TRƯỜNG ĐẠI HỌC NGOẠI NGỮ-TIN HỌC TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

**KHÓA LUẬN TỐT NGHIỆP**

**PHÂN TÍCH CẢM XÚC DỰA THEO LUẬT**

GIẢNG VIÊN HƯỚNG DẪN: TS. Trần Khải Thiện

SINH VIÊN THỰC HIỆN: Đinh Phú Cường – 20DH111722

*TP. Hồ Chí Minh, tháng 05 năm 2024*

# LỜI CẢM ƠN

Em xin dành lời cảm ơn đến với thầy, cô khoa công nghệ thông tin trong trường Đại học Ngoại Ngữ Tin Học Thành phố Hồ Chí Minh đã tạo cơ hội cho em được học tập, rèn luyện, tích luỹ kiến thức và kỹ năng để em có thể hoàn thành khoá luận tốt nghiệp này.

Em xin gửi lời cảm ơn đến thầy hướng dẫn TS. Trần Khải Thiện đã dành nhiều thời gian và công sức theo dỗi và đưa ra những hướng dẫn, lời khuyên bổ ích cho em hoàn thành khoá luận.

Do kiến thức của bản thân còn hạn chế và thiếu kinh nghiệm nên nội dung khoá luận không thể tránh khỏi thiếu sót. Em rất mong nhận được sự góp ý từ thầy, cô để em có thể cải thiện hơn.

Em xin chân thành cảm ơn!

# LỜI CAM ĐOAN

Em xin cam đoan đề tài “Phân tích cảm xúc dựa theo luật” là bài của cả nhân em, dựa vào sự cố gắng vả nỗ lực của mình dưới sự hướng dẫn của giảng viên hướng dẫn TS. Trần Khải Thiện.

Em xin hoàn toàn chịu trách nhiệm về tính trung thực của nội dung trong đề tài của mình.

TP. HCM, ngày….tháng….năm….

Đinh Phú Cường

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc167371198)

[LỜI CAM ĐOAN ii](#_Toc167371199)

[MỤC LỤC iii](#_Toc167371200)

[DANH MỤC BẢNG BIỂU v](#_Toc167371201)

[DANH MỤC HÌNH ẢNH vi](#_Toc167371202)

[DANH MỤC TỪ VIẾT TẮT vii](#_Toc167371203)

[CHƯƠNG 1. GIỚI THIỆU CHUNG 1](#_Toc167371204)

[1.1. Mở đầu 1](#_Toc167371205)

[1.2. Phạm vi của khoá luận 4](#_Toc167371206)

[1.3. Mục đích của khoá luận 4](#_Toc167371207)

[1.4. Cấu trúc của khoá luận 4](#_Toc167371208)

[CHƯƠNG 2. SƠ LƯỢC CÁC PHƯƠNG PHÁP PHÂN TÍCH CẢM XÚC 5](#_Toc167371209)

[2.1 Phân tích cảm xúc 5](#_Toc167371210)

[2.1.1 Phân tích cảm xúc là gì? 5](#_Toc167371211)

[2.1.2 Tại sao phải phân tích cảm xúc? 5](#_Toc167371212)

[2.1.3 Những thách thức đối với bài toán phân tích cảm xúc 6](#_Toc167371213)

[2.1.4 Phân tích cảm xúc mức văn bản 7](#_Toc167371214)

[2.2 Phương pháp tiếp cận học máy 7](#_Toc167371215)

[2.1.1 Phân lớp Naïve Bayes 8](#_Toc167371216)

[2.1.2 Support Vector Machine (SVM) 10](#_Toc167371217)

[2.3 Phương pháp tiếp cận dựa trên học sâu 12](#_Toc167371218)

[2.3.1. Convolutional Neural Network (CNN) 13](#_Toc167371219)

[2.3.2. Recurrent Neural Network (RNN) 16](#_Toc167371220)

[2.4 Phương pháp tiếp cận dựa trên từ điển 20](#_Toc167371221)

[2.5 Kết luận chương 22](#_Toc167371222)

[CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT 23](#_Toc167371223)

[3.1. Mô hình đề xuất 23](#_Toc167371224)

[3.2. Xây dựng các từ điển cảm xúc 26](#_Toc167371225)

[3.3. Tập dữ liệu 29](#_Toc167371226)

[3.4. Thực nghiệm 30](#_Toc167371227)

[3.4.1. Cấu hình 30](#_Toc167371228)

[3.4.2. Tiền xử lý dữ liệu 30](#_Toc167371229)

[3.4.3. Thực hiện mô hình phân tích cảm xúc bằng phương pháp dựa trên từ điển. 35](#_Toc167371230)

[3.4.4. Đánh giá thực nghiệm 36](#_Toc167371231)

[CHƯƠNG 4. KẾT LUẬN 39](#_Toc167371232)

[TÀI LIỆU THAM KHẢO 41](#_Toc167371233)

# DANH MỤC BẢNG BIỂU

[Bảng 1. Từ điển cảm xúc 25](#_Toc168125236)

[Bảng 2. Từ điển phủ định 26](#_Toc168125237)

[Bảng 3. Từ điển sửa lỗi 27](#_Toc168125238)

[Bảng 4. Từ điển token 28](#_Toc168125239)

[Bảng 5. Từ điển stop word 28](#_Toc168125240)

[Bảng 6. Tập dữ liệu UIT 29](#_Toc168125241)

[Bảng 7. Tập dữ liệu UIT\_Final 35](#_Toc168125242)

[Bảng 8. Bảng đánh giá thực nghiệm 37](#_Toc168125243)

# DANH MỤC HÌNH ẢNH

[Hình 2.1. Phân tích cảm xúc bằng phương pháp Naive Bayes [10] 10](#_Toc167352598)

[Hình 2.2. Mô tả siêu phẳng trong không gian hai chiều 11](#_Toc167352599)

[Hình 2.3. Mô tả về margin trong SVM 12](#_Toc167352600)

[Hình 2.4. Convolution với bộ lọc 3x3 [14] 13](#_Toc167352601)

[Hình 2.5. Sơ đồ hoạt đồng CNN 14](#_Toc167352602)

[Hình 2.6. Cấu trúc mạng CNN trong xử lý ngôn ngữ tự nhiên. [15] 15](#_Toc167352603)

[Hình 2.7. Sơ đồ RNN [16] 16](#_Toc167352604)

[Hình 2.8. Các loại bài toán trong RNN [17] 17](#_Toc167352605)

[Hình 2.9. Mô-đun lặp lại trong RNN tiêu chuẩn chứa một lớp duy nhất [18] 18](#_Toc167352606)

[Hình 2.10. Mô-đun lặp lại LSTM chứa 4 lớp [18] 18](#_Toc167352607)

[Hình 2.11. Ký hiệu trong LSTM [18] 19](#_Toc167352608)

[Hình 2.12. Trạng thái tế bào [18] 19](#_Toc167352609)

[Hình 2. 13. Cổng quên nhân với vector [18] 20](#_Toc167352610)

[Hình 2. 14. LSTM-CNN [8] 20](#_Toc167352611)

[Hình 2.15. Positive 21](#_Toc167352612)

[Hình 2.16. Negative 21](#_Toc167352613)

[Hình 3.1. Mô hình phân tích cảm xúc dựa trên từ điển 24](#_Toc167352410)

# DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Viết tắt** | **Đầy đủ** |
| AI | Artificial Intelligence |
| NLP | Natural language processing |
| BoW | Bag of Word |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| SVM | Support Vector Machine |
| NB | Naïve Bayes |
| ME | Maximum Entropy |
| CNN | Convolutional Neural Network |
| RNN | Recurrent neural network |
| LSTM | Long short-term memory |

# CHƯƠNG 1. GIỚI THIỆU CHUNG

1. **Mở đầu**

Trong giai đoạn đại dịch COVID-19, các hoạt động ngoài trời đều bị phong toả, cách duy nhất mà con người có thể kết nối với nhau đó là thông qua mạng xã hội. Chính phủ sử dụng mạng xã hội như cầu nối với người dân. Tất cả thông tin về dịch bệnh, vùng cách ly, tiêm phòng, tuyên truyền chống dịch, cảnh báo,…mọi người dân đều dễ dàng nắm bắt được. Quan tâm đến cảm xúc của người dân trên mạng xã hội cũng là điều mà chính phủ quan tâm, ví dụ như trong giai đoạn cách ly có rất nhiều thông tin sai lệch, bịa đặt làm ảnh hưởng đến công tác phòng chống dịch. Mọi người khi đối mặt với các thông tin trên cảm thấy lo lắng, bất an. Sau khi biết được cảm xúc của mọi người, chính phủ đã lên tiếng trấn an cũng như điều tra về những thông tin bịa đặt, không có chứng cứ.

Sau khi đại dịch kết thúc, lượng người sử dụng mạng xã hội không giảm đi mà còn có xu hướng tăng liên tục qua từng năm. Theo thống kê 2019 [1] có khoảng 58 triệu người sử dụng mạng xã hội, chiếm 60% dân số Việt Nam, đây là giai đoạn trước khi đại dịch COVID-19. Vào năm 2023, sau một năm lệnh cách ly được gỡ bỏ, trong báo cáo năm 2023 lượng người sử dụng mạng xã hội đã tăng lên thêm 12 triệu người chiếm 71% dân số Việt Nam [2]. Nắm bắt được tâm lý của phần lớn của những người sử dụng mạng xã hội là một điều tốt cho công tác quản lý của chính phủ. Các doanh nghiệp mong muốn hiểu được khách hàng của mình, trải nghiệm của khách hàng về sản phẩm hay dịch vụ, từ đó cải thiện sản phẩm, xác định được xu hướng và chiến lược trong tương lai. Các nhà quảng cáo và marketing có thể nắm bắt được sở thích để tạo ra những nội dung quảng cáo phù hợp.

Những ứng dụng mạng xã hội, mua hàng online, xe ôm công nghệ hay giao đồ ăn đã trở nên phổ biến đối với mọi người. Những dịch vụ này có một tiêu chuẩn đánh chung đó là đánh giá ‘like’ là thích hay ‘dislike’ là không thích hay từ một sao đến năm sao, năm sao là cao nhất thể hiện thái độ hài lòng của khách hàng trong khi một sao là thấp nhất thể hiện sự không hài lòng của khách hàng. Ngoài ra, người dùng cũng có thể nêu ra ý kiến của mình thông qua những bình luận. Cách này giúp cho các doanh nghiệp có thể quản lý được chất lượng dịch vụ của mình. Trong những ứng dụng mua hàng, ngoài việc xem giá, thông tin của sản phẩm thì người dùng còn quan tâm đến những đánh giá của những khách hàng đã mua và trải nghiệm sản phẩm, dịch vụ. Nếu sản phẩm có nhiều lượt đánh giá tích cực thì sản phẩm, dịch vụ sẽ được nhiều người biết đến nhiều hơn ngược lại thì sẽ không ai quan tâm nữa. Trong một số trường hợp ví dụ như các tài xế xe ôm công nghệ, các tài xế giao hàng khi nhận quá nhiều lượt đánh giá tiêu cực còn có thể khiến công việc của mình bị ảnh hưởng, họ có thể bị giảm thu nhập hay thậm chí mất việc .

Hiện nay, mọi thứ đang dần được hiện đại hoá, mọi vấn đề đều có thể xử lý bởi công nghệ. Trong những năm gần đây nổi trội nhất là sự phát triển của trí tuệ nhân tạo. Trí tuệ nhân AI (Artificial Intelligence) hay còn gọi là AI ra đời nhằm thay thế con người trong việc xử lý các vấn đề một cách tự động với hiệu quả cao và chính xác lớn.

Trong lĩnh vực trí tuệ nhân tạo có một nhánh được gọi là xử lý ngôn ngữ tự nhiên hay NLP (Natural language processing). Xử lý ngôn ngữ tự nhiên cho phép máy tính hiểu và xử lý ngôn ngữ con người. Ứng dụng của xử lý ngôn ngữ tự nhiên trong việc xử lý, phân loại hay phân tích tài liệu, sử dụng chatbox để trả lời khách hàng tự động. ChatGPT những năm gần đây đã trở nên phổ biến khi cho phép người dùng nhập một câu hỏi bất kì ChatGPT đều có thể trả lời được.

Sự phát triển của của học máy và xử lý ngôn ngữ tự nhiên (NLP) từ những năm 2000 đã mở ra cơ hội cho phân tích cảm xúc. Từ đó đến nay đã có hàng loạt các phương pháp học máy như Support Vector Machine (SVM), Naïve Bayes, các mô hình học sâu như Recurrent neural network (RNN), Convolutional Neural Networks (CNN), Long short-term memory (LSTM) và rất nhiều phương pháp khác được áp dụng cho phân tích dữ liệu văn bản. Từ năm 2010 đến nay phân tích cảm xúc đã trở thành một lĩnh vực nghiên cứu và ứng dụng phát triển mạnh mẽ trong nhiều lĩnh vực khác nhau, một số ứng dụng thực tế của phân tích cảm xúc là phân tích ý kiến từ các bình luận trên các mạng xã hội, dự đoán cảm xúc của khách hàng. Cho đến nay đã có rất nhiều đã có nhiều nghiên cứu và phát triển công nghệ NLP ở Việt Nam, cụ thể là các doanh nghiệp và tổ chức tập trung vào nghiên cứu phân tích cảm xúc khách hàng, các nhóm cộng đồng NLP ở Việt Nam ngày càng phát triển cùng với đó là các công cụ và thư viện NLP được hỗ trợ bởi người Việt như PyVi, VnCoreNLP,….

Việc sử dụng mạng xã hội đã trở nên phổ biến, mọi người bình luận hay tương tác trên mạng xã hội ngày càng nhiều, đánh giá của khách hàng về sản phẩm hay dịch vụ nào đó ngày một tăng. Do đó, việc hiểu được cảm xúc của người dùng là vấn đề rất quan trọng . Phân tích cảm xúc dựa theo quy luật giúp hiểu sâu về con người, cách họ phản ứng đối với một tình huống hay một sự kiện nào đó.

Phân tích cảm xúc cụ thể là phân tích cảm xúc của mọi người trên mạng xã hội là điều không thể thiếu đối với cá nhân hay bất kỳ tổ chức nào. Lượng lớn thông tin được cập nhật hàng ngày trên mạng xã hội đòi hỏi phải có một phương pháp giúp phân tích được cảm xúc trong một câu hoặc một đoạn văn bản. Hầu hết các phương pháp phân tích cảm xúc dựa trên luật đều dựa vào nhiều ngôn ngữ khác nhau, tuy nhiên đối với tiếng Việt thì còn nhiều hạn chế. Hiểu được người dân, hiểu được khách hàng là mục tiêu mà bất kỳ cá nhân hay tổ chức nào cũng nhắm đến để có thể phát triển trong tương lai.

Đề tài tập trung vào việc sử dụng quy tắc ngôn ngữ và ngữ pháp để đánh giá và phân loại cảm xúc trong văn bản. Thay vì dựa vào máy học, phân tích cảm xúc dựa trên luật sử dụng các quy tắc ngôn ngữ được xây dựng thủ công để hiểu và đánh giá cảm xúc, giúp phân loại văn bản thành các danh mục như tích cực, tiêu cực, hoặc trung tính dựa trên ngữ cảnh và cấu trúc ngôn ngữ.

Trong bài luận này sẽ tập trung vào phân tích cảm xúc bằng phương pháp ngữ nghĩa. Đây là kỹ thuật được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP) để xác định cảm xúc của một đoạn văn bản. Phương pháp này sử dụng từ hoặc cụm từ trong từ vựng hoặc từ điển được gán nhãn cảm xúc khác nhau ví dụ như tích cực, tiêu cực, trung tính.

1. **Phạm vi của khoá luận**

Phân tích cảm xúc của con người thông qua văn bản là một thách thức đặc biệt đối với ngôn ngữ tiếng Việt còn khá nhiều hạn chế. Phạm vi đề tài của em sẽ tập trung vào bài toán nhỏ phân tích cảm xúc bằng phương pháp dựa trên từ điển.

1. **Mục đích của khoá luận**

Mục đích của khoá luận là trình bày về phân tích cảm xúc, tìm hiểu về các phương pháp phân tích cảm xúc như phương pháp học máy, học sâu và phương pháp dựa vào từ điển. Trong khoá luận, em sử dụng phương pháp dựa trên từ điển để giải quyết bài toán phân tích cảm xúc tích cực hay tiêu cực cho tập dữ liệu đánh giá, nhận xét của sinh viên Việt Nam.

1. **Cấu trúc của khoá luận**

Bài luận gồm có bốn chương. Trong đó, chương một giới thiệu chung những vấn đề hiện nay, sự phát triển của phân tích cảm xúc trong NLP về phân tích cảm xúc. Chương hai trình bày sơ lược về phương pháp phân tích cảm xúc, tìm hiểu về phân tích cảm xúc, tại sao lại cần phải phân tích cảm xúc, thách thức và một số phương pháp phân tích cảm xúc. Chương tiếp theo, đưa ra phương pháp và mô hình đề xuất, xây dựng từ điển cảm xúc và tập dữ liệu, cuối cùng là đưa ra kết quả và đánh giá thực nghiệm

# CHƯƠNG 2. SƠ LƯỢC CÁC PHƯƠNG PHÁP PHÂN TÍCH CẢM XÚC

## Phân tích cảm xúc

### Phân tích cảm xúc là gì?

Phân tích cảm xúc là một kỹ thuật NLP nhằm xác định các mẫu thông tin và đặc trưng trong một văn bản lớn. Tập trung phân tích thái độ, quan điểm, nhận xét, yêu cầu, ý kiến, câu hỏi và sở thích được tác giả truyền đạt trong văn bản dựa trên cảm xúc mà không bị ảnh hưởng bởi lý do hoặc lập luận logic. Cảm xúc của tác giả thường thông qua các bài đăng trên blog, đánh giá sản phẩm, diễn đàn trực tuyến, bài phát biểu, tài liệu, nguồn cơ sở dữ liệu, mạng xã hội. Trong đó gồm có ba yếu tố:

* Ý kiến hoặc cảm xúc: ý kiến được dựa trên tính phân cực (polarity), trong đó cảm xúc hạnh phúc, vui vẻ, buồn bã, tức giận, ngạc nhiên,… được xác định là tiêu cực, tích cực hay trung tính. Ví dụ: “Tôi không thích bộ quần áo này”, “Tôi ghét xem phim này”, “Món này ăn rất ngon” hoặc đánh giá cảm xúc bằng thang đo điểm ví dụ như là đánh giá một dịch vụ thông qua số sao hay xếp hạng một bộ phim theo thang điểm từ một đến mười.
* Chủ đề: một chủ đề có thể được đề cập và thảo luận theo nhiều khía cạnh khác nhau.
* Người đưa ra ý kiến: dùng để chỉ tác giả hoặc người đưa ra ý kiến

### Tại sao phải phân tích cảm xúc?

* Trước nhất là để hiểu quan điểm và thái độ của người dùng: phân tích cảm xúc giúp hiểu rõ quan điểm, thái độ của người dùng đối với một sản phẩm, dịch vụ hay sự kiện. Điều này giúp các doanh nghiệp, tổ chức hiểu rõ được khách hàng và đưa ra phương hướng phát triển và chiến lược kinh doanh phù hợp.
* Phát hiện xu hướng và ý kiến chung: phân tích cảm xúc thông qua các trang mạng xã hội, diễn đàn trực tuyến về những lượt bình luận thì phát hiện xu hướng và ý kiến chung của một cộng đồng về một chủ đề.
* Tăng cường trải nghiệm người dùng: hiểu và nắm bắt được tâm lý người dùng giúp các doanh nghiệp, tổ chức phát triển, cải thiện và tối ưu hoá trải nghiệm người dùng. Ví dụ như một ứng dụng về đặt đồ ăn thông qua tìm hiểu khách hàng thì phát triển thêm tính năng mới như định vị được quán ăn được đánh giá tốt trong một bán kính cụ thể , cải thiện về chất lượng nhân sự khi giao đồ ăn được nhanh hơn hay tối ưu hoá về giao diện ứng dụng dễ sử dụng phù hợp cho những người không rành về công nghệ.
* Quản lý hình ảnh và thương hiệu: theo dõi và quản lý hình ảnh thương hiệu trên các nền tảng trực tuyến giúp nắm bắt ý kiến và cảm xúc của khách hàng phản ứng kịp thời đối với những điều tiêu cực và phát triển hình ảnh đối với những phản hồi tích cực.
* Dự báo và đánh giá xu hướng thị trường: phân tích cảm xúc có thể dự báo và đánh giá xu hướng thị trường, từ đó giúp các doanh nghiệp, tổ chức có thể phát triển chiến lược kinh doanh trong tương lai.

### Những thách thức đối với bài toán phân tích cảm xúc

* Những nghiên cứu về phân tích cảm xúc bằng tiếng Việt còn rất hạn chế.
* Trong những tài liệu văn bản, giọng điệu và ngữ cảnh rất khó có thể hiện được trong văn bản. Điều này có thể ảnh hưởng đến phân tích cảm xúc. Ví dụ như: “Trời hôm nay đẹp quá”, trong câu này ta có thể dễ dàng nhận định đây là một câu tích cực nhưng câu này cũng có thể xem đây là một câu tiêu cực nếu có thể xét yếu tố giọng điệu nhấn mạnh vào “đẹp quá” và bối cảnh là một bầu trời âm u.
* Trên các trang mạng xã hội khi người dùng bình luận trên đó cũng gây ra một số thách thức đối với phân tích cảm xúc như các nội dung bình luận, tán gẫu thường rất ngắn và cảm xúc người dùng thường bị ảnh hưởng bởi nội dung mà họ đọc hoặc nghe. Ngoài ra, việc viết tắt hay sai chính tả chẳng hạn như từ “Không” nguời dùng có thể viết thành “ko” hay “o” hoặc “k”. Xử lý những vấn đề này là một điều thách thức lớn.
* Trên mạng xã hội còn có các biểu tượng cảm xúc (emoji), các ảnh động (gif), ảnh chế (meme) để thể hiện cảm. Nắm bắt được cảm xúc của người sử dụng cũng là một vấn đề không nhỏ.

### Phân tích cảm xúc mức văn bản

Trong ‘Một số khái niệm và hướng tiếp cận phân tích cảm xúc - Áp dụng cho tiếng Việt’ đề cập đến phân tích cảm xúc mức văn bản [3], sự thống trị của các phương pháp học máy đã thống trị hầu hết các bài toán phân tích cảm xúc, đặc biệt là bài toán mức văn bản bởi biểu diễn đặc trưng có tác động lớn hiệu năng của giải thuật học máy [4]. Trong công trình [5] [6], thực nghiệm với ba phương pháp học mày truyền thống như Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME) cho ra kết quả tốt với phân lớp nhị phân hay phân lớp nhiều cấp độ. D. Nguyen và cộng sự [7] sử dụng mạng CNN để phân lớp 25.991 nhận xét sản phẩm tiếng Việt đạt kết quả 84.40%. Tác giả Q.Vo và cộng sự [8] nhận ra mạng kết hợp LSTM – CNN hoạt động tốt khi phân lớp 17.500 nhận xét tiếng Việt theo ba mực khen, chê, trung tính. Lý do có kết quả tốt là dựa vào giải thuật biểu diễn (representation learning) hay còn gọi là học đặc trưng (feature learning) như các hướng tiếp cận theo học sâu, kỹ thuật tự động phân biệt và giải thích các biểu diện văn bản từ dữ liệu.

## Phương pháp tiếp cận học máy

Học máy (Machine Learning) [9] là một lĩnh vực trong trong AI, sử dụng thuật toán để huấn luyện trên các tập dữ liệu cho phép máy móc thực hiện các nhiệm vụ mà lẽ ra con người mới có thể thực hiện được. Một số loại thuật toán học máy phổ biến:

1. Học máy có giám sát (Supervised learning): là học từ dữ liệu được gán nhãn sẵn để dự đoán dữ liệu mới.
2. Học máy không giám sát (Unsupervised learning): không sử dụng tập dữ liệu được gán nhãn sẵn, thuật toán học không giám sát nhóm các dữ liệu dựa trên những điểm tương đồng .
3. Học máy bán giám sát (Unsupervised learning): kết hợp giữa hai phương pháp học có giám sát và học không giám sát.
4. Học tăng cường (Reinforcement learning): là một loại kỹ thuật mà hệ thống tự tìm hiểu và cải thiện hành vi của mình thông qua tương tác với môi trường

Các mô hình học máy được huấn luyện để phân loại các lớp cảm xúc tích cực, tiêu cực hay trung tính như một số thuật toán phân loại Support Vector Machine (SVM), Naïve Bayes,…

Để thực hiện phương pháp này gồm có các bước:

* Thu thập dữ liệu: Thu thập dữ liệu từ các trang mạng xã hội, diễn đàn,….Để thực hiện bước này có thể sử dụng cơ sở dữ liệu của trang web hoặc dùng Crawler để tự động thu thập dữ liệu.
* Tiền xử lý dữ liệu: Sau khi thực hiện thu thập dữ liệu sẽ tiến hành xử lý dữ liệu. Bước này giúp tách từ (tokenizing), loại bỏ những từ (stop\_word) hay kí tự thừa thải và khoảng trống không cần thiết. Chuẩn hoá dữ liệu bằng cách biến tất cả chữ hoa thành chữ thường, sửa các câu bình luận sai chính tả.
* Trích xuất đặc trưng: Dữ liệu sau khi được tiền xử lý sẽ được biểu diễn dưới dạng vector hoá. Đối với văn bản có các phương pháp như BoW (Bag of Word), TF-IDF (Term Frequency-Inverse Document Frequency)….
* Huấn luyện mô hình: Đầu tiên sẽ chon mô hình huấn luyện có rất nhiều thuật toán có thể áp dụng cho bài toán phân tích cảm xúc như Naïve Bayes, SVM,…Sau khi đã chọn được mô hình sẽ chia tập dữ liệu ra thành hai phần là tập huấn luyện và tập kiểm tra và tiến hành huấn luyện mô hình.
* Đánh giá mô hình: Đánh giá mô hình bằng các chỉ số accuracy, precision, recall, F1-score.
* Triển khai thực tiễn: Mô hình được triển khai trong môi trường thực tế và đánh giá hiệu suất khi sử dụng với dữ liệu mới.

### Phân lớp Naïve Bayes

Phân lớp Naïve Bayes (NB) [10] sử dụng xác suất có điều kiện giữa các đặc trưng và nhãn để dự đoán xác suất nhãn của một văn bản cần phân loại. Các đặc trưng được đưa vào mô hình được độc lập với nhau.

Phân lớp Naïve Bayes dựa trên định lý Bayes có công thức tổng quát:

Trong đó:

* P(y|X): Xác suất (có điều kiện) của sự kiện y xảy ra, nếu biết sự kiện X đã xảy ra.
* P(X|y): Xác suất (có điều kiện) của sự kiện X xảy ra, nếu biết sự kiện B đã xảy ra.
* P(y): Xác suất sự kiện y xảy ra.
* P(X): Xác suất sự kiện X xảy ra.

X là vector các đặc trưng, có viết dưới dạng:

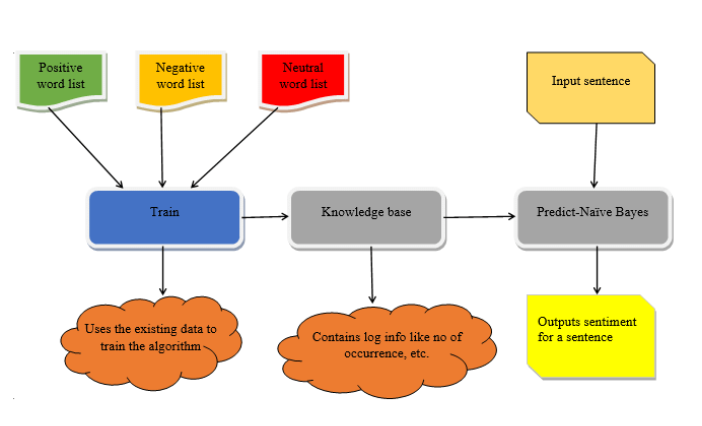
Khi đó đẳng thức Bayes trở thành:

Ưu điểm:

* Đơn giản, dễ triển khai
* Dự đoán đa lớp với độ chính xác cao

Nhược điểm:

* Nếu một đặc trưng chưa xuất hiện trong dữ liệu huấn luyện thì xác suất dự đoán là 0
* Nếu bộ dữ liệu mất cân bằng về các lớp phân loại có thể ảnh hưởng đến mất cân bằng dữ liệu và ảnh hưởng đến hiệu suất của phân loại.Trong trường hợp phân tích cảm xúc nếu bộ dữ liệu nếu một lớp có số lượng mẫu nhiều hơn lớp còn lại thì mô hình sẽ nghiêng về phía có số lượng mẫu cao hơn.

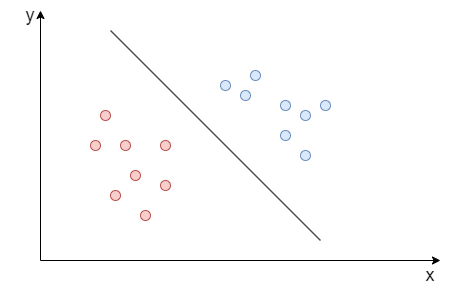


Hình 2.. Phân tích cảm xúc bằng phương pháp Naive Bayes [11]

Naïve Bayes là một thuận toán phổ biến trong phương pháp học máy có giám sát. Ứng dụng của thuật toán phân lớp Naïve Bayes trong phân tích cảm xúc để dự đoán hoặc phân loại cảm xúc của một văn bản ví dụ như phân loại bình luận trên mạng xã hội [12], phân loại các lượt đánh giá [13].

### Support Vector Machine (SVM)

Support Vector Machine (SVM) hay máy vector hỗ trợ là thuật toán thuộc nhóm học có giám sát. Mục tiêu của phương pháp này là tìm một siêu phẳng (hyperplane) để phân tách các điểm dữ liệu thành các miền khác nhau với mỗi miền là một loại dữ liệu. Khoảng cách từ siêu phẳng tới điểm gần nhất được gọi là margin. Khi có một vector mới thì khoảng cách từ siêu phẳng đến vector đó sẽ được tính toán và phân loại phù hợp.

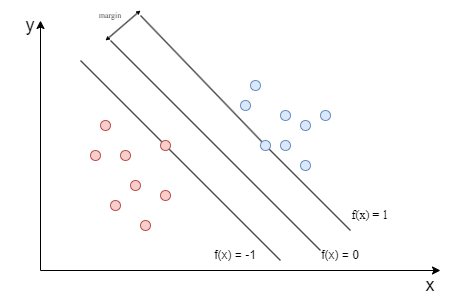


Hình 2.. Mô tả siêu phẳng trong không gian hai chiều

Hàm tuyến tính phân loại của SVM:

Trong đó:

* x là vector đặc trưng của điểm dữ liệu cần phân loại
* w là trọng số của đặc trưng
* b (bias) là độ lệch
* là giá trị đầu ra của hàm cho điểm dữ liệu x



Hình 2.. Mô tả về margin trong SVM

Trong SVM, tồn tại rất nhiều đường phân cách thì đường phân cách tốt nhất là đường có khoảng cách margin bằng nhau và lớn nhất có thể. Ví dụ hình 2.3 có 2 lớp đỏ và xanh dương thì đường phân cách tốt nhất là margin của hai lớp bằng nhau và lớn nhất.

Ưu điểm:

* Có thể xử lý trên không gian nhiều chiều
* Tiết kiệm bộ nhớ
* Tính linh hoạt

Nhược điểm:

* Chưa thể hiện rõ tính xác suất
* Trường hợp số lượng thuộc tính của tập dữ liệu nhiều hơn so với lượng dữ liệu thì SVM cho kết quả không tốt.

## Phương pháp tiếp cận học sâu

Học sâu là một nhánh học máy dựa vào mạng nơ-ron với các thuật toán dựa theo khả năng tư duy và suy nghĩ giống con người. Học sâu có cấu trúc phức tạp hơn học máy. Trong phân tích cảm xúc thường sử dụng các mô hình học sâu như Convolutional Neural Network (CNN), Recurrent Neural Netwok (RNN), Long short-term memory (LSTM),…

### Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) [14] hay mạng nơ-ron tích chập được sử dụng rộng rãi trong lĩnh vực Computer Vision (thị giác máy tính) để phân loại và nhận dạng hình ảnh. Ngoài ra, CNN còn được sử dụng trong nhận dạng giọng nói, văn bản, âm thanh,….

Convolutional là một cửa sổ trượt (sliding window) trên một ma trận:



Hình 2.. Convolution với bộ lọc 3x3 [15]

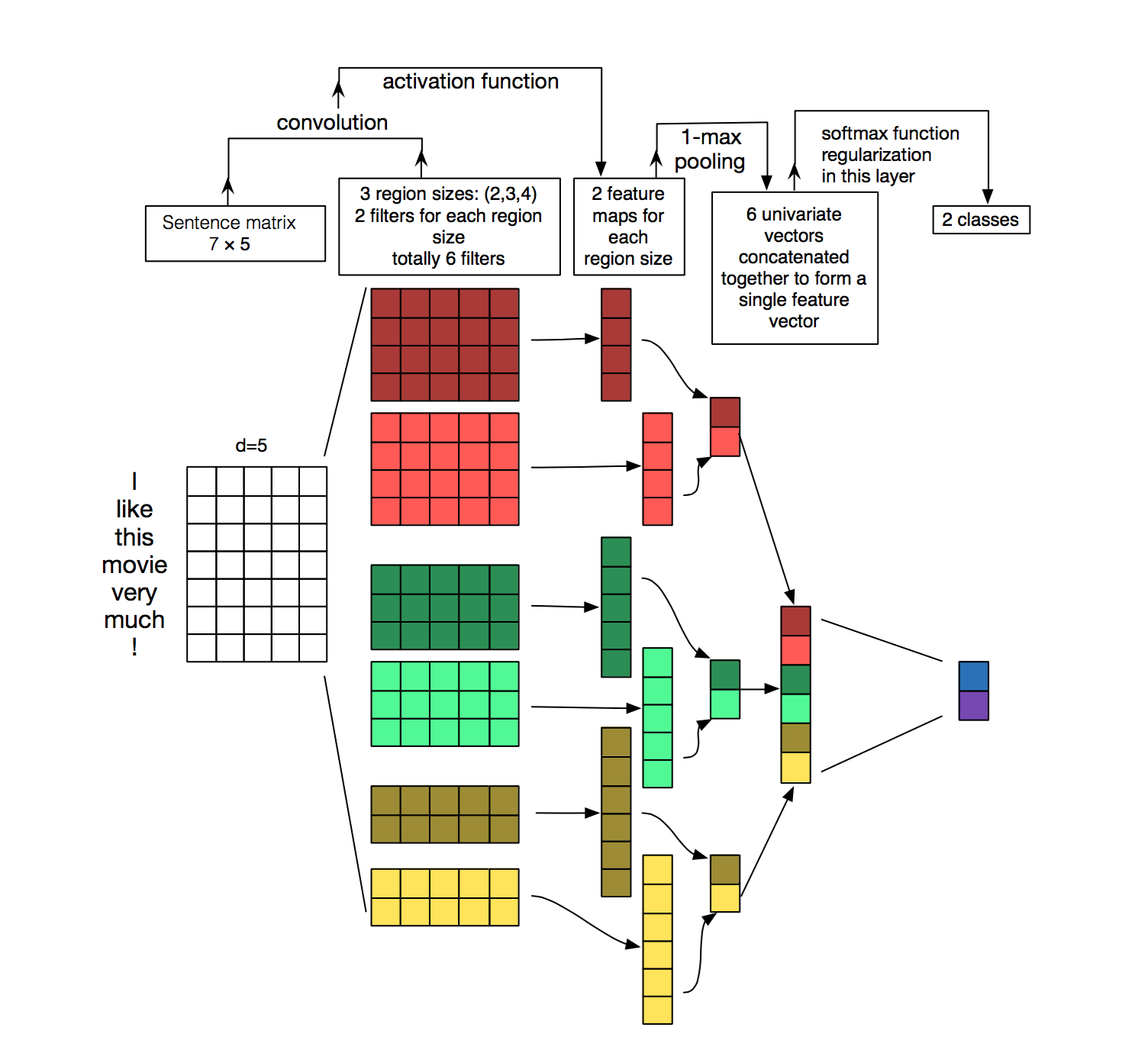
Trong hình 2.4, bên trái là một ảnh đen trắng với mỗi ô là một điểm ảnh có giá trị là 0 tượng trưng cho màu đen và 1 cho màu trắng (0-255 theo thang đo ảnh xám). Cửa sổ trượt được gọi là kernel, bộ lọc (filter), feature detector là ma trận màu vàng có kích thước 3x3 với chữ số đỏ. Bên phải là một ma trận 3x3 là ma trận bộ lọc nhân với ma trận gốc và tính tổng lại với nhau. Liên tục thực hiện trượt đến hết ma trận.



Hình 2.. Sơ đồ hoạt đồng CNN

Mạng CNN là tập hợp các lớp Convolution chồng lên nhau, các lớp Pooling, Fully-Connected,….và sử dụng hàm kích hoạt như activation như ReLU, tanh để tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo.

Trong xử lý ngôn ngữ tự nhiên, đối với xử lý văn bản đầu vào là các mệnh đề, câu văn được biểu diễn dưới dạng ma trận, mỗi dòng của ma trận sẽ ứng với một từ, có thể là một kí tự. Biểu diễn này được tạo ra từ kỹ thuật word embedding.



Hình 2.. Cấu trúc mạng CNN trong xử lý ngôn ngữ tự nhiên. [16]

Hình 2.6 có đầu vào một mệnh đề ‘I like this movie very much’ là ma trận có kích thước 7x5. Có ba region sizes: (2,3,4) với mỗi vùng có hai bộ lọc tổng cộng có 6 bộ lọc. Mỗi bộ lọc thực hiện tính chập trên ma trận câu và tạo ra feature map (variable-length). Sau đó thực hiện 1-max pooling trên mỗi map để tìm giá trị lớn nhất trong từng feature map. Kết quả sẽ có được 1 vector với kích thước 1x1 từ 6 map, kết hợp tất cả để tạo thành vector đặc trưng cuối cùng. Softmax sẽ nhận vector đặc trưng làm đầu vào và sử dụng để phân loại câu.

### Recurrent Neural Network (RNN)

Recurrent Neural Netwok (RNN) [17] là một mô hình phổ biến sử dụng để giải quyết các bài toán liên quan tới dạng chuỗi (sequence). Ý tưởng của RNN là sử dụng các chuỗi thông tin, tất cả đầu vào và đầu ra đều độc lập với nhau. RNN được gọi là “recurrent” vì có thể thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi và đầu ra của mỗi phần tử phụ thuộc vào các phép tính trước đó. RNN có một ‘bộ nhớ’ có thể ghi lại các thông tin đã tính toán. Trên lý thuyết, RNN có thể xử lý được các chuỗi dài nhưng thực tế khả năng chỉ nhớ được một vài bước. Đây là vấn đề của phụ thuộc xa (long-term dependencies), điều này xảy ra do tính mất đạo hàm (gradient vanishing) khi khoảng cách càng lớn.



Hình 2.. Sơ đồ RNN [17]

Trong hình 2.7, tại từng thời điểm t sẽ đưa vào một mô hình các đầu vào:

* là phần tử của đầu vào thứ t.
* là trạng thái ẩn ở bước t. Đây là bộ nhớ của mạng, được dự đoán dựa trên trạng thái ẩn phía trước và đầu vào tại bước đó:

)

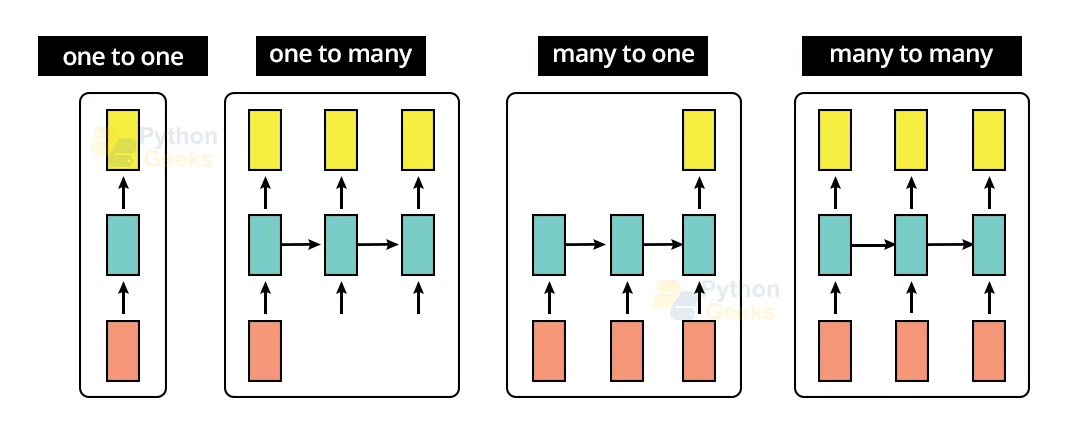
Hàm thường là một hàm phi tuyến như tanh hay ReLU. để tính toán trạng thái ẩn đầu tiên thì khởi tạo giá trị là 0.

* là đầu ra tại bước t. Nếu muốn dự đoán từ tiếp theo có thể xuất hiện trong câu thì chính là một vector xác suất trong danh sách từ vựng:

)

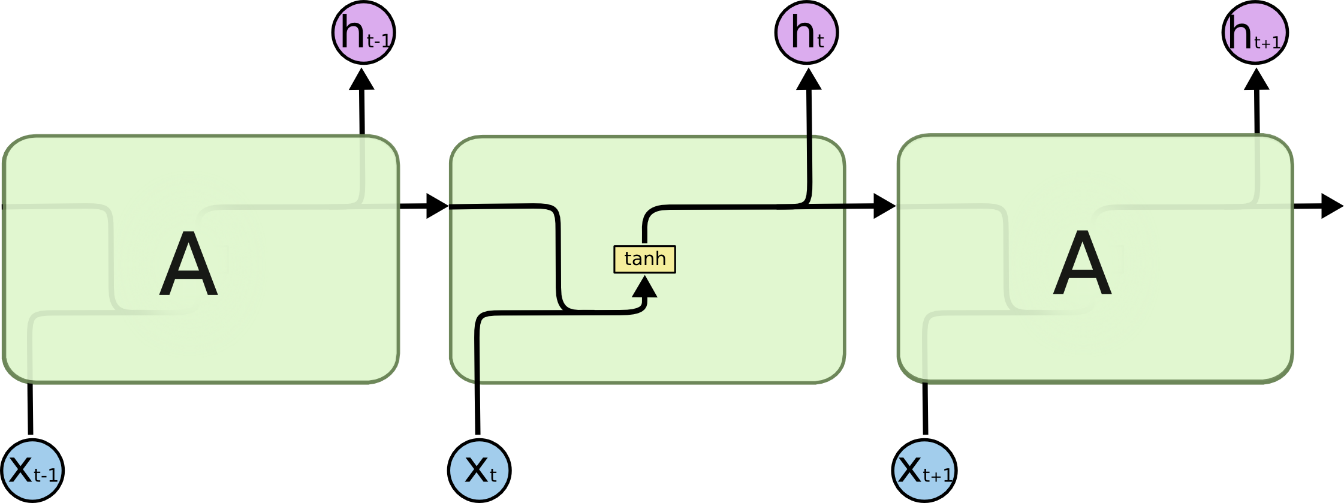
Phân loại bài toán RNN:

* One to one: còn được gọi là Vanilla Neural Network đây loại RNN cơ bản nhất để giải quyết các vấn đề tổng quát về học máy. Trong one to one chỉ có một lớp đầu vào và đầu ra.
* One to many: thường giải quyết các bài toán thêm mô tả cho ảnh. Gồm một lớp đầu vào và nhiều lớp đầu ra.
* Many to one: là bài toán có nhiều đầu ra và một đầu vào. RNN này phù hợp với các phân tích cảm xúc để phân loại cảm như xúc tích cực hay tiêu cực.
* Many to Many: phức tạp hơn các bài toán khác. Có nhiều đầu vào và đầu ra.



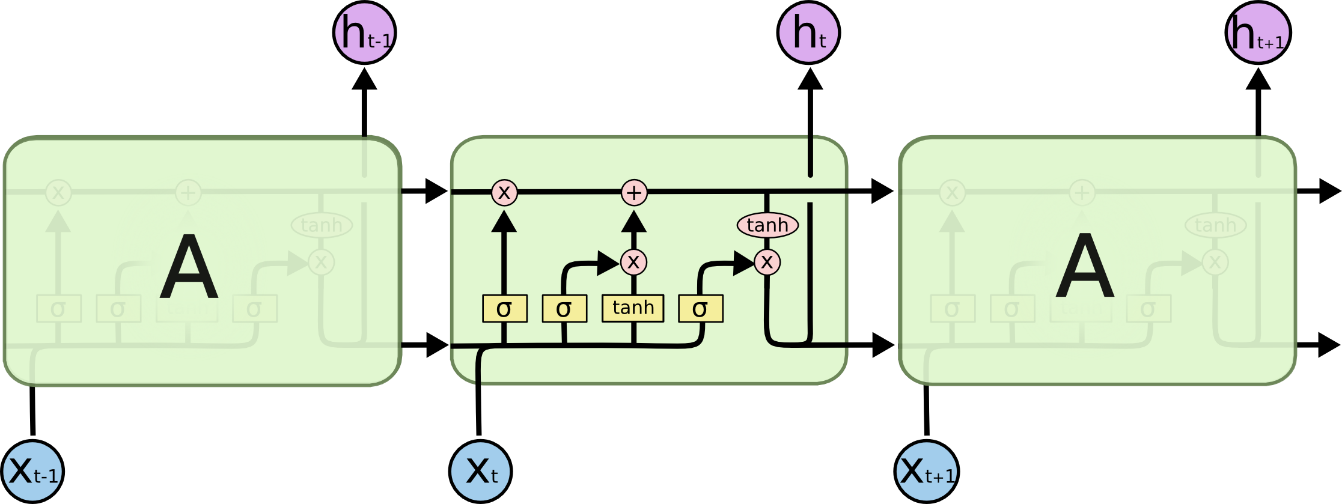
Hình 2.. Các loại bài toán trong RNN [18]

Long-Short Term Memory (LSTM) [19] là một mạng đặc biệt, cải tiến từ RNN có khả năng học được các phụ thuộc xa nhằm khắc phục việc mất đạo hàm khi xử lý chuỗi ở RNN.



Hình 2.. Mô-đun lặp lại trong RNN tiêu chuẩn chứa một lớp duy nhất [19]

LSTM cũng có kiến trúc giống như trên thay vì một tầng thì nó có bốn tầng tương tác với nhau.



Hình 2.. Mô-đun lặp lại LSTM chứa 4 lớp [19]

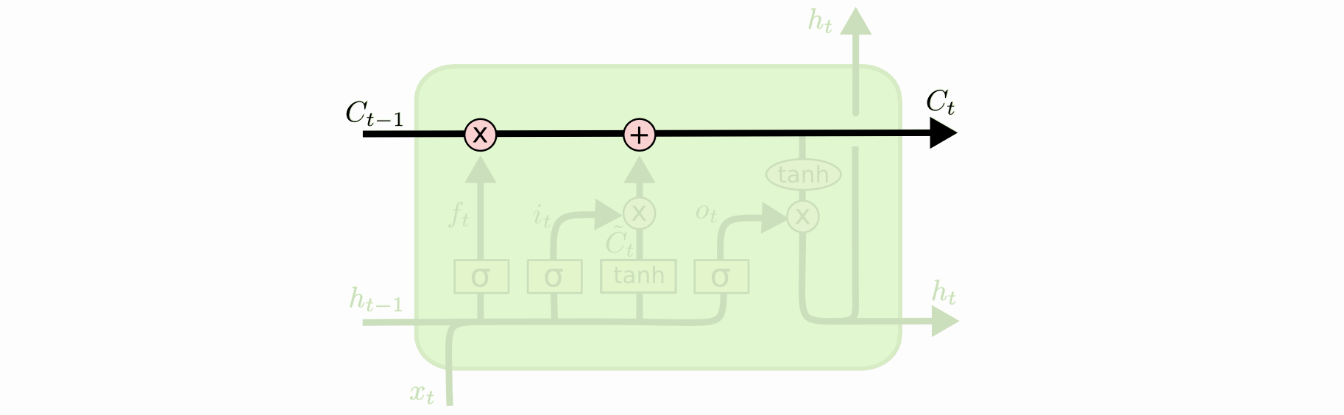
Đây là một vài ký hiệu được sử dụng trong sơ đồ trên :



Hình 2.. Ký hiệu trong LSTM [19]

* Hình chữ nhật màu vàng: các hàm phi tuyến tính, với sơ đồ trên thì là hàm *tanh* và *sigmoid*.
* Hình tròn màu hồng: các phép toán như cộng vector, nhân vector.
* Mũi tên (Vector Transfer): các vector.
* Mũi tên gộp lại (Concatenate): biểu diễn việc kết hợp.
* Mũi tên rẽ nhánh (Copy): nội dung được sao chép và rẽ sang nhánh khác.

Ý tưởng của LSTM là trạng thái tế bào (cell state) chứa các thông tin quan trọng từ các bước trước đó và truyền qua các bước tiếp theo trong quá trình xử lý chuỗi dữ liệu.



Hình 2.. Trạng thái tế bào [19]

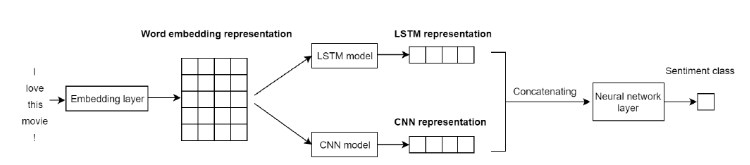
Cổng quên (forget gate) giúp loại bỏ thông tin không cần thiết ở trạng thái trước đó. Hàm *sigmoid* có đầu vào là [0,1] mô tả thông tin có thể được thông qua. Đầu ra là 0 nghĩa là không cho thông tin nào qua cả, còn 1 là thông tin được đi qua.

LSTM gồm có 3 cổng là cổng vào (input gate), cổng ra (output gate) và cổng quên (forget gate).



Hình 2. . Cổng quên nhân với vector [19]

LSTM có thể kết hợp với CNN để giải quyết một số bài toán như phân lớp nhận xét tiếng Việt theo ba mức khen, chê, trung tính. Trong [8], CNN hoạt động tốt trong việc bắt được các mối quan hệ lân cận nhau, LSTM với cơ chế nhớ quên có thể xử lý được các mối phụ thuộc ở khoảng cách xa trong văn bản.



Hình 2. . LSTM-CNN [8]

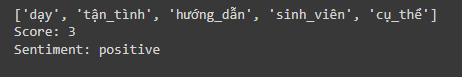
## Phương pháp tiếp cận dựa trên từ điển

Phương pháp dựa trên danh sách các từ vựng hoặc từ điển được gán nhãn cảm xúc như tích cực, tiêu cực, trung tính. Các từ được gán nhãn cảm xúc sẽ được tính điểm. Thông qua cách thức tính điểm sẽ đưa ra được dự đoán cảm xúc của một văn bản. Ví dụ trong tập dữ liệu UIT [20] một từ hoặc cụm từ trong từ điển điều có một số điểm nhất định. Số điểm được quy định là 1 nếu như là tích cực và -1 nếu như là tiêu cực.

|  |  |
| --- | --- |
| Word | Score |
| tận\_tình | 1 |
| hướng\_dẫn | 1 |
| cụ\_thể | 1 |
| ít | -1 |
| chưa | -1 |
| liều | -1 |
| ủng\_hộ | 1 |
| vắng\_mặt | -1 |
| vui | 1 |
| xoá | -1 |

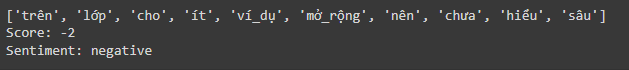
Bảng 1. Tập dữ liệu UIT

Trong câu: “dạy tận tình , hướng dẫn sinh viên cụ thể .” khi tách từ trong câu thấy được có những từ hay cụm từ được gán nhãn như 'dạy', 'tận\_tình', 'hướng\_dẫn', 'sinh\_viên', 'cụ\_thể' trong đó từ ‘tần\_tình’, ‘hướng\_dẫn’, ‘cụ\_thể’ là 1. Kết luận đây là một câu tích cực vì có tổng số điểm là 3



Hình 2.. Positive

Trong một ví dụ khác có câu: “trên lớp cho ít ví dụ mở rộng nên chưa hiểu sâu .” Có ‘ít’ và ‘chưa’ là -1. Vậy thì tổng điểm của câu là -2



Hình 2.. Negative

Những từ được gán nhãn được quy định bởi chính người tạo ra từ điển. Vậy nên đôi khi sẽ có những từ không được gán nhãn. Phương pháp này cũng có một vài hạn chế là không thể xét đến yếu tố ngữ cảnh, điều này có thể ảnh hưởng đến kết quả cuối cùng khi phân tích cảm xúc.

## Kết luận chương

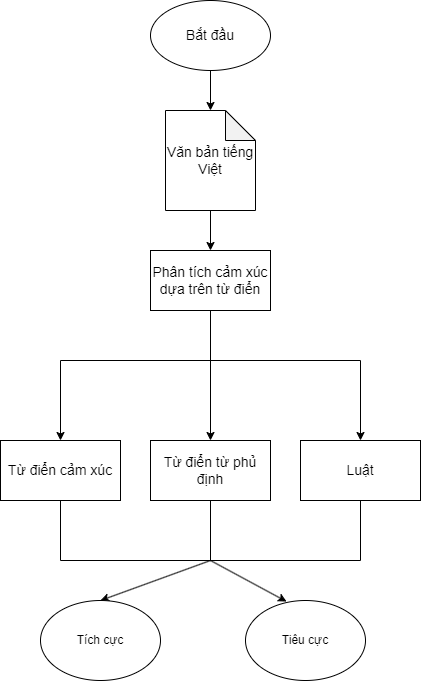
Trong chương này, giới thiệu tổng quan về phân tích cảm xúc, tìm hiểu nguyên nhân vì sao cần phải phân tích xúc và những thách thức gặp phải. Tìm hiểu tổng quan về một số phương pháp phổ biến được sử dụng để phân loại cảm xúc trong văn bản.

# CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

Trong chương này sẽ trình bày về mô hình của phương pháp dựa trên từ điển được đề xuất để thực hiện phân tích cảm xúc

## Mô hình đề xuất

Mô hình được đề xuất là phương pháp dựa trên từ điển và các hệ thống luật để đưa ra dự đoán cảm xúc. Đây là là mô hình sử dụng tập dữ liệu đã được tiền xử lý. Mô hình có thể tính toán các giá trị cảm xúc của văn bản và đưa ra dự đoán dựa trên giá trị cảm xúc đó.



Hình 3.. Mô hình phân tích cảm xúc dựa trên từ điển

Trong mô hình này sử dụng hai bộ từ điển đó là từ điển cảm xúc và từ điển phủ định và hệ thống luật để tính giá trị cảm xúc:

* Từ điển cảm xúc: là tập từ điển đã được gán nhãn với giá trị là 1 thể hiển cảm xúc tích cực, giá trị -1 mang cảm xúc tiêu cực. Mục đích của việc gán giá trị này là để tính giá trị cảm xúc của văn bản.

|  |  |
| --- | --- |
| Word | Score |
| đảo\_lộn | -1 |
| đào\_sâu | 1 |
| đáp ứng | 1 |
| đạt | 1 |
| đau | -1 |
| kỹ | 1 |
| kỹ\_càng | 1 |
| kỳ\_vọng | 1 |
| khá | 1 |
| khá\_hay | 1 |

Bảng . Từ điển cảm xúc

* Từ điển phủ định: bao gồm các từ sẽ khiến cho các giá trị cảm xúc của từ trong từ điển cảm xúc bị thay đổi giá trị. Ví dụ từ ‘lưu loát’ chỉ cách diễn đạt trôi chảy không ngập ngừng và có giá trị là 1 nhưng khi từ ‘lưu loát’ đi chung với từ ‘không’ trong tập dữ liệu phủ định thì ý nghĩa của từ ‘không lưu loát’ trái ngược lại với ‘lưu loát’ và thay đổi giá trị thành -1.

|  |
| --- |
| **Word** |
| bất |
| chưa |
| chưa\_có |
| khỏi |
| không |
| không\_có |
| không\_còn |
| không\_cần |
| không\_thấy |
| mất |
| thiếu |

Bảng . Từ điển phủ định

* Hệ thống luật giúp tính toán các giá trị cảm xúc bằng cách tách các từ trong văn bản để có các ‘token’, sử dụng từ điển tìm ra giá trị của các ‘token’ từ đó áp dụng các luật để xác định cảm xúc cho văn bản. Kết quả cuối cùng cho ra tổng số điểm cảm xúc của văn bản. Dựa vào số điểm đã tính để đưa ra kết luận đó là một văn bản tích cực hay tiêu cực.

## Xây dựng các từ điển cảm xúc

* Từ điển cảm xúc gồm 1798 từ hoặc cụm từ được gán nhãn là 1 hay -1, trong đó 1 thể hiện cho cảm xúc tích cực, -1 thể hiện cho tiêu cực.
* Tập từ điển phủ định: là tập từ điển mà trong đó là tập hợp các từ trái nghĩa nhau ví dụ như ‘không\_có’ đối lập với ‘có’.
* Tập từ điển sửa lỗi: là tập từ điển được xây dựng để sửa lỗi chính tả khi dữ liệu thu thập thường là những bình luận, đánh giá với cách viết tự do nên dễ xảy ra việc sai chính tả, viết tắt. Việc sửa lỗi chính tả giúp cho việc tính toán chính xác và dự đoán cảm xúc tốt hơn.

|  |  |
| --- | --- |
| incorrect\_Word | correct\_Word |
| áp đặp | áp đặt |
| bải tập | bài tập |
| bào giảng | bài giảng |
| bài giảng | bài tập |
| bỗ trợ | bổ trợ |
| bổ xung | bổ sung |
| buốn ngủ | buồn ngủ |
| cận kẽ | cặn kẽ |
| có lẻ | có lẽ |
| cọc lốc | cộc lốc |

Bảng . Từ điển sửa lỗi

* Tập từ điển token: đây là tập dữ liệu nhằm mục đích biến các từ có trong một câu văn bản thành các token giúp thuận tiện cho việc xác định các từ trong từ điển cảm xúc.

|  |  |
| --- | --- |
| Word | Token |
| ẩm vừa | ẩm\_vừa |
| cực tốt | cực\_tốt |
| khá hay | khá\_hay |
| hơi lạ | hơi\_lạ |
| ít khi | ít\_khi |
| kiểm soát | kiểm\_soát |
| không no | không\_no |
| mưa lớn | mưa\_lớn |
| quay lại | quay\_lại |
| tươi mới | tươi\_mới |

Bảng . Từ điển token

* Tập từ điển stop word [21]: tập từ điển này gồm các từ không mang ý nghĩa quan trọng trong việc phân tích ngôn ngữ tự nhiên và thường bị loại bỏ trong quá trình xử lý ngôn ngữ tự nhiên.

|  |
| --- |
| stop\_Word |
| a\_lô |
| bấy\_nhiêu |
| bển |
| chung\_quy |
| ủa |
| à\_ơi |
| vâng\_dạ |
| tiện\_thể |
| thì\_là |
| sao |

Bảng . Từ điển stop word

## Tập dữ liệu

Tập dữ liệu UIT [20] gồm các nhận xét đánh giá của sinh viên. Những nhận xét tập trung vào giảng viên, cơ sở vật chất, chương trình giảng dạy. Đối với những nhận xét dành cho giảng viên chủ yếu là hành động, lời nói, cử chỉ của giảng viên trong quá trình giảng dạy, đồng thời đóng góp ý kiến giúp giảng viên tốt hơn. Trong vấn đề chương trình giảng dạy sinh viên trình bày quan điểm về môn học, bài học, bài tập. Về cơ sở vật chất sinh viên đề cập chất lượng cơ sở vật chất có ảnh hưởng gì đến quá trình giảng dạy, học tập. Ngoài ra, trong tập dữ liệu còn có những phản hồi mà không thuộc một chủ đề nào cả.Những phản hồi của sinh viên được việt tự do, thường rất ngắn, trong đó có sử dụng một số kí tự đặc biệt để thể hiện cảm xúc như "<3", ":-)", ":D, ":v", ":)", "=)",…đôi khi có những lỗi chính tả. Trong tập dữ liệu được chia thành hai cột, cột ‘Reviews’ chứa những phản hồi của sinh viên và cột ‘Sentiment’ là phân loại cảm xúc ‘positive’ nghĩa là tích cực và ‘negative’ nghĩa là tiêu cực của phản hồi.

|  |  |
| --- | --- |
| Reviews | Sentiment |
| phương pháp dạy hiệu quả . | positive |
| thầy hiền vui tính , dạy bài tập dễ hiểu . | positive |
| cô tạo được một không khí gần gũi , không áp lực , học rất thoải mái . | positive |
| giảng bài sinh động , có nhiều cuộc thảo luận trên lớp . | positive |
| tận tình giúp đỡ sinh viên . | positive |
| giảng viên không sử dụng moodle để sinh viên nộp bài tập mà yêu cầu gửi qua mail cá nhân . | negative |
| bài tập cho quá khó . | negative |
| phòng thực hành nóng . | negative |
| cô còn giảng hơi nhanh . | negative |
| nhiều khi đang làm thực hành mà không có mạng để tìm hiểu thứ mình cần . | negative |

Bảng . Tập dữ liệu UIT

## Thực nghiệm

* + 1. **Cấu hình**
* Môi trường thực nghiệm trên Google Colab có cấu hình:
* Operating System: Linux-6.1.58+-x86\_64-with-glibc2.35
* Python Version: 3.10.12
* TensorFlow Version: 2.15.0
* GPU: T4 GPU
* CPU Cores: 2
* Thư viện sử dụng:
* Re
* Emoji
* Pandas
* CSV
  + 1. **Tiền xử lý dữ liệu**

Trước khi bắt đầu tiền xử lý, kiểm tra tập dữ liệu stop word có từ trùng với từ trong từ điển khác hay không. Nếu có thì loại bỏ từ đó ra khỏi stop word.

def remove\_stop\_words(input\_file, stop\_word\_file):

    # Đọc dữ liệu từ file

    with open(opposite\_file, 'r', newline='', encoding='utf-8') as word\_csv:

        word\_reader = csv.reader(word\_csv)

        words = [row[0] for row in word\_reader]

    # Đọc dữ liệu từ file stop\_word.csv

    with open(stop\_word\_file, 'r', newline='', encoding='utf-8') as stop\_word\_csv:

        stop\_word\_reader = csv.reader(stop\_word\_csv)

        stop\_words = [row[0] for row in stop\_word\_reader]

    # Loại bỏ các từ trong stop\_words nếu có trong từ điển

    cleaned\_stop\_words = [word for word in stop\_words if word not in words]

    with open(stop\_word\_file, 'w', newline='', encoding='utf-8') as stop\_word\_csv:

        stop\_word\_writer = csv.writer(stop\_word\_csv)

        for word in cleaned\_stop\_words:

            stop\_word\_writer.writerow([word])

#opposite\_Word

remove\_stop\_words('/content/drive/MyDrive/KLTN/dictionary/opposite\_word.csv', '/content/drive/MyDrive/KLTN/dictionary/stop\_word.csv')

#UIT

remove\_stop\_words('/content/drive/MyDrive/KLTN/dictionary/VN\_UIT\_dict.csv', '/content/drive/MyDrive/KLTN/dictionary/stop\_word.csv')

**Bước 1: Loại bỏ những kí tự đặc biệt**

Trong tập dữ liệu là những phản hồi của sinh viên và viết tự do. Những phản hồi này có các kí tự đặc biệt và có những emoji cần phải xử lý.

def remove\_special\_text(tweet):

  #Loại bỏ các ký tự và đường dẫn

  tweet = re.sub(r'(http\S+)|(@\S+)|RT|\#|!|:|\.|,', ' ', tweet)

  # Chuyển emoji sang text

  tweet = emoji.demojize(tweet)

  #Thêm dấu khoảng trắng trước và sau dấu text emoji

  tweet = re.sub(r'(:[a-zA-Z0-9\_-]+:)', r' \1 ', tweet)

  tweet = tweet.replace('-', '\_')

  tweet = tweet.replace(':', '\_')

  characters\_to\_remove = [' ̉ ', ' ̃ ', ' ̣ ', ' ́ ', ' ̀ ']

  for char in characters\_to\_remove:

    tweet = tweet.replace(char.strip(), '')

  return tweet

**Bước 2: Xử lý từ lỗi**

Sau khi loại bỏ được những kí tự đặc biệt, dựa vào từ điển error\_word kiểm tra xem các từ trong một tweet có đúng chính tả hay không, nếu có thì sửa lại.

Trong đoạn code, nhận vào một dòng tweet duyệt xem trong dòng tweet đó có những từ sai chính tả hay không. Nếu dòng tweet có dòng sai chính tả ‘incorrect\_word’ thì thay thế từ sai đó bằng từ tương ứng trong cột ‘correct\_word’.

def correct\_Word(tweet):

    errorWord\_dict = pd.read\_csv('/content/drive/MyDrive/KLTN/dictionary/error\_word.csv')

    for index, row in errorWord\_dict.iterrows():

        if row['incorrect\_Word'] in tweet:

            tweet  = tweet.replace(row['incorrect\_Word'], row['correct\_Word'])

    return tweet

**Bước 3: Tách từ**

Thực hiện việc tách từ bằng từ điển token. Vòng lặp duyệt qua từng từ trong từ điển, nếu tìm thấy được từ có trong từ điển token sẽ thay thế bằng ‘token’.

def tokenization(tweet):

  token\_dict = pd.read\_csv('/content/drive/MyDrive/KLTN/dictionary/token\_word.csv')

  # Duyệt qua từng dòng trong từ điển token

  for index, row in token\_dict.iterrows():

    # Nếu từ trong từ điển được tìm thấy trong tweet

    if row['Word'] in tweet:

    # Thay thế từ bằng token tương ứng

      tweet = tweet.replace(row['Word'], row['Token'])

  return tweet

**Bước 4: Loại bỏ những từ không quan trọng**

Sử dụng từ điển stop word để loại bỏ các từ không quan trọng, giảm nhiễu và cái thiện mô hình. Vòng lặp sẽ duyệt qua từ điển stop word và sẽ loại bỏ những từ trong tweet nếu có trong từ điển stop word.

def remove\_stopWord(tweet):

    # Chuẩn hóa từ theo từ điển

    stopWord\_dict = pd.read\_csv('/content/drive/MyDrive/KLTN/dictionary/stop\_word.csv')

    for index, row in stopWord\_dict.iterrows():

        stop\_word = row['stop\_Word']

        # Loại bỏ stop word nếu nó xuất hiện trong tweet

        tweet = ' '.join([word for word in tweet.split() if word != stop\_word])

    return tweet

Ví dụ cho loại bỏ stop word:

* Input: là một tweet ‘'cô còn giảng hơi nhanh .'
* Output: [‘giảng’, ‘hơi’, ‘nhanh’]

Trong từ điển đã loại bỏ những từ ‘cô’, ‘còn’ và chỉ để lại ‘giảng’,’hơi’, ‘nhanh’. Vậy thì từ ‘giảng’ là một từ không có trong từ điển cảm xúc và ‘hơi’. ‘nhanh’ là từ có trong từ điển với score của hai từ đều là -1. Kết luận câu này là negative.

**Bước 5: Hàm pre\_process**

Hàm pre\_process nhận đầu vào là một tweet và thực hiện một loạt các bước xử lý từ các bước trước với đầu ra là tokens.

def pre\_process(tweet):

    tweet = remove\_special\_text(tweet)

    tweet = correct\_Word(tweet)

    tweet = tokenization(tweet)

    tweet = remove\_stopWord(tweet)

    tokens = ViTokenizer.tokenize(tweet)

    tokens = tokens.split()

    return tokens

**Bước 6: Thực hiện quá trình tiền xử lý cho tập dữ liệu UIT**

Trong process\_dataset đầu vào nhận vào một file và kiểm tra xem file đó có tồn tại hay không, nếu có thì sẽ lấy nội dụng của cột ‘Reviews’ làm ‘tweet’ và thực hiện tiền xử lý. Khi đã thực hiện xong tiền xử lý và có được ‘tokens’ thì tạo ra một Dataframe mới với ‘Sentiment’ và ‘Reviews’ được lấy từ dữ liệu đầu vào, ‘Tokens’ là tokens sau khi thực hiện tiền xử lý. Đầu ra sẽ là một file csv.

def process\_dataset(input\_file, output\_file):

    try:

        data = pd.read\_csv(input\_file, encoding='utf-8')

    except FileNotFoundError:

        print("File input không tồn tại!")

        return

    print("Input length:", len(data))

    rows = []

    for index, row in data.iterrows():

        tweet = row['Reviews']

        tokens = pre\_process(tweet)

        sentiment = row['Sentiment']

        rows.append([sentiment, tokens, tweet])

    df = pd.DataFrame(rows, columns=['Sentiment', 'Tokens', 'Reviews'])

    df.to\_csv(output\_file, index=False, encoding='utf-8')

process\_dataset('/content/drive/MyDrive/KLTN/input/UIT.csv', '/content/drive/MyDrive/KLTN/output/UIT\_Final\_input.csv')

Tập dữ liệu UIT sau khi thực hiện tiền xử lý:

|  |  |  |
| --- | --- | --- |
| Sentiment | Tokens | Reviews |
| positive | ['giáo\_viên', 'tận\_tình', 'chuẩn', 'học', 'chu\_đáo'] | giáo viên tận tình , chuẩn bị bài học chu đáo . |
| positive | ['thầy', 'hiền', 'vui', 'dạy', 'tập'] | thầy hiền vui tính , dạy bài tập dễ hiểu . |
| positive | ['không\_khí', 'gũi', 'không', 'áp\_lực', 'học', 'thoải\_mái'] | cô tạo được một không khí gần gũi , không áp lực , học rất thoải mái . |
| positive | ['giảng', 'sinh\_động', 'thảo\_luận', 'lớp'] | giảng bài sinh động , có nhiều cuộc thảo luận trên lớp . |
| positive | ['tận\_tình', 'giúp\_đỡ', 'sinh\_viên', '.'] | tận tình giúp đở sinh viên . |
| negative | ['giảng\_viên', 'không', 'sử\_dụng', 'moodle', 'sinh\_viên', 'nộp', 'tập', 'yêu\_cầu', 'gửi', 'mail', 'cá\_nhân'] | giảng viên không sử dụng moodle để sinh viên nộp bài tập mà yêu cầu gửi qua mail cá nhân . |
| negative | ['nội\_dung', 'học', 'cần', 'cải\_thiện'] | nội dung học cần cải thiện . |
| negative | ['phòng', 'thực\_hành', 'nóng'] | phòng thực hành nóng . |
| negative | ['giảng', 'hơi'] | cô còn giảng hơi nhanh . |
| negative | ['thực\_hành', 'không', 'mạng', 'cần'] | nhiều khi đang làm thực hành mà không có mạng để tìm hiểu thứ mình cần . |

Bảng . Tập dữ liệu UIT\_Final

* + 1. **Thực hiện mô hình phân tích cảm xúc bằng phương pháp dựa trên từ điển.**

Hệ thống luật dựa vào hai yếu tố đó là từ điển cảm xúc và từ điển phủ định. Trong đó đầu vào là tập dữ liệu đã được tiền xử lý, duyệt từng token kiểm tra xem có trong từ điển cảm xúc hay không. Nếu có bất kì token nào có trong từ điển thì sẽ cộng hoặc trừ một điểm vào score. Nếu một token xuất hiện sau một hay hai từ có trong từ điển phủ định thì điểm số của sẽ được nhân -1 vào score, nếu không thì điểm số của từ sẽ đảo ngược. Cuối cùng dựa vào số điểm đã được tính toán đưa ra dự đoán cho văn bản đó.

def final\_score(tokens):

    score = 0

    for i in range(len(tokens)):

        word = tokens[i]

        if word in vndict:

            \_score = vndict[word][0]

            if i-1 >= 0 and tokens[i-1] in oppositeWord\_dict:

                \_score \*= -1

            elif i-2 >= 0 and tokens[i-2] in oppositeWord\_dict:

                \_score \*= -1

            score += \_score

    return score

Sau khi có được điểm số cuối cùng được đánh giá nếu như lớn hơn 0 sẽ là ‘positive’ ngược lại là ‘negative’.

def define\_sentiment(score):

    if score > 0:

        return 'positive'

    else:

        return 'negative'

Cuối cùng thực hiện predict cho tập dữ liệu.

def predict\_dataset(input\_file, output\_file):

    try:

        data = pd.read\_csv(input\_file, encoding='utf-8')

    except FileNotFoundError:

        print("File input không tồn tài!")

        return

    print("Input length:", len(data))

    rows = []

    for index, row in data.iterrows():

        tokens = eval(row['Tokens'])

        score = final\_score(tokens)

        pre\_sentiment = define\_sentiment(score)

        rows.append([score, pre\_sentiment])

    df = pd.DataFrame(rows, columns=['Score', 'pre\_Sentiment'])

    existing\_df = pd.read\_csv(input\_file, encoding='utf-8')

    combined\_df = pd.concat([df, existing\_df], axis=1)

    combined\_df.to\_csv(output\_file, index=False, encoding='utf-8')

vndict = load\_dictionary('/content/drive/MyDrive/KLTN/dictionary/VN\_UIT\_dict.csv')

oppositeWord\_dict = load\_special\_word('/content/drive/MyDrive/KLTN/dictionary/opposite\_word.csv')

* + 1. **Đánh giá thực nghiệm**

Sau khi tính toán và đưa ra dự đoán cho tập dữ liệu. So sánh tập dữ liệu đã phân tích cảm xúc so với lúc chưa phân tích cảm xúc.

def compare\_sentiments(input\_file):

    # Đọc dữ liệu từ tệp CSV

    data = pd.read\_csv(input\_file, encoding='utf-8')

    # Tính số lượng dòng giống nhau giữa hai cột

    total\_rows = len(data)

    matching\_rows = len(data[data['pre\_Sentiment'] == data['Sentiment']])

    # Tính phần trăm giống nhau

    percentage\_matching = (matching\_rows / total\_rows) \* 100

    return percentage\_matching

input\_file = '/content/drive/MyDrive/KLTN/output/UIT\_No\_stopword\_Final.csv'

percentage\_matching = compare\_sentiments(input\_file)

print(f"Phần trăm giống nhau giữa 'pre\_Sentiment' và 'Sentiment' là {percentage\_matching:.2f}%")

So sánh các trường hợp khi thực nghiệm:

* Sử dụng đầy bộ đủ từ điển.
* Không sử dụng từ điển phủ định khi tính điểm.

|  |  |
| --- | --- |
| Trường hợp | Độ chính xác |
| Sử dụng đầy đủ bộ từ điển | 84.70% |
| Không sử dụng từ điển phủ định khi tính điểm. | 82.03% |

Bảng . Bảng đánh giá thực nghiệm

Khi xét đến trường hợp sử dụng và không sử dụng từ điển phủ định thì trường hợp có sử dụng từ điển phủ có kết quả cao hơn. Như vậy, tập dữ liệu phủ định đóng góp một vai trò lớn trong bài phân tích cảm xúc này.

Khi sử dụng stop word để loại bỏ các từ không cần thiết để tối ưu mô hình là một điều tốt. Tuy nhiên, có trường hợp sử dụng từ điển stop word vô tình đã loại bỏ những từ có trong từ điển cảm xúc làm cho ảnh hưởng đến độ chính xác khi phân tích cảm xúc. Không chỉ ở từ điển stop word mà bất kì từ điển nào dùng trong bài phân tích đều có thể bị tương tự. Trong từ điển phủ định cũng có hiện tượng này nhưng tập từ điển này rất ít và có thể dễ dàng kiểm soát hơn. Cho nên việc xây dựng tập từ điển chặt chẽ cũng là một thách thức khi sử dụng phương pháp dựa trên từ điển.

# CHƯƠNG 4. KẾT LUẬN

Phương pháp phân tích cảm xúc dựa trên từ vựng là một trong những phương tiện quan trọng trong xử lý ngôn ngữ tự nhiên để hiểu và đánh giá cảm xúc trong văn bản. Trong phương pháp này, mỗi từ trong văn bản được gán một điểm số tương ứng với cảm xúc mà từ đó biểu đạt, dựa trên một bộ từ điển cảm xúc được xây dựng trước. Tuy nhiên, mặc dù có thể làm việc tốt với các trường hợp cơ bản, phương pháp này phải đối mặt với những thách thức như phủ định và chuyển dịch cảm xúc.

Khi xử lý phủ định, phương pháp phải nhận biết và đánh giá các từ tiêu cực được phủ định bằng các từ phủ định như "không", "không phải", "không có gì", vv. Ví dụ, trong câu "Không có gì tệ bằng việc mất mát", từ "tệ" ban đầu mang ý nghĩa tiêu cực, nhưng với sự xuất hiện của từ "không", nó trở thành tích cực.

Một vấn đề khác là chuyển dịch cảm xúc, khi cảm xúc của một từ có thể thay đổi trong các ngữ cảnh khác nhau. Ví dụ, từ "cô đơn" có thể biểu thị sự buồn bã trong một ngữ cảnh, nhưng trong một bài thơ về sự tự do, nó có thể mang ý nghĩa tích cực của sự độc lập. Để xử lý chuyển dịch cảm xúc này, phương pháp cần phải hiểu và đánh giá ngữ cảnh xung quanh từ.

Trong tất cả các trường hợp, việc sử dụng các ví dụ cụ thể có thể làm sáng tỏ. Ví dụ, trong câu "Tôi không phải là một người thích giải toán, nhưng bài toán này thực sự thú vị!", từ "không" phủ định ý nghĩa tiêu cực của "thích" và dẫn đến một cảm xúc tích cực. Trong trường hợp của chuyển dịch cảm xúc, câu "Cô ấy cảm thấy buồn bã trong những ngày mưa, nhưng với một chiếc ô và một cuốn sách, cô ấy lại tìm thấy niềm vui" là một ví dụ điển hình, với từ "buồn bã" chuyển từ cảm xúc tiêu cực sang tích cực dưới sự ảnh hưởng của "tìm thấy niềm vui".

Tóm lại, phương pháp phân tích cảm xúc dựa trên từ vựng có thể đối mặt với các thách thức như phủ định và chuyển dịch cảm xúc, nhưng với sự linh hoạt và sự hiểu biết về ngôn ngữ, nó vẫn có thể đem lại những kết quả đáng tin cậy trong việc đánh giá cảm xúc trong văn bản.

Phương hướng pháp triển:

* Xây dựng thêm và phát triển bộ từ điển.
* Phát triển ứng dụng, trang web cho phân tích cảm xúc.
* Phân tích cảm xúc đa lớp (vui, buồn, tức giận,…)
* So sánh thêm các phương pháp khác

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | S. Kemp, "Digital 2019: Vietnam," 31 1 2019. [Online]. Available: https://datareportal.com/reports/digital-2019-vietnam. [Accessed 3 2024]. |
| [2] | S. Kemp, "Digital 2023: Vietnam," 13 2 2023. [Online]. Available: https://datareportal.com/reports/digital-2023-vietnam. [Accessed 3 2024]. |
| [3] | Trần Khải Thiện, Tiếu Phùng Mai Sương, "Một số khái niệm và hướng tiếp cận phân tích cảm xúc - Áp dụng cho tiếng Việt," *HJS,* vol. 6, p. 82, tháng 3 2022. |
| [4] | Yann LeCun, Yoshua Bengio & Geoffrey Hinton, "Deep learning," *Nature,* vol. 521, p. 436–444, May 2015. |
| [5] | Nguyen Thi Duyen, Ngo Xuan Bach & Tu Minh Phuong, "An Empirical Study on Sentiment Analysis for Vietnamese," *The 2014 International Conference on Advanced Technologies for Communications (ATC'14),* pp. 309-314, 2014. |
| [6] | Thien Khai Tran, Tuoi Thi Phan, "Multi-class opinion classification for Vietnamese hotel reviews," *International Journal of Intelligent Technologies and Applied Statistics,* vol. 9, p. 1:7, Mar 2016. |
| [7] | Dinh Nguyen, Khuong Vo, Dang Pham, Mao Nguyen and Tho Quan, "A Deep Architecture for Sentiment Analysis of News Articles," *Cham,* pp. 129-140, 2018. |
| [8] | Quan Hoang Vo, Quan Hoang Vo, Bac Le and Minh Le Nguyen, "Multi-channel LSTM-CNN model for Vietnamese," *9th International Conference on Knowledge and Systems Engineering (KSE),* pp. 24-29, 2017. |
| [9] | Staff, "What Is Machine Learning? Definition, Types, and Examples," [Online]. Available: https://www.coursera.org/articles/what-is-machine-learning. [Accessed 4 2024]. |
| [10] | H. A. Cong, "NGHIÊN CỨU LÝ THUYẾT NAIVE BAYES VÀ ỨNG DỤNG PHÂN LOẠI TÀI LIỆU TIẾNG VIỆT TRONG THƯ VIỆN SỐ," *VNU,* pp. 457-468, 2020. |
| [11] | Priya Chakriswaran, Durai Raj Vincent, Kathiravan Srinivasan, Vishal Sharma, Vishal Sharma and Daniel Gutiérrez Reina, "Emotion AI‐Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues," *Applied Sciences,* vol. 9, p. 5462, 2019. |
| [12] | Huma Parveen and Shikha Pandey, "Sentiment Analysis on Twitter Data-set using Naive," *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),* pp. 416-419, 2016. |
| [13] | Arif Abdurrahman Farisi et al, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *Journal of Physics: Conference Series,* vol. 1192, 2019. |
| [14] | Denny, "Understanding Convolutional Neural Networks for NLP," 11 7 2015. [Online]. Available: https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/. [Accessed 4 2024]. |
| [15] | U. Tutorial, "Feature Extraction Using Convolution," 12 5 2017. [Online]. Available: http://ufldl.stanford.edu/tutorial/supervised/FeatureExtractionUsingConvolution/. [Accessed 4 2024]. |
| [16] | Ye Zhang and Byron C. Wallace, "A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification," 2015. |
| [17] | Denny, "Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs," 17 9 2015. [Online]. Available: https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-1/. [Accessed 4 2024]. |
| [18] | "Recurrent Neural Network – RNN in Machine Learning," 2023. [Online]. Available: https://pythongeeks.org/recurrent-neural-network/. [Accessed 4 2024]. |
| [19] | Colah, "Understanding LSTM Networks," 27 8 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 4 2024]. |
| [20] | Kiet Van Nguyen,Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong and Ngan Luu Thuy Nguyen, "UIT-VSFC: Vietnamese Students’ Feedback Corpusfor Sentiment Analysis," *2018 10th International Conference on Knowledge and Systems Engineering (KSE),* pp. 19-24, 2021. |
| [21] | duyet, "vietnamese-stopwords," 2022. [Online]. Available: https://github.com/stopwords/vietnamese-stopwords. [Accessed 4 2024]. |