|  |
| --- |
| **ĐẠI HỌC ĐÀ NẴNG**  **TRƯỜNG ĐẠI HỌC CNTT VÀ TT VIỆT – HÀN**  **KHOA KHOA HỌC MÁY TÍNH**    Trường Đại học Công nghệ Thông tin và Truyền thông Việt Hàn - Đại học Đà  Nẵng  **BÁO CÁO**  **ĐỒ ÁN CHUYÊN NGÀNH 2**  **ĐỀ TÀI: PHÂN TÍCH TÌNH CẢM THEO KHÍA CẠNH CẢM TRONG VĂN BẢN**  **( chủ đề Covid19 )**  Sinh viên thực hiện: Đinh Đức Anh – 20IT553  Phan Thành Nhân - 20IT739  Võ Văn Việt - 20IT527    Giảng viên hướng dẫn: TS.NGUYỄN ĐỨC HIỂN  ***Đà Nẵng, tháng 5 năm 2023*** |
| **ĐẠI HỌC ĐÀ NẴNG**  **TRƯỜNG ĐẠI HỌC CNTT VÀ TT VIỆT – HÀN**  **KHOA KHOA HỌC MÁY TÍNH**    Trường Đại học Công nghệ Thông tin và Truyền thông Việt Hàn - Đại học Đà  Nẵng  **BÁO CÁO**  **ĐỒ ÁN CHUYÊN NGÀNH 2**  **ĐỀ TÀI: PHÂN TÍCH TÌNH CẢM THEO KHÍA CẠNH CẢM TRONG VĂN BẢN**  **( chủ đề Covid19 )**  Sinh viên thực hiện: Đinh Đức Anh – 20IT553  Phan Thành Nhân - 20IT739  Võ Văn Việt - 20IT527  Giảng viên hướng dẫn: **TS.NGUYỄN ĐỨC HIỂN**  ***Đà Nẵng, tháng 10 năm 2023*** |

# LỜI CẢM ƠN

Trong suốt thời gian làm đồ án, nhóm chúng em nhận được rất nhiều sự giúp đỡ. Chúng em xin chân thành cảm ơn Khoa khoa học máy tính - Trường ĐH CNTT & TT Việt Hàn - ĐHĐN đã tạo điều kiện thuận lợi cho em thực hiện đề tài đồ án chuyên ngành này. Xin cảm ơn quý Thầy Cô bộ môn của Khoa khoa học máy tính - Trường ĐH CNTT & TT Việt Hàn - ĐHĐN đã tận tình giảng dạy, trang bị cho chúng em những kiến thức quý báu trong năm học vừa qua.

Chúng em xin chân thành tỏ lòng biết ơn sâu sắc đến giảng viên TS. NGUYỄN ĐỨC HIỂN đã tận tình hướng dẫn em trong suốt quá trình thực hiện đồ án chuyên ngành lần này.

Cảm ơn các bạn trong lớp, trong trường, trong Khoa khoa học máy tính - Trường ĐH CNTT & TT Việt Hàn - ĐHĐN đã hỗ trợ, giúp đỡ, chia sẻ kinh nghiệm và kiến thức cho nhóm trong quá trình thực hiện đề tài.

Trong quá trình làm đồ án, do kiến thức về lập trình chưa thực sự sâu sắc nên đồ án của chúng em không thể tránh khỏi những sai sót. Mong thầy cô thông cảm và góp ý thêm cho chúng em để đồ án được hoàn chỉnh hơn.

Chúng em xin chân thành cảm ơn!

**LỜI NÓI ĐẦU**

Cùng với sự phát triển không ngừng của các ngành khoa học kỹ thuật, các ngành công nghiệp cũng phát triển nhanh chóng. Việc áp dụng những công nghệ hiện đại vào nhiều lĩnh vực trong cuộc sống đó là điều tất yếu. Trong đó cảm xúc là một phần quan trọng trong trải nghiệm con người, đóng vai trò quan trọng trong việc tạo ra một loạt phản ứng và hành vi.

Sự ra đời của công nghệ thông tin là sự tích hợp đồng thời các tiến bộ về công nghệ và tổ chức thông tin, đem đến nhiều ảnh hưởng tích cực cho sự phát triển của xã hội. Công nghệ thông tin và đặc biệt là sự phát triển của internet mở ra một kho kiến thức vô cùng đa dạng và phong phú cho mọi người, giúp cho việc tìm hiểu kiến thức đơn giản hơn rất nhiều. Hiện nay việc áp dụng các thành tựu khoa học kĩ thuật và trí tuệ nhân tạo vào phân tích cảm xúc của con người, việc áp dụng công nghệ vào phân tích cảm xúc nó mang đến nhiều thành tựu đáng. Qua cái nhìn tổng thể nhóm chúng em quyết định chọn đề tài “ PHÂN TÍCH TÌNH CẢM THEO KHÍA CẠNH CẢM TRONG VĂN BẢN” .Ứng dụng sẽ giúp người dùng tính toán được phần trăm mà người dùng có thể mắc các bệnh nghiêm trọng về da. Có thể ứng dụng sẽ không được hoàn thiện 100% nhưng nó sẽ là bước ngoặt lớn trong sự phát triển ý tưởng của nhóm chúng em.Phân tích cảm xúc của con người qua một đoạn văn bản có thể thực hiện bằng cách xem xét các yếu tố như từ ngữ, ngữ cảnh, và các phần tử biểu thị cảm xúc.Phân tích cảm xúc trong văn bản giúp chúng ta hiểu thêm về trạng thái tâm lý và trải nghiệm của con người trong tình huống cụ thể và tạo ra sự kết nối giữa người viết và độc giả.

Bằng sự cố gắng nỗ lực và đặc biệt là sự giúp đỡ tận tình, chu đáo của giảng viên hướng dẫn TS.NGUYỄN ĐỨC HIỂN, nhóm chúng em đã hoàn thành đồ án môn học đúng thời hạn. Do thời gian làm đồ án có hạn và trình độ còn nhiều hạn chế nên không thể tránh khỏi những thiếu sót. Em rất mong nhận được sự đóng góp ý kiến của các thầy cô cũng như là của các bạn sinh viên để bài đồ án này hoàn thiện hơn nữa.

# NHẬN XÉT

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

*Đà nẵng, ngày … tháng … năm 2023*

Chữ ký của giảng viên hướng dẫn

Mục lục

[**LỜI CẢM ƠN 3**](#_Toc154630780)

[**NHẬN XÉT 5**](#_Toc154630781)

[**Chương 1: Tổng quan 7**](#_Toc154630782)

[**1. Lý do chọn đề tài 7**](#_Toc154630783)

[**2. Đối tượng và phạm vi nghiên cứu 7**](#_Toc154630784)

[**3.Kết quả nghiên cứu: 8**](#_Toc154630785)

[**Chương 2: Khái niệm lý thuyết 9**](#_Toc154630786)

[**1. Phân tích tình cảm trong văn bản là gì? 9**](#_Toc154630787)

[**2. Các loại phân tích tình cảm chính 9**](#_Toc154630788)

[**3. Giới thiệu về các thuật toán tìm kiếm ngữ nghĩa (CSS) theo ngữ cảnh 10**](#_Toc154630789)

[**Chương 3: Phân tích và thiết kế 11**](#_Toc154630790)

[**1. Dataset 11**](#_Toc154630791)

[**2. Xây dựng model 12**](#_Toc154630792)

[**3. Kết quả và mục tiêu phát triển 23**](#_Toc154630793)

[**4. Xây dựng ứng dụng Web app để phân tích tình cảm trong văn bản 24**](#_Toc154630794)

[**4.1 Tổng quan về Web app 25**](#_Toc154630795)

[**4.2 Chức năng phân tích đoạn text 26**](#_Toc154630796)

[**4.3 Chức năng tông quan về dataset 28**](#_Toc154630797)

[**4.4 Chức năng phân tích theo CSV 29**](#_Toc154630798)

[**4.5 Chức năng trực quan hóa dữ liệu 31**](#_Toc154630799)

[**4.6 Chức năng google dịch 31**](#_Toc154630800)

[**Tài liệu tham khảo 32**](#_Toc154630801)

**Danh mục ảnh**

[Hình 1: Tiền xử lý dữ liệu 15](#_Toc154635047)

[Hình 2: Model 1 15](#_Toc154635048)

[Hình 3: Kết quả 15](#_Toc154635049)

[Hình 4: Model 2 16](#_Toc154635050)

[Hình 5: Kết quả 16](#_Toc154635051)

[Hình 6: Đồ thị 17](#_Toc154635052)

[Hình 7: Model 3 17](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635053)

[Hình 8: Kết quả 17](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635054)

[Hình 9: Model 4 18](#_Toc154635055)

[Hình 10: Kết quả 18](#_Toc154635056)

[Hình 11: Lưu model 19](#_Toc154635057)

[Hình 12: Xác định trạng thái dự đoán 19](#_Toc154635058)

[Hình 13: Dataset 19](#_Toc154635059)

[Hình 14: Dataset 20](#_Toc154635060)

[Hình 15: Import dữ liệu 20](#_Toc154635061)

[Hình 16: Tiền xử lý dữ liệu 20](#_Toc154635062)

[Hình 17: Model 21](#_Toc154635063)

[Hình 18: Kết quả 21](#_Toc154635064)

[Hình 19: Thông tin về Model 21](#_Toc154635065)

[Hình 20: Đồ thị 22](#_Toc154635066)

[Hình 21: Lưu model 22](#_Toc154635067)

[Hình 22: Kiểm tra dự đoán 22](#_Toc154635068)

[Hình 23: Sơ lược về Data 23](#_Toc154635069)

[Hình 24: Thiết kế model 23](#_Toc154635070)

[Hình 25: Thiết kế model 24](#_Toc154635071)

[Hình 26: Thông tin model 24](#_Toc154635072)

[Hình 27: Kết quả 25](#_Toc154635073)

[Hình 28: Đồ thị 25](#_Toc154635074)

[Hình 29: Input & test 26](#_Toc154635075)

[Hình 30: Streamlit 27](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635076)

[Hình 31: Pycharm 27](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635077)

[Hình 32: Jupyter 28](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635078)

[Hình 33: Lưu model 28](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635079)

[Hình 34: Web app 29](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635080)

[Hình 35: Phân tích đoạn text 29](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635081)

[Hình 36: Label 30](#_Toc154635082)

[Hình 37: Điểm Label 30](#_Toc154635083)

[Hình 38:Điểm từ khóa 30](#_Toc154635084)

[Hình 39: Phân loại cảm xúc 31](#_Toc154635085)

[Hình 40: Tổng quan Dataset 31](#_Toc154635086)

[Hình 41: Tổng quan Dataset 32](#_Toc154635087)

[Hình 42: Input Dataset 32](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635088)

[Hình 43: Input Dataset 32](#_Toc154635089)

[Hình 44: Chọn cột chứa văn bản cần phân tích 33](#_Toc154635090)

[Hình 45: Chọn ngôn ngữ 33](#_Toc154635091)

[Hình 46: Kết quả 33](#_Toc154635092)

[Hình 47: Trực quan hóa 34](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635093)

[Hình 48: Kiểm tra kết quả dịch 34](file:///D:\New%20folder%20(2)\BaoCao.docx#_Toc154635094)

**Chương 1: Tổng quan**

### Lý do chọn đề tài

Cảm xúc là một phần quan trọng trong trải nghiệm con người, đóng vai trò quan trọng trong việc tạo ra một loạt phản ứng và hành vi. Chúng ta trải qua một loạt cảm xúc hàng ngày, từ niềm vui, sự hạnh phúc và yêu thương đến sự buồn bã, sợ hãi và lo lắng. Cảm xúc không chỉ ảnh hưởng đến trạng thái tâm lý cá nhân mà còn tác động đến hành vi, quan điểm và quyết định của chúng ta. Cảm xúc của con người cũng được ảnh hưởng bởi các yếu tố nội tại như suy nghĩ, quan điểm và giá trị cá nhân. Những suy nghĩ và quan điểm của chúng ta về một tình huống cụ thể có thể tạo ra phản ứng cảm xúc khác nhau. Đơn giản như trong một đoạn văn bản cũng có thể chứa đựng muôn vàn cảm xúc của con người được biểu hiện qua từ ngữ, ngữ cảnh và các phân tử biểu hiện cảm xúc.

Vậy nên để hiểu rõ hơn về tâm lý con người để ứng dụng trong lĩnh vực tình cảm và tâm lý, ứng dụng trong khoa học dữ liệu và trí tuệ nhân tạo giúp cải thiện chất lượng cuộc sống và nghiên cứu phát triển trong lĩnh vực mới đó là lý do nhóm chúng em thực hiện đề tài này.

### Đối tượng và phạm vi nghiên cứu

#### 2.1. Đối tượng nghiên cứu

Trong phân tích cảm xúc trong đoạn văn bản, phạm vi đối tượng nghiên cứu tập trung vào các yếu tố cảm xúc được diễn đạt trong văn bản. Đối tượng nghiên cứu trong trường hợp này là các đoạn văn bản hoặc câu văn được viết bởi con người. Phân tích cảm xúc trong đoạn văn bản có thể áp dụng cho các ngữ cảnh và loại văn bản khác nhau, bao gồm nhưng không giới hạn:

* Bình luận và đánh giá sản phẩm: Phân tích cảm xúc trong đoạn văn bản có thể tập trung vào việc đánh giá sự hài lòng, không hài lòng, hoặc sự phản hồi cảm xúc khác của người viết đối với sản phẩm, dịch vụ hoặc trải nghiệm.
* Bài viết và tin tức: Phân tích cảm xúc trong đoạn văn bản có thể áp dụng để xác định và đo lường các yếu tố cảm xúc như tin tức vui mừng, lo lắng, phẫn nộ, hoặc sự đồng cảm trong các bài viết, tin tức, bài phê bình, và các nguồn thông tin khác.
* Mạng xã hội và bình luận trực tuyến: Phân tích cảm xúc trong đoạn văn bản có thể áp dụng để phân tích và đánh giá các bình luận, trạng thái, bài đăng trên mạng xã hội và các diễn đàn trực tuyến. Điều này giúp hiểu rõ hơn về ý kiến và cảm xúc của người dùng với các vấn đề, sự kiện, hoặc chủ đề cụ thể.
* Tương tác chatbot và email: Phân tích cảm xúc trong đoạn văn bản có thể áp dụng để đánh giá và cải thiện tương tác giữa người dùng và chatbot hoặc qua email. Điều này giúp xác định tâm trạng và nhu cầu của người dùng để cung cấp phản hồi và hỗ trợ phù hợp hơn.

#### 2.2. Phạm vi nghiên cứu:

Phạm vi nghiên cứu của đối tượng trong trường hợp phân tích cảm xúc trong đoạn văn bản bao gồm:

* Phân loại cảm xúc: Nghiên cứu có thể tập trung vào việc phân loại cảm xúc trong đoạn văn bản thành các loại như vui mừng, buồn bã, tức giận, sợ hãi, bất ngờ, hoặc đồng cảm. Mục tiêu là xác định và đánh giá cảm xúc mà tác giả hoặc người viết muốn truyền tải thông qua văn bản.
* Nhận dạng tông giọng: Nghiên cứu có thể tập trung vào việc nhận dạng tông giọng cảm xúc trong đoạn văn bản, tức là xác định sự biểu đạt cảm xúc qua ngôn ngữ và cú pháp. Ví dụ, phân tích có thể tập trung vào việc xác định sự biểu đạt của sự hài lòng, không hài lòng, hoặc phẫn nộ trong văn bản.
* Phân tích ngữ cảnh và biểu hiện: Nghiên cứu có thể tập trung vào việc phân tích ngữ cảnh và các biểu hiện cảm xúc trong đoạn văn bản. Điều này bao gồm việc xác định các từ khóa, cấu trúc câu, ngữ pháp, và các yếu tố khác để hiểu ngữ cảnh và cảm xúc được diễn đạt trong văn bản.
* Đánh giá cảm xúc: Nghiên cứu có thể tập trung vào việc đánh giá mức độ cảm xúc trong đoạn văn bản. Điều này có thể bao gồm việc xác định mức độ cảm xúc tích cực hay tiêu cực trong văn bản, hoặc xác định mức độ mạnh hay nhẹ của cảm xúc được diễn đạt.

Tóm lại, phạm vi nghiên cứu của đối tượng trong phân tích cảm xúc trong đoạn văn bản bao gồm việc phân loại cảm xúc, nhận dạng tông giọng, phân tích ngữ cảnh và biểu hiện, cùng việc đánh giá mức độ cảm xúc trong văn bản.

### 3.Kết quả nghiên cứu:

Kết quả của đề tài phân tích cảm xúc trong đoạn văn bản có thể là khả năng phân loại chính xác cảm xúc, hiểu ngữ cảnh và biểu hiện cảm xúc, đánh giá và phân tích tương tác xã hội, cũng như cải thiện tương tác và hỗ trợ dựa trên cảm xúc của người dùng.

# Chương 2: Khái niệm lý thuyết

### Phân tích tình cảm trong văn bản là gì?

Phân tích tình cảm là một ứng dụng của trí tuệ nhân tạo, nó sử dụng các thuật toán rất phức tạp để xử lý ngôn ngữ tự nhiên của con người và xác định những đặc điểm cảm xúc tiêu cực/tích cực tại một thời điểm thông qua văn bản hoặc lời nói. Các nguồn dữ liệu được phân tích khá phổ biến như Social media, Blog, Website đánh giá sản phẩm, tổng đài Contact center,… Hiện nay, nhờ sự tiến bộ của các công nghệ thuộc thế hệ mới, các thuật toán phân tích quan điểm ngày càng được nâng cấp với độ chính xác ngày càng cao, từ đó hỗ trợ trong các sản phẩm thông minh như trợ lý ảo tích hợp trên xe hơi, căn hộ… cải thiện chất lượng cuộc sống của con người. Đây chính là câu trả lời cho **Sentiment Analysis** là gì.

* Lợi ích của Kỹ thuật phân tích quan điểm

Việc nhận biết được cảm xúc của người dùng trong cuộc trò chuyện giúp doanh nghiệp phân tích được mức độ quan tâm của khách hàng đối với thương hiệu, sản phẩm/ dịch vụ của mình. Đây là nguồn thông tin có giá trị cao dành cho doanh nghiệp để điều chỉnh chiến lược về sản phẩm của họ, kinh doanh, marketing và đặc biệt là các dịch vụ tư vấn phù hợp. Thông qua cảm xúc của khách hàng thể hiện trong các cuộc trò chuyện bằng tin nhắn hoặc cuộc gọi điện, doanh nghiệp có thể dễ dàng xác định được thái độ và sự hài lòng của khách hàng đối với hoạt động tư vấn. Từ đó, việc đánh giá chất lượng nhân viên trở nên dễ dàng hơn và quản lý hiệu quả hơn.

### Các loại phân tích tình cảm chính

#### 2.1. Phân tích chi tiết

Mô hình [phân tích tình cảm](https://en.wikipedia.org/wiki/Sentiment_analysis) này giúp xác định độ chính xác của các đặc điểm. Các đặc điểm chính được phân loại là rất tích cực, tích cực, trung tính, tiêu cực hoặc rất tiêu cực. Phân tích chi tiết như thế này rất tốt để đánh giá các cuộc trò chuyện.

Bạn có thể coi đó là thang đánh giá từ 1 đến 5, với 1 là rất tiêu cực và 5 là rất tích cực. Trên thang điểm từ 1 đến 10, bạn có thể cho điểm từ 1 đến 2 là rất tiêu cực và từ 9 đến 10 là rất tích cực.

#### 2.2. Cơ sở khía cạnh:

Phân tích chuyên sâu xác định khuynh hướng cảm xúc tổng thể của khách hàng trong cuộc trò chuyện, trong khi phân tích dựa trên khía cạnh đi sâu hơn và xác định các khía cạnh cụ thể của ngôn ngữ.

Ví dụ, khi một khách hàng nói, “Ánh sáng nhân tạo làm cho máy ảnh trở nên kỳ lạ.” Phân tích dựa trên khía cạnh không chỉ có thể xác định đây là tâm lý tiêu cực mà còn xác định rằng người dùng đã đưa ra nhận xét tiêu cực về đối tượng “máy ảnh”.

#### 2.3 Nhận dạng cảm xúc

Cảm xúc ở đây bao gồm các sắc thái giận dữ, buồn bã, hạnh phúc, thất vọng, sợ hãi, lo lắng và hoảng loạn. Một số trình phân loại nâng cao sử dụng thuật toán máy học (ML) mạnh mẽ.

#### 2.4 Phân tích ý định

Xác định chính xác ý định của người tiêu dùng giúp doanh nghiệp tiết kiệm thời gian, tiền bạc và công sức. Bởi vì các doanh nghiệp có thể ít quan tâm đến những khách hàng tiềm năng chưa có ý định mua thứ gì đó. Phân tích ý định chính xác có thể giải quyết nhiều vấn đề cho các công ty như vậy.

Phân tích tâm lý thị trường giúp doanh nghiệp xác định ý định của người tiêu dùng – liệu khách hàng có ý định mua hay chỉ đang duyệt qua. Bạn có thể nhắm mục tiêu. Nếu người tiêu dùng chưa sẵn sàng mua, bạn có thể tiết kiệm thời gian và nguồn lực bằng cách không nhắm mục tiêu vào họ.

### 3. Giới thiệu về các thuật toán tìm kiếm ngữ nghĩa (CSS) theo ngữ cảnh

Điều quan trọng nhất trong việc thu thập thông tin hữu ích từ khách hàng của bạn là hiểu các khía cạnh của thương hiệu mà khách hàng của bạn đang nói đến. Ví dụ: Amazon muốn tách các thông báo cho những thứ như giao hàng trễ, vấn đề thanh toán, yêu cầu khuyến mại và đánh giá sản phẩm. Mặt khác, Starbucks muốn phân loại các thông điệp dựa trên việc liệu chúng có liên quan đến hành vi của nhân viên, hương vị cà phê mới, phản hồi về vệ sinh, đặt hàng trực tuyến, tên và địa điểm cửa hàng,… Nhưng làm thế nào một máy tính có thể phân loại những thứ này?

Giải pháp cho vấn đề này là Xác định Ngữ nghĩa Ngữ cảnh (CSS), một thuật toán tìm kiếm thông minh. CSS hoạt động bằng cách lấy hàng nghìn thông báo và khái niệm làm đầu vào và lọc ra tất cả các thông báo có liên quan chặt chẽ đến một khái niệm cụ thể. Ví dụ, trong khái niệm về giá, thuật toán thực hiện tìm kiếm từ khóa về giá và các từ có liên quan chặt chẽ khác như (giá, phí, đô la, thanh toán).

Tuy nhiên, hạn chế của phương pháp này là không thể tìm thấy tất cả các từ khóa có liên quan và các biến thể của chúng. Kỹ thuật phân tích phối cảnh được áp dụng vào Giải pháp trung gian trí tuệ nhân tạo – Tổng đài AI. Tổng đài AI có thể tạo cuộc gọi với phản hồi nhanh, chính xác và hội thoại tự nhiên nhờ ứng dụng công nghệ xử lý giọng nói (TTS) và xử lý ngôn ngữ tự nhiên (NLU). Phản hồi nhanh hơn với nội dung được cá nhân hóa cho từng khách hàng, khả năng xử lý lưu lượng cuộc gọi lớn, hỗ trợ 24/7 và mở rộng không giới hạn, tối ưu hóa thông qua tự động hóa một số hoạt động của trung tâm cuộc gọi như điều hướng và kết nối khách hàng đồng thời cải thiện trải nghiệm người dùng với hiệu suất và chi phí vận hành được tối ưu hóa.

# Chương 3: Phân tích và thiết kế

## **1. Dataset**

* Bộ dữ liệu Twitter\_samples của thư viện NLTK :

Bộ dữ liệu Twitter\_samples của thư viện NLTK là một tập dữ liệu gồm các tweet được thu thập từ Twitter. Bộ dữ liệu này chứa hơn 1 triệu tweet, được phân loại theo cảm xúc tích cực, tiêu cực, hoặc trung lập.

Bộ dữ liệu Twitter\_samples được sử dụng để đào tạo và đánh giá các mô hình phân tích cảm xúc trong văn bản. Bộ dữ liệu này có nhiều ưu điểm, chẳng hạn như:

Kích thước lớn: Bộ dữ liệu này chứa hơn 1 triệu tweet, cung cấp một lượng dữ liệu lớn để đào tạo các mô hình học máy.

Diversity: Bộ dữ liệu này bao gồm các tweet từ nhiều nguồn khác nhau, đại diện cho nhiều loại cảm xúc.

Availability: Bộ dữ liệu này có sẵn miễn phí và có thể dễ dàng tải xuống.

Dưới đây là một số thông tin chi tiết về bộ dữ liệu Twitter\_samples:

Kích thước: 1,085,947 tweet

Ngôn ngữ: Tiếng Anh

Cảm xúc: Tích cực, tiêu cực, trung lập

Thư mục: nltk\_data/twitter\_samples

Link data: https://realpython.com/python-nltk-sentiment-analysis/

* Dataset tự tạo:

Với data chúng em thu thập từ các review tiếng việt về các quán ăn ở trên internet. Với mục đích chính là thử xây dựng một model ngôn ngữ riêng cho tiếng việt và biết thêm kiếm thức về thu thập dữ liệu và đánh nhãn cho dữ dạng văn bản.

Linkdata:

https://drive.google.com/drive/folders/1oUy6AEY7UgBAjueP3y-\_6YKmgJoH26xJ?usp=sharing

* Dataset từ một bài báo nghiên cứu khoa học về lĩnh vực phân tích tình cảm

Ở bài báo này nhóm tác giả đã sử dụng mô hình Bert để phân tích và cho kết quả rất tốt. Chúng em lấy dataset này về với mục đích là sử dụng dataset chuẩn và tự xây dựng riêng một model khác để tăng thêm kiến thức về môn học này thay vì sử dụng các model đã xây dựng sẵn

Bộ dữ liệu thu tập từ các dòng tweet về COVID-19 trên nên tảng mạng xã hội Twitter. Dùng API Scrapper Twitter được sử dụng để trích xuất dữ liệu có hashtag (#Covid2019 HOẶC#Covid19 HOẶC corona&virus) từ ngày 20 tháng 1 năm 2020 đến ngày 25 tháng 4 năm 2020

2020

Link data:

[GitHub - savan77/EmotionDetectionBERT: Multi Emotion Detection from COVID-19 Text using BERT](https://github.com/savan77/EmotionDetectionBERT)

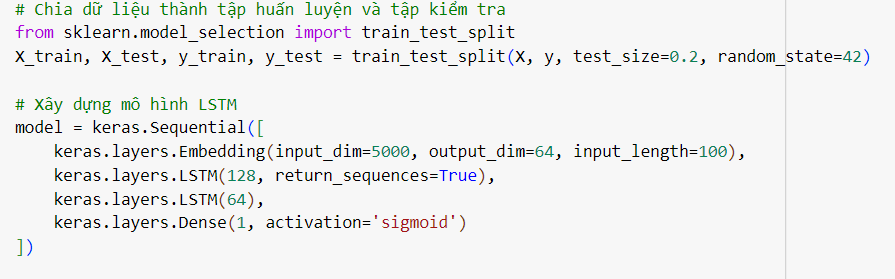
## **2.** **Xây dựng model**

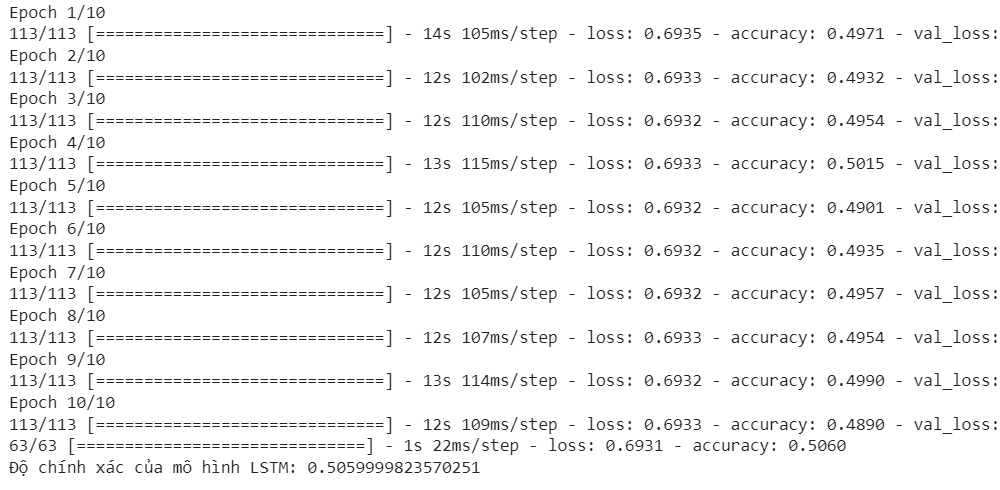
Model 1:



Hình 1: Tiền xử lý dữ liệu

Ở đây chúng em tạo một một thư viện gồm 5000 token và nếu số từ vượt quá 5000 thì sẽ đánh giấu là 00V. Độ dài của một token là 100, nếu câu ngắn hơn 100 thì phía sau sẽ được tự động thêm các sô 0 và nếu độ dài chuỗi vượt qua 100 thì sẽ bị cắt về đúng 100 từ.



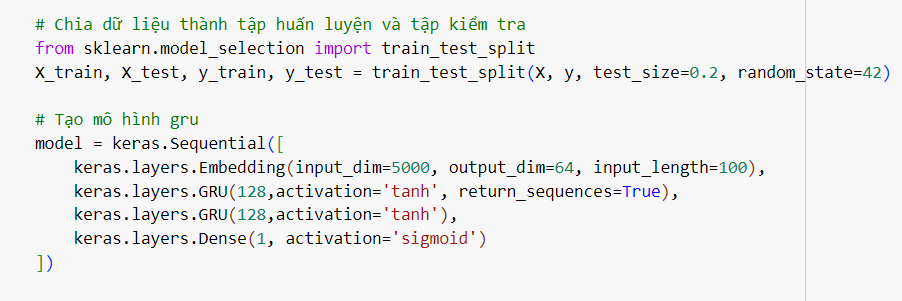
Hình 2: Model 1

Hình 3: Kết quả

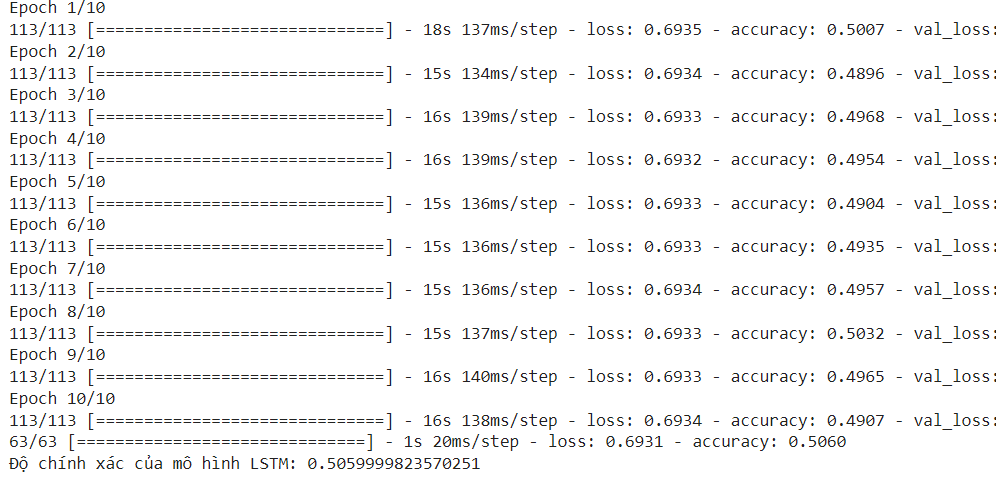
Vì dataset là một data chuẩn nên nếu kết quả mô hình có độ chính xác không cao thì khả năng cao là cho chúng em thiết kế chưa phù hợp. Vì vậy chúng em tiếp tục xây dựng lại một model khác để kiểm tra.

Model 2:

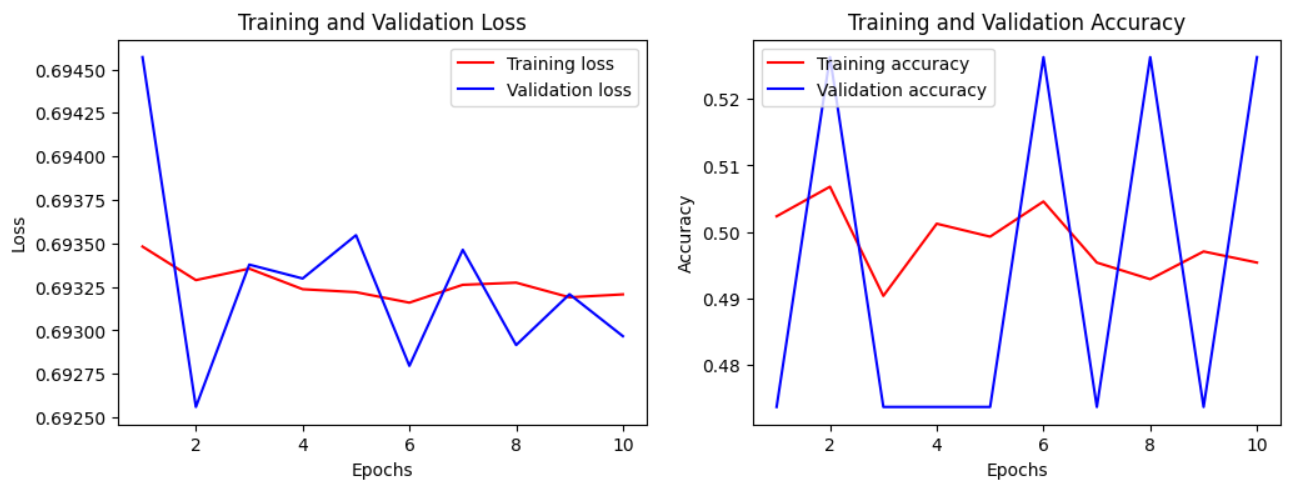
Ở model 2 này chúng em đã thay LSTM bằng GRU để kiểm tra bên cạnh đó cũng tinh chỉnh các siêu tham sô để lọc ra kết quả tốt nhất. Tuy nhiên sau khi thử nhiều lần thì vẫn cho kết quả độ chính xác vẫn rất thấp



Hình 4: Model 2



Hình 5: Kết quả



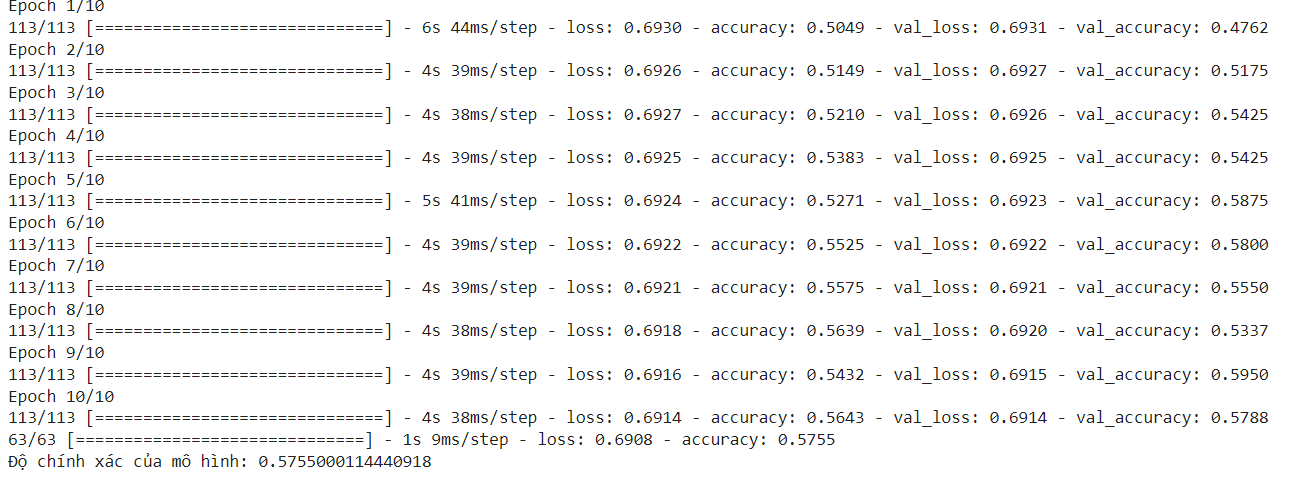
Hình 6: Đồ thị

* Từ đồ thị có thể thấy giá trị thực tế và giá trị dự đoán hoàn toàn không trùng nhau và có thể kết luận hai model trên hoàn toàn không phù hơp.

Model 3:

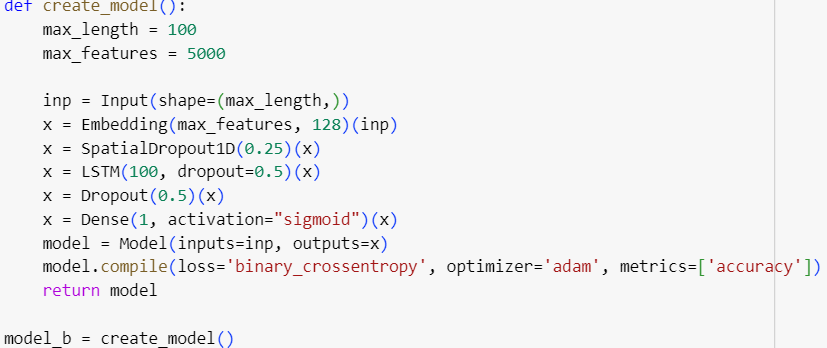
* Ở model này chúng em vẫn sử dụng dataset cũng và thiết kết lại cấu trúc model, thêm nhiều lớp và có thêm cả Dropout, thì cho kế quả khá tốt khoảng 57%

Hình 7: Model 3



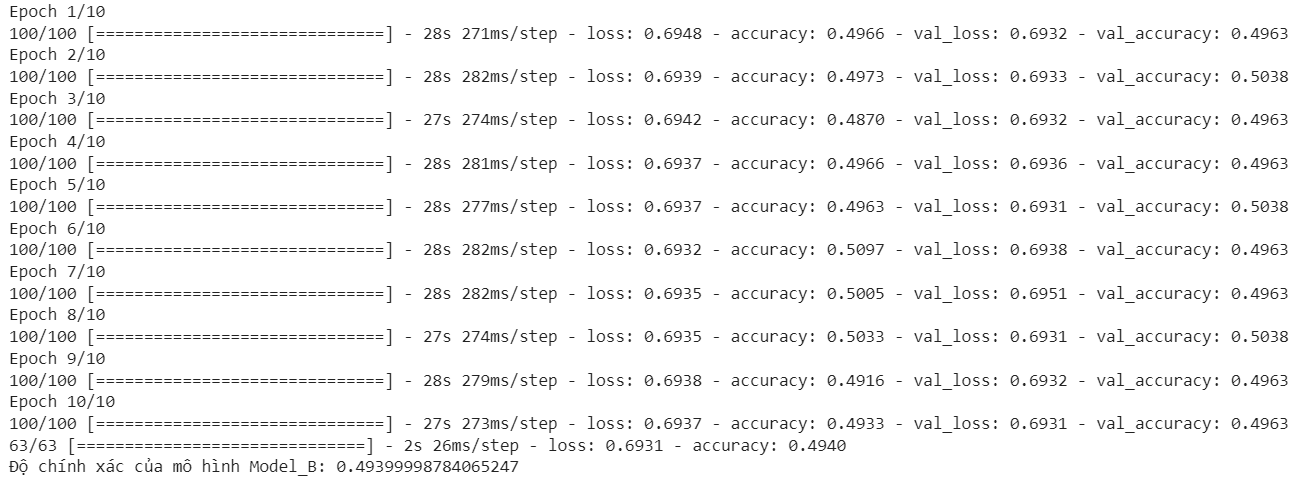
Hình 8: Kết quả

Model 4:



Hình 9: Model 4

Kết quả



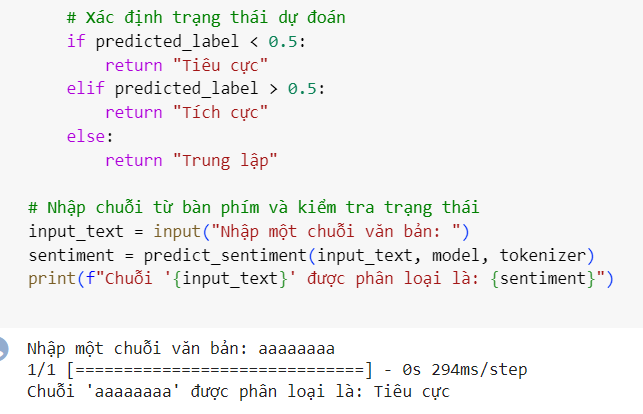
Hình 10: Kết quả

=> Sau khi thiết kết thì chỉ có được một model có độ chính xác cao nhất là hơn 57% tuy nhiên dựa vào đồ thị có thể thấy model này không phù hợp với dataset.

Lưu model và test một chuỗi bất kỳ nhập vào từ bàn phím

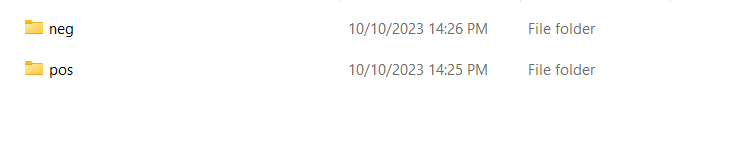


Hình 11: Lưu model



Hình 12: Xác định trạng thái dự đoán

Sau khi thử thiết kế nhiều model khác nhau nhưng không cho ra kết quả như ý. Chúng em đã chuyển hướng thay vì thiết kế model thì sẽ tự thiết kế một dataset riêng.

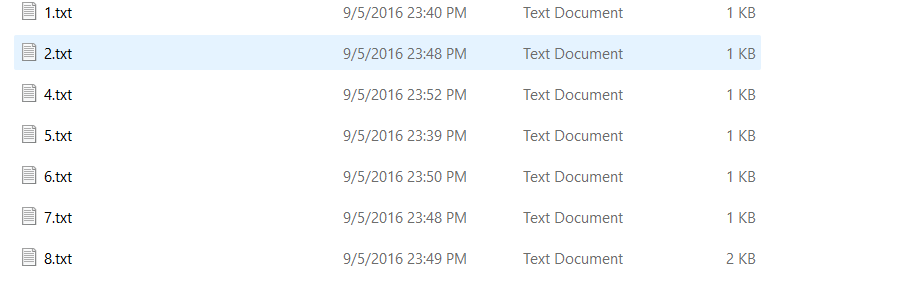


Hình 13: Dataset

Sử dụng một bộ dữ liệu tự tạo, thu thập trên internet về các bình luận của các nhà hàng bằng tiến việt.

Chúng em sẽ thu thập dữ liệu trên internet và chia làm hai tệp

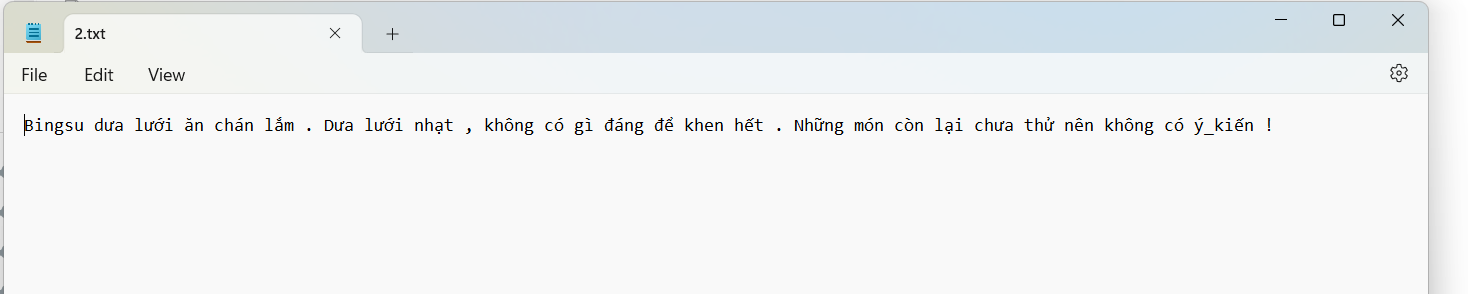
Mỗi tệp sẽ chứa các file .txt liên quan đến tiêu cực và tích cực



Hình 14: Dataset

Trong mỗi file sẽ chứa các đoạn text

Tiến hành Import dữ liệu và tiền xử lý dữ liệu



Hình 15: Import dữ liệu

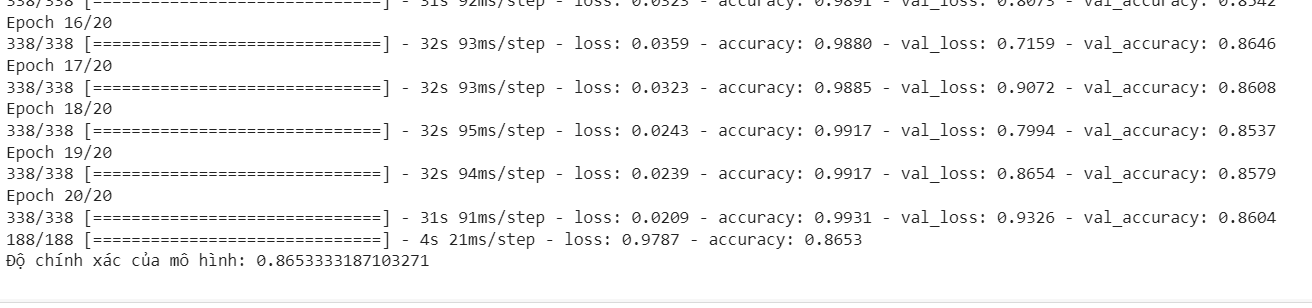


Hình 16: Tiền xử lý dữ liệu

Model



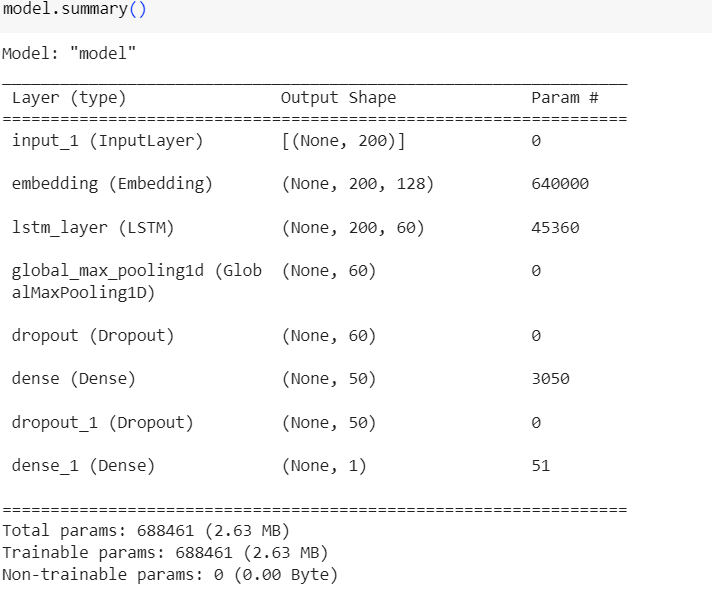
Hình 17: Model

Kết quả

Hình 18: Kết quả

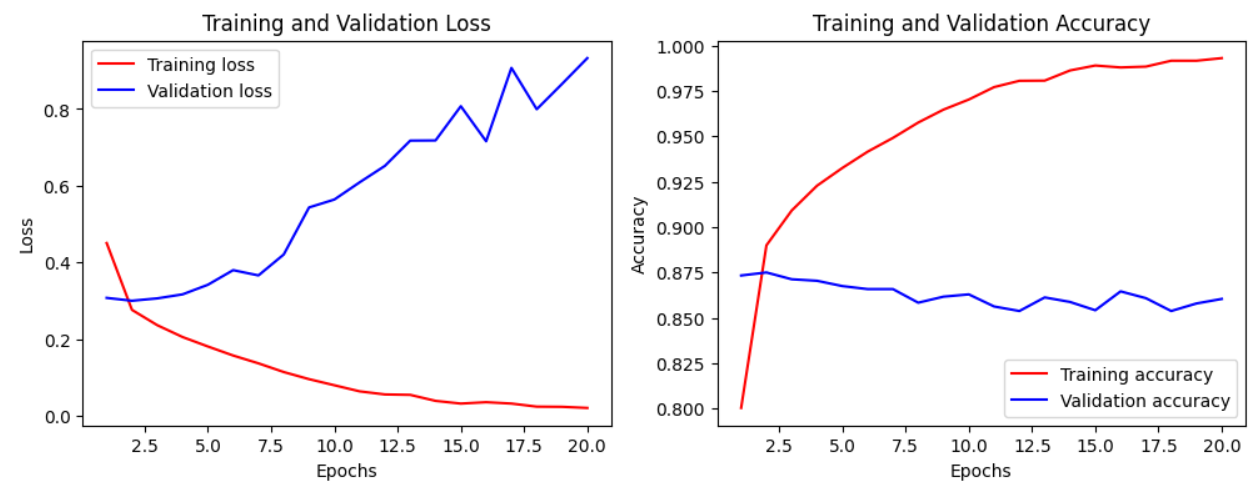
Từ kết quả trên có thể thấy model cho kết quả khá cao hơn 86%

Thông tin về model



Hình 19: Thông tin về Model

Đồ thị

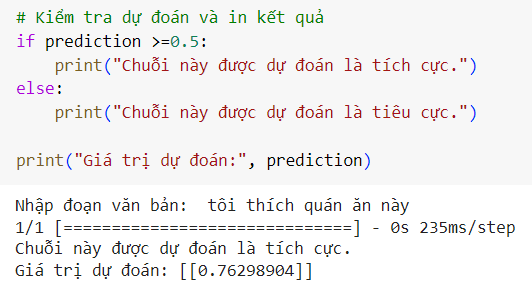


Hình 20: Đồ thị

Tuy nhiên khi nhìn vào đồ thị đo độ chính xác và hàm loss khi training và thực tế thì lại hoàn toàn trái ngược nhau, hàm loss thực tế tăng nhưng training lại giảm và accuracy cũng vậy

Lưu model và test với một chuỗi văn bản bất kỳ nhập vào từ bàn phím

Hình 21: Lưu model

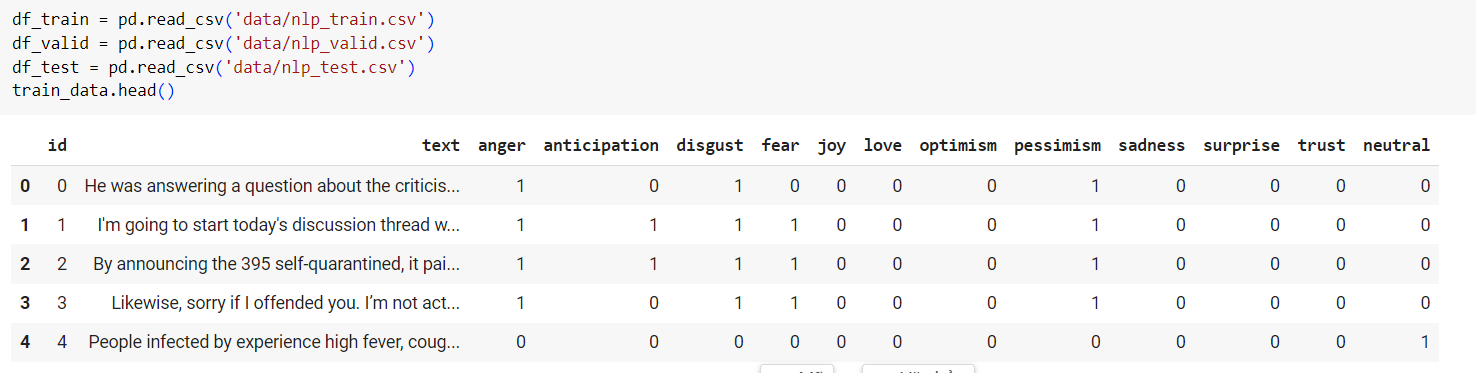


Hình 22: Kiểm tra dự đoán

=> Tuy đã thay đổi model và dataset tuy nhiên kết quả cho ra vẫn chưa quá chính xác và vẫn còn sai nhiều

Vì vậy nhóm chúng em quyết định tìm hiểu và chọn lại một bộ dataset tốt hơn để xử dụng. Qua quá trình tìm hiểu thì chúng em có tìm được một bài báo về phân tích tình cảm văn bản về covid19 sử dụng mô hình Bert, đặc biệt là nhóm nghiên cứu này có cung cấp dataset đã được gắn nhãn rất cụ thể.

Sơ lược về data



Hình 23: Sơ lược về Data

Có thể thấy bộ dữ liệu này có chia ra rất nhiều loại cảm xúc khác nhau. Và có gián nhãn theo khía cạnh, Một câu có thể có nhiều loại cảm xúc khác nhau. Bộ dữ liệu này rất đa dạng và được thiết kế rất tốt nên chúng em quyết định sử dụng dataset này

Phần tiền xử lý dữ liệu cơ bản vẫn không có gì thay đổi, chỉ thay đổi một giá trị maxlen từ 100 lên 250 và tiến hành gán nhãn cho từng class, ở đây có 12 class tưng ứng với 12 loại tình cảm khác nhau

Sau đó tiến hành thiết kế model và huấn luyện. Ở đây chúng em sử dụng lại cấu trúc model đã thiết kế ở các phần trước mà có độ chính xác cao nhất để tiến hành test và thay đổi siêu tham số cho phù hợp

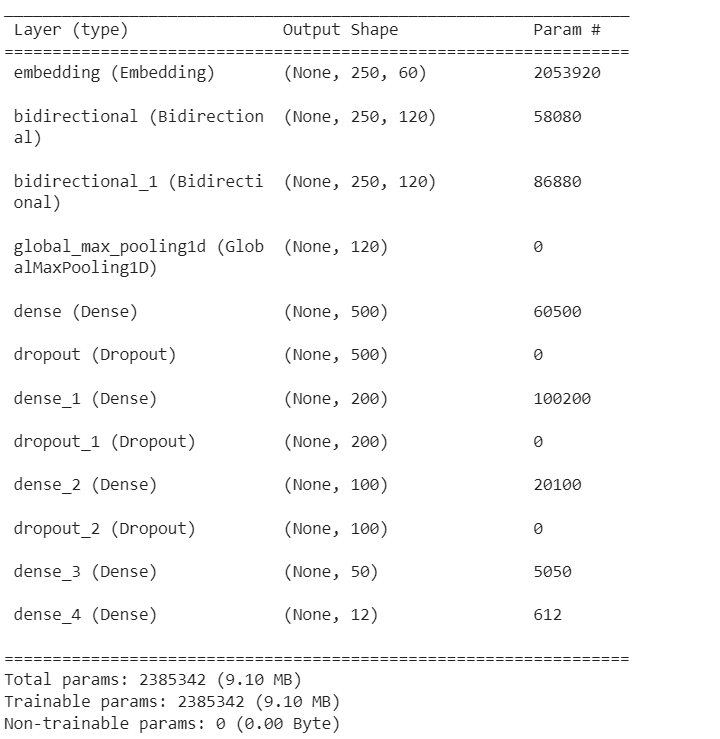


Hình 24: Thiết kế model



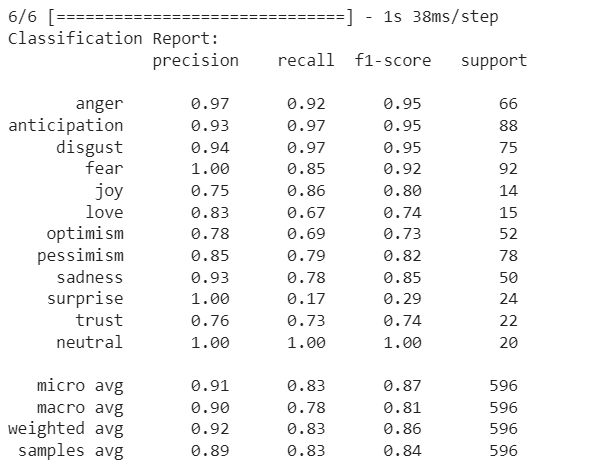
Hình 25: Thiết kế model

Thông tin về model

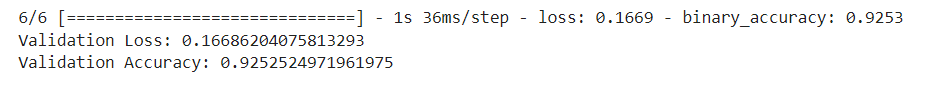


Hình 26: Thông tin model

Kết quả



Hình 27: Kết quả



Kết qua cho ra rất tốt có độ chính xác lên tới hơn 92%. Có thể thấy model mà chúng em thiết kế khá là phù hợp với bộ dataset này, tuy còn thấp hơn so với mô hình Bert mà nhóm nghiên cứu trong bài báo là 98%. Nhưng bù lại là mô hình của chúng em nhỏ gọn là training nhanh hơn, có thể xem đây là một kết quả khá tốt.

Đồ thị:



Hình 28: Đồ thị

Tiến hành kiểm tra một chuỗi văn bản bất kỳ được nhập vào từ bàn phím



Hình 29: Input & test

Sau khi test nhiều lần thì chúng em thấy model hoạt động khá hiệu quả.

## **3. Kết quả và mục tiêu phát triển**

- Kết Quả:

Qua quá trình nghiện cứu chúng em đã tìm được một bộ dataset và xây dựng được một model phụ hợp với bài toán phân loại cảm xúc trong văn bản, từ đó tiếp tục phát triển cho mục tiêu sắp tới.

* Mục tiêu phát triển

Từ model và bộ dataset đã phát triển nhóm chũng em sẽ áp dụng để giải quyết các bài toán thực tế trong kinh doanh.

* Giới thiệu chung về hoạt động kinh doanh liên quan đến đề tài:

COVID-19 là một đại dịch toàn cầu đã gây ra những tác động to lớn đến các hoạt động kinh doanh trên toàn thế giới. Các doanh nghiệp đã phải đối mặt với nhiều thách thức, bao gồm giảm doanh số bán hàng, gián đoạn chuỗi cung ứng và thay đổi hành vi của người tiêu dùng.

Trong bối cảnh này, việc phân tích tình cảm trong các bình luận về COVID-19 có thể giúp các doanh nghiệp hiểu rõ hơn về tâm lý của người tiêu dùng và đưa ra các quyết định kinh doanh phù hợp.

Ví dụ một doanh nghiệp có thể sử dụng phân tích tình cảm để:

+ Xác định những sản phẩm hoặc dịch vụ nào đang được người tiêu dùng quan tâm

+ Tìm hiểu về những mối quan tâm và lo lắng của người tiêu dùng về COVI-19

+ Tạo ra các thông điệp tiếp thị phù hợp với tâm lý của người tiêu dùng

* **Dựa trên mục đích của đề tài, các câu hỏi cần trả lời bao gồm:**

Tâm lý của người tiêu dùng về COVID-19 là gì?

+ Đánh giá về dịch vụ cụ thể trước và sau covid?

+ Người quan tâm đang muốn quan tâm những sản phẩm nào trong thời kỳ này?

+ Người tiêu dùng quan tâm đến những sản phẩm hoặc dịch vụ nào trong thời kỳ COVID-19?

- Tập trung vào 1 vài sản phẩm cụ thể và phân tích khía cạnh của nó?

**Kết luận:**

Phân tích tình cảm trong các bình luận về COVID-19 có thể là một công cụ hữu ích cho các doanh nghiệp trong việc hiểu rõ hơn về tâm lý của người tiêu dùng và đưa ra các quyết định kinh doanh phù hợp.

Dưới đây là một số ý tưởng cụ thể về cách sử dụng phân tích tình cảm để hỗ trợ hoạt động kinh doanh:

- Tạo ra các thông điệp tiếp thị phù hợp với tâm lý của người tiêu dùng: Các doanh nghiệp có thể sử dụng phân tích tình cảm để xác định những từ và cụm từ mà người tiêu dùng sử dụng khi nói về COVID-19. Từ đó, các doanh nghiệp có thể tạo ra các thông điệp tiếp thị phù hợp với tâm lý của người tiêu dùng và thu hút sự chú ý của họ.

- Thay đổi chiến lược sản phẩm và dịch vụ: Phân tích tình cảm có thể giúp các doanh nghiệp xác định những sản phẩm hoặc dịch vụ nào đang được người tiêu dùng quan tâm. Từ đó, các doanh nghiệp có thể điều chỉnh chiến lược sản phẩm và dịch vụ của mình để đáp ứng nhu cầu của người tiêu dùng.

- Cải thiện dịch vụ khách hàng: Phân tích tình cảm có thể giúp các doanh nghiệp hiểu rõ hơn về những mối quan tâm và lo lắng của người tiêu dùng. Từ đó, các doanh nghiệp có thể cải thiện dịch vụ khách hàng của mình để đáp ứng nhu cầu của người tiêu dùng.

## **4. Xây dựng ứng dụng Web app để phân tích tình cảm trong văn bản**

Ứng dụng của chúng em xây dựng bằng ngôn ngữ python và dựng trên mã nguồn mở của thự viện Streamlit để xây dựng web app

* Stremlit: Streamlit là công cụ được xây dựng với mục đích dành cho Machine Learning Engineer, tạo ra giao diện web như Jupyter notebook. Điểm đặc biệt khác với Jupyter notebook là Streamlit không phải hiển thị code, giúp cho bạn có thể tạo ra sản phẩm có tính hoàn thiện cao.

Hình 30: Streamlit

* Pycham: Pycharm là một nền tảng kết kết hợp được JetBrains phát triển như một IDE (Môi trường phát triển tích hợp) để phát triển các ứng dụng cho lập trình trong Python. Một số ứng dụng lớn như Tweeter, Facebook, Amazon và Pinterest sử dụng Pycharm để làm IDE Python của họ.

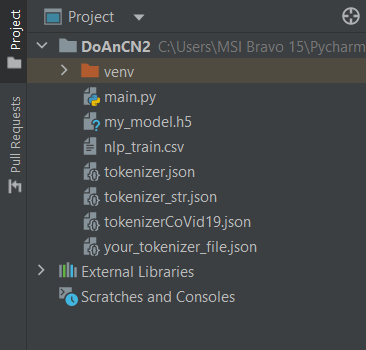


Hình 31: Pycharm

* Jupyter notebook: Jupyter Notebook là 1 mức sử dụng thực hiện Python thả hoàn hảo nhất mang lại câu hỏi tính tân oán công nghệ (Scientific computing) cùng phân tích dữ liệu (data analysis) với phần đông thư viện nhỏng numpy, pandas với mathplotlib.

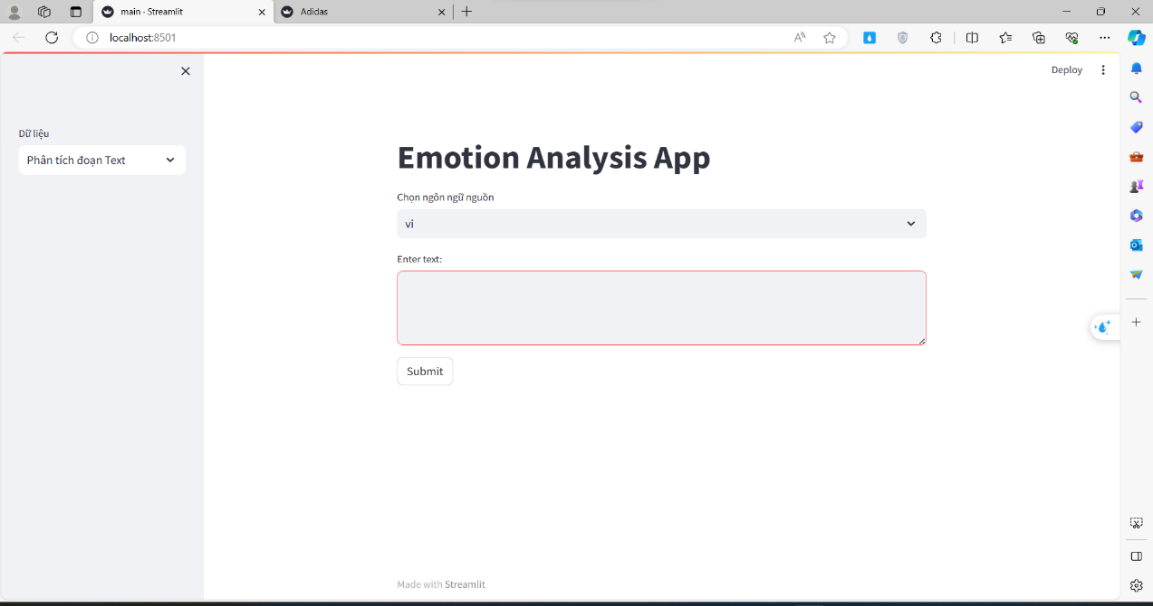
Hình 32: Jupyter

### 4.1 Tổng quan về Web app

Sau khi training xong model chúng em lưu model dưới dạng là My\_model.h5 và lưu tokenizer thành file tokenizer.josn để đưa vào Web app để xử dụng.

Hình 33: Lưu model

Web app của chúng em sử dụng model đã training ở phần trước để dựng đoán. Vì model sử dụng bộ dữ liệu tiếng anh để thiết kế nên chỉ có thể dự đoán bằng ngôn ngữ tiếng anh. Để phân tích tình cảm trên một ngôn ngữ khác đặc biệt là tiếng Việt thì phải huấn luyện model bằng dataset tiếng việt mà dataset tiếng viết chúng em không có nên chúng em quyết định dùng phương pháp dùng API của google dịch để dịch một văn bản sang tiếng anh rồi mới đưa vào model để dự đoán.



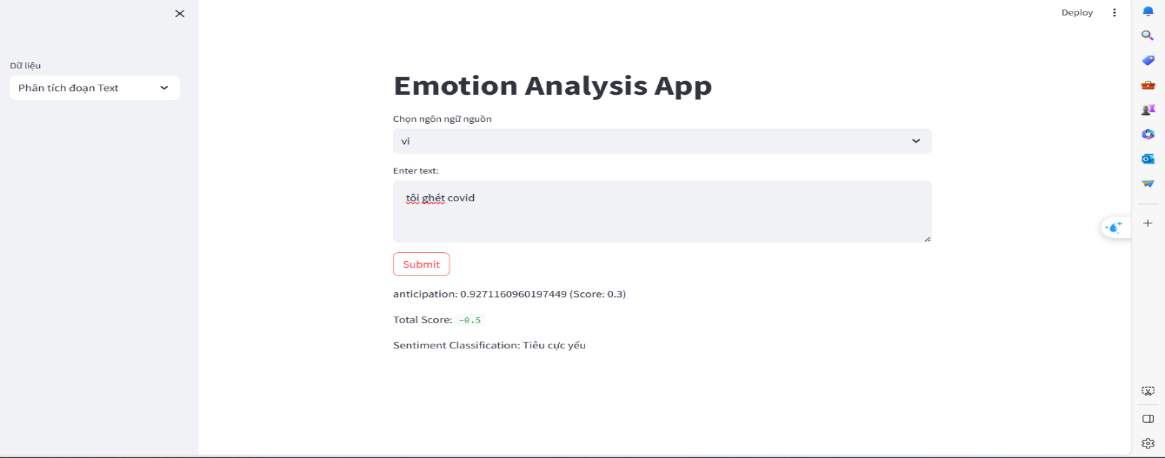
Hình 34: Web app

**4.2 Chức năng phân tích đoạn text**

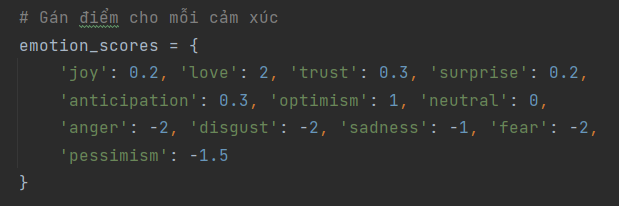
Ở tính năng này chúng em cho người dùng nhập vào một đoạn text sau khi bấm submit thì sẽ tiến hành phân tích tình cảm của văn bản đó là cực kỳ tích cực, tích cực trung bình, tích cực yếu, trung tính, cực kỳ tiêu cực, tiêu cực trung bình hặc là tiêu cực yếu

Chúng em còn cho người dùng chọn ngôn ngữ trước khi dự đoán mục đích là để dịch ngôn ngữ đó sang tiếng anh

Dataset gồm có 12 label là

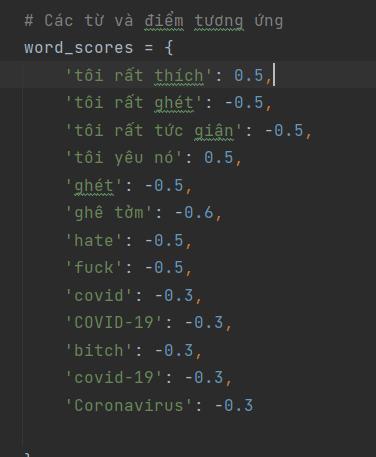
Hình 36: Label

Hình 35: Phân tích đoạn text



Hình 37: Điểm Label

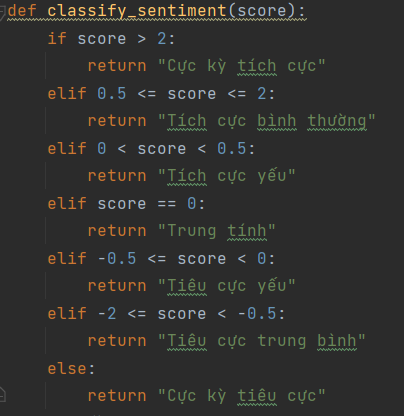
Kết hợp với phương pháp từ khóa, nếu trong câu văn có các từ khóa chỉ định thì sẽ được cộng hoặc trừ điểm tương ứng.



Hình 38:Điểm từ khóa

Tổng điểm (score) nếu:

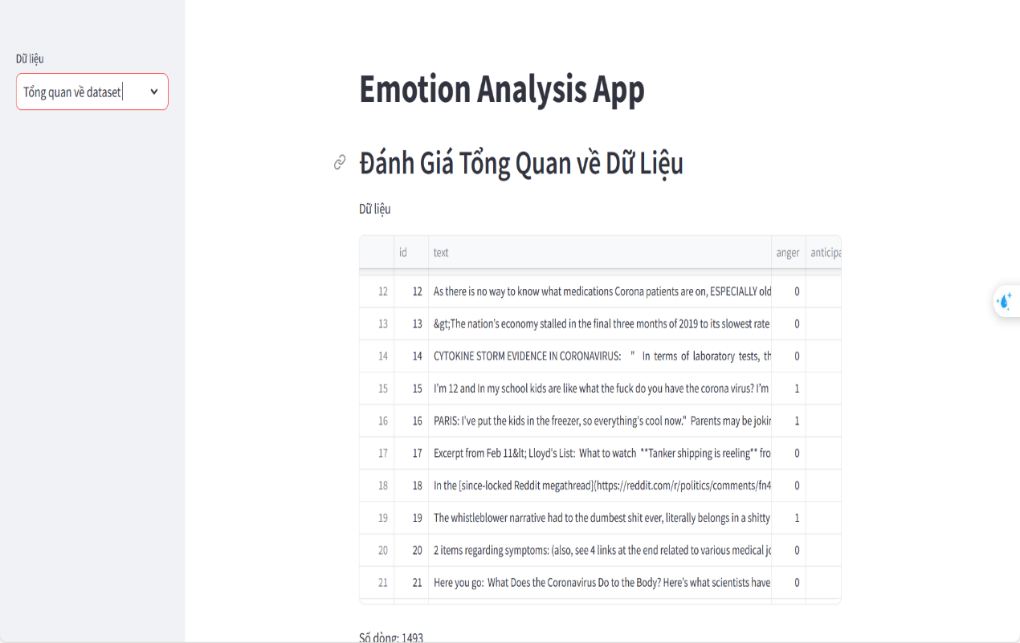
* Lớn hơn 2 sẽ là cực kỳ tích cực
* Trong khoảng từ 0.5 đến 2 là tích cực trung bình
* Trong khoảng từ 0 đến 0.5 là tích cực yếu
* Bằng 0 là trung tính
* Trong khoảng từ -0.5 đến 0 là tiêu cực yếu
* Trong khoảng từ -2 đến -0.5 là tiêu cực trung bình
* Dưới -2 là cực kỳ tiêu cực



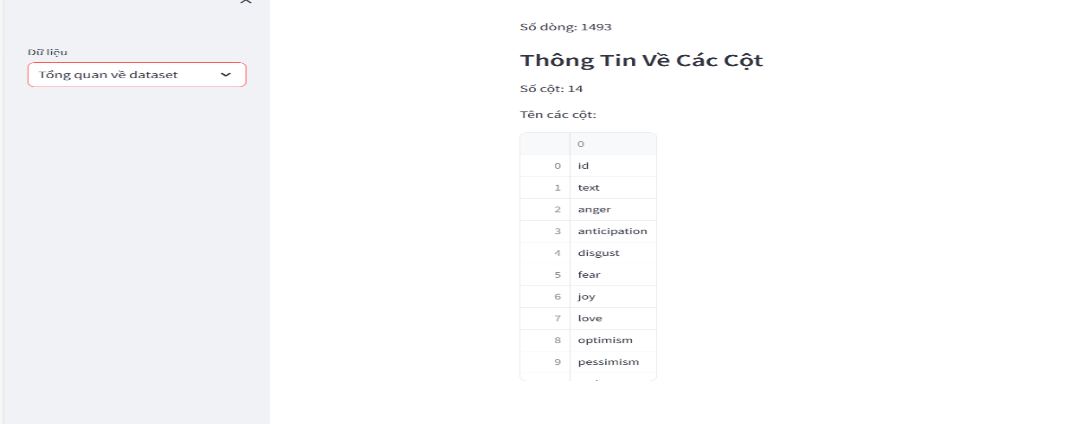
Hình 39: Phân loại cảm xúc

**4.3 Chức năng tông quan về dataset**

Với tính năng này sẽ cho người dùng thấy những thông tin cơ bản về dataset đã sử dụng như số dòng, số cột...

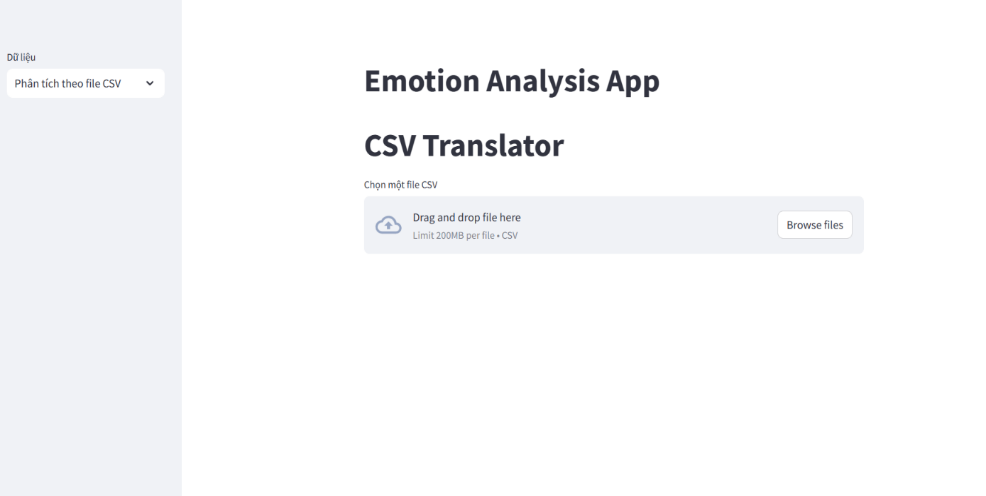


Hình 40: Tổng quan Dataset



Hình 41: Tổng quan Dataset

### 4.4 Chức năng phân tích theo CSV

Chức năng này sẽ cho người dùng nhập 1 file csv có chứ các văn bản vào để tiến hành phân tích tình cảm

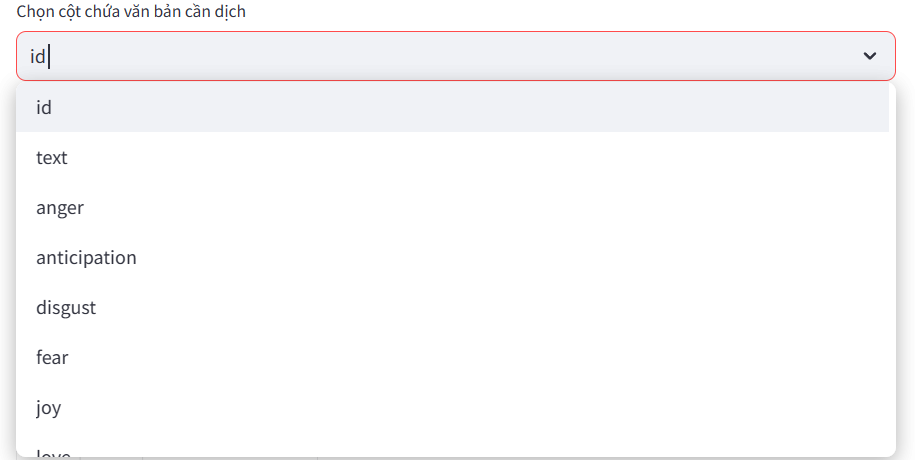
Hình 42: Input Dataset



Hình 43: Input Dataset

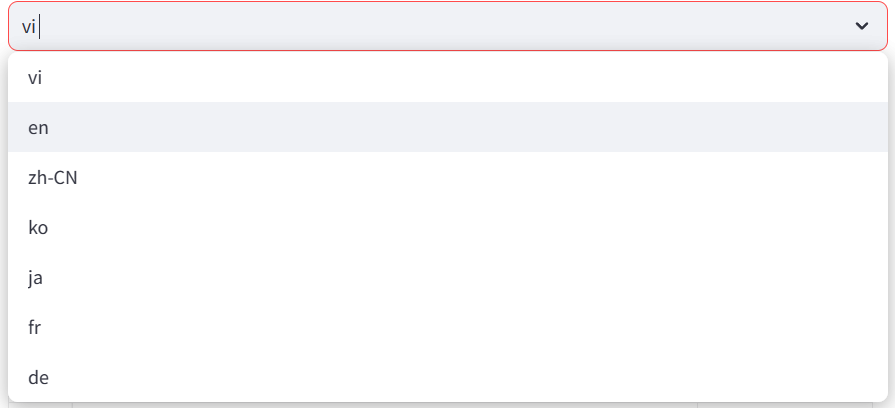
Ứng dụng sẽ cung cấp thông tin cơ bản của file csv

Người dùng sẽ chọn cột chứa văn bản cần phân tích



Hình 44: Chọn cột chứa văn bản cần phân tích

Chọn ngôn ngữ sử dụng



Hình 45: Chọn ngôn ngữ



Hình 46: Kết quả

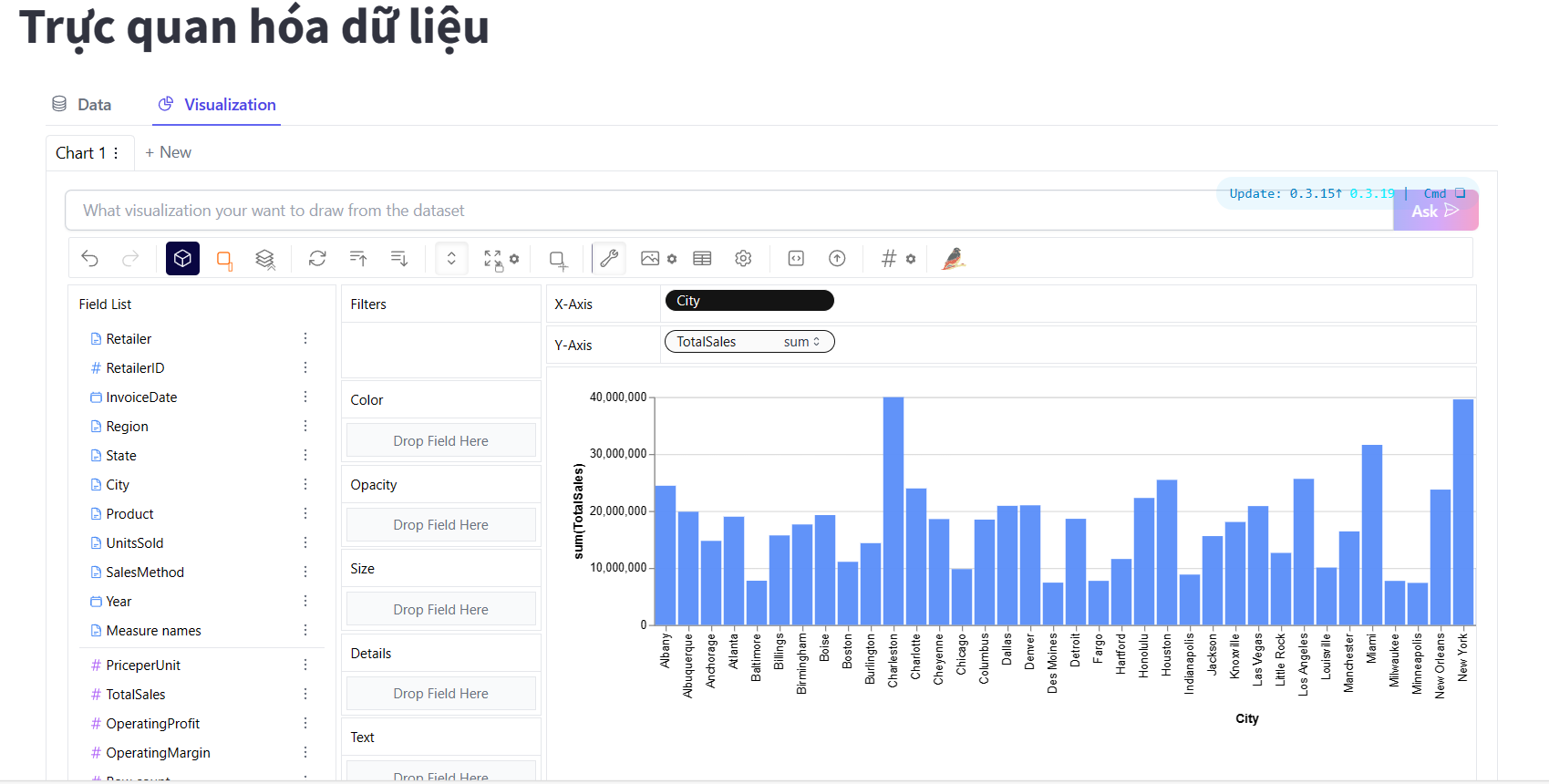
Sau đó thì app sẽ tự động cho ra kết quả tương ứng. Người dùng hoàn toàn nó thể tải bảng dữ liệu này về với chứng năng download khi đưa chuột vào mỗi bảng tương ứng

## **4.5 Chức năng trực quan hóa dữ liệu**

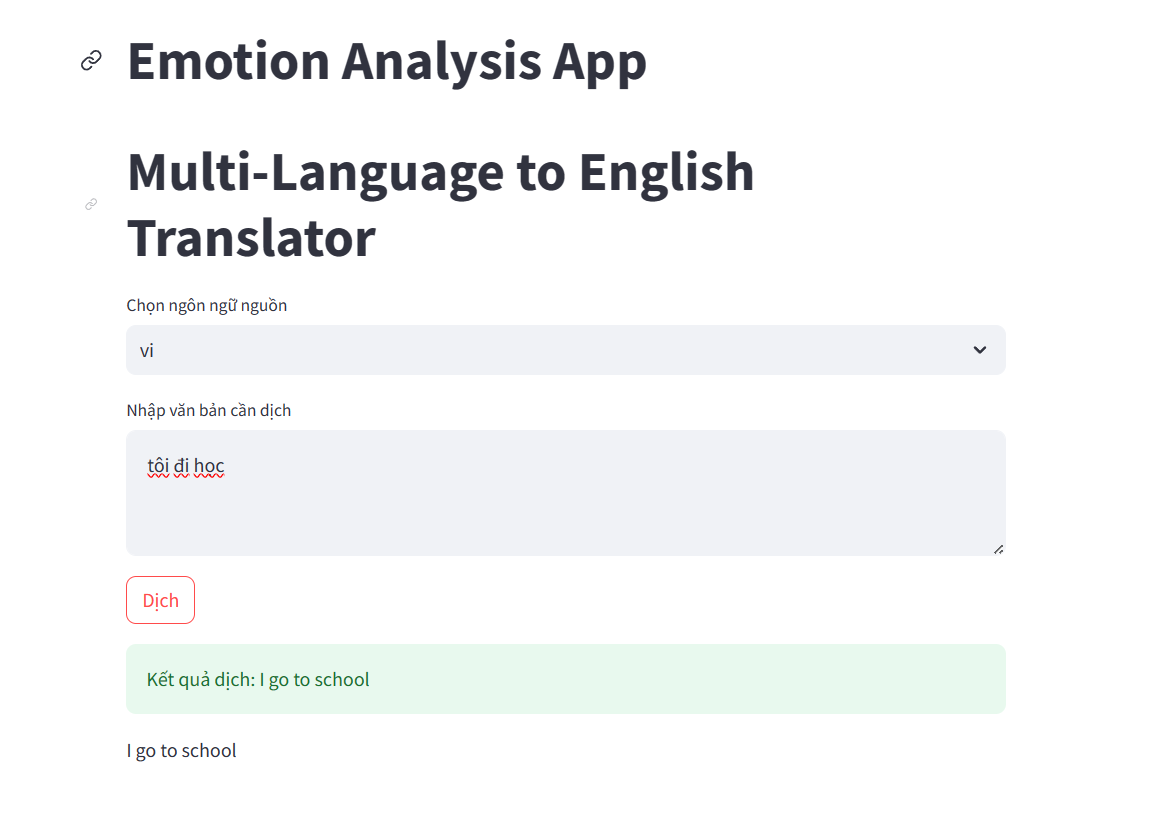
Với tính năng này sẽ giúp người dùng vẽ các biểu đồ thống kê với data bất kỳ giúp người dùng có cái nhìn trục quan nhất về dữ liệu mà mình đang sử dụng.

Chức năng cúng cho phép thiết lập và vẽ nhiều biểu đồ khác nhau, cho phép lưu biểu đồ nếu người dùng có nhu cầu

Hình 47: Trực quan hóa



### 4.6 Chức năng google dịch

Người dùng có thể dùng chức năng này để kiểm tra sẽ ứng dụng đã dịch đúng hay chưa

Hình 48: Kiểm tra kết quả dịch

# Tài liệu tham khảo

* + - 1. Nghiên cứu Sentiment analysis on the impact of coronavirus in social life using the BERT model của nhóm tác giả Mrityunjay Singh, Amit Kumar Jakhar1, Shivam Pandey2
      2. <https://docs.streamlit.io/>
      3. <https://www.tensorflow.org/tutorials/keras/classification?hl=vi>
      4. <https://www.nltk.org/>
      5. <https://textblob.readthedocs.io/en/dev/>