graduation]= -0.018

Cosine\_Similarity[Embedding\_virus, Embedding\_ graduation]= -0.04

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ graduation]= 0.071

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_flu]= 0.469

Cosine\_Similarity[Embedding\_virus, Embedding\_ flu]= 0.638

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ flu]= 0.56

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_happy]= 0.106

Cosine\_Similarity[Embedding\_virus, Embedding\_ happy]= 0.073

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ happy]= -0.008

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_pandemic]= 0.373

Cosine\_Similarity[Embedding\_virus, Embedding\_pandemic]= 0.513

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ pandemic]= 1.0

Ta sẽ lấy giá trị lớn nhất của 5 từ của caption với các từ “Coronavirus”, “Virus”, “Pandemic” và tính trung bình.

Average\_similarity = ( 1.0 + 0.071 + 0.638 + 0.106 + 1.0 ) / 5 = 0.563

Tiếp tục với các từ khác như “virus”, “pandemic”.

## **3.5. Giá trị tương đồng lớn nhất giữa hastags với các từ phổ biến liên quan đến Covid 19**

Ví dụ: Có 1 danh sách các từ trong hashtagss = [“god”, “godmother”, “virus”, “us”, “coronavirus”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”.

Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Embedding\_coronavirus = Embedding[“coronavirus”]

Embedding\_virus = Embedding[“virus”]

Embedding\_pandemic = Embedding[“pandemic”]

Embedding\_virus = Embedding[“virus’”]

Embedding\_god = Embedding[“god”]

Embedding\_godmother = Embedding[“godmother”]

Embedding\_us = Embedding[“us”]

Embedding\_coronavirus = Embedding[“coronavirus”]

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_god]= 0.045.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_godmother]= 0.099.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_us]= -0.020.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_coronavirus]= 1.000.

Ta sẽ chọn giá trị tương đồng lớn nhất đó là 1.000 của 2 từ Coronavirus.

Tiếp tục với các từ khác như “virus”, “pandemic” và chọn ra giá trị tương đồng lớn nhất.

## **3.6. Giá trị tương đồng trung bình giữa hashtag với các từ phổ biến liên quan đến Covid 19**

Ví dụ: Có 1 danh sách các từ trong hashtagss = [“god”, “godmother”, “virus”, “us”, “coronavirus”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”.

Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_god]= 0.045.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_godmother]= 0.099.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_us]= -0.020.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_coronavirus]= 1.000.

Tiếp theo tôi sẽ tính giá trị tương đồng trung bình của năm cặp từ trên.

Average\_Similarity = (0.045 + 0.099 + 0.457 - 0.020 + 1) / 5 = 0.3162.

Tiếp tục với các từ khác như “virus”, “pandemic” và chọn ra giá trị tương đồng trung bình.

## **3.7. Phần trăm các giá trị tương đồng của các từ trong hashtags với các từ phổ biến liên quan đến Covid 19 chiếm trên 20% trên tổng độ dài hashtags**

Ở đây tôi tính toán giá trị tương đồng giữa hashtags và các từ phổ biến và chỉ lấy những giá trị tương đồng lớn hơn mức ngưỡng mà tôi đặt ra là 0.2.

\*Chúng tôi đã giải thích tại sao chúng tôi chọn ngưỡng là 0.2 ở phần 3.3 nên tôi không giải thích lại.

Cũng giống như caption thay vì tính trung bình ở đây tôi sẽ tính phầm trăm của các từ. Thông qua việc tính phần trăm của các từ có giá trị tương đồng lớn hơn 0.2 ta sẽ biết được chúng chiêm bao nhiêu phần trăm trên tổng số từ của hashtag.

Ví dụ: Có 1 danh sách các từ trong hashtags = [“god”, “godmother”, “virus”, “us”, “coronavirus”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”.

Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_god]= 0.045

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_godmother]= 0.099

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_us]= -0.020

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_coronavirus]= 1.000

Đầu tiên chúng tôi lấy mốc là 0.2, thì có hai giá trị tương đồng lớn hơn 0.2 là 0.574, 1.000 chúng chiếm 40% và các giá trị dưới 0.3 chiếm 60%.

Tiếp tục với các từ khác như “virus”, “pandemic”.

## **3.8. Giá trị tương đồng trung bình lớn nhất giữa hashtags với ba từ Coronavirus, Virus, Pandemic**

Ở đây chúng tôi sẽ lấy từng từ của hashtags với tính giá trị tương đồng với 3 từ là “Coronavirus”, “Virus”, “Pandemic” và lấy giá trị lớn nhất. Tiếp tục như vậy với từ thứ 2 trong caption và cũng lấy giá trị lớn nhất. Sau khi tính hết tất cả các từ trong hashtags thì ta sẽ tính trung bình các giá trị lớn nhất đó lại.

Ví dụ: Có 1 danh sách các từ trong hashtags = [“god”, “godmother”, “virus”, “us”, “coronavirus”] và 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic”. Đầu tiên sử dụng mô hình Embedding để đưa các từ về dạng Vector mà vẫn mang đầy đủ ý nghĩa của chúng.

Các ví dụ trên giống phần trên nên tôi không viết lại.

Tiếp theo sử dụng Cosine\_Similarity để tính toán sự tương đồng giữa các Vector.

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_virus]= 0.574

Cosine\_Similarity[Embedding\_virus, Embedding\_virus]= 1.0

Cosine\_Similarity[Embedding\_pandemic, Embedding\_virus]= 0.51

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_god]= 0.045

Cosine\_Similarity[Embedding\_virus, Embedding\_ god]= 0.055

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ god]= 0.048

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_godmother]= 0.099

Cosine\_Similarity[Embedding\_virus, Embedding\_ godmother]= -0.002

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ godmother]= 0.058

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_us]= -0.02

Cosine\_Similarity[Embedding\_virus, Embedding\_ us]= 0.074

Cosine\_Similarity[Embedding\_pandemic, Embedding\_ us]= 0.103

Cosine\_Similarity[Embedding\_coronavirus, Embedding\_Coronavirus]= 1.0

Cosine\_Similarity[Embedding\_virus, Embedding\_Coronavirus]= 0.574

Cosine\_Similarity[Embedding\_pandemic, Embedding\_Coronavirus]= 0.373

Ta sẽ lấy giá trị lớn nhất của 5 từ của caption với các từ “Coronavirus”, “Virus”, “Pandemic” và tính trung bình.

Average\_similarity = ( 1.0 + 0.055 + 0.058 + 0.103 + 1.0 ) / 5 = 0.443.

Tiếp tục với các từ khác như “virus”, “pandemic”.

# **IV. THỰC NGHIỆM**

## **4.1.Tập dữ liệu và cách thu thập dữ liệu**

### **4.1.1.Tập dữ liệu**

Bộ dữ liệu gồm 9730 dòng, mỗi dòng là một bài viết có Hashtags liên quan đến Covid 19. Bộ dữ liệu gồm có Caption thể hiện nội dung bài viết và Hashtags thể hiện Hashtag của bài viết.

Bộ dữ liệu được lấy trong 4 năm, số lượng bài viết được lấy ngẫu nhiên.

+ Năm 2020 gồm 330 bài viết.

+ Năm 2021 gồm 800 bài viết.

+ Năm 2022 gồm 2100 bài viết.

+ Năm 2023 gồm 6500 bài viết.

Có thể thấy bộ dữ liệu đang bị mất cân bằng giữa các năm nên chúng tôi đã giới hạn lại số bài viết của từng năm để số lượng bài viết giữa các năm là bằng nhau.

Bộ dữ liệu mới gồm 1280 bài viết, được chia thành 4 năm, các bài viết được lấy ngẫu nhiên.

+ Năm 2020 gồm 320 bài viết.

+ Năm 2021 gồm 320 bài viết.

+ Năm 2022 gồm 320 bài viết.

+ Năm 2023 gồm 320 bài viết.

### **4.1.2.Cách thu thập dữ liệu**

Bộ dữ liệu của chúng tôi được thu thập bằng hai cách đó là sử dụng Selenium [3], Instagrapi [4] và sử dụng công cụ APIFY, APIFY là một công cụ để trích xuất dữ liệu từ Website nhanh và tiện ích hơn.

Đầu tiên chúng tôi sử dụng Selenium để tìm kiếm các bài viết với các 3 từ phổ biến liên quan đến Covid 19 là “coronavirus”, “virus”, “pandemic” và lấy tên của người dùng đăng bài và liên kết của bài viết. Vì số lượng bài viết cần lấy rất nhiều sử dụng Selenium lấy có thể gây mất thời gian nên chúng tôi chỉ lấy 2 thông tin trên.

Từ đó chúng tối sẽ sử dụng Instagrapi để lấy thêm các thông tin khác như số lượng like, comment, caption, hashtag …

Vì Selenium không phù hợp để lấy nhiều dữ liệu nên chúng tôi sử dụng phương pháp khác là công cụ APIFY và chúng tôi gộp hai bộ dữ liệu lấy theo hai phương pháp trên thành một bộ duy nhất.

## **4.2. Tiền Xử lý dữ liệu**

### **4.2.1.Loại bỏ những dữ liệu bị thiếu (Missing data)**

Ký tự bị thiếu là những giá trị hiển thị NaN, Null, N/A hoặc rỗng …

Missing data xuất hiện khi:

+ Dữ liệu bị mất trong quá trình chuyển đổi dữ liệu.

+ Lỗi chương trình.

+ Người dùng quên điền.

…

Dữ liệu bị thiếu có thể gây nhiều vấn đề trong quá trình phân tích dữ liệu.

+ Có thể gây chênh lệch kết quả phân tích.

+ Ảnh hướng đến độ chính xác của kết quả.

+ Không thể chạy được các chương trình.

…

Ở đây chúng tôi sử dụng thư viện Pandas [5] để loại bỏ những dữ liệu bị thiếu.

### **4.2.2.Loại bỏ những ký tự đặc biệt và những từ không phải là tiếng anh**

Trong văn bản sẽ tồn tại nhiều ký tự đặc biệt như hình trái tim, mặt cười, dấu câu “ ? , ! … và những từ không phải tiếng anh, ở đây chúng tôi sử dụng tiếng anh để phân tích nên chúng tôi sẽ loại bỏ những ngôn ngữ khác. Loại bỏ những điều trên sẽ giảm kích thước của bộ dữ liệu và tăng tốc quá trình xử lý.

Có một thư viện để sử lý những vấn đề trên và cũng là thư viện chúng tôi đã sử dụng đó là thư viện Natural Language Toolkit (NLTK) [6].

### **4.2.3.Loại bỏ Stop words**

Stop Words thường xuất hiện nhiều lần và không đóng góp nhiều vào ý nghĩa câu, chúng có thể gây nhiễu, trong tiếng anh các từ này có thể kể đến như “The”, “is”, “at”, “on”, “what” …Các từ này sẽ được loại bỏ để giảm kích thước của bộ từ vựng.

Thư viện Natural Language Toolkit (NLTK) cũng cung cấp một danh sách các Stop Words trong tiếng anh và chúng tôi đã sử dụng chúng để loại bỏ khỏi bộ dữ liệu của chúng tôi.

### **4.2.4.Tách từ ghép**

Trong bộ dữ liệu của chúng tôi có nhiều từ ghép ví dụ như “gradstudent”, “copperhaircolor”, “bodybuilding”… Chúng tôi đã sử dụng thư viện Splitter [7] để tách các từ ghép thành từ đơn như “gradstudent” thành “grad” và “student”.

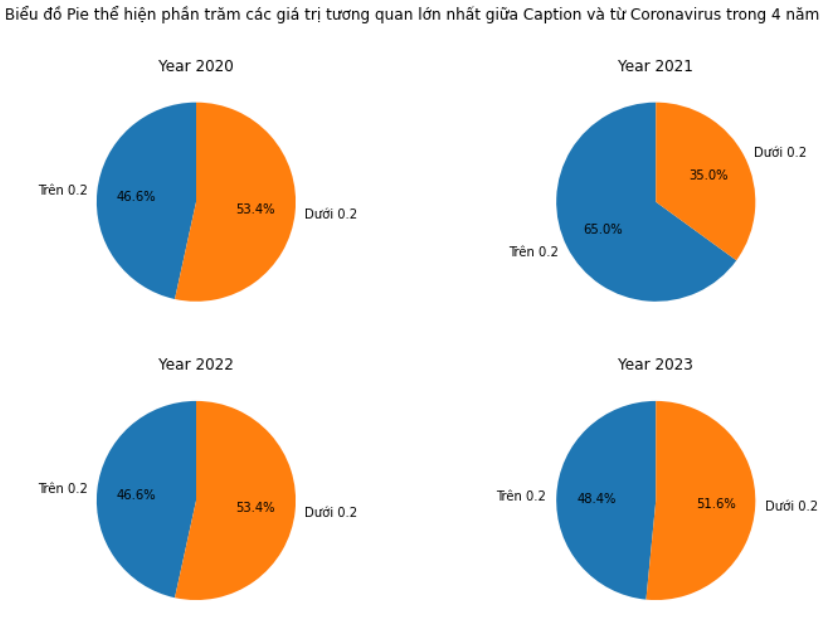
# **V. KẾT QUẢ**

## 5.1. Biểu đồ kết quả

A group of colored bars

Description automatically generated with medium confidence

*Hình 4. Biểu đồ Histogram thể hiện các giá trị tương đồng lớn nhất giữa Caption và từ Coronavirus trong 4 năm*

**

*Hình 5. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan lớn nhất giữa Caption và từ Coronavirus trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence.*

*Hình 6. Biểu đồ Histogram thể hiện các giá trị tương quan trung bình giữa Caption và từ Coronavirus trong 4 năm*

*A diagram of a number of circles with numbers and a blue triangle

Description automatically generated with medium confidence*

*Hình 6. Biểu đồ Pie thể hiện phần trăm của các giá trị tương quan trung bình giữa caption và từ Coronavirus lớn hơn 0.2 trong 4 năm*

*A group of pie charts

Description automatically generated*

*Hình 7 .Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Caption và từ Coronavirus lớn hơn 0.2 chiếm trên 20% trên tổng số từ của Caption trong 4 năm*

*A group of colored bars

Description automatically generated with medium confidence*

*Hình . Biểu đồ Histogram các giá trị tương đồng lớn nhất giữa Caption và từ Virus trong 4 năm*

*A diagram of a number of pies

Description automatically generated with medium confidence*

*Hình . Biểu đồ Pie thể hiện phần trăm các giá trị tương quan lớn nhất giữa Caption và từ Virus trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence*

*Hình 9. Biểu đồ Histogram thể hiện các giá trị tương đồng trung bình giữa caption và từ Virus trong 4 năm*

*A diagram of a number of circles with numbers and a blue circle

Description automatically generated with medium confidence*

*Hình 10. Biểu đồ Pie thể hiện phần trăm của các giá trị tương đồng trung bình giữa caption và từ Virus lớn hơn 0.2 trong 4 năm*

*A screenshot of a graph

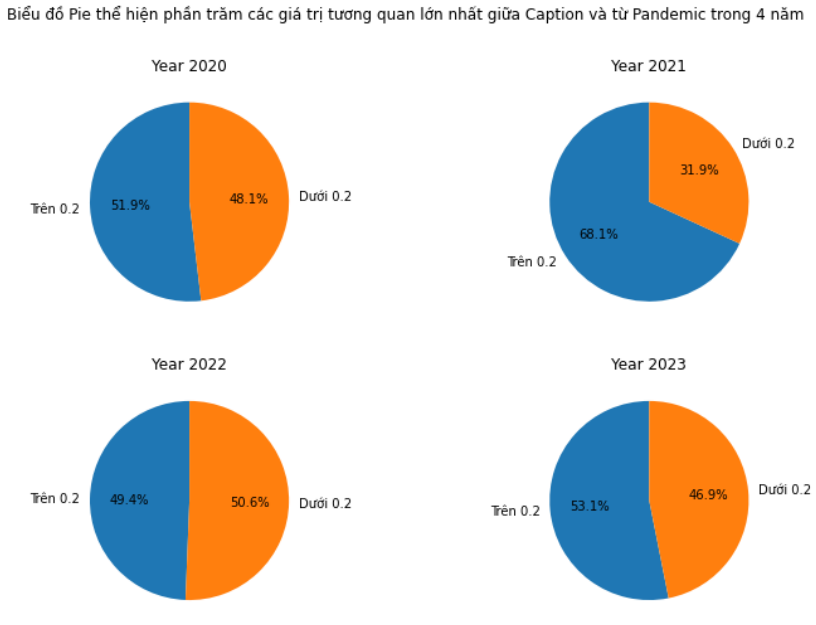
Description automatically generated*

*Hình 11. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Caption và từ Virus lớn hơn 0.2 chiếm trên 20% trên tổng số từ của Caption trong 4 năm*

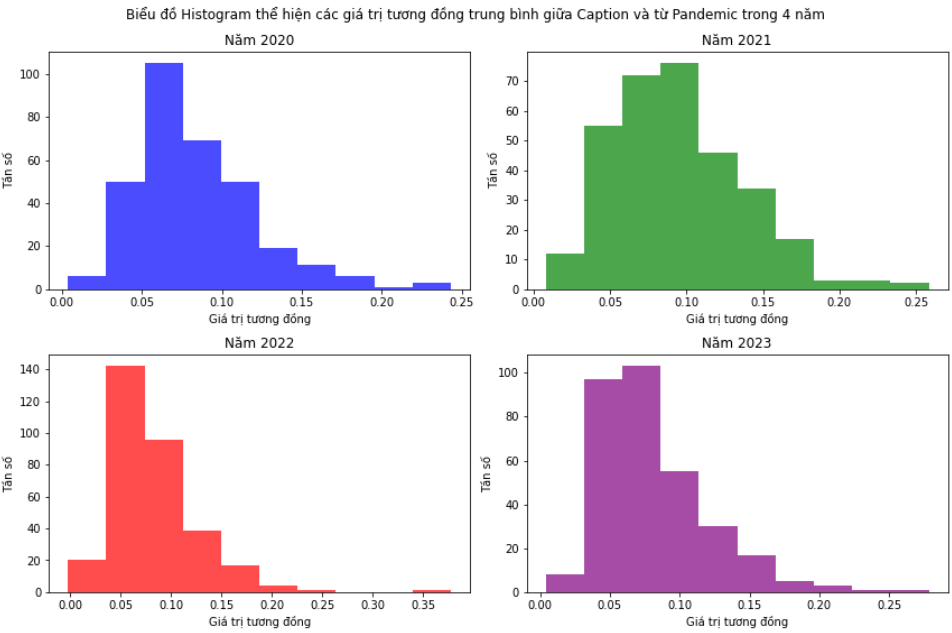
*A group of colored bars

Description automatically generated with medium confidence*

*Hình 12. Biểu đồ Histogram thể hiện các giá trị tương quan lớn nhất giữa Caption và từ Pandemic trong 4 năm*

**

*Hình . Biểu đồ Pie thể hiện phần trăm các giá trị tương quan lớn nhất giữa Caption và từ Pandemic trong 4 năm*

**

*Hình 13. Biểu đồ Histogram thể các giá trị tương quan trung bình giữa Caption và từ Pandemic trong 4 năm*

*A diagram of a number of circles

Description automatically generated with medium confidence*

*Hình 14. Biểu đồ Pie thể hiện phần trăm của các giá trị tương quan trung bình giữa caption và từ Pandemic lớn hơn 0.2 trong 4 năm*

*A diagram of a number of pies

Description automatically generated with medium confidence*

*Hình 15. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Caption và từ Pandemic lớn hơn 0.2 chiếm trên 20% trên tổng số từ của Caption trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence*

*Hình 16. Biểu đồ Histogram thể hiện các giá trị tương quan trung bình lớn nhất giữa Caption và ba từ CoronaVirus, Virus, Pandemic trong 4 năm*

*A screenshot of a graph

Description automatically generated*

*Hình 17.* *Biểu đồ Histogram thể hiện tần suất xuất hiện các giá trị tương quan lớn nhất giữa Hashtag và từ Coronavirus trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence*

*Hình 18. Biểu đồ Histogram thể hiện của các giá trị tương quan trung bình giữa Hashtag và từ Coronavirus trong 4 năm*

*A diagram of a number of pies

Description automatically generated with medium confidence*

*Hình 19. Biểu đồ Pie thể hiện phần trăm của các giá trị tương quan trung bình giữa Hashtag và từ Coronavirus trong 4 năm*

*A blue and orange pie chart

Description automatically generated*

*Hình 20. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Hashtag và từ Coronavirus lớn hơn 0.2 chiếm trên 20% trên tổng số từ Hashtags trong 4 năm*

*A group of colorful bars

Description automatically generated with medium confidence*

*Hình 21. Biểu đồ Histogram thể hiện các giá trị tương quan lớn nhất giữa Hashtags và từ Virus trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence*

*Hình 22. Biểu đồ Histogram thể hiện các giá trị tương quan trung bình giữa hashtags và từ Virus trong 4 năm*

*A diagram of a number of pies

Description automatically generated with medium confidence*

*Hình 23. Biểu đồ Pie thể hiện phần trăm của các giá trị tương quan trung bình giữa hashtags và từ Virus trong 4 năm*

*A blue and orange pie chart

Description automatically generated*

*Hình 24. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Hashtags và từ Virus lớn hơn 0.2 chiếm trên 20% trên tổng số từ của Hashtag*

*A group of colored bars

Description automatically generated with medium confidence*

*Hình 25 .Biểu đồ Histogram thể hiện các giá trị tương quan lớn nhất giữa Hashtag và từ Pandemic trong 4 năm*

*A group of colored graphs

Description automatically generated with medium confidence*

*Hình 26. Biểu đồ Histogram thể hiện của các giá trị tương quan trung bình giữa hashtags và từ Pandemic trong 4 năm*

*A diagram of a pie chart

Description automatically generated*

*Hình 27. Biểu đồ Pie thể hiện phần trăm của các giá trị tương quan trung bình giữa Hashtag và từ Pandemic trong 4 năm*

*A blue and orange pie chart

Description automatically generated2*

*Hình 28. Biểu đồ Pie thể hiện phần trăm các giá trị tương quan của các từ trong Hashtags và từ Pandemic lớn hơn 0.2 chiếm trên 20% trên tổng số từ của Hashtag*

*A group of colored bars

Description automatically generated with medium confidence*

*Hình 29. Biểu đồ Histogram thể hiện các giá trị tương quan trung bình lớn nhất giữa caption và ba từ CoronaVirus, Virus, Pandemic trong 4 năm*

# 

## **5.2. Nhận xét**

* *Nhận xét*

Từ các biểu đồ và bảng kết quả thì ta có thể dễ dàng nhận ra rằng giá trị tương đồng giữa Caption với các từ liên quan đến Covid 19 thấp hơn rất nhiều so với giá trị tương đồng giữa Hashtag với các từ liên quan đến Covid 19.

Có thể thấy giá trị tương đồng trung bình giữa Caption và các từ liên quan đến Covid 19 xuất hiện nhiều nhất ở khoảng từ 0.05 đến 0.1. Giá trị tương đồng giữa Hashtag và các từ liên quan đến Covid 19 xuất hiện nhiều nhất ở khoảng từ 0.05 đến 0.4, lớn hơn rất nhiều lần so với Caption.

Tiếp đến là phần trăm các giá trị tương đồng trung bình trên 0.2. Với Caption thì chiếm khoảng từ 1% đến 18%, chúng chiếm khá thấp. Còn Hashtag tăng lên cao nhất là từ 10% đến 56% trên tổng các giá trị tương đồng trung bình của Hashtags.

Về phần trăm các giá trị tương đồng của các từ trong Caption và từ Coronavirus lớn hơn 0.2 chiếm trên 20% tổng số từ của Caption chiếm từ 8% đến 18%. Còn đối với Hashtag thì chiếm từ 72% đến 99%, hầu như chiếm gần hết.

Cuối cùng ta xem xét đến các giá trị tương đồng trung bình lớn nhất giữa Caption và ba từ “Coronavirus”, “Virus”, “Pandemic” xuất hiện nhiều nhất từ 0.1 đến 0.25, đối với Hashtag thì xuất hiện nhiều nhất từ 0.2 đến 0.5.

Từ những kết quả và chứng minh trên thì ta có thể khẳng định rằng caption và hastags hầu như là không liên quan nhiều đến nhau. Có khả năng là người dùng Instagram đang sử dụng hashtags để tiếp cận đối tượng rộng hơn, ngay cả khi hashtag đó không liên quan trực tiếp đến nội dung của caption.

# **VI. TÀI LIỆU THAM KHẢO**

[1] Scikit-Learn. Cosine Similarity, 2022.

[2] Radim Rehurek. Gensim Library, Provides tools for creating language models, analyzing text, and processing natural language, 2022.

[3] Selenium, Selenium WebDriver, Provides a suite of tools for automating web browser interaction, 2023.

[4] Instagram API.

[5] McKinney. Pandas Library, Provides data structures and analysis tools designed to make working with "relational" or "labeled" data both easy and intuitive, 2023.

[6] Natural Language Toolkit (NLTK), Provides a comprehensive introduction to natural language processing with Python, 2023.