TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN &

TRUYỀN THÔNG VIỆT HÀN

Khoa Khoa Học Máy Tính



BÁO CÁO TOÀN VĂN

SENTIMENT ANALYSIS IN BUSINESS

**THÁI ĐỘ CỦA NHÀ ĐẦU TƯ TRÊN MẠNG XÃ HỘI VỀ BIẾN ĐỘNG GIÁ CỦA BITCOIN**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | Lê Văn Phước | 20IT487 |
|  | Hồ Luận | 20IT669 |
|  | Võ Văn Nguyên | 20IT325 |
|  | Nguyễn Đỗ Quang Minh | 20IT832 |
|  | Phan Thị Kim Ngân | 20IT225 |
| Giảng viên hướng dẫn: | TS. Đặng Trung Thành | |

Đà Nẵng, 05 tháng 01 năm 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN &

THÔNG VIỆT HÀN

Khoa Khoa Học Máy Tính



BÁO CÁO TOÀN VĂN

SENTIMENT ANALYSIS IN BUSINESS

**THÁI ĐỘ CỦA NHÀ ĐẦU TƯ TRÊN MẠNG XÃ HỘI VỀ BIẾN ĐỘNG GIÁ CỦA BITCOIN**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | Lê Văn Phước | 20IT487 |
|  | Hồ Luận | 20IT669 |
|  | Võ Văn Nguyên | 20IT325 |
|  | Nguyễn Đỗ Quang Minh | 20IT832 |
|  | Phan Thị Kim Ngân | 20IT225 |
| Giảng viên hướng dẫn: | TS. Đặng Trung Thành | |

Đà Nẵng, 05 tháng 01 năm 2023

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

|  |  |
| --- | --- |
|  | Chữ ký của giảng viên  *(Ký và ghi rõ họ tên)* |

**LỜI CẢM ƠN**

Đồ án môn học “Chuyên đề 3(IT) AD Sentiment analysis in Business” lần này được hoàn thành tại trường Đại học Công nghệ thông tin và Truyền thông Việt-Hàn. Trong quá trình học tập, triển khai và hoàn thành đồ án với đề tài Thái độ của nhà đầu tư trên mạng xã hội về biến động giá của bitcoin, chúng em đã nhận được rất nhiều sự giúp đỡ, vì vậy chúng em muốn viết lời cảm ơn này đến:

Chúng em xin gửi lời cảm ơn chân thành tới thầy cô trường Đại học Công nghệ thông tin và Truyền thông Việt Hàn đã giảng dạy và nâng đỡ cho chúng em thời gian qua. Đặc biệt là TS. Đặng Trung Thành là người trực tiếp chỉ bảo và hướng dẫn chúng em hoàn thành đồ án này.

Bên cạnh đó để hoàn thành đồ án này chúng em đã nhận được những lời khuyên sự giúp đỡ tận tình từ anh chị em bạn bè, chúng em xin cảm ơn.

Mặc dù đã vận dụng tất cả kiến thức đã được học tập và kinh nghiệm thực tế từ bản thân để hoàn thành đồ án này, song có thể còn có những mặt hạn chế, thiếu sót. Chúng em rất mong nhận được ý kiến đóng góp và sự chỉ dẫn của các thầy cô giáo. Chúng em xin chân thành cảm ơn.

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 6](#_Toc155761562)

[DANH MỤC CÁC TỪ VIẾT TẮT 7](#_Toc155761563)

[MỞ ĐẦU 8](#_Toc155761564)

[1. Lý do chọn đề tài 8](#_Toc155761565)

[2. Mục tiêu của đề tài 8](#_Toc155761566)

[3 Đối tượng và phạm nghiên cứu 8](#_Toc155761567)

[4. Cấu trúc đề tài 8](#_Toc155761568)

[Chương 1. GIỚI THIỆU CHUNG VỀ ĐỀ TÀI VÀ CÁC BÀI TOÁN KINH DOANH 9](#_Toc155761569)

[1. Giới thiệu chung về hoạt động kinh doanh liên quan đến đề tài 9](#_Toc155761570)

[2. Thái độ của nhà đầu tư bitcoin trên mạng xã hội 9](#_Toc155761571)

[3. Các câu hỏi kinh doanh cần trả lời 10](#_Toc155761572)

[Chương 2: CƠ SỞ LÝ THUYẾT VỀ PHÂN TÍCH TÌNH CẢM TRONG KINH DOANH 11](#_Toc155761573)

[1. Khái niệm về phân tích tình cảm 11](#_Toc155761574)

[2. Ứng dụng của phân tích tình cảm 12](#_Toc155761575)

[3. Trích xuất khía cạnh 13](#_Toc155761576)

[3. Phân lớp tình cảm của khía cạnh bằng phương pháp từ vựng 13](#_Toc155761577)

[4. Word Embedding 15](#_Toc155761578)

[5. LSTM 15](#_Toc155761579)

[Chương 3: KẾT QUẢ 17](#_Toc155761580)

[3.1 Quá trình thu nhập dữ liệu 17](#_Toc155761581)

[3.2 Làm sạch dữ liệu 18](#_Toc155761582)

[3.3 Trích xuất khía cạnh 19](#_Toc155761583)

[3.4 Phân tích tình cảm cấp khía cạnh bằng phương pháp từ vựng 20](#_Toc155761584)

[3.5 Phân tích tình cảm cấp khía cạnh bằng phương pháp học sâu 22](#_Toc155761585)

[KẾT LUẬN 25](#_Toc155761586)

[PHÂN CÔNG CÔNG VIỆC 26](#_Toc155761587)

[DANH MỤC TÀI LIỆU THAM KHẢO 27](#_Toc155761588)

# DANH MỤC HÌNH ẢNH

[Hình 1. Word Embedding 15](#_Toc155755596)

[Hình 2. Cấu trúc của LSTM 16](#_Toc155755597)

[Hình 3. Thu nhập dữ liệu từ API Reddit 18](#_Toc155755598)

[Hình 4. Trích xuất khía cạnh 20](#_Toc155755599)

[Hình 5. Phân tích tình cảm bằng phương pháp từ vựng 21](#_Toc155755600)

[Hình 6. Dữ liệu đã được gán nhãn 22](#_Toc155755601)

[Hình 7. Cấu trúc mô hình phân ltích cảm xúc 23](#_Toc155755602)

[Hình 8. Độ chính xác 24](#_Toc155755603)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| Cụm từ viết tắt | Cụm từ đầy đủ |
| BTC | Bitcoin |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory |
| CNTT&TT | Công nghệ thông tin và truyền thông |

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Bitcoin là một loại tiền ảo có giá trị biến động mạnh, gây ra nhiều tranh cãi trong cộng đồng. Một số nhà đầu tư lạc quan về tiềm năng của bitcoin và tin rằng nó có thể trở thành một loại tiền tệ chính thống trong tương lai. Ngược lại, một số nhà đầu tư khác thì thận trọng hơn về bitcoin và lo ngại về rủi ro của loại tiền ảo này.

Thái độ của nhà đầu tư trên mạng xã hội về bitcoin có thể phản ánh những biến động giá của bitcoin. Khi giá bitcoin tăng cao, các nhà đầu tư lạc quan thường đăng tải các bài viết và bình luận tích cực về bitcoin. Ngược lại, khi giá bitcoin giảm, các nhà đầu tư thận trọng thường trở nên tích cực hơn.

Vì vậy, việc nghiên cứu thái độ của nhà đầu tư trên mạng xã hội về biến động giá của bitcoin là cần thiết để hiểu rõ hơn về tâm lý của các nhà đầu tư, từ đó có thể đưa ra các dự báo chính xác hơn về giá bitcoin trong tương lai.

## 2. Mục tiêu của đề tài

Mục tiêu của nghiên cứu này là:

* Khảo sát thái độ của nhà đầu tư trên mạng xã hội về biến động giá của bitcoin.
* Xác định các yếu tố ảnh hưởng đến thái độ của nhà đầu tư về bitcoin.
* Đưa ra các khuyến nghị cho các nhà đầu tư và các nhà quản lý.

## 3 Đối tượng và phạm nghiên cứu

Đối tượng nghiên cứu của đề tài này là các nhà đầu tư bitcoin trên mạng xã hội. Phạm vi nghiên cứu là các bài viết, bình luận trên các trang mạng xã hội như Facebook, Twitter,...

## 4. Cấu trúc đề tài

Báo cáo được trình bày trong 3 chương, cụ thể như sau:

- *Chương 1*: Giới thiệu về bài toán kinh doanh.

- *Chương 2*: Cơ sở lý thuyết phân tích tình cảm trong kinh doanh.

- *Chương 3*: Kết quả.

# Chương 1. GIỚI THIỆU CHUNG VỀ ĐỀ TÀI VÀ CÁC BÀI TOÁN KINH DOANH

## 1. Giới thiệu chung về hoạt động kinh doanh liên quan đến đề tài

Hoạt động kinh doanh liên quan đến bitcoin là hoạt động kinh doanh sử dụng bitcoin, một loại tiền ảo, làm phương tiện thanh toán hoặc đầu tư. Hoạt động kinh doanh này có thể được chia thành hai nhóm chính:

- Nhóm hoạt động cung cấp sản phẩm, dịch vụ sử dụng bitcoin:

* Mua bán bitcoin: Đây là nhóm hoạt động chính trong lĩnh vực này, bao gồm việc mua bán bitcoin giữa các cá nhân, tổ chức.
* Sử dụng bitcoin để thanh toán: Các doanh nghiệp có thể chấp nhận bitcoin làm phương tiện thanh toán cho các sản phẩm, dịch vụ của mình.
* Sử dụng bitcoin để đầu tư: Các cá nhân, tổ chức có thể đầu tư vào bitcoin để thu lợi nhuận.

- Nhóm hoạt động liên quan đến công nghệ blockchain:

* Phát triển ứng dụng trên blockchain: Các doanh nghiệp có thể phát triển các ứng dụng trên blockchain, sử dụng bitcoin làm phương tiện thanh toán hoặc trao đổi.
* Nghiên cứu và phát triển công nghệ blockchain: Các doanh nghiệp có thể đầu tư vào nghiên cứu và phát triển công nghệ blockchain, nhằm ứng dụng công nghệ này trong các lĩnh vực khác nhau.

Hoạt động kinh doanh liên quan đến bitcoin có thể mang lại nhiều lợi ích cho doanh nghiệp, bao gồm:

* Tăng cường tính thanh khoản: bitcoin có thể được sử dụng để thanh toán cho các sản phẩm, dịch vụ, giúp doanh nghiệp tăng cường tính thanh khoản.
* Giảm chi phí: bitcoin có thể giúp doanh nghiệp giảm chi phí thanh toán, chi phí chuyển tiền,...
* Mở rộng thị trường: bitcoin có thể giúp doanh nghiệp mở rộng thị trường ra ngoài phạm vi quốc gia, đến với các khách hàng sử dụng bitcoin.

Tuy nhiên, hoạt động kinh doanh liên quan đến bitcoin cũng tiềm ẩn nhiều rủi ro như:

* Volatility: Giá bitcoin có thể biến động mạnh, gây ra rủi ro cho doanh nghiệp.
* Tính pháp lý: bitcoin chưa được công nhận là phương tiện thanh toán hợp pháp ở nhiều quốc gia, gây ra rủi ro cho doanh nghiệp khi sử dụng bitcoin.
* Tính an toàn: bitcoin sử dụng công nghệ blockchain, có thể bị tấn công bởi tin tặc, gây ra rủi ro mất tiền cho doanh nghiệp.

Doanh nghiệp cần cân nhắc kỹ lưỡng các lợi ích và rủi ro trước khi tham gia vào hoạt động kinh doanh liên quan đến bitcoin.

## 2. Thái độ của nhà đầu tư bitcoin trên mạng xã hội

Thái độ của nhà đầu tư trên mạng xã hội về bitcoin khá phân cực. Một số nhà đầu tư rất lạc quan về tiềm năng của bitcoin và tin rằng nó có thể trở thành một loại tiền tệ chính thống trong tương lai. Họ lập luận rằng bitcoin có nhiều ưu điểm hơn so với các loại tiền tệ fiat truyền thống, chẳng hạn như tính bảo mật, tính minh bạch và khả năng truy cập toàn cầu.

Các nhà đầu tư khác thì thận trọng hơn về bitcoin. Họ lập luận rằng bitcoin là một tài sản rủi ro cao, với giá trị biến động mạnh. Họ cũng lo ngại về tính pháp lý của bitcoin ở nhiều quốc gia.

Thái độ của nhà đầu tư trên mạng xã hội cũng phản ánh những biến động giá của bitcoin. Ví dụ, khi giá bitcoin tăng cao, các nhà đầu tư lạc quan thường đăng tải các bài viết và bình luận tích cực về bitcoin. Ngược lại, khi giá bitcoin giảm, các nhà đầu tư thận trọng thường trở nên tích cực hơn.

Nhìn chung, thái độ của nhà đầu tư trên mạng xã hội về bitcoin khá tích cực. Tuy nhiên, cũng có một số nhà đầu tư thận trọng về loại tiền ảo này.

## 3. Các câu hỏi kinh doanh cần trả lời

Dựa trên giới thiệu chung về hoạt động kinh doanh liên quan đến đề tài, đề tài này sẽ trả lời các câu hỏi sau:

* Những từ nào thường được sử dụng để thể hiện cảm xúc tiêu cực, tích cực về Bitcoin?
* Các khía cạnh nào được liên kết nhiều nhất với các ý kiến tích cực, tiêu cực?
* Thái độ của nhà đầu tư về thị trường bitcoin nói chung và các khía cạnh theo thời gian?
* Thái độ của nhà đầu tư về thị trường bitcoin với khía cạnh “Elon Musk” theo thời gian?
* Thái độ của nhà đầu tư về biến động giá của bitcoin thể hiện như thế nào trên mạng xã hội?
* Các yếu tố nào ảnh hưởng đến thái độ của nhà đầu tư trên mạng xã hội?
* Thái độ của nhà đầu tư trên mạng xã hội có tác động như thế nào đến thị trường tài chính?

# Chương 2: CƠ SỞ LÝ THUYẾT VỀ PHÂN TÍCH TÌNH CẢM TRONG KINH DOANH

## 1. Khái niệm về phân tích tình cảm

Phân tích tình cảm (sentiment analysis) hay khai thác ý kiến (opinion mining ) là một lĩnh vực nghiên cứu phân tích ý kiến, tình cảm, đánh giá, thái độ và cảm xúc đối với các đối tượng và thuộc tính của chúng thể hiện trên văn bản.

Các đối tượng: sản phẩm, dịch vụ, tổ chức, cá nhân, sự kiện, chủ đề,...

Có nhiều tên gọi và tác vụ khác nhau:

* + Phân tích/ Khai thác cảm xúc (sentiment analysis/ mining)
  + Khai thác/ Phân tích/ Trích xuất ý kiến (opinion mining/ analysis/ extraction).
  + Phân tích tính chủ quan (subjectivity analysis).
  + Phân tích tác động (affect analysis).
  + Phân tích cảm xúc (emotion analysis).
  + Khai thác bài đánh giá (review mining)Tạo cơ sở dữ liệu cho server gốc.

**1.1 Các dạng phân tích**

- Phân tích chi tiết (Fine-Grained): từ rất tích cực đến trung tính đến rất tiêu cực, người dùng có thể trải nghiệm bất kỳ thuộc tính nào liên quan đến thương hiệu của bạn và những thuộc tính này có thể có hình dạng hữu hình dưới dạng xếp hạng (ví dụ - dựa trên sao) và tất cả những gì mô hình của bạn cần làm là khai thác các dạng khác nhau này xếp hạng từ nhiều nguồn khác nhau.

- Dựa trên khía cạnh (Aspect-Based): người dùng thường chỉ ra một số điều tốt hoặc xấu trong các bài đánh giá của họ ngoài xếp hạng và bày tỏ cảm xúc. Ví dụ: một bài đánh giá về doanh nghiệp du lịch của bạn có thể đề cập đến, “Hướng dẫn thực sự hữu ích và chỉ cho chúng tôi tất cả các địa điểm trong khu vực và thậm chí còn giúp chúng tôi lên chuyến bay”. Nhưng, nó cũng có thể là, “Người phụ trách bàn du lịch cực kỳ thô lỗ và lờ đờ. Chúng tôi đã phải đợi một giờ trước khi có hành trình trong ngày”.

- Phát hiện cảm xúc (Emotion Detection): ví dụ, nếu họ mua quần áo từ cửa hàng thương mại điện tử của bạn, họ có thể hài lòng với thủ tục gửi hàng, chất lượng quần áo hoặc nhiều lựa chọn của bạn hoặc thất vọng với chúng.

- Phân tích ý định (Intent Analysis): các loại thông số phân tích tình cảm khác nhau:

* Polarity- tập trung vào các đánh giá mà thương hiệu của bạn nhận được trực tuyến (tích cực, trung lập và tiêu cực)
* Cảmxúc- tập trung vào cảm xúc của sản phẩm hoặc dịch vụ của bạn trong tâm trí khách hàng (vui, buồn, thất vọng, vui mừng và hơn thế nữa)
* Khẩn cấp- tập trung vào tính tức thời của việc sử dụng thương hiệu của bạn hoặc tìm ra giải pháp hiệu quả cho các vấn đề của khách hàng (khẩn cấp và có thể chờ đợi)
* Ý định- tập trung vào việc tìm hiểu xem người dùng của bạn có quan tâm đến việc sử dụng sản phẩm hoặc thương hiệu của bạn hay không

**1.2 Lĩnh vực liên quan**

* Khai phá dữ liệu (data mining),
* Khai phá web (web mining)
* Truy xuất thông tin (information retrieval)

**1.3 Dữ liệu**

* Đa phần là dữ liệu văn bản; có thể kết hợp với hình ảnh, video và audio.
* Nguồn dữ liệu: web và truyền thông xã hội (social media).

## 2. Ứng dụng của phân tích tình cảm

**2.1 Ý kiến có vai trò quan trọng**

* Doanh nghiệp, tổ chức cần biết ý kiến của khách hàng và công chúng về sản phẩm dịch vụ của họ
* Cơ quan nhà nước cần biết ý kiến của người dân về các chính sách
* Cơ quan đối ngoại cần biết quan điểm và tình cảm của người dân nước khác về các vấn đề và sự kiện của nước họ và quốc tế
* Bên mua và sử dụng sản phẩm, dịch vụ cần biết ý kiến từ bên đã từng mua và sử dụng trước đó

**2.2 Sử dụng dữ liệu truyền thông xã hội để ra quyết định**

Truyền thông xã hội đã thay thế phương tiện truyền thông truyền thống để

cung cấp thông tin cho cá nhân, tổ chức ra quyết định:

* Cá nhân: tham khảo các bài đánh giá và thảo luận để lựa chọn sản phẩm, dịch vụ.
* Doanh nghiệp/ tổ chức: thu thập ý kiến của công chúng và người tiêu dùng công khai có sẵn trên mạng xã hội.
* Cơ quan nhà nước: theo dõi phương tiện truyền thông mạng xã hội để thu thập ý kiến của công chúng về chính sách và biến động ở các quốc gia khác.

Thách thức: sự bùng nổ của các trang mạng xã hội → nhiều nguồn, lượng bài viết lớn → Khó khăn cho người dùng thông thường trong xác định đủ các nguồn tin và trích xuất, tổng hợp các ý kiến → Cần có các hệ thống phân tích tình cảm tự động.

**2.3 Sử dụng dữ liệu nội bộ doanh nghiệp:**

* Dữ liệu nội bộ: bao gồm phản hồi của khách hàng qua email và trung tâm hỗ trợ khách hàng (call centers) và kết quả khảo sát khách hàng, thị trường.
* Cần thiết phân tích tình cảm để xác định các vấn đề với sản phẩm, dịch vụ và tổng hợp ý kiến khách hàng.

**2.4 Lĩnh vực ứng dụng**

Phân tích tình cảm có nhiều ứng dụng trong nhiều ngành nghề như:

* Sản phẩm tiêu dùng: điện tử, điện máy, thực phẩm, thời trang,...
* Chăm sóc sức khỏe: bệnh viện, bảo hiểm y tế,...
* Du lịch và lữ hành: khách sạn, nhà hàng, vận tải hành khách,...
* Dịch vụ tài chính: ngân hàng, đầu tư, chứng khoán,...
* Sự kiện chính trị - xã hội: bầu cử, lễ hội, ....

Qua đó, ta có thể thấy đây là một đề tài quan trọng, nhiều công ty tham gia phát triển dịch vụ phân tích tình cảm với quy mô và mức độ khác nhau và có cơ hội để thực hiện các nghiên cứu ứng dụng.

## 3. Trích xuất khía cạnh

Trích xuất khía cạnh là quá trình xác định và tách các khía cạnh cụ thể từ dữ liệu văn bản. Trong ngữ cảnh phân tích tình cảm, một “khía cạnh” thường là một sản phẩm, dịch vụ, chức năng, hoặc bất kỳ thuộc tính nào mà người dùng có thể có ý kiến. Ví dụ, trong câu “Màn hình của điện thoại này rất sáng nhưng pin không bền lắm”, “màn hình” và “pin” là các khía cạnh đã được trích xuất. Quá trình này giúp chúng ta hiểu rõ hơn về những gì mà người dùng thực sự quan tâm và phản hồi về sản phẩm hay dịch vụ.

Quá trình trích xuất khía cạnh từ dữ liệu phân tích tình cảm thường bao gồm các bước sau:

1. **Tiền xử lý dữ liệu:** Dữ liệu đầu vào, thường là văn bản, được làm sạch và chuẩn hóa. Các công việc cụ thể có thể bao gồm loại bỏ dấu câu, chuyển đổi tất cả văn bản thành chữ thường, loại bỏ các từ không có nhiều ý nghĩa (như “và”, “là”, “một”), và chia văn bản thành từng từ riêng biệt (quá trình này được gọi là “tokenization”).
2. **Phân tích cú pháp**: Các từ riêng biệt sau đó được phân tích để xác định vai trò ngữ pháp của chúng trong câu. Điều này giúp xác định các khía cạnh cụ thể mà người dùng đang nói về.
3. **Trích xuất khía cạnh:** Dựa trên thông tin ngữ pháp và ngữ nghĩa, các khía cạnh cụ thể được trích xuất từ văn bản. Một “khía cạnh” có thể là một sản phẩm, dịch vụ, chức năng, hoặc bất kỳ thuộc tính nào mà người dùng có thể có ý kiến.
4. **Phân loại cảm xúc:** Cuối cùng, dựa trên ngữ cảnh và từ vựng, hệ thống sẽ xác định xem người dùng có cảm xúc tích cực, tiêu cực hay trung lập đối với mỗi khía cạnh đã được trích xuất.

## 3. Phân lớp tình cảm của khía cạnh bằng phương pháp từ vựng

**3.1 Các cách tiếp cận tính toán tình cảm đối với mục tiêu**

* Sử dụng hàm tổng hợp tình cảm có tính đến khoảng cách giữa từ ngữ diễn đạt tình cảm và đối tượng/ khía cạnh mục tiêu trong câu.
* Tìm phạm vi áp dụng của mỗi từ ngữ diễn đạt tình cảm trong câu để xác định xem nó có bao hàm đối tượng/khía cạnh mục tiêu hay không: thường thực hiện thông qua khai thác các mối quan hệ cú pháp.

Ta cũng có thể tiếp kết hợp hai cách tiếp cận trên.

**3.2 Phương thức chung**

Trước hết, ta cần chuẩn bị :

* Một bộ từ vựng của các từ ngữ (và quy tắc) diễn đạt tình cảm
* Một bộ các luật để nắm bắt các cấu trúc về ngôn ngữ và loại câu
* Một hàm tổng hợp tình cảm, hoặc tập các mối quan hệ giữa tình cảm và mục tiêu sinh ra từ cây đoán nhận

Để xác định chiều tình cảm đối với mỗi khía cạnh, ta cần thực hiện 4 bước sau, giả sử rằng các mục tiêu và khía cạnh là đã biết hoặc đã được trích xuất từ trước.

1. Đánh dấu các từ ngữ diễn đạt tình cảm:

* Đánh dấu tất cả các từ ngữ diễn đạt tình cảm, cho điểm tình cảm cho mỗi từ ngữ là 1 nếu diễn đạt tích cực và -1 nếu diễn đạt tiêu cực

2. Áp dụng các bộ chuyển đổi tình cảm

* Các bộ chuyển đổi tình cảm là các từ ngữ có thể làm thay đổi chiều hướng tình cảm:
* Từ phủ định: not, never, none, nobody, nowhere, neither, cannot,... (Lưu ý ngoại lệ, nhu: not only ... but also)

3. Kiểm soát các mệnh đề nhưng (but-clause)

* Cần đặc biệt quan tâm các từ ngữ diễn đạt sự tương phản để nhận biết, điều chỉnh điểm tình cảm
* Các từ diễn đạt phổ biến: but, however, except that, except for, with the exception of,..

4. Tổng hợp điểm tình cảm

* Sử dụng hàm tổng hợp điểm tình cảm để xác định tình cảm về mỗi khía cạnh trong câu
* Cho câu s chứa một tập các khía cạnh {a1, ..., an} và một tập các từ ngữ diễn đạt tình cảm với điểm tình cảm được xác định từ các bước 1-3

Nếu điểm tổng hợp là dương thì tình cảm là tích cực, là âm thì tình cảm là tiêu cực, còn lại là trung lập

**3.3 Cải thiện hiệu quả**

Xác định chính xác phạm vi của mỗi từ ngữ diễn đạt tình cảm thay vì sử dụng khoảng cách từ. Cách thức thực hiện là khám phá mối quan hệ giữa từ ngữ diễn đạt tình cảm và mục tiêu của chúng, gồm 3 cách:

Các phụ thuộc về cú pháp: Đa phần là các mối quan hệ phụ thuộc tính từ - danh từ và động từ - trạng từ.

Ví dụ: “This camera takes great pictures” → Quan hệ danh từ và tính từ.

“I can install this software easily” → Quan hệ động từ và trạng từ.

Từ diễn đạt tình cảm chính là khía cạnh mục tiêu: nhiều từ tình cảm bản thân nó đã đề cập đến khía cạnh đánh giá.

Ví dụ: expensive đề cập về giá, beautiful đề cập đến hình thức,...

Các mối quan hệ ngữ nghĩa: Khó để xác định vì phụ thuộc vào nghĩa và/hoặc cách sử dụng từng từ/ngữ riêng lẻ

**3.4 Ưu và nhược điểm**

Về ưu điểm:

* Hiệu năng tốt trên nhiều ứng dụng
* Độc lập với lĩnh vực ứng dụng
* Không cần tốn công sức gắn nhãn dữ liệu
* Linh hoạt, dễ mở rộng và cải tiến

Về nhược điểm:

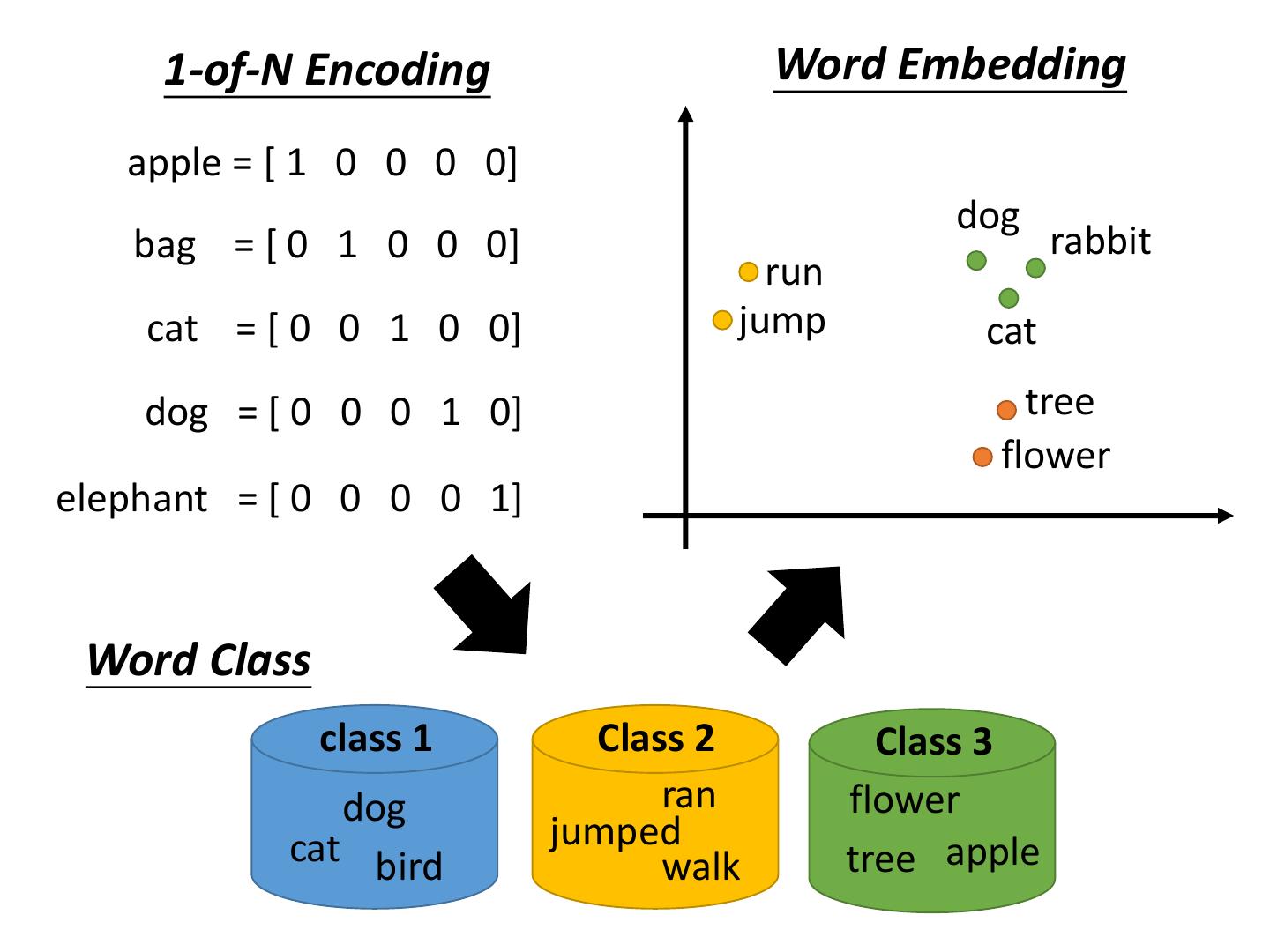
* Cần đầu tư nhiều thời gian và công sức để xây dựng cơ sở tri thức về từ vựng, mẫu và quy luật
* Đối với từng lĩnh vực vẫn cần quan tâm đến thuật ngữ của lĩnh vực

## 4. Word Embedding

Word embedding là một kỹ thuật trong xử lý ngôn ngữ tự nhiên mà từ ngữ được biểu diễn dưới dạng vectơ số. Mục tiêu của word embedding là để nắm bắt các mối quan hệ ngữ nghĩa giữa các từ, như đồng nghĩa, tương phản, và các mối quan hệ khác.

Một trong những cách phổ biến nhất để tạo word embedding là sử dụng mô hình như Word2Vec hoặc GloVe. Những mô hình này được huấn luyện trên một lượng lớn văn bản và học cách dự đoán một từ dựa trên các từ xung quanh nó trong câu. Kết quả là một không gian vectơ nhiều chiều (thường là hàng trăm chiều) nơi mà các từ ngữ có mối quan hệ ngữ nghĩa tương tự nhau sẽ có các vectơ gần nhau.

Word embedding đã trở thành một phần quan trọng trong nhiều ứng dụng xử lý ngôn ngữ tự nhiên, từ phân loại văn bản và phân tích tình cảm đến dịch máy và hỏi đáp tự động.

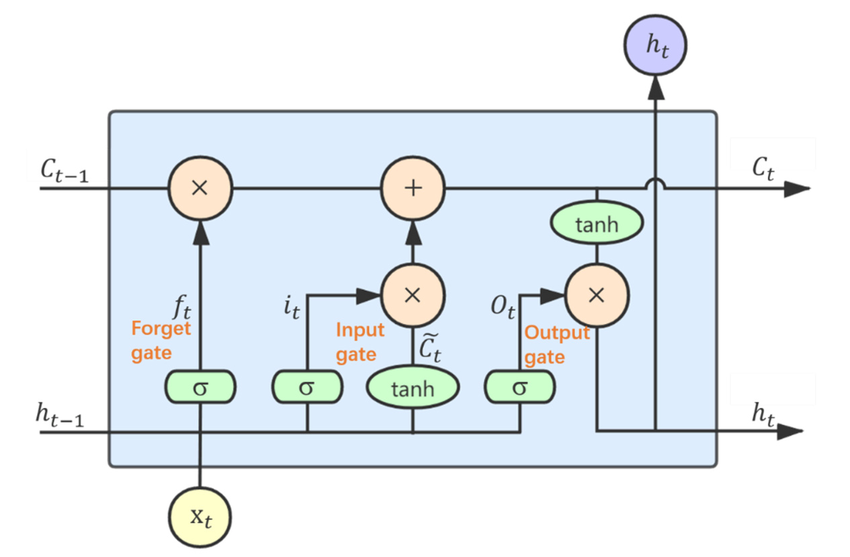


Hình . Word Embedding

## 5. LSTM

LSTM là một kiến trúc mạng thần kinh hồi quy (RNN) nhân tạo, được sử dụng để xử lý các dữ liệu tuần tự, như lời nói, văn bản, video, hoặc chuỗi thời gian.

Lớp LSTM có một tế bào (cell), một cổng vào (input gate), một cổng ra (output gate), và một cổng quên (forget gate). Tế bào ghi nhớ các giá trị trong các khoảng thời gian bất kỳ, và ba cổng điều chỉnh luồng thông tin ra/vào tế bào. Các cổng sử dụng các hàm kích hoạt như sigmoid hoặc tanh để quyết định xem nên giữ hay bỏ đi những thông tin nào.



Hình . Cấu trúc của LSTM

Cách thức hoạt động của LSTM có thể được minh họa như sau:

- Cho một chuỗi đầu vào X = (x1, x2, …, xn), mỗi phần tử xi là một vector đặc trưng của một đơn vị tuần tự, như một từ, một âm thanh, hoặc một khung hình.

- Lớp LSTM tính toán các trạng thái ẩn (hidden states) hi (i = 1, 2, …, n) theo công thức:

* , đây là cổng quên, quyết định xem nên giữ lại bao nhiêu thông tin từ trạng thái ẩn trước đó và tế bào trước đó.
* , đây là cổng vào, quyết định xem nên thêm vào bao nhiêu thông tin mới từ đầu vào hiện tại và trạng thái ẩn trước đó.
* , đây là tế bào, cập nhật giá trị của tế bào dựa trên cổng quên và cổng vào.
* , cổng ra, quyết định xem nên đưa ra bao nhiêu thông tin từ tế bào hiện tại và trạng thái ẩn trước đó.
* , trạng thái ẩn, đại diện cho thông tin đã học được từ các đầu vào trước đó từ trái sang phải.

# Chương 3: KẾT QUẢ

## 3.1 Quá trình thu nhập dữ liệu

Dữ liệu được thu nhập trên nền tảng Reddit thông qua api do Reddit cung cấp. Đầu tiên, ta cần tải thư viện praw để sử dụng trong Python thông qua lệnh:

pip install praw

Khi tạo tài khoản api, reddit sẽ cung cấp cho chúng ta 3 khóa để sử dụng tương tự như tài khoản và mật khẩu.

reddit = praw.Reddit(

    client\_id="khóa client id",

    client\_secret="Khóa client secret",

    user\_agent="script by r/TenNguoiDung"

)

Thực hiện chương trình sau để thu nhập dữ liệu bình luận về giá Bitcoin trên nền tảng Reddit:

def contains\_link(text):

url\_regex = re.compile(r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-\_@.&+] |[!\*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+'

    )

    if url\_regex.search(text):

        return True

    return False

comments = []

posts = reddit.subreddit("Bitcoin").search("Daily Discussion", sort="comments", limit=1000)

for post in posts:

    post.comments.replace\_more(limit=0)

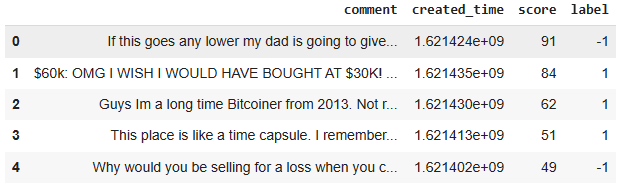
    for i, cmt in enumerate(post.comments.list()):

        if i > 300:

            break

        if not contains\_link(cmt.body):

            comments.append([cmt.body.replace("\n", " "), cmt.created\_utc, cmt.score])



Hình . Thu nhập dữ liệu từ API Reddit

## 3.2 Làm sạch dữ liệu

Làm sạch dữ liệu là bước quan trọng trong quá trình làm việc với văn bản. Dữ liệu sạch sẽ thì mô hình mới có thể học được những thông tin hữu ích.

Ở bài toán này, chúng tôi thực hiện làm sạch dữ liệu bao gồm:

* Loại bỏ ký tự đặc biệt và chuẩn hóa chữ thường.
* Thay thế các từ viết tắt.
* Loại bỏ số và ký tự đi kèm theo số.
* Loại bỏ những icon.
* Loại bỏ các từ quá dài.

#Từ điển chứa các từ viết tắt và nghĩa tương ứng

standfor = {"im" : "i am", "youre" : "you are","hes" : "he is",

            "shes" : "she is", "its" : "it is", "were" : "we are",

            "theyre" : "they are", "dont" : "do not",

            "doesnt" : "does not",  "didnt" : "did not",

            "cant" : "cannot","couldnt" : "could not",

            "shouldnt" : "should not", "wont" : "will not",

            "wouldnt" : "would not", "isnt" : "is not",

            "arent" : "are not", "hasnt" : "has not",

            "havent" : "have not", "hadnt" : "had not",

            "wasnt" : "was not", "werent" : "were not",

            "aint" : "be not", "y'all": "you all", "cuz": "because",

            "ill": "i will","hodl": "hold","u": "you",

            "itll": "it will","idk": "i do not know",

            "ive": "i have"}

# Hàm này loại bỏ giá trị nan

def remove\_nan(df):

    df = df.dropna(subset=['comment'])

    return df

#Hàm này thực hiện việc thay thế các từ viết tắt trong chuỗi s bằng nghĩa tương ứng trong từ điển standfor

def replace\_words(s):

    pattern = re.compile(r'\b(' + '|'.join(standfor.keys()) + r')\b')

    result = pattern.sub(lambda x: standfor[x.group()], s)

    return result

# Làm sạch dữ liệu

def clean\_data(sentences):

    new = []

    for sen in sentences:

        # Loại bỏ ký tự đặc biệt + chuẩn hóa chữ thường

        c =  re.sub(r"[^\w\s.,]", '', sen.lower())

        c = replace\_words(c)

        # i thành I

        c= re.sub(r'\bi\b', 'I', c)

        # Loại bỏ số và ký tự đi kèm theo số: vd như 30k

        c = re.sub('\w\*\d\w\*', '', c)

        # Loại bỏ những icon

        c = ''.join(s for s in c if s <= '\uFFFF')

        c =  c.encode('ascii', 'ignore').decode('ascii')

        # Loại bỏ các từ quá dài

        c = " ".join([w for w in c.split() if len(w) < 10])

        new.append(c)

    return pd.DataFrame(new, columns=["comments"])

## 3.3 Trích xuất khía cạnh

Chúng tôi thực hiện trích xuất khía cạnh bằng phương pháp luật văn bản. Đầu tiên, ta sử dụng hàm pos\_tag trong thư viện nltk để phân loại từ và hàm dependencies\_parse của thư viện StandfordCoreNLP để phân tích cú pháp.

tokens = word\_tokenize(sen) # Chia câu ra từng từ

pos = pos\_tag(tokens) # Phân loại pos tag cho các từ

dep = nlp.dependency\_parse(sen)

Tiếp theo, chúng tôi sử dụng những luật văn bản sau để trích xuất khía cạnh:

* Luật 1: Nếu một từ là danh từ và đứng sau một danh từ khác thì hai từ đó là khía cạnh.
* Luật 2: Nếu một từ là danh từ có mối quan hệ vị ngữ "obj" với 1 động từ thì từ đó là khía cạnh
* Luật 3: Nếu một từ là danh từ có mối quan hệ chữ ngủ "nsubj" với 1 tính từ thì từ đó là khía cạnh
* Luật 4: Nếu một từ là danh từ có mối quan hệ bổ ngữ "nmod" với 1 động từ thì từ đó là khía cạnh

def get\_dep(dependencies, idx):

    for dep in dependencies:

        if dep[2] == (idx + 1):

            return dep

    return None

aspects\_arr = []

for i in range(len(tokens)):

if len(tokens) != len(dep):

        break

    if pos[i][1].startswith("N"):

    # Luật 1::

    if i > 0 and pos[i - 1][1].startswith("N"):

    aspects\_arr.append(tokens[i - 1] + " " + tokens[i])

    # Luật 2::

    elif get\_dep(dep, i)[0] == "obj" and pos[get\_dep(dep, i)[1] - 1][1].startswith("V"):

        aspects\_arr.append(tokens[i])

    # Luật 3::

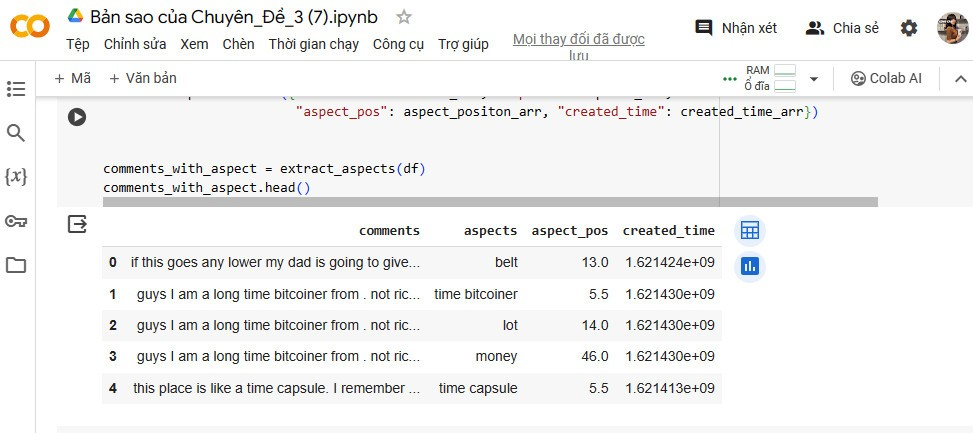
    elif get\_dep(dep, i)[0] == "nsubj" and pos[get\_dep(dep, i)[1] - 1][1].startswith("J"):

        aspects\_arr.append(tokens[i])

    # Luật 4::

    elif get\_dep(dep, i)[0] == "nmod" and pos[get\_dep(dep, i)[1] - 1][1].startswith("V"):

        aspects\_arr.append(tokens[i])



Hình . Trích xuất khía cạnh

## 3.4 Phân tích tình cảm cấp khía cạnh bằng phương pháp từ vựng

Sau khi đã trích xuất được khía cạnh, gán điểm tình cảm và thực hiện kiểm soát các mệnh để *nhưng*, chúng tôi thực hiện phân tích tình cảm bằng công thức sau:

Trong đó:

* sej là từ ngữ diễn đạt tình cảm trong câu s
* 𝑑𝑖𝑠𝑡(𝑠𝑒𝑗, 𝑎𝑖) là khoảng cách từ giữa khía cạnh ai với sej trong câu s
* 𝑠𝑒𝑗.𝑠𝑠 là điểm tình cảm của sej

Nếu điểm tổng hợp là dương thì tình cảm là tích cực, là âm thì tình cảm là tiêu cực, còn lại là trung lập.

import numpy as np

import math

import textblob

# hàm tìm tất cả tính từ trong dataset

def find\_words\_feeling(df):

    words\_feeling = []

    for index in range(len(df)):

        sen = df['comments'][index]

        sen = re.sub(r'\.', '', sen)

        tokens = word\_tokenize(sen)

        pos = pos\_tag(tokens)

        #Tìm các tính từ chỉ tình cảm

        for i in range(len(tokens)):

            if pos[i][1].startswith("J"):

              if[pos[i][0]] not in words\_feeling:

                  words\_feeling += [pos[i][0]]

    return words\_feeling

# hàm thực hiện phân tích cảm xúc

def sentiment\_analysis\_1(sentence, aspects, aspects\_position):

    score = 0

    words = sentence.split()

    for index, word in enumerate(sentence.split()):

      if(word in word\_score.keys()):

        score = score + word\_score[word]/(math.floor(abs(aspects\_position-index))+0.001)

    return score

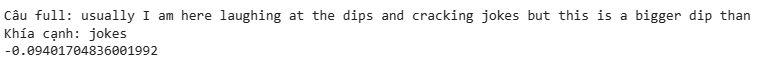
words\_feeling = find\_words\_feeling(comments\_with\_aspect) # Các tính từ trong df

word\_score = {} # Bộ từ vựng chỉ cảm xúc và điểm tương ứng

for word in words\_feeling:

  blob = textblob.TextBlob(word)

  word\_score[word] = blob.polarity

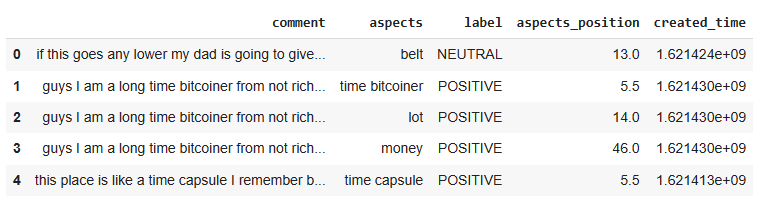


Hình . Phân tích tình cảm bằng phương pháp từ vựng

## 3.5 Phân tích tình cảm cấp khía cạnh bằng phương pháp học sâu

**3.5.1 Tiền xử lý dữ liệu**

Dữ liệu bao gồm các bình luận và các khía cạnh đã được gán nhãn theo phương pháp từ vựng.



Hình . Dữ liệu đã được gán nhãn

Chúng tôi tiến hành thay đổi nhãn thành định dạng số để dễ dàng xử lý

## thay đổi nhãn thành số nguyên cho giai đoạn huấn luyện

df['label'][df["label"] == 'NEGATIVE'] = 0

df['label'][df["label"] == 'POSITIVE'] = 1

df['label'][df["label"] == 'NEUTRAL'] = 2

Tiếp theo, chúng ta sẽ tạo ra 1 tokenizer để xử lý văn bản, loại bỏ các ký tự được chỉ định trong filters và chuyển đổi tất cả các ký tự thành chữ thường.

tokenizer = Tokenizer(num\_words=NUM\_WORDS,

                      filters='!"#$%&()\*+,-./:;<=>?@[\\]^\_`{|}~\t\n',

                      lower=True, )

## Cập nhật từ vựng nội bộ dựa trên danh sách comment.

tokenizer.fit\_on\_texts(list(X\_train.comment))

text\_X\_train\_tokenized = tokenizer.texts\_to\_sequences(X\_train.comment) #

Aspect\_X\_train\_tokenized = tokenizer.texts\_to\_sequences(X\_train['aspects'])

text\_X\_test\_tokenized = tokenizer.texts\_to\_sequences(X\_test.comment

Aspect\_X\_test\_tokenized = tokenizer.texts\_to\_sequences(X\_test['aspects'

Thực hiện padding nhằm đưa các câu về cùng 1 độ dài trước khi training.

##Maximum length of sentences

Max\_Len =  max([len(one\_title) for one\_title in text\_X\_train\_tokenized])

## Padding sequences

text\_X\_train\_padded = pad\_sequences(text\_X\_train\_tokenized, maxlen=Max\_Len)

aspect\_X\_train\_padded = pad\_sequences(Aspect\_X\_train\_tokenized, maxlen=1)

text\_X\_test\_padded = pad\_sequences(text\_X\_test\_tokenized, maxlen=Max\_Len)

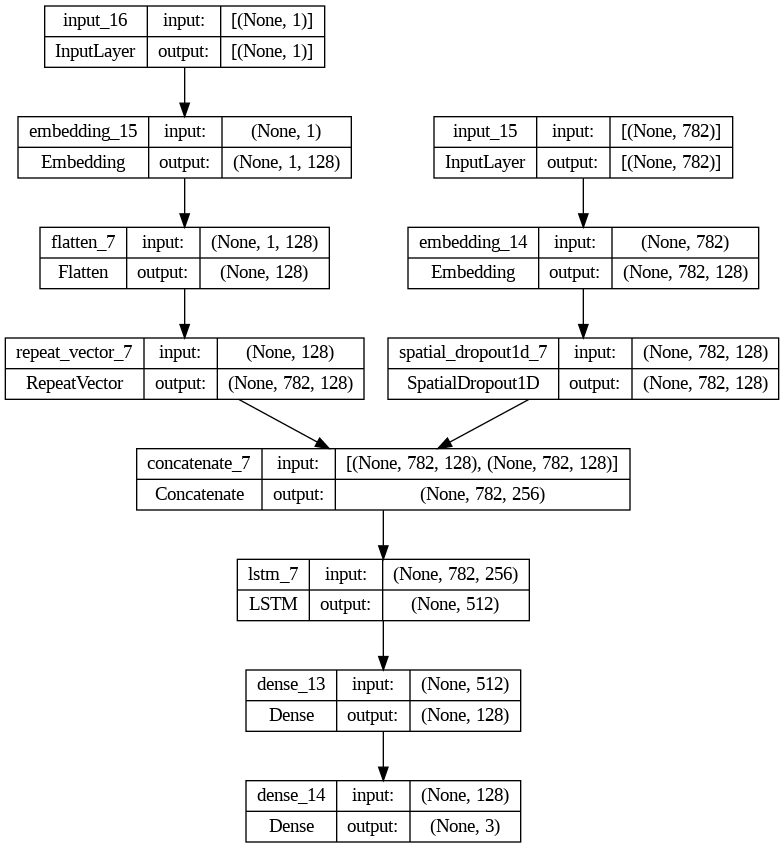
aspect\_X\_test\_padded = pad\_sequences(Aspect\_X\_test\_tokenized, maxlen=1)

**3.5.2 Xây dựng mô hình**

Mô hình sẽ xử lý câu bình luận và từ khía cạnh riêng lẻ và sau đó kết hợp chúng lại với nhau:

* Câu bình luận sẽ được biến đổi thành vector thông qua lớp Embedding. Sau đó, đầu ra sẽ được đưa qua lớp SpatialDropout giúp giảm thiểu hiện tượng overfiting.
* Từ khía cạnh được đưa vào lớp Embedding và biến đổi thành vector, sau đó được làm phẳng tại lớp Flatten. Tiếp tục, đầu ra sẽ được đưa vào lớp RepeatVector, lặp lại một số lần nhất định để có cùng kích thước với câu bình luận.
* Kết hợp câu bình luận và từ khía cạnh thông qua lớp Concatenate.
* Đưa vector kết hợp qua 1 lớp LSTM và lớp Fully Connector để trích xuất đặc trưng và dự đoán kết quả.

Cấu trúc mô hình được thể hiện thông qua hình dưới:



Hình . Cấu trúc mô hình phân ltích cảm xúc

def ae\_lstm(lstm\_units = 512):

    # Định nghĩa đầu vào cho văn bản và phần aspect

    input\_text = Input(shape=(Max\_Len,))

    input\_aspect = Input(shape=(1,),)

    # Tạo lớp embedding cho văn bản

    word\_embedding = Embedding(NUM\_WORDS, EMBEDDING\_SIZE, input\_length=Max\_Len)

    text\_embed = SpatialDropout1D(0.2)(word\_embedding(input\_text))

    # Tạo lớp embedding cho phần aspect

    asp\_embedding = Embedding(NUM\_WORDS, EMBEDDING\_SIZE, input\_length=Max\_Len)

    aspect\_embed = asp\_embedding(input\_aspect)

    # Đưa aspect về dạng vector 2 chiều

    aspect\_embed = Flatten()(aspect\_embed)

    repeat\_aspect = RepeatVector(Max\_Len)(aspect\_embed

    # Kết hợp văn bản và aspect

    input\_concat = concatenate([text\_embed, repeat\_aspect], axis=-1)

    # Xử lý LSTM cho dữ liệu kết hợp

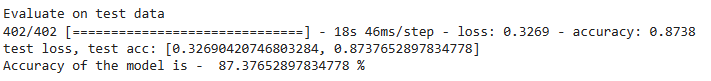
    hidden = LSTM(lstm\_units)(input\_concat)

    Dense\_layer  = Dense(128, activation='tanh')(hidden)

    output\_layer = Dense(3, activation='softmax')(Dense\_layer)

    return Model([input\_text, input\_aspect], output\_layer)

Độ chính xác trên tập test đạt được 87%.



Hình . Độ chính xác

# KẾT LUẬN

**1. Kết quả đạt được**

Qua quá trình nỗ lực, cả nhóm đã đạt được những thành tựu sau:

* Làm sạch dữ liệu.
* Xác định được các khía cạnh
* Xác định được tình cảnh theo khía cạnh
* Trả lời được các câu hỏi kinh doanh đã đặt ra
* Xây dựng được mô hình học sâu dự đoán tình cảm theo khía cạnh
* Hoàn thành báo cáo

Tuy nhiên, chúng tôi nhận thấy vẫn còn vài hạn chế như:

* Gán nhãn dữ liệu chưa chính xác cao
* Dữ liệu huấn luyện vẫn còn ít
* Chưa cải thiện được độ chính xác của phương pháp từ vựng

**2. Hướng phát triển**

Qua những hạn chế còn tồn đọng đã nêu trên, hướng phát triển dự kiến sẽ là:

* Thực hiện gán nhãn dữ liệu bằng tay hoặc sử dụng các phương pháp nâng cao hơn
* Thu nhập dữ liệu từ nhiều mạng xã hội khác nhau như: reddit, facebook, twitter,…
* Nghiên cứu cải thiện độ chính xác của phương pháp từ vựng

# PHÂN CÔNG CÔNG VIỆC

***Thành viện:***

* Lê Văn Phước, MSSV: 20IT487
* Hồ Luận, MSSV: 20IT669
* Võ Văn Nguyên, MSSV: 20IT325
* Phan Thị Kim Ngân, MSSV: 20IT225
* Nguyễn Đỗ Quang Minh, MSSV: 20IT832

***Công việc:***

|  |  |  |
| --- | --- | --- |
| Thành Viên | Công Việc | Mức độ đóng góp |
| Lê Văn Phước | + Phân nhiệm vụ và hướng dẫn thực hiện cho các thành viên có mặt tại các buổi hẹn  + Clean data  + Trích xuất các khía cạnh  + Trích xuất, gắn nhãn cho các từ chỉ tình cảm  + Gắn nhãn cho các bình luận theo khía cạnh  + Trực quan hóa dữ liệu, trả lời các câu hỏi kinh doanh  + Tìm hiểu và trainning các mô hình LSTM, LSTM-ATTENSION đưa ra nhận xét  + Đọc hiểu code | 30% |
| Hồ Luận | + Thu thập dữ liệu  + Clean data  + Trích xuất các khía cạnh  + Trích xuất, gắn nhãn cho các từ chỉ tình cảm  + Gắn nhãn cho các comment theo khía cạnh  + Làm slide, báo cáo  + Đọc hiểu code | 23% |
| Võ Văn Nguyen | + Làm đề xuất, báo cáo  + Trích xuất các khía cạnh  + Trích xuất, gắn nhãn cho các từ chỉ tình cảm  + Gắn nhãn cho các comment theo khía cạnh  + Đọc hiểu code | 16% |
| Phan Thị Kim Ngân | + Làm đề xuất, báo cáo  + Trích xuất, gắn nhãn cho các từ chỉ tình cảm  + Gắn nhãn cho các comment theo khía cạnh  + Trả lời các câu hỏi  + Đọc hiểu code | 16% |
| Nguyễn Đỗ Quang Minh | + Làm đề xuất, báo cáo  + Trích xuất, gắn nhãn cho các từ chỉ tình cảm  + Gắn nhãn cho các comment theo khía cạnh  + Đọc hiểu code | 16% |

# DANH MỤC TÀI LIỆU THAM KHẢO

[1] Slides Phân tích tình cảm trong kinh doanh của TS. Đặng Trung Thành

[2] Mô hình học sâu trên [www.kaggle.com](http://www.kaggle.com)

[3] Erik Cambria Dipankar Das Sivaji Bandyopadhyay Antonio Feraco Editors A Practical Guide to Sentiment Analysis

[4] Aakanksha Sharaff (editor), G R Sinha (editor), Surbhi Bhatia (editor) - New Opportunities for Sentiment Analysis and Information Processing-Engineering Science Reference (2021)

[5] Blog “Aspect Based Sentiment Analys is python” trên enjoyalgorithms.com