**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**Lê Thế Thao**

**PHÂN TÍCH CẢM XÚC KHÁCH HÀNG QUA CÁC BÌNH LUẬN TRÊN TRANG WEB BÀN HÀNG**

**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Truyền thông và mạng máy tính**

**HÀ NỘI – 2017**

**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**Lê Thế Thao**

**PHÂN TÍCH CẢM XÚC KHÁCH HÀNG QUA CÁC BÌNH LUẬN TRÊN TRANG WEB BÀN HÀNG**

**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY**

**Ngành: Truyền thông và mạng máy tính**

**Cán bộ hướng dẫn: TS. Nguyễn Văn Nam**

*(ký tên)*

**HÀ NỘI – 2017**

TÓM TẮT

**Tóm tắt:** Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của Internet các trang web bán hàng thương mại điện tử cũng ngày càng phát triển và phổ biến ở mọi nơi mang lại rất nhiều lợi ích cho mọi người trong việc trao đổi mua bán sản phẩm, hàng hóa. Chính bởi sự phát triển nhanh chóng như vậy đã nảy sinh thêm nhiều khó khăn và thách thức trong việc quản lý , đánh giá một khối lượng lớn các sản phẩm ngày một tăng lên. Trong đó vấn đề được chú ý đến nhiều là phân tích cảm xúc bình luận trên các sản phẩm của khách hàng. Lượng thông tin, dữ liệu có giá trị trong đó còn chưa được khai thác hiệu quả. Đây vừa là cơ hội vừa là thách thức để xây dựng một hệ thống phân tích cảm xúc, quan điểm để khách hàng dễ dàng hơn trong việc lựa chọn sản phẩm cũng như các cửa hàng, doanh nghiệp nắm bắt được tình hình, ý kiến của khách hàng qua đó đưa ra những giải quyết, xử lý kịp thời.

***Từ khóa:*** *phân tích cảm xúc*

LỜI CẢM ƠN

Tôi xin bày tỏ lòng biết ơn sâu sắc tới TS. Nguyễn Văn Nam, thầy đã truyền đạt kiến thức, kĩ năng trong học tập và nghiên cứu, hướng dẫn tận tâm giúp tôi hoàn thành khóa luận này.

Và tôi cũng xin chân thành cảm ơn các thầy, cô, các cán bộ làm việc tại trường Đại học Công nghệ, Đại Học Quốc Gia Hà Nội và đặc biệt là các thầy, cô trong khoa Công nghệ thông tin đã trang bị nền tảng kiến thức giúp tôi trong quá trình học tập và làm khóa luận.

Cuối cùng, tôi xin gửi lời cảm ơn tới gia đình, người thân, bạn bè đã quan tâm, động viên, tạo điều kiện cho tôi trong quá trình học tập.

|  |  |
| --- | --- |
|  | Hà Nội, ngày tháng năm  Sinh viên  Lê Thế Thao |

LỜI CAM ĐOAN

Tôi xin cam đoan đã thực hiện quá trình làm khóa luận một cách khoa học, trung thực, đúng đắn dưới sự hướng dẫn của TS.Nguyễn Văn Nam.

Các tài liệu, bài báo khoa học, giáo trình giảng dạy nếu được sử dụng trong khóa luận đều được trích dẫn đầy đủ. Nếu có điều gì sai trái tôi xin chịu trách nhiệm trước hội đồng về kết quả khóa luận của mình.

|  |  |
| --- | --- |
|  | Hà Nội, ngày tháng năm  Sinh viên  Lê Thế Thao |

MỤC LỤC

[CHƯƠNG 1 : Tổng quan về phân tích cảm xúc 3](#_Toc499547316)

[1.1 Giới thiệu chung 3](#_Toc499547317)

[1.1.1 Phân tích cảm xúc là gì? 3](#_Toc499547318)

[1.1.2 Mục đích của phân tích cảm xúc 3](#_Toc499547319)

[1.1.3 Ý nghĩa của phân tích cảm xúc 3](#_Toc499547320)

[1.1.4 Một số ứng dụng của phân tích cảm xúc 4](#_Toc499547321)

[1.2 Các kỹ thuật phân tích cảm xúc 4](#_Toc499547322)

[1.2.1 Phân lớp Naïve Bayes 5](#_Toc499547323)

[1.2.2 Mô hình phân lớp SVM 6](#_Toc499547324)

[1.2.3 Thuật toán Neural Network 7](#_Toc499547325)

[1.3 Đánh giá và kết luận 17](#_Toc499547326)

[CHƯƠNG 2: XÂY DỰNG HỆ THỐNG PHÂN TÍCH CẢM XÚC DỰA TRÊN NEURAL NETWORK 19](#_Toc499547327)

[2.1 Mô hình xử lý dữ liệu cho bài toán phân loại cảm xúc 19](#_Toc499547328)

[2.2 Mô hình hệ thống 20](#_Toc499547329)

[2.2.1 Thu thập dữ liệu 21](#_Toc499547330)

[2.2.2 Tiền xử lý dữ liệu 22](#_Toc499547331)

[2.2.3 Trích xuất đặc trưng 23](#_Toc499547332)

[2.2.4 Phân loại dữ liệu 23](#_Toc499547333)

[2.2.5 Huấn luyện mô hình 24](#_Toc499547334)

[2.3 Kết quả phân lớp và đánh giá 27](#_Toc499547335)

[CHƯƠNG 3: CÀI ĐẶT HỆ THỐNG 29](#_Toc499547336)

[3.1 Mô hình cài đặt 29](#_Toc499547337)

[3.1.1 Python 29](#_Toc499547338)

[3.1.2 Scrapy 30](#_Toc499547339)

[3.1.3 SQLite 30](#_Toc499547340)

[3.1.4 Khung làm việc Tensorflow 30](#_Toc499547341)

[3.1.5 Django 31](#_Toc499547342)

[3.2 Cài đặt hệ thống 31](#_Toc499547343)

[3.2.1 Môi trường cài đặt 31](#_Toc499547344)

[3.2.2 Cài đặt cụ thể 32](#_Toc499547345)

[3.3 Kết quả và đánh giá 32](#_Toc499547346)

[CHƯƠNG 4: ĐÁNH GIÁ 33](#_Toc499547347)

[4.1 Hoạt động của hệ thống 33](#_Toc499547348)

[4.2 Đánh giá hiệu quả 38](#_Toc499547349)

[4.3 Kết luận 38](#_Toc499547350)

[KẾT LUẬN 39](#_Toc499547351)

[TÀI LIỆU THAM KHẢO 40](#_Toc499547352)

BẢNG KÍ HIỆU VIẾT TẮT

|  |  |
| --- | --- |
| SVM | Support Vector Machine |
| CNNs | Convolutional Neural Networks |
| OM | Opinion Mining |
| SA | Sentiment Analysis |
| ANN | Artificial Neural Network |
| NLP | Natural Language Processing |
| RELU | Rectified Linear Unit |

MỤC LỤC BẢNG BIỂU

MỤC LỤC HÌNH VẼ

**MỞ ĐẦU**

Ngày nay, các trang thương mại điện tử như Lazada.vn, Tiki.vn, Adayroi.vn không chỉ còn là nơi cung cấp thông tin sản phẩm và bán hàng mà còn là nơi người dùng bày tỏ cảm xúc, chia sẻ, trao đổi ý kiến và góp ý. Hiện nay mới chỉ có hệ thống rating đơn giản để thể hiện đánh giá của khách hàng nhưng như thế vẫn chưa đủ đáp ứng được nhu cầu của người dùng cũng như người bán hàng. Người dùng thì luôn muốn biết bình luận đánh giá của các khách hàng đã mua như thế nào để đánh giá sản phẩm đó có tốt không và điều này ảnh hưởng rất lớn đến quyết định mua hàng hay không của họ. Người bán muốn nắm bắt được những phản hồi, góp ý trực tiếp từ khách hàng để kịp thời xử lí và điều chỉnh thông tin, dịch vụ chăm sóc cho phù hợp, tránh các thông không chính xác hay mất khách hàng. Thông qua thu thập, phân tích bình luận trên những kênh bán hàng không chỉ giúp người dùng, người bán hàng xử lý kịp thời các vấn đề mà các công ty và doanh nghiệp thông qua đó cũng biết được sự phổ biến của thương hiệu, sản phẩm của mình và hiệu quả của các chiến dịch quảng cáo trên đó. Vì lẽ đó, việc đánh giá cảm xúc của khách hàng qua những phản hồi ý kiến, đóng góp về sản phẩm là vô cùng quan trọng, hữu ích và có nhu cầu rất lớn. Chính vì thế, chúng tôi quyết định chọn đề tài luận văn của mình là “Phân tích cảm xúc khách hàng qua các bình luận trên trang web bán hàng”. Luận văn tập trung vào xây dựng và thực nghiệm mô hình phân lớp quan điểm bình luận của khách hàng sau đó phân tích đánh giá kết quả đó.

**Một số giải pháp hiện tại:** Hiện nay, bài toán phân loại cảm xúc được sử dụng trong rất nhiều lĩnh vực khác nhau và cũng có nhiều hướng tiếp cận, giải quyết khác nhau như: sử dụng bộ từ vựng, sử dụng quy luật theo ngữ nghĩa và sử dụng học máy ... Trong số đó nổi bật lên là phương pháp sử dụng học máy đạt hiệu quả và cho kết quả tốt đặc biệt với những bài toán có lượng dữ liệu lớn. Các thuật toán học máy phổ biến được sử dụng trong giải quyết bài toán phân lớp quan điểm là: Naïve Bayes, Support Vector Machine, Neural Network. Và trong vài năm gần đây, ứng dụng thành công của học sâu (Deep Learning) vào trong bài toán phân loại quan điểm đạt được kết quả cao [1]. Tuy nhiên đa phần là áp dụng cho tiếng Anh, việc phân loại cảm xúc người dùng cho tiếng Việt còn nhiều hạn chế. Sau quá trình tìm hiểu và thực nghiệm chúng tôi quyết định sử dụng thuật toán Neural Network, cụ thể là mô hình thuật toán Convolutional Neural Networks (CNNs) áp dụng vào giải quyết bài toán này.

**Những đóng góp và kết quả đạt được:** Sau quá trình xây dựng và thực nghiệm mô hình phân lớp quan điểm sử dụng mạng CNNs chúng tôi đã xây dựng được hệ thống phân loại quan điểm dành cho tiếng Việt đạt độ chính xác rất khả quan và sử dụng kết quả đó để xây dựng trang web đánh giá cảm xúc bình luận của khách hàng.

**Nội dung của luận văn:** Ngoài phần giới thiệu, kết luận và tài liệu tham khảo, nội dung của luận văn bao gồm 4 chương:

Chương 1: Tổng quan về phân tích cảm xúc

Chương này sẽ giới thiệu phân tích cảm xúc là gì, mục đích, ý nghĩa của phân tích các xúc. Tiếp theo là các kỹ thuật phân tích cảm xúc.

Chương 2: Xây dựng hệ thống phân tích cảm xúc dựa trên neural network

Chương này sẽ đề xuất mô hình phân lớp quan điểm đồng thời trình bày chi tiết các pha cũng như các bước trong mô hình, huấn luyện và chạy mô hình.

Chương 3: Cài đặt hệ thống

Chương này trình bày mô hình cài đặt hệ thống, chi tiết cài đặt các công cụ phần mềm cần thiết để chạy hệ thống.

Chương 4: Đánh giá

Chương này trình bày quá trình chạy thử và hoạt động của hệ thống sau đó đánh giá hiệu quả và đưa ra kết luận.

# CHƯƠNG 1 : Tổng quan về phân tích cảm xúc

Chương này sẽ giới thiệu tổng quan về phân tích cảm xúc: phân tích cảm xúc là gì? Mục đích và ý nghĩa của phân tích cảm xúc. Các kỹ thuật phân tích cảm xúc

## Giới thiệu chung

Phân tích cảm xúc là gì?

Phân tích cảm xúc (sentiment analysis) hay khai phá quan điểm (opinion mining) là việc tính toán, nghiên cứu về các ý kiến, đánh giá, thái độ và cảm xúc của con người đối với các thực thể, cá nhân, vấn đề, sự kiện, chủ đề và các thuộc tính của những đối tượng đó [2].

Nhiệm vụ cơ bản của phân tích quan điểm ​​là phân loại phân cực (polarity classification). Phân loại phân cực xảy ra khi một mẩu văn bản đưa ra ý kiến ​​về một vấn đề duy nhất được phân loại là một trong hai cảm xúc đối lập. Ví dụ như: các bài đánh giá như "like" so với "dislike" là những ví dụ phân loại phân cực. Phân loại phân cực cũng xác định cách diễn tả là tốt hay không tốt trong các bài đánh giá trực tuyến và giúp đánh giá sản phẩm đáng tin cậy hơn [3].

Mục đích của phân tích cảm xúc

Với sự phát triển bùng nổ của các phương tiện truyền thông xã hội (các bài đánh giá, thảo luận trên diễn đàn, blog và mạng xã hội) trên web, các cá nhân và tổ chức ngày càng sử dụng các ý kiến ​​công khai trong các phương tiện truyền thông này để đưa ra quyết định. Tuy nhiên, tìm kiếm và giám sát các trang bình luận trên web và trích lọc thông tin chứa trong chúng vẫn là một thách thức to lớn vì sự gia tăng của các trang web ngày càng đa dạng. Mỗi trang web thường chứa một khối lượng lớn các văn bản có ý kiến ​​mà không dễ dàng giải mã trong các diễn đàn và blog. Người đọc bình thường sẽ gặp khó khăn trong việc xác định các thông tin và ý kiến ​​có trong đó. Hơn nữa, phân tích của mọi người về thông tin có sự sai khác nhau, ví dụ như mọi người thường chú ý nhiều hơn tới các ý kiến ​​phù hợp với sở thích của họ. Mọi người cũng gặp khó khăn, do những hạn chế về tâm lý và thể chất khi lượng thông tin được xử lý lớn. Do đó cần có các hệ thống khai thác phân tích và tổng hợp ý kiến ​​tự động, vì những ý kiến chủ quan và những hạn chế về tâm lý có thể được khắc phục bằng hệ thống phân tích cảm xúc khách quan [2].

Ý nghĩa của phân tích cảm xúc

Ý nghĩa của việc phân tích cảm xúc:

* Xác định chính xác quan điểm được đưa ra trong một tài liệu
* Đưa ra các kết quả phân tích một cách khách quan
* Trích lọc được những thông tin quan điểm có giá trị
* Dựa vào đó có thể đưa ra các quyết định phù hợp và chính xác

Một số ứng dụng của phân tích cảm xúc

Khi người tiêu dùng đưa ra lựa chọn sản phẩm, một thông tin quan trọng là độ tín nhiệm của sản phẩm đó được lấy từ ý kiến ​​của những người tiêu dùng khác và phân tích cảm xúc có thể cho biết những người khác nghĩ gì về một sản phẩm. Ứng dụng đầu tiên của phân tích cảm xúc là đưa ra chỉ dẫn và khuyến nghị trong việc lựa chọn sản phẩm theo số đông người tiêu dùng.. Phân tích cảm xúc có thể tập hợp ý kiến và đánh giá xếp hạng về các khía cạnh khác nhau của sản phẩm.

Một ứng dụng khác của phân tích cảm xúc là dành cho các công ty muốn biết ý kiến ​​của khách hàng về sản phẩm của họ. Sau đó, họ có thể cải thiện các vấn đề mà khách hàng thấy không hài lòng.

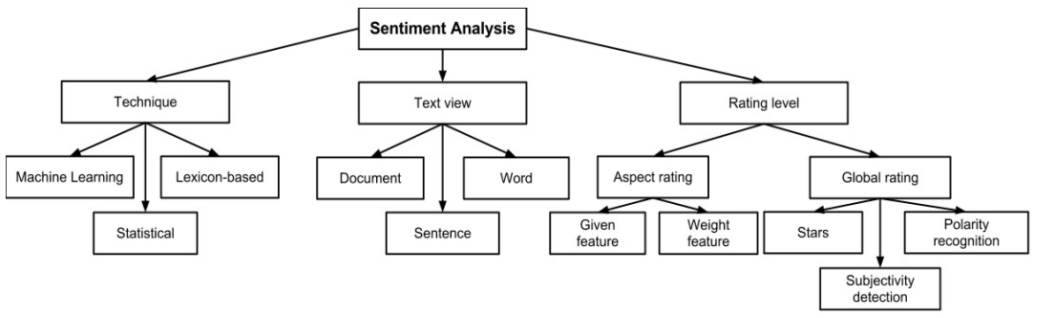
## Các kỹ thuật phân tích cảm xúc

Các kỹ thuật phân tích cảm xúc có thể được phân loại theo nhiều quan điểm khác nhau: kỹ thuật được sử dụng, quan điểm của văn bản, mức độ chi tiết của phân tích văn bản, mức độ đánh giá [4] … hình 1.1

Theo đặc điểm cấu trúc văn bản có thể chia theo 3 mức độ khác nhau: mức tài liệu, mức câu và mức từ, đặc trưng. Phân loại mức tài liệu nhằm tìm ra một phân cực cảm xúc cho toàn bộ, trong khi phân loại mức câu hoặc mức từ thể hiện sự phân cực đối với mỗi câu và thậm chí cho mỗi từ của một bài đánh giá. Hầu hết các phương pháp có xu hướng tập trung vào phân loại mức tài liệu [4].

Từ quan điểm kỹ thuật ta có thể chia thành 4 cách tiếp cận [4]:

* Cách tiếp cận dựa trên từ điển (lexicon-based): bao gồm việc tính toán phân cực (polarity) cảm xúc sử dụng theo hướng ngữ nghĩa của các từ hoặc câu để xác định tính chủ quan và ý kiến ​​trong văn bản.
* Cách tiếp cận dựa trên quy tắc (rule-based): tìm kiếm các từ chứa quan điểm ​​trong một văn bản và sau đó phân loại nó dựa trên số từ tích cực và tiêu cực., sử dụng các quy tắc phân loại khác nhau như từ điển phân cực, từ phủ định, từ tăng cường, thành ngữ, biểu tượng cảm xúc, ý kiến ​​hỗn hợp …
* Các mô hình thống kê (statistical models): thể hiện mỗi đánh giá như một sự kết hợp của các khía cạnh ẩn và xếp hạng. Giả định rằng các khía cạnh và xếp hạng có thể được đại diện bởi các phân bố đa thức và cố gắng gộp từ đầu vào các khía cạnh và cảm xúc thành xếp hạng
* Phương pháp học máy (machine learning) sử dụng một số thuật toán học để xác định cảm xúc bằng cách huấn luyện trên một tập dữ liệu đã biết.



Hình 1.1 Các kỹ thuật phân tích cảm xúc

Theo quan điểm kỹ thuật, trong phương pháp học máy lại có các thuật toán học máy có giám sát được sử dụng phổ biến trong phân lớp quan điểm như: Naïve Bayes, Support Vector Machine (SVM), Neural Network. Trong phần này chúng tôi sẽ trình bày về các thuật toán này rồi từ đó đưa ra những ưu nhược điểm của từng thuật toán và quyết định lựa chọn thuật toán phù hợp.

Phân lớp Naïve Bayes

Bộ phân lớp quan điểm Naïve Bayes được xây dựng dựa trên lý thuyết Bayes về xác suất có điều kiện:

Mục tiêu là tìm được phân lớp c\* sao cho là lớn nhất hay xác suất của tài liệu d thuộc lớp c\* là lớn nhất ⟺ P(c).P(d|c) lớn nhất.

Để có thể xấp xỉ giá trị của P(d|c), thuật toán Naïve Bayes giả sử rằng: các vector đặc trưng fi của một tài liệu khi đã biết phân lớp là độc lập với nhau. Từ đó ta có công thức:

Trong đó f là các vector đặc trưng cho tài liệu d.

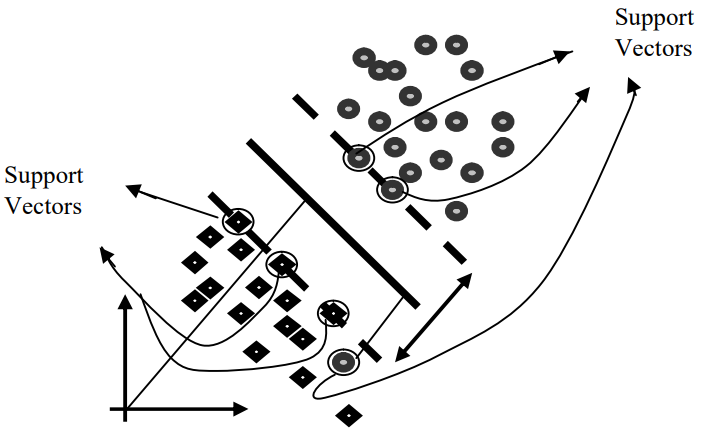
Khi tiến hành huấn luyện, thuật toán sử dụng phương pháp xấp xỉ hợp lý cực đại MLE (Maximum Likelihood Estimation) để xấp xỉ P(c) và . Ta có:

Trong đó Nc là số văn bản được phân loại vào lớp c; N là tổng số văn bản trong tập huấn luyện.

Trong đó là số lần xuất hiện của vector đặc trưng i trong tài liệu thuộc phân lớp c.

Mô hình phân lớp SVM

SVM là mô hình học có giám sát. Mô hình này được kết hợp với một thuật toán học phân tích dữ liệu và xác định mẫu phân loại. Khái niệm thuật toán SVM dựa trên siêu phẳng quyết định để xác định ranh giới của việc quyết định. Một siêu phẳng quyết định chia tách một nhóm thành các lớp khác nhau. Ví dụ: xem xét một thể hiện thuộc về lớp Circle hay lớp Diamond. Có một đường chia tách được coi là ranh giới (Hình 1.2). Tất cả các thể hiện bên phải đường ranh giới này là Circle và tất cả phía bên trái là Diamond.[5].



Hình 1.2 Nguyên tắc của SVM

Một tập huấn luyện được biểu diễn trong không gian vector với mỗi điểm là biểu diễn của một dữ liệu. SVM sẽ tìm ra một siêu phẳng f mà có khoảng cách lề (margin) giữa hai lớp là lớn nhất để chia các điểm trên không gian này thành hai lớp riêng biệt, tương ứng là lớp “1” và lớp “-1”. Khoảng cách lề càng lớn thì siêu phẳng quyết định càng tốt và chất lượng phân lớp càng cao.

Có một số phần mở rộng làm cho SVM mạnh mẽ và tương thích hơn với các vấn đề thực tế. Những phần mở rộng bao gồm:

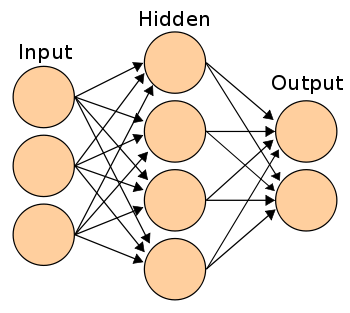
* Phân chia lề mềm (Soft Margin Classification)
* Phân chia không tuyến tính (Non-linear Classification)
* Phân chia đa lớp (Multiclass SVM)

Thuật toán Neural Network

Kiến trúc mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) là một mô hình toán học hay mô hình tính toán được xây dựng dựa trên các mạng nơ-ron sinh học.

Một mạng nơ-ron là một nhóm các nút nối với nhau, mô phỏng mạng nơ ron thần kinh của não người. Mạng nơ ron nhân tạo được thể hiện thông qua ba thành phần cơ bản: mô hình của nơ ron, cấu trúc và sự liên kết giữa các nơ ron.Trong nhiều trường hợp, mạng nơ ron nhân tạo là một hệ thống thích ứng, tự thay đổi cấu trúc của mình dựa trên các thông tin bên ngoài hay bên trong chạy qua mạng trong quá trình học.



Hình 1.3 Kiến trúc mạng nơ-ron nhân tạo

Kiến trúc chung của một ANN gồm 3 thành phần đó là Input Layer, Hidden Layer và Output Layer (Hình 1.3)

Trong đó, lớp ẩn (Hidden Layer) gồm các nơ-ron, nhận dữ liệu đầu vào từ các nơ-ron ở lớp trước đó và chuyển đổi các đầu vào này cho các lớp xử lý tiếp theo. Trong một mạng ANN có thể có nhiều Hidden Layer.

ANN được huấn luyện theo ba kỹ thuật cơ bản đó là: học có giám sát (supervised learning) và học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning)

Mạng nơron nhân tạo đã được sử dụng để giải quyết nhiều bài toán thuộc nhiều lĩnh vực, các ngành khác nhau. Các nhóm ứng dụng mà mạng nơ-ron nhân tạo đã được áp dụng rất có hiệu quả là:

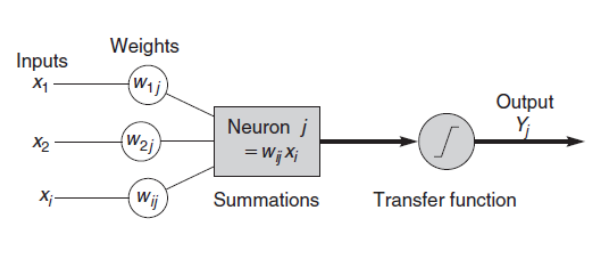
* Bài toán phân lớp (classification): Phân loại các đối tượng quan thành thành các nhóm. Ví dụ: phân loại chữ viết, văn bản, nhận diện hình ảnh ...
* Bài toán dự đoán (preditive): Mạng nơ-ron nhân tạo đã được ứng dụng thành công trong việc xây dựng các mô hình dự báo sử dụng tập dữ liệu trong quá khứ để dự đoán số liệu trong tương lai. Ví dụ: dữ báo thiên tai, dự báo chứng khoán ...

Lợi thế lớn nhất của các mạng ANN là khả năng được sử dụng như một cơ chế xấp xỉ hàm tùy ý mà “học” được từ các dữ liệu quan sát. Tuy nhiên, sử dụng chúng không đơn giản như vậy, một số các đặc tính và kinh nghiệm khi thiết kế một mạng nơ-ron ANN.

* **Chọn mô hình**: Điều này phụ thuộc vào cách trình bày dữ liệu và các ứng dụng. Mô hình quá phức tạp có xu hướng dẫn đến những thách thức trong quá trình học.
* **Thuật toán học:** Có hai vấn đề cần học đối với mỗi mạng ANN, đó là học tham số của mô hình (parameter learning) và học cấu trúc (structure learning). Học tham số là thay đổi trọng số của các liên kết giữa các nơ-ron trong một mạng, còn học cấu trúc là việc điều chỉnh cấu trúc mạng bằng việc thay đổi số lớp ẩn, số nơ-ron mỗi lớp và cách liên kết giữa chúng. Hai vấn đề này có thể được thực hiện đồng thời hoặc tách biệt.

Nếu các mô hình, hàm chi phí và thuật toán học được lựa chọn một cách thích hợp, thì mạng ANN sẽ cho kết quả có thể vô cùng mạnh mẽ.

Hoạt động của mạng nơ-ron nhân tạo



Hình 1.8 Quá trình xử lý thông tin của một mạng nơ-ron nhân tạo

**Inputs**: Mỗi Input tương ứng với 1 đặc trưng của dữ liệu. Ví dụ như trong ứng dụng của ngân hàng xem xét có chấp nhận cho khách hàng vay tiền hay không thì mỗi input là một thuộc tính của khách hàng như thu nhập, nghề nghiệp, tuổi, số con,…

**Outputs**: Kết quả của một ANN là một giải pháp cho một vấn đề, ví dụ như với bài toán xem xét chấp nhận cho khách hàng vay tiền hay không thì output là yes hoặc no.

**Connection Weights** (Trọng số liên kết) : Đây là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng, độ mạnh của dữ liệu đầu vào đối với quá trình xử lý thông tin chuyển đổi dữ liệu từ lớp này sang lớp khác. Quá trình học của ANN thực ra là quá trình điều chỉnh các trọng số weight của các dữ liệu đầu vào để có được kết quả mong muốn.

**Summation Function** (Hàm tổng): Tính tổng trọng số của tất cả các input được đưa vào mỗi nơ-ron. Hàm tổng của một nơ-ron đối với n input được tính theo công thức sau:

**Transfer Function** (Hàm chuyển đổi): Hàm tổng của một nơ-ron cho biết khả năng kích hoạt của nơ-ron đó còn gọi là kích hoạt bên trong. Các nơ-ron này có thể sinh ra một output hoặc không trong ANN, nói cách khác rằng có thể output của 1 Nơ-ron có thể được chuyển đến layer tiếp theo trong mạng Nơ-ron hoặc không. Mối quan hệ giữa hàm tổng và kết quả output được thể hiện bằng hàm chuyển đổi.

Việc lựa chọn **Transfer Function** có tác động lớn đến kết quả của ANN. Hàm chuyển đổi phi tuyến được sử dụng phổ biến trong ANN là sigmoid (logical activation) function.

)

Kết quả của Sigmoid Function thuộc khoảng [0,1] nên còn gọi là hàm chuẩn hóa (Normalized Function).

Kết quả xử lý tại các nơ-ron (Output) đôi khi rất lớn, vì vậy transfer function được sử dụng để xử lý output này trước khi chuyển đến layer tiếp theo. Đôi khi thay vì sử dụng Transfer Function người ta sử dụng giá trị ngưỡng (Threshold value) để kiểm soát các output của các neuron tại một layer nào đó trước khi chuyển các output này đến các Layer tiếp theo. Nếu output của một neuron nào đó nhỏ hơn Threshold thì nó sẻ không được chuyển đến Layer tiếp theo.

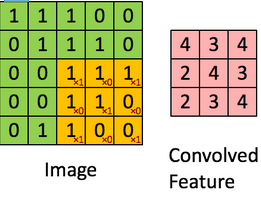
Mô hình Convolutional Neural Networks

CNNs là một trong những thuật toán Deep Learning cho kết quả tốt nhất hiện nay trong hầu hết các bài toán về thị giác máy như phân lớp, nhận dạng, …

Hiện nay, CNNs đã được áp dụng vào các vấn đề trong xử lý ngôn ngữ tự nhiên (NLP - Natural Language Processing) và nhận được kết quả tốt. Trong phần này tôi sẽ tóm tắt CNNs là gì và cách chúng được sử dụng trong NLP.

* **Tích chập là gì**

Tích chập (Convolution) là một cửa sổ trượt (Sliding Windows) trên một ma trận như mô tả hình dưới:

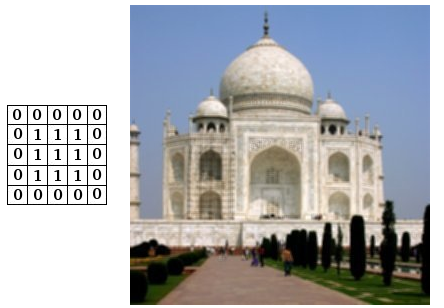


Hình 1.9 Cửa sổ tích chập

Ma trận bên trái là một hình ảnh trắng đen được số hóa. Ma trận có kích thước 5x5 và mỗi điểm ảnh có giá trị 1 hoặc 0 là giao điểm của dòng và cột. Sliding Window hay còn gọi là kernel, filter là một ma trận có kích thước nhỏ như trong ví dụ trên là 3x3. Convolution hay tích chập là nhân từng phần tử bên trong ma trận 3x3  với ma trận bên trái. Kết quả được một ma trận gọi là Convoled Feature được sinh ra từ việc nhận ma trận filter với ma trận ảnh 5x5 bên trái.

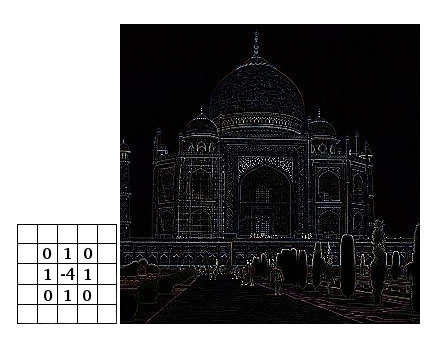
Công dụng của tích chập:

* Nếu lấy giá trị trung bình xung quanh cho vị trí điểm ảnh trung tâm sẽ làm mờ bức ảnh ban đầu.



Hình 1.10 Làm mờ bức ảnh bằng tích chập

* Nếu tính tích phân sẽ phát hiện được các biên cạnh giữa các điểm ảnh lân cận bằng cách thêm vào hoặc hủy các giá trị bằng 0 trong ma trận. Nếu có một đường sắc nét nghĩa là giá trị điểm ảnh tại vị trí đó cao thì quá trình biến đổi sẽ tạo ra một sự tách biệt giữa các giá trị màu trắng và đen khi đó sẽ phát hiện ra các đường biên rõ.



Hình 1.11 Phát hiện các đường biên của bức ảnh bằng tích chập

* **Mạng Convolutional Neural**

CNNs về cơ bản chỉ là một vài lớp của tích chập và sử dụng các hàm kích hoạt phi tuyển (nonlinear activation) như ReLU và tanh để áp dụng vào kết quả. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural đầu vào cho mỗi neural đầu ra trong các lớp tiếp theo. Mô hình này gọi là mạng kết nối đầy đủ (fully connected layer) hay mạng toàn vẹn (affine layer). Còn trong mô hình CNNs thì không làm vậy mà sử dụng convolutions trên lớp đầu vào để tính đầu ra. Layer tiếp theo sử dụng kết quả convolution từ layer trước đó, điều này dẫn đến kết nối cục bộ.

Mỗi lớp đều áp dụng các bộ lọc khác nhau, thường là hàng trăm hoặc hàng ngàn và kết hợp các kết quả của chúng lại. Ngoài ra có một số layer khác như pooling (subsampling) layer dùng để chắt lọc lại các thông tin hữu ích hơn (loại bỏ các thông tin nhiễu). Trong quá trình huấn luyện mạng (traning)  CNNs tự động học các giá trị qua các lớp bộ lọc (filter) dựa vào cách thức mà bạn thực hiện. Ví dụ: trong phân loại hình ảnh CNNs có thể tìm hiểu để phát hiện các cạnh từ các pixel thô trong lớp đầu tiên, sau đó sử dụng các cạnh để phát hiện các hình dạng đơn giản trong lớp thứ hai và sau đó sử dụng các hình dạng này để ngăn các đặc trưng mức cao hơn, chẳng hạn như hình dạng khuôn mặt ở các lớp cao hơn. Lớp cuối cùng là phân loại sử dụng những đặc trưng này.

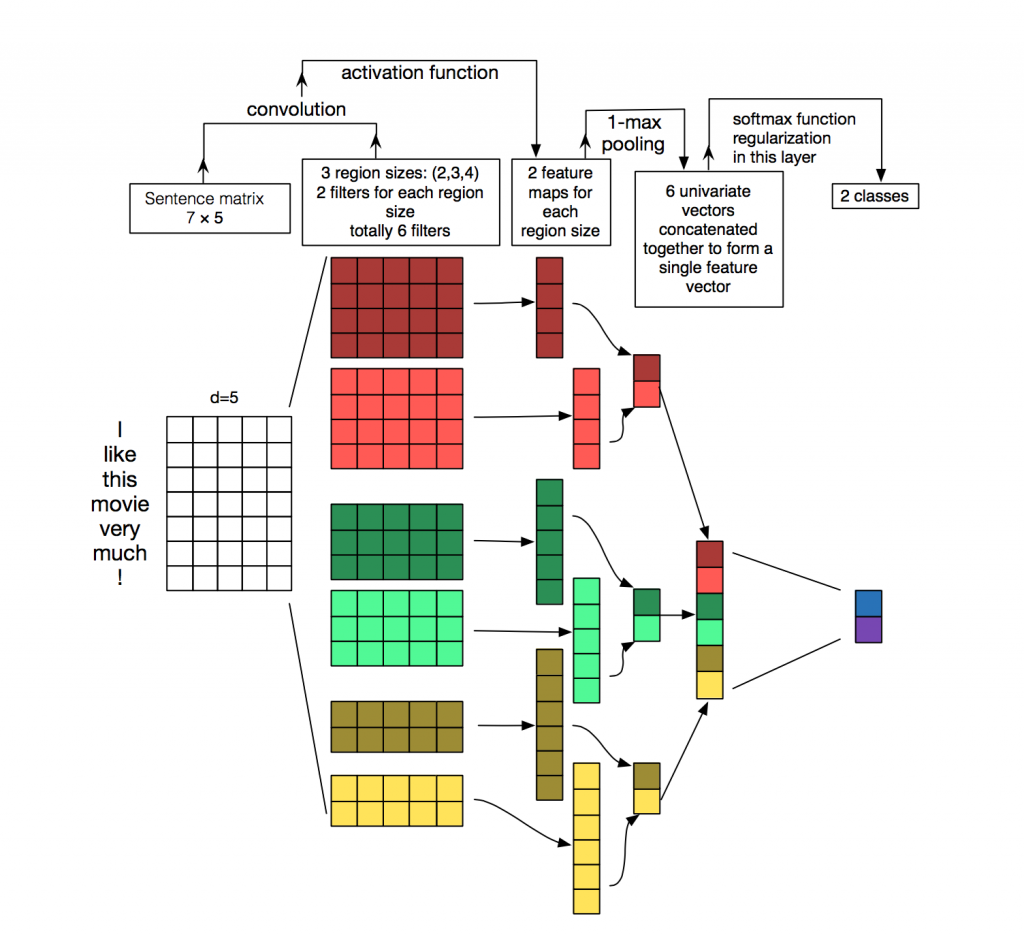


Hình 1.12 Các lớp trong mô hình CNNs

Trong mô hình CNNs có 2 khía cạnh cần quan tâm là Location Invariance (tính  bất biến) and Compositionality (tính kết hợp). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các gốc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Pooling layer sẽ cho bạn tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các bộ lọc filter. Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. cũng giống như cách con người nhận biết các vật thể trong tự nhiên.

Thay vì các điểm ảnh, đầu vào cho hầu hết các tác vụ NLP là các câu hoặc tài liệu được biểu diễn dưới dạng ma trận. Mỗi hàng của ma trận tương ứng với một từ hoặc nó có thể là một ký tự. Mỗi hàng là vector đại diện cho một từ. Thông thường, những vector này là words embeddings (hay word vector) như word2vec, nhưng chúng cũng có thể là các vector với chỉ mục từ trong một bộ từ vựng (vocabulary). Đối với một câu 10 từ sử dụng 100 chiều sẽ có một ma trận 10×100 là đầu vào như một hình ảnh.

Trong thị giác máy tính, các bộ lọc trượt qua các vùng cục bộ của một hình ảnh, nhưng trong NLP thường sử dụng bộ lọc trượt qua hàng đầy đủ của ma trận từ. Do đó, "chiều rộng" của bộ lọc của thường bằng với chiều rộng của ma trận đầu vào. Chiều cao, hoặc kích thước vùng, có thể thay đổi, nhưng cửa sổ trượt trên 2-5 từ một lần là điển hình. Kết hợp tất cả những điều trên với nhau, CNNs cho NLP có thể trông như sau [6]:

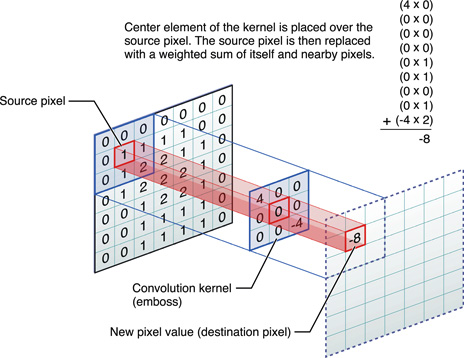


Hình 1.13 Mô hình CNNs cho phân loại văn bản

* **Các lớp trong mô hình**

**Lớp Convolution:**

Lớp này chính là nơi thể hiện tư tưởng ban đầu của mạng nơ-ron tích chập. Thay vì kết nối toàn bộ điểm ảnh, lớp này sẽ sử dụng một bộ các bộ lọc filters có kích thước nhỏ so với ảnh (thường từ 2 đến 5) áp vào một vùng trong ảnh và tiến hành tính tích chập giữa bộ lọc filter và giá trị điểm ảnh trong vùng cục bộ đó. Bộ lọc filter sẽ lần lượt được dịch chuyển theo một giá trị bước trượt (stride) chạy dọc theo ảnh và quét toàn bộ ảnh.



Hình 1.14 Tính tích chập với bộ lọc filter

Như vậy nếu với một bức ảnh 32×32 và một filter 3×3, ta sẽ có kết quả là một tấm ảnh mới có kích thước 32×32 (với điều kiện đã thêm padding vào ảnh gốc để tính tích chập cho các trường hợp filter quét ra các biên cạnh) là kết quả tích chập của filter và ảnh. Với bao nhiêu filter trong lớp này thì ta sẽ có bấy nhiêu ảnh tương ứng mà lớp này trả ra và được truyền vào lớp tiếp theo. Các trọng số của filter ban đầu sẽ được khởi tạo ngẫu nhiên và sẽ được học dần trong quá trình huấn luyện mô hình.

**Lớp RELU – Rectified Linear Unit:**

Lớp này thường được cài đặt ngay sau lớp Convolutional. Lớp này sử dụng hàm kích hoạt . Nói một cách đơn giản, lớp này có nhiệm vụ chuyển toàn bộ giá trị âm trong kết quả lấy từ lớp Convolutional thành giá trị 0. Ý nghĩa của cách cài đặt này chính là tạo nên tính phi tuyến cho mô hình. Tương tự như trong mạng truyền thẳng, việc xây dựng dựa trên các phép biến đổi tuyến tính sẽ khiến việc xây dựng đa tầng đa lớp trở nên vô nghĩa. Có rất nhiều cách để khiến mô hình trở nên phi tuyến như sử dụng các hàm kích hoạt sigmoid, tanh, … nhưng hàm dễ cài đặt, tính toán nhanh mà vẫn hiệu quả.

**Lớp Pooling:**

Lớp này sử dụng một cửa sổ trượt quét qua toàn bộ ảnh dữ liệu, mỗi lần trượt theo một bước trượt (stride) cho trước. Khác với lớp Convolutional, lớp Pooling không tính tích chập mà tiến hành lấy mẫu (subsampling). Khi cửa sổ trượt trên ảnh, chỉ có một giá trị được xem là giá trị đại diện cho thông tin ảnh tại vùng đó (giá trị mẫu) được giữ lại. Các phương thức lấy phổ biến trong lớp Pooling là MaxPooling ( lấy giá trị lớn nhất), MinPooling (lấy giá trị nhỏ nhất) và AveragePooling (lấy giá trị trung bình).

Lớp Pooling có vai trò giảm kích thước dữ liệu. Với một bức ảnh kích thước lớn qua nhiều lớp Pooling sẽ được thu nhỏ lại tuy nhiên vẫn giữ được những đặc trưng cần cho việc nhận dạng (thông qua cách lấy mẫu). Việc giảm kích thước dữ liệu sẽ làm giảm lượng tham số, tăng hiệu quả tính toán và góp phần kiểm soát hiện tượng quá khớp (overfitting).

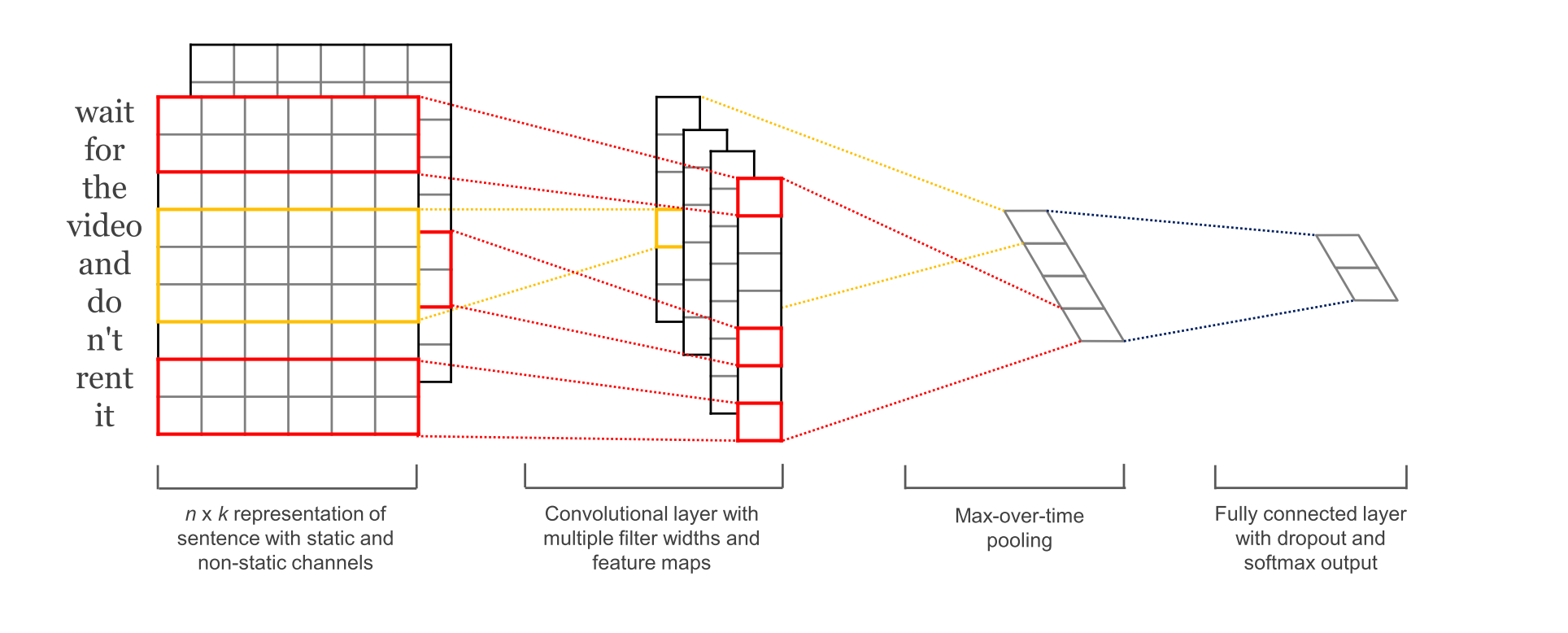
**Lớp FC – fully connected:**

Sau khi ảnh được xử lý và rút trích đặc trưng từ các lớp trước đó, dữ liệu ảnh sẽ không còn quá lớn so với mô hình truyền thẳng nên ta có thể sử dụng mô hình truyền thẳng để tiến hành nhận dạng. Tóm lại, lớp fully-connected đóng vai trò như một mô hình phân lớp và tiến hành dựa trên dữ liệu đã được xử lý ở các lớp trước đó.

* **Hoạt động của mô hình**

Một mạng nơ-ron tích chập được hình thành bằng cách ghép các lớp nêu trên lại với nhau. Mô hình bắt đầu với lớp Convolutional. Lớp RELU thường luôn được cài đặc ngay sau lớp Convolutional hoặc thậm chí kết hợp cả hai lớp này thành một lớp. Các lớp tiếp theo có thể là Convolutional hay Pooling tùy theo kiến trúc mà ta muốn xây dựng. Cuối cùng sẽ là lớp fully-connected để tiến hành phân lớp.Để xem mô hình này hoạt động như thế nào ta có thể xét một kiến trúc sau đây:

Conv1 (với RELU) – Pooling – FC



Hình 1.15 Hoạt động của mô hình CNNs

Có thể coi đầu vào ở đây là một hình ảnh với kích thước AxB sẽ được đưa vào lớp Conv1 (Convolutional kết hợp RELU) gồm các cửa sổ có kích thước thường từ 2 đến 5, mỗi filter sẽ được dùng để tính tích chập với ảnh và cho ra một ảnh kết quả tương ứng. Nếu 4 filter ta sẽ có 4 ảnh kết quả như sau:

Mỗi ảnh trên đều có kích thước tương ứng là 1×B. Sau đó, cả 4 ảnh này đều được cho qua lớp Pooling và kết quả trả ra sẽ là 4 ảnh có kích thước 1x1.  
Với kích thước đủ nhỏ như vậy, lớp Fully-connected tiếp theo sẽ xử lý và đưa ra kết quả phân lớp hay kết quả nhận dạng.

Tương tự như mạng nơ-ron truyền thẳng, mạng nơ-ron tích chập cũng là một mô hình học cho nên khởi tạo ban đầu cho các trọng số trong mạng là ngẫu nhiên và sẽ được điều chỉnh thông qua quá trình học. Thuật toán học cho mạng nơ-ron tích chập cũng tương tự như mạng nơ-ron truyền thẳng là thuật toán lan truyền ngược sử dụng gradient descent; chỉ khác nhau ở chỗ mạng tích chập không liên kết đầy đủ nên độ lỗi ở mỗi lớp chỉ tính dựa vào một phần các node trong lớp tiếp theo chứ không phải toàn bộ.

## Đánh giá và kết luận

Ưu nhược điểm của các phương pháp:

Phân loại Naïve Bayes:

* Ưu điểm: mô hình đơn giản, dễ cài đặt, bộ phân lớp chạy nhanh và tốn ít bộ nhớ
* Nhược điểm: thiếu chính xác do giả định về các thuộc tính độc lập, có thể không nhất thiết là có giá trị.

Phân loại SVM:

* Ưu điểm: kết quả thực nghiệm rất tốt, cho độ chính xác cao, phụ thuộc ít vào số chiều của tập dữ liệu
* Nhược điểm: thời gian huấn luyện lâu, không gian bộ nhớ sử dụng lớn, khó giải thích mô hình kết quả

Phân loại Neural Network:

* Ưu điểm: được sử dụng như một hàm xấp xỉ tùy ý , có thể học được mỗi và mọi mối quan hệ giữa các biến đầu vào và đầu ra, cho độ chính xác cao.
* Nhược điểm: cần nhiều thời gian để thực hiện hơn so với các kỹ thuật khác vì cần đủ dữ liệu huấn luyện

Qua những ưu nhược điểm trên, chúng tôi quyết định lựa chọn phương pháp Neural Network để áp dụng cho bài toán phân loại cảm xúc.

# CHƯƠNG 2: XÂY DỰNG HỆ THỐNG PHÂN TÍCH CẢM XÚC DỰA TRÊN NEURAL NETWORK

Trong phần này, luận văn sẽ giới thiệu về mô mình xử lý dữ liệu trong bài toán phân loại cảm xúc, sau đó đề xuất một mô hình giải quyết bài toán, trình bày chi tiết hướng giải quyết của từng pha trong mô hình.

## Mô hình xử lý dữ liệu cho bài toán phân loại cảm xúc

Mặc dù có những hướng tiếp cận để giải quyết bài toán khác nhau, tuy nhiên các bài toán phân loại cảm xúc đều có một mô hình chung như sau:

Hình 1.2 Mô hình xử lý dữ liệu cho bài toán phân loại cảm xúc

* Thu thập dữ liệu

Bước này sẽ thu thập các trang web chứa bình luận người dùng. Để thực hiện bước này thường sử dụng các công cụ trích xuất dữ liệu một cách tự động . Kết quả bước này sẽ thu thập được các dữ liệu cần lấy ở các định dạng dữ liệu HTML hay TXT, dữ liệu này sẽ được đi qua xử lý ở các bước sau.

* Chuẩn hóa dữ liệu

Dữ liệu thu được ở bước thứ nhất ở định dạng HTML hoặc TXT có nhiều dữ liệu dư thừa, nhiễu do đó bước chuẩn hóa dữ liệu sẽ lọc lấy những thông tin cần thiết cho mục đích xử lý như thời gian, tiêu đề bài viết cùng với bình luận người dùng… Ngoài ra, những bình luận người dùng thu thập được có thể không thỏa mãn những yêu cầu về ngữ pháp hoặc ngữ nghĩa. Bước này cũng sẽ loại bỏ những mẫu bình luận không phù hợp hoặc sửa đổi (như thêm dấu với Tiếng Việt) để đảm bảo dữ liệu thu thập được phù hợp với việc Gán nhãn hay Trích chọn đặc trưng ở bước sau.

* Gán nhãn dữ liệu

Tại bước này dữ liệu sau khi làm sạch, chuẩn hóa sẽ được gán nhãn câu, gán nhãn từ và từ loại hay những biểu tưởng cảm xúc trong câu là tùy theo từng hướng tiếp cận và yêu cầu cụ thể của mỗi bài toán.

* Trích chọn đặc trưng

Dựa trên dữ liệu được gán nhãn, đặc trưng để phân lớp là trọng số hướng ngữ nghĩa của bình luận với cách sử dụng bộ từ vựng. Các thuật toán học máy sẽ sử dụng đặc trưng ngôn ngữ như n-grams sau khi tách từ hay gán nhãn từ loại để huấn luyện và kiểm thử với bộ dữ liệu chuẩn được chọn ra từ các bước trước đó.

* Phân lớp dữ liệu

Dữ liệu đầu vào đã qua các bước tiền xử lý sẽ qua bộ phân lớp sử dụng các thuật toán học máy. Với cách tiếp cận bằng bộ từ vựng, bước này sẽ áp dụng các luật ngữ pháp hay các quy tắc thay đổi ngữ nghĩa để tính ra trọng số cuối cùng sau đó quyết định hướng ngữ nghĩa của bình luận.

## Mô hình hệ thống

Qua quá trình khảo sát dữ liệu, nghiên cứu và tham khảo các công trình khác, luận văn sẽ áp dụng mô hình phân lớp quan điểm theo các bước sau:

**Bước 1:** Thu thập dữ liệu bình luận của sản phẩm từ trang web bán hàng (lazada) sử dụng khung làm việc Scrapy.

**Bước 2:** Tiền xử lý dữ liệu thu thập được: làm sạch và chuẩn hóa dữ liệu.

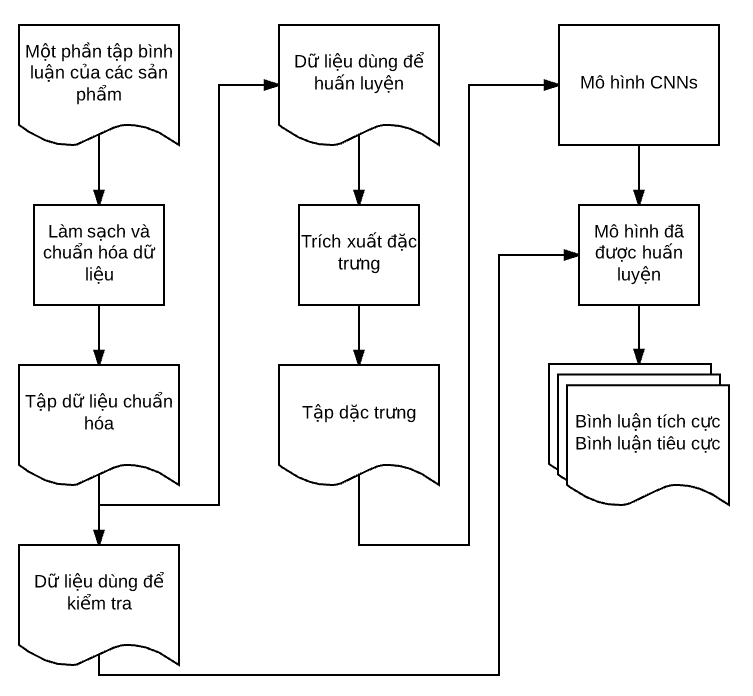
**Bước 3:** Nhận dạng thủ công từng bình luận trong bộ dữ liệu mẫu, phân vào các lớp positive (tích cực), negative (tiêu cực).

**Bước 4:** Lựa chọn, trích xuất đặc trưng và huấn luyện mô hình.

**Bước 5:** Chạy bộ phân lớp, so sánh kết quả phân lớp tự động và phân lớp thủ công.

**Bước 6:** Chạy phân loại cho toàn bộ bình luận thu thập được

Các bước trên được sơ đồ hóa với mô hình phân lớp được thể hiện trong hình:



Hình 2.1 Sơ đồ mô hình hệ thống

### Thu thập dữ liệu

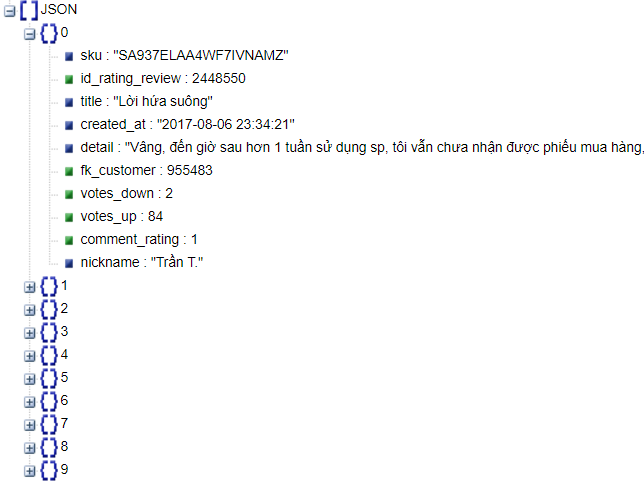
Trong bước này tôi có hai nhiệm vụ chính:

* **Nhiệm vụ 1:** Thu thập thông tin của các sản phẩm trên trang lazada, cụ thể trong nhóm sản phẩm điện tử.
  + Thu thập danh sách đường dẫn của các mục con trong nhóm sản phẩm điện tử như: điện thoại, máy tính bảng, phụ kiện điện thoại, tivi …
  + Sử dụng khung làm việc Scrapy để lấy danh sách sản phẩm trong các mục con đó

Bảng 2.1 Danh sách mục thu thập dữ liệu

|  |  |
| --- | --- |
| Danh mục sản phẩm | Các mục con |
| Điện tử | Điện thoại |
| Máy tính bảng |
| Phụ kiện điện thoại |
| Phụ kiện máy tính bảng |
| Laptop |
| Máy tính để bàn |
| Phụ kiện máy tính |
| Phụ kiện máy ảnh |
| Tivi |
| Phụ kiện tivi |
| Thiết bị âm thanh |
| Game |
| Thiết bị đeo công nghệ |
| Thiết bị đeo |

* **Nhiệm vụ 2:** Lấy bình luận của từng sản phẩm trong danh sách sản phẩm đã lấy được ở bước trên

****

Hình 2.2 Dữ liệu bình luận sản phẩm thu thập được

### Tiền xử lý dữ liệu

Vì dữ liệu bình luận là nhiễu và sai chính tả nhiều do đó việc làm sạch và chuẩn hóa là rất quan trọng. Dữ liệu chúng tôi cần là nội dung của các bình luận nên chúng tôi cần phải thực hiện cắt bỏ các dữ liệu dư thừa như: thời gian, tên người viết, tiêu đề… Thêm vào đó tập dữ liệu có hiện tượng các bình luận lặp lại làm ảnh hưởng đến độ chính xác khi đánh giá vì vậy thực hiện loại bỏ các dữ liệu dư thừa và các bình luận lặp lại.

### Trích xuất đặc trưng

Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép có trong câu. Vấn đề này tưởng đơn giản với con người nhưng đối với máy tính lại là bài toán khó giải quyết.

Chính vì lý do đó tách từ được xem là bước xử lý quan trọng đối với các hệ thống xử lý ngôn ngữ tự nhiên, đặc biệt là đối với các ngôn ngữ theo loại hình ngôn ngữ đơn lập, ví dụ: tiếng Trung Quốc, tiếng Nhật, tiếng Thái, và tiếng Việt. Với các ngôn ngữ thuộc loại hình này, ranh giới từ không chỉ đơn giản là những khoảng trắng như trong tiếng Anh…, mà có sự liên hệ chặt chẽ giữa các tiếng với nhau, một từ có thể cấu tạo bởi một hoặc nhiều tiếng. Vì vậy đối với tiếng Việt, vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ.

Bởi vì các lý do trên, trước khi đưa vào mô hình huấn luyện và trả lời câu hỏi chúng tôi đã thực hiện tách từ tiếng Việt và sử dụng công cụ vnTokenizer của tác giả Lê Hồng Phương cùng các cộng sự với độ chính xác tới 95%-97% [7].

### Phân loại dữ liệu

Sau khi thu thập và tiền xử lí chúng tôi thu được tập dữ liệu có 5797 bình luận. Tập này được tách thành tập dữ liệu huấn luyện và tập dữ liệu test với thành phần nhãn của các tập dữ liệu cụ thể như sau:

Bảng 2.2 Tập dữ liệu huấn luyện

|  |  |
| --- | --- |
| Tích cực | Tiêu cực |
| 3237 | 992 |

Bảng 2.3 Tập dữ liệu kiểm tra

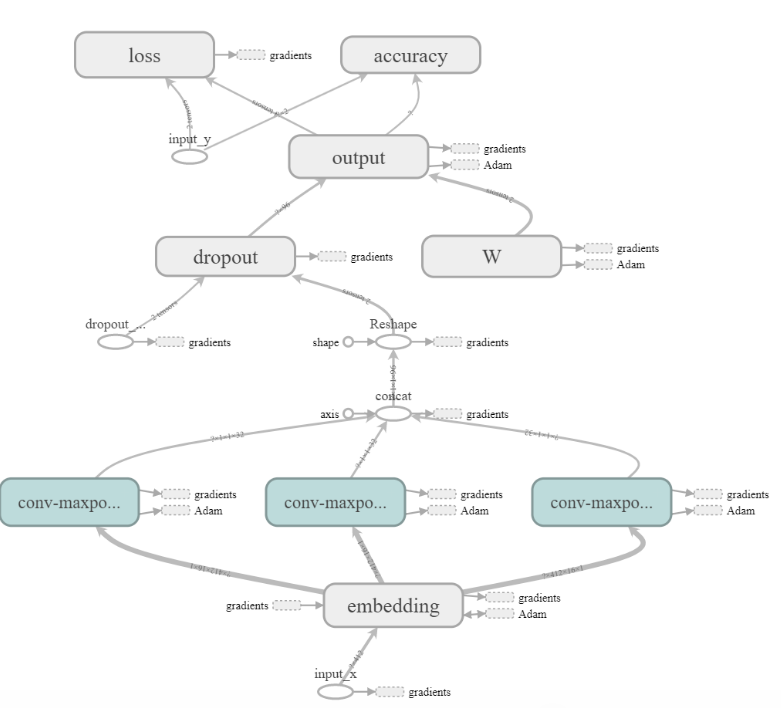
|  |  |
| --- | --- |
| Tích cực | Tiêu cực |
| 809 | 248 |

### Huấn luyện mô hình

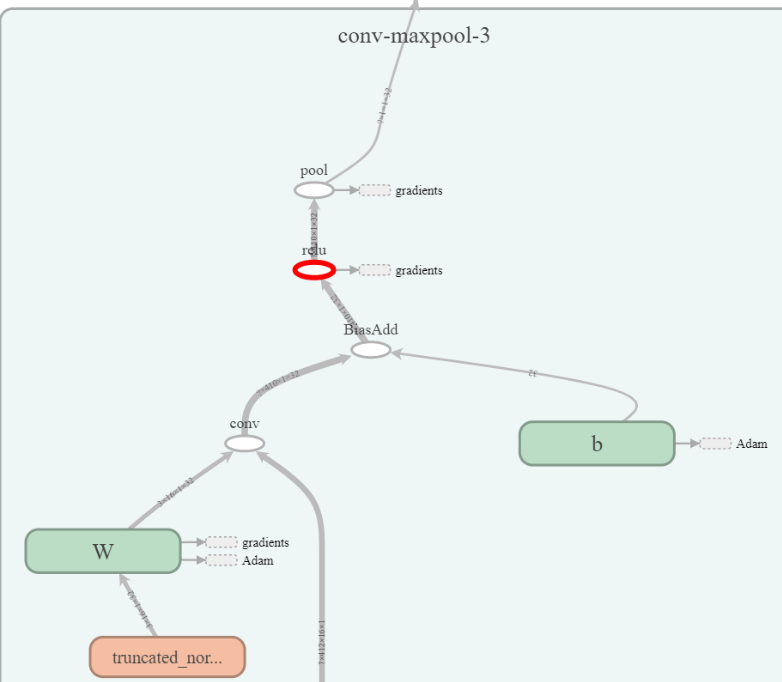
Quá trình huấn luyện sử dụng tập dữ liệu 4229 bình luận. Tập dữ liệu này lại được chia ra thành hai phần là tập dữ liệu để huấn luyện (train) và tập dữ liệu dùng để tinh chỉnh (dev) theo tỉ lệ 9:1 như sau:

Hình 2.4 Phân chia dữ liệu tập huấn luyện và tập phát triển

Mô hình sau khi cài đặt bằng tensorflow gồm các quá trình xử lý như hình sau:

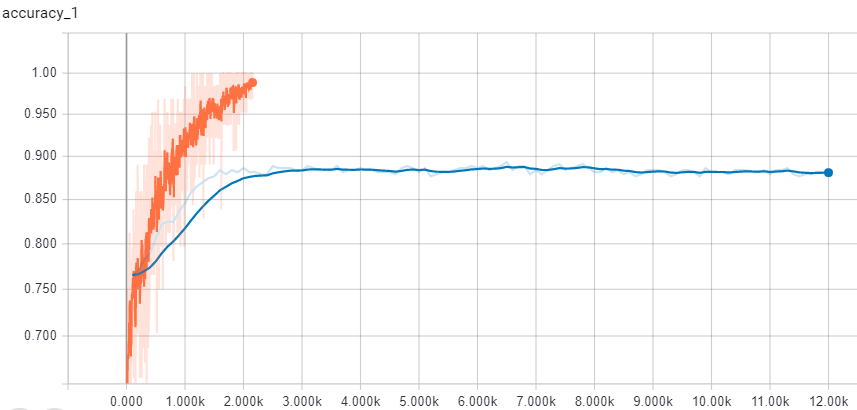


Hình 2.5 Mô hình cài đặt bằng tensorflow



Hình 2.6 Quá trình xử lý của chi tiết của lớp tích chập

Mô hình sau khi chạy huấn luyện đạt kết quả rất khả quan: 88,15%



Hình 2.7 Kết quả chạy mô hình khi huấn luyện

## Kết quả phân lớp và đánh giá

Các chỉ số đo kiểm chất lượng bộ phân lớp

Mô hình được đánh giá dựa trên bộ ba tiêu chí đánh giá: độ chính xác (Precision), độ bao phủ (recall) và độ đo F1 là giá trị trung hòa giữa hai giá trị độ chính xác và độ bao phủ.

* Độ chính xác (precision)

Độ chính xác của bộ phân lớp được định nghĩa như sau:

* Độ bao phủ (recall)

Độ bao phủ của bộ phân lớp được định nghĩa như sau:

* F­1

Độ đo F1 của bộ phân lớp được định nghĩa như sau:

Đồ án tiến hành phân lớp quan điểm đối với tập dữ liệu trên Lazada gồm 5797 bình luận với 4229 bình luận dữ liệu huấn luyện và 1057 bình luận dữ liệu test. Thực nghiệm trên bộ phân lớp CNNs với các đặc trưng cũng như biểu diễn dữ liệu đã được trình bày ở trên.

Đồ án thực hiện phân lớp dữ liệu thành hai lớp trong bảng sau:

Bảng 2.3 Kí hiệu các lớp dữ liệu

|  |  |  |
| --- | --- | --- |
| Kí hiệu | Tên lớp tiếng Anh | Tên lớp tiếng Việt |
| 1 | Positive | Tích cực |
| 0 | Negative | Tiêu cực |

Kết quả phân lớp trên bộ dữ liệu kiểm tra với mô hình như sau:

Chính xác/Accurancy: 85,24%

Bảng 2.4 Kết quả thực nghiệm phân lớp sử dụng CNNs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lớp | Số bình luận | Độ chính xác(%) | Độ bao phủ(%) | F1(%) |
| Positive | 809 | 91,49 | 89 | 90,23 |
| Negative | 248 | 67,04 | 72,98 | 69,88 |
| All | 1057 | 85,24 | 85,24 | 85,24 |

Từ bảng kết quả trên chúng ta có thể thấy, kết quả của bộ phân lớp tính theo tiêu chí độ chính xác của các nhãn positive, negative lần lượt là 91,49%, 67,04% . Các giá trị này xấp xỉ với kết quả tính theo độ bao phủ, lần lượt là 89%, 72,98%. Có thể thấy độ chính xác của các nhãn negative thấp hơn các nhãn positive là do lượng dữ liệu huấn luyện của negative ít hơn hẳn nên độ chính xác khi kiểm tra cũng thấp hơn. Điều này cho thấy, bộ phân lớp tương đối ổn định khi đánh giá theo hai tiêu chí trên, kết quả là giá trị F1 theo từng nhãn cũng xấp xỉ nhau.

# CHƯƠNG 3: CÀI ĐẶT HỆ THỐNG

Trong chương này, chúng tôi sẽ giới thiệu các công nghệ được sử dụng trong luận văn sau đó trình bày việc sử dụng các công nghệ để cài đặt các chức năng của hệ thống.

## Mô hình cài đặt

Chúng tôi chia hệ thống thành 4 phần, tướng ứng với mỗi phần sẽ là những công nghệ được cài đặt như sau:

Hình 3.1 Sơ đồ mô hình cài đặt hệ thống

Python

Python là một ngôn ngữ lập trình thông dịch do Guido van Rossum tạo ra năm 1990. Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý. Python có những ưu nhược điểm sau:

* Ưu điểm:
  + Cú pháp đơn giản, có cấu trúc rõ ràng giúp người lập trình dễ dàng đọc và tìm hiểu
  + Có thể biên dịch và chạy trên hầu hết các nền tảng
  + Có các phần mềm, thư viện phong phú, tài liệu rõ ràng trong nhiều lĩnh vực, đặc biệt là học máy
  + Cộng đồng phát triển lớn
* Nhược điểm:
  + Không có các thuộc tính như: protected,private hay public, không có vòng lặp do…while và switch….case.
  + Python có tốc độ thực hiện chậm hơn nhiều lần so với các ngôn ngữ biên dịch như Fortran, C

Bởi những ưu điểm đã nêu ở trên và đặc biệt là việc python có rất nhiều thư viện, phần mềm hỗ trợ rất tốt cho học máy như trích xuất dữ liệu, cơ sở dữ liệu, hỗ trợ tính toán khoa học, học máy nên tôi quyết định sử dụng python.

### Scrapy

Scrapy là một framework được viết bằng Python, cung cấp sẵn 1 cấu trúc tương đối hoàn chỉnh để thực hiện việc thu thập (crawl) và trích xuất dữ liệu (data extract) từ website một cách nhanh chóng và dễ dàng. Nếu muốn lấy dữ liệu từ các website nhưng dữ liệu đó quá lớn để sao chép rồi dán vào cơ sở dữ liệu, scrapy hỗ trợ ta làm điều đó. Việc lấy dữ liệu website hoàn toàn tự động nhanh chóng và việc sử dụng scrapy cũng rất đơn giản giúp tiếp kiệm được nhiều thời gian và công sức.

### SQLite

SQLite là phần mềm quản lý cơ sở dữ liệu tương tự như Mysql, PostgreSQL... Đặc điểm của SQLite là gọn, nhẹ, đơn giản. Chương trình gồm 1 file duy nhất vỏn vẹn chưa đến 400kB, không cần cài đặt, không cần cấu hình hay khởi động mà có thể sử dụng ngay. Dữ liệu Database cũng được lưu ở một file duy nhất. Không có khái niệm user, password hay quyền hạn trong SQLite Database.

SQLite không thích hợp với những hệ thống lớn nhưng ở quy mô vừa tầm thì SQLite phát huy uy lực và không hề yếu kém về mặt chức năng hay tốc độ. Với các đặc điểm trên SQLite được sử dụng nhiều trong việc phát triển, thử nghiệm …

### Khung làm việc Tensorflow

Tensorflow là một thư viện phần mềm nguồn mở cho tính toán số sử dụng biểu đồ luồng dữ liệu. TensorFlow ban đầu được phát triển bởi các nhà nghiên cứu và kỹ sư làm việc trong nhóm Brain Google trong tổ chức nghiên cứu của Google nhằm mục đích tiến hành học máy và sâu nghiên cứu các mạng nơ-ron thần kinh, nhưng hệ thống là đủ nói chung có thể áp dụng trong một loạt các lĩnh vực khác như tốt.

Tensorflow là một hệ thống học máy hoạt động ở quy mô lớn và trong môi trường phức tạp. TensorFlow sử dụng đồ thị luồng dữ liệu Dataflow để đại diện cho sự tính toán, chia sẻ trạng thái, và các hoạt động biến đổi trạng thái đó. Nó ánh xạ các nút của một đồ thị dataflow trên nhiều máy trong một cluster, và bên trong một máy trên nhiều thiết bị tính toán, bao gồm CPU, GPU đa lõi, các chíp ASIC tùy biến được gọi là tenxơ Processing Units (TPUs). Kiến trúc này rất linh hoạt cho phép cho các nhà phát triển ứng dụng: trong khi trước đây "tham số máy chủ" thiết kế quản lý chia sẻ trạng thái (shared state) được xây dựng sẵn trên hệ thống, TensorFlow cho phép các nhà phát triển để thử nghiệm các tối ưu hoá mới và các thuật toán huấn luyện.

TensorFlow hỗ trợ một loạt các ứng dụng, với sự hỗ trợ đặc biệt mạnh mẽ cho việc huấn luyện và suy luận trên các mạng học sâu Deep Learning. Google đã phát hành TensorFlow như là một dự án mã nguồn mở, và nó đã trở thành sử dụng rộng rãi cho các nghiên cứu học máy.

Trong bài luận văn này, chúng tôi sử dụng TensorFlow để huấn luyện và tạo ra các mô hình đối thoại cho tiếng Việt, một kết quả rất khả quan khi sử dụngTensorFlow là chúng tôi đạt được những mô hình có chất lượng tốt.

### Django

Django là một web framework miễn phí mã nguồn mở được viết bằng Python. Django sử dụng mô hình Model-View-Control (MVC). Django được phát triển bởi Django Software Foundation(DSF) – một tổ chức phi lợi nhuận độc lập.

Django đơn giản hóa việc tạo các website phức tạp có sử dụng cơ sở dữ liệu. Django tập trung vào tính năng “có thể tái sử dụng” và “có thể tự chạy” của các component, tính năng phát triển nhanh, không làm lại những gì đã làm.

## Cài đặt hệ thống

### Môi trường cài đặt

Bảng 3.1 Cấu hình phần cứng hệ thống:

|  |  |
| --- | --- |
| **Thành phần** | **Chỉ số** |
| Bộ vi xử lý | Intel Core i5 (2.6GHz) |
| RAM | 4 GB |
| Hệ điều hành | Windows 10 Education 64bit |
| Bộ nhớ | 232 GB |

Bảng 3.2 Danh sách các công cụ phần mềm:

|  |  |  |
| --- | --- | --- |
| STT | Tên công cụ | Phiên bản |
| 1 | Python | 2.7.14 và 3.5.3 |
| 2 | Scrapy | 1.4.0 |
| 3 | Sqlite | 3 |
| 4 | Tensorflow | 1.3.0 |
| 5 | Django | 1.7 |

### Cài đặt cụ thể

Chúng tôi sử dụng 2 phiên bản python khác nhau là python 2.7.14 cho Scrapy, Django và python 3.5.3 cho Tensorflow. Vì vậy chúng tôi dùng phần mềm Virtualenv để tạo môi trường ảo để tách biệt các môi trường project khác nhau giúp việc quản lý, cài đặt dễ dàng hơn. Chi tiết cài đặt như sau:

Tải 2 phiên bản python về tại đường dẫn :

<https://www.python.org/downloads/>

Sau khi cài đặt xong python tại các vị trí thu mục C:\python27, C:\python35, tiếp tục cài Virtualenv bằng câu lệnh:

pip install virtualenv

Tạo môi trường ảo, cài Scrapy và Django:

virtualenv --python=C:\Python27\python.exe env27

env27\Scripts\activate

pip install scrapy

pip install Django

deactivate

Tạo môi trường ảo và cài Tensorflow:

virtualenv --python=C:\Python27\python.exe env35

env35\Scripts\activate

pip install tensorflow

* Cài đặt Scrapy để trích xuất dữ liệu:

## Kết quả và đánh giá

# CHƯƠNG 4: ĐÁNH GIÁ

Trong chương này chúng tôi sẽ trình bày về quá trình chạy của hệ thống, đưa ra các tính năng và đánh giá hiệu quả của các tính năng này.

## Hoạt động của hệ thống

Sử dụng Django chạy server trên localhost bằng câu lệnh sau:

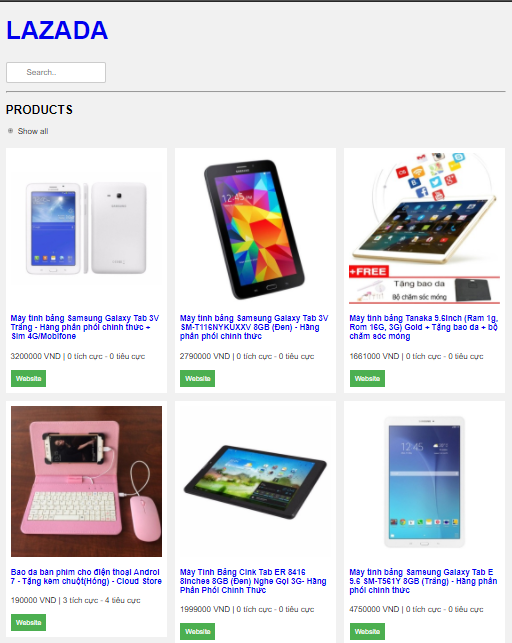
python manage.py runserver

Server sẽ chạy tại địa chỉ: <http://127.0.0.1:8000/>

Truy cập theo đường dẫn sau để truy cập trang web:

<http://127.0.0.1:8000/lazada>

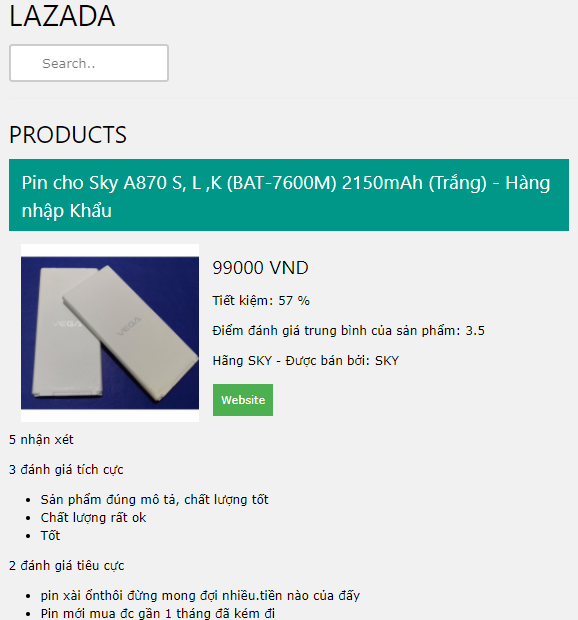
Giao diện trang chủ:



Hình 4.1 Giao diện hiển thị của trang web đánh giá

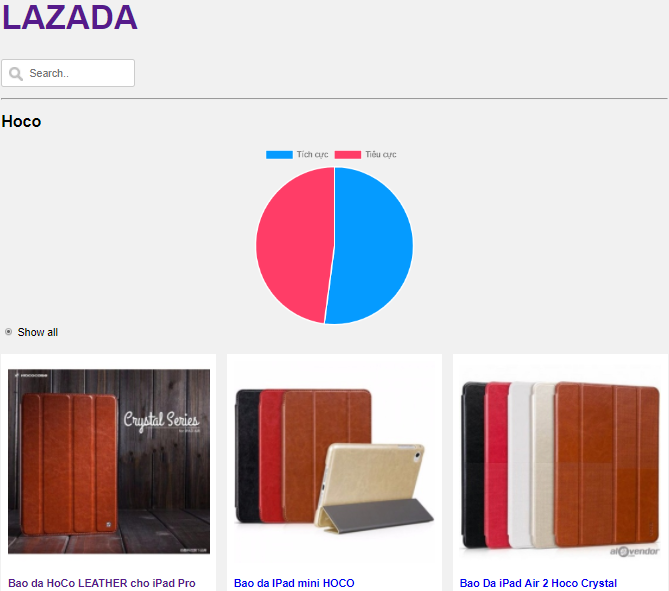
Trang web chúng tôi xây dựng có các tính năng sau:

* Tại trang chủ: hiển thị danh sách sản phẩm với các thông tin: tên sản phẩm, giá sản phẩm khi đã giảm giá, thông tin đánh giá sản phẩm tích cực hay tiêu cực
* Truy cập sản phẩm trên trang web chính thức
* Bấm vào để xem chi tiết sản phẩm với các thông tin như: hãng sản xuất, cửa hàng bán, giá, giảm giá bao nhiêu, có những bình luận tích cực, tiêu cực nào.



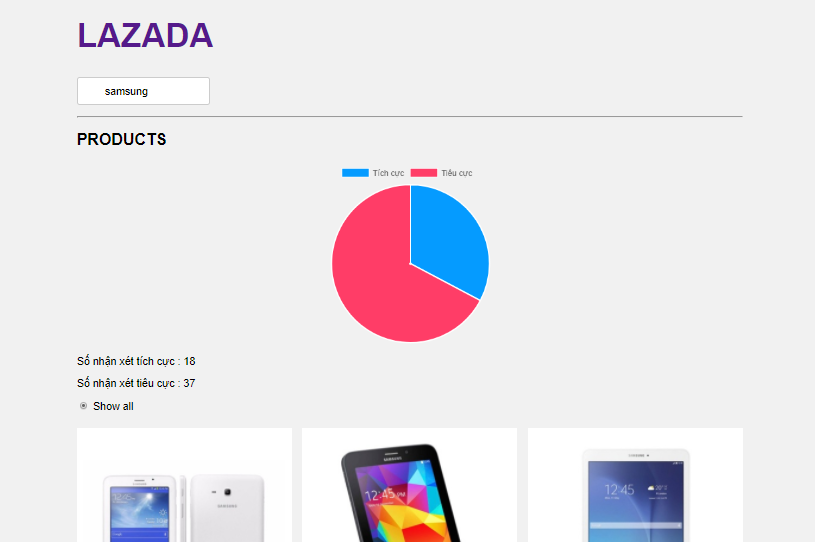
Hình 4.2 Thông tin cụ thể của sản phẩm

* Xem đánh giá tỉ lệ bình luận tích cực với tiêu cực của các nhà cung cấp và các sản phẩm của nhà cung cấp đó:



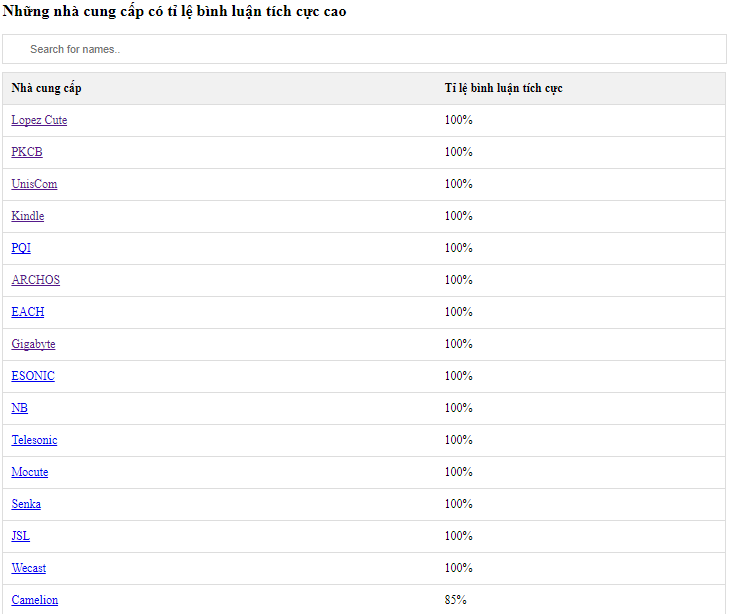
Hình 4.3 Thông tin tỉ lệ bình luận và sản phẩm của nhà cung cấp

* Tìm kiếm sản phẩm, cửa hàng, hãng sản xuất để xem các thông tin sản phẩm và đánh giá trực quan.



Hình 4.4 Kết quả tìm kiếm sản phẩm với đánh giá theo tỉ lệ phần tram

* Xem danh sách các nhà cung cấp và sản phẩm có tỉ lệ bình luận tích cực cao:



Hình 4.5 Các nhà cung cấp có tỉ lệ bình luận tích cực cao

## Đánh giá hiệu quả

## Kết luận

Với những

# KẾT LUẬN

# TÀI LIỆU THAM KHẢO

1. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification

2.Bing Liu, Lei Zhang, “A Survey of Opinion Mining and Sentiment Analysis” (<https://link.springer.com/chapter/10.1007%2F978-1-4614-3223-4_13>)

3. Cambria, E; Schuller, B; Xia, Y; Havasi, C (2013). "New avenues in opinion mining and sentiment analysis". IEEE Intelligent Systems. 28 (2): pp. 15–21

(<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.688.1384&rep=rep1&type=pdf>)

4. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation

(http://liris.cnrs.fr/Documents/Liris-6508.pdf)

5. Singh, Pravesh Kumar, and Mohd Shahid Husain. "Methodological Study Of Opinion Mining And Sentiment Analysis Techniques."International Journal on Soft Computing 5.1, 2014.]

(<https://www.researchgate.net/profile/Mohd_Shahid_Husain/publication/269598010_Methodological_Study_Of_Opinion_Mining_And_Sentiment_Analysis_Techniques/links/548fe59a0cf214269f264148.pdf>)

6. Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification.

7. Le-Hong, P., T M H. Nguyen, A. Roussanaly, and T V. Ho (2008), “A hybrid approach to word segmentation of Vietnamese texts”, pp. 240-249

13. Tang, D., Qin, B. and Liu, T. (2015), “Deep Learning for Sentiment

Analysis: Successful Approaches and Future Challenges.”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(6), pp. 292-303.

14. Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification.