BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH

KHOA ĐÀO TẠO CHẤT LƯỢNG CAO

BỘ MÔN CƠ ĐIỆN TỬ

___oOo___

BÁO CÁO

MÔN HỌC TRÍ TUỆ NHÂN TẠO

GVHD**: PGS.TS Nguyễn Trường Thịnh**

SVTH**: Trần Triệu Vĩ** MSSV**: 19146301**

Thành phố Hồ Chí Minh, ngày 20 tháng 6 năm 2022

NHẬN XÉT CỦA GIẢNG VIÊN	

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Chữ ký giảng viên

MỤC LỤC

Chương 1: TỔNG QUAN ĐỀ TÀI

1.1. Lí do chọn đề tài

Cùng với sự phát triển vượt bậc của Internet và công nghệ thông tin, các sàn thương mại điện tử, những website bán hàng ra đời tạo cho khách hàng một phương tiện hữu ích để tham khảo sản phẩm, mua sắm trực tuyến cũng như chia sẻ những đánh giá, trải nghiệm về sản phẩm, dịch vụ, hậu mãi của sản phẩm trong quá trình mua và sử dụng. Thông qua website bán hàng, fanpage, các trang mạng xã hội của cửa hàng hoặc doanh nghiệp, khách hàng có thể để lại các đánh giá tiêu cực hoặc tích cực về mọi mặt chứ không chỉ riêng về sản phẩm. Có thể là thái độ phục vụ của nhân viên, quy trình bảo hành, bảo dưỡng, chính sách đổi trả và những điểm mạnh điểm yếu hoặc đưa ra những ý kiến để cải thiện dịch vụ, sản phẩm.

Vậy làm sao chúng ta có thể phân tích được ý kiến, cảm xúc của khách hàng thông qua website để thấu hiểu họ, giải quyết vấn đề cho họ cũng nhưng tiếp thu ý kiến để cải thiện dịch vụ, sản phẩm mà doanh nghiệp, cá nhân đang cung cấp? Nếu để một cá nhân, tập thể nào đó phân tích hàng ngàn, hàng triệu bình luận của khách hàng thì rất lãng phí nhân lực, mất rất nhiều thời gian để đưa ra kết quả. Vì vậy không thể đáp ứng kịp thời của mong muốn, những ý kiến đóng góp của khách hàng. Để giải quyết vấn đềnày, nghiên cứu đề xuất khai thác các đánh giá của khách hàng thông qua website, fanpage, mạng xã hội, sử dụng các thuật toán học máy để dự đoán cảm xúc đánh giá của khách. Từ đó, giúp cá nhân, cửa hàng, doanh nghiệp đưa ra được kế hoạch cải thiện chất lượng sản phẩm, dịch vụ đang cung cấp.

Hình Minh họa việc phân tích cảm xúc đánh giá khách hàng

1.2. Giới hạn đề tài

Đề tài được giới hạn như sau:

- Mô hình dự đoán cảm xúc của khách hàng về các sản phẩm điện thoại thông minh.
- Dự đoán trả về 2 trạng thái cảm xúc của bình luận khách hàng: tiêu cực và tích cực.

1.3. Mục tiêu và nhiệm vụ nghiên cứu

Mục tiêu:

- Xây dựng mô hình dự đoán cảm xúc đánh giá của khách hàng là tiêu cực hay tích cực về sản phẩm điện thoại thông minh với độ chính xác trên 80%
- Thiết kế giao diện trực quan để người dùng có thể dễ dàng dự đoán cảm xúc đánh giá khách hàng của họ

Nhiệm vụ:

- Thu thập dữ liệu những đáng giá của khách hàng để xây dựng mô hình
- Nghiên cứu về các thuật toán xử lý ngôn ngữ tự nghiên (NLP)
- Thử nghiệm các thuật toán Deep Learning để xây dựng mô hình có độ chính xác cao
- Xây dựng, thử nghiệm và đánh giá mô hình
- Thiết kế giao diện website trực quan để dễ dàng sử dụng mô hình

1.4. Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học:

- Đề tài góp phần bổ sung lý luận về xử lý ngôn ngữ tự nhiên, giải quyết vấn đề trong việc dự đoán cảm xúc thông qua các đánh giá
- Đưa ra mô hình dự đoán đã xây dựng góp phần nghiên cứu nghiên cứu mô hình dự đoán cảm xúc thông qua đánh giá cho người nghiên cứu

Ý nghĩa thực tiễn:

- Xây dựng mô hình dự đoán cảm xúc của khách hàng thông qua các bình luận một cách nhanh chóng giúp cá nhân, doanh nghiệp kịp thời tiếp thu ý kiến để cải thiện chất lượng sản phẩm dịch vụ của mình đang cung cấp

- Có thể đánh giá nhanh về sản phẩm Demo sau khi tung ra thị trường bằng việc dự đoán cảm xúc thông qua bình luận của khách hàng từ đó kịp thời cải tiến để đưa ra sản phẩm chính thức
- Thấu hiểu khách hàng thông qua những đánh giá từ đó đưa ra những kế hoạc kinh doanh phù hợp

Chương 2: CƠ SỞ LÍ THUYẾT

2.1. Xử lý ngôn ngữ tự nhiên

2.1.1. Tổng quan

Hình Minh họa Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người, dưới dạng tiếng nói (speech) hoặc văn bản (text). Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói. Có thể chia Xử lý ngôn ngữ tự nhiên thành hai nhánh lớn, không hẳn là độc lập hoàn toàn, bao gồm xử lý văn bản (text processing) và xử lý tiếng nói (speech processing).

Xử lý ngôn ngữ tự nhiên bao gồm hiểu ngôn ngữ tự nhiên (Natural Language Understanding – NLU) và sinh ngôn ngữ tự nhiên (Natural Language Generation – NLG). Xử lý văn bản trong xử lý ngôn ngữ tự nhien bao gồm các bước chính:

- Phân tích hình vị: phân tích, nhận biết, miêu tả cấu trúc của hình vị trong một ngôn ngữ cho trước và các đơn vị ngôn ngữ khác. Nhận diện được từ chính, từ gốc, từ biên, phụ tố, từ loại,.. Trong xử lý ngôn ngữ tự nhiên tiếng Việt, có hai bài toán điển hình là tách từ (Word Segmentation) và gán nhãn từ loại (Part-of-speech Tagging).
- Phân tích cú pháp: quy trình phân tích một chuỗi các biểu tượng, ở dạng ngôn ngữ tự nhiên hoặc ngôn ngữ máy tính, tuân theo văn phạm hình thức. Văn phạm hình thức thường dùng trong phân tích cú pháp của ngôn ngữ tự nhiên bao gồm: văn phạm phi ngữ cảnh (Context Free Grammar – CFG), văn phạm danh mục kết nối (Combinatory Categorial Grammar – CCG) và Văn phạm phụ thuộc (Dependency Grammar – DG).
- Phân tích ngữ nghĩa: quá trình liên hệ cấu trúc ngữ nghĩa, từ cấp độ cụm từ, mệnh đề, câu và đoạn đến cấp độ toàn bài viết, với ý nghĩa độc lập của chúng. Nói cách khác, việc này nhằm tìm ra ngữ nghĩa của đầu vào ngôn từ. Phân tích ngữ nghĩa bao gồm hai mức độ: ngữ nghĩa từ vựng biểu hiện các ý nghĩa của những từ thành phần, và phân biệt nghĩa của từ. Ngữ nghĩa thành phần liên quan đến cách thức các từ liên kết để hình thành những nghĩa rộng hơn.
- Phân tích diễn ngôn: phân tích văn bản có xét tới mối quan hệ giữa ngôn ngữ và ngữ cảnh sử dụng (Context Of Use). Phân tích diễn ngôn, do đó, được thực hiện ở mức độ đoạn văn hoặc toàn bộ văn bản thay vì chỉ phân tích riêng ở mức câu.

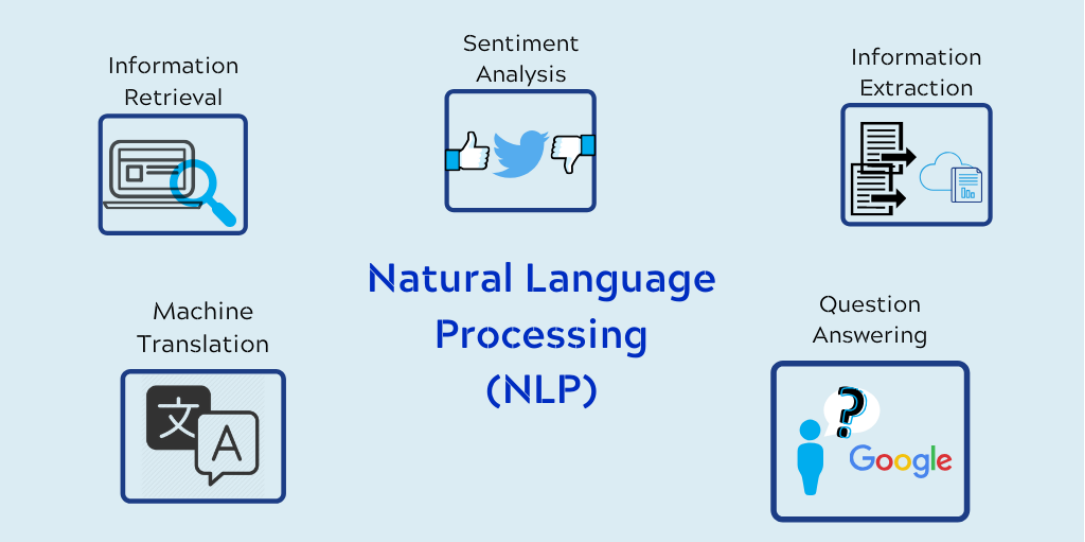
2.1.2. Ứng dụng

Xử lý ngôn ngữ tự nhiên ngày càng được ứng dụng nhiều trong đời sống hiện nay. Một số ứng dụng điển hình có thể kể đến:

- Nhận dạng tiếng nói (Automatic Speech Recognition – ASR hoặc Speech To Text – STT), chuyển đổi ngôn ngữ từ dạng tiếng nói sang dạng văn bản, thường được ứng dụng trong các chương trình điều khiển qua giọng nói.
- Tổng hợp tiếng nói (Speech Synthesis hoặc Text to Speech – TTS) chuyển đổi ngôn ngữ từ dạng văn bản sang tiếng nói, thường được dùng trong đọc văn bản tự động.

- Truy xuất thông tin (Information Retrieval – IR) có nhiệm vụ tìm các tài liệu dưới dạng không có cấu trúc (thường là văn bản) đáp ứng nhu cầu về thông tin từ những nguồn tổng hợp lớn. Những hệ thống truy xuất thông tin phổ biến nhất bao gồm các công cụ tìm kiếm như Google, Yahoo, hoặc Bing search. Những công cụ này cho phép tiếp nhận một câu truy vấn dưới dạng ngôn ngữ tự nhiên làm đầu vào và cho ra một danh sách các tài liệu được sắp xếp theo mức độ phù hợp.
- Trả lời câu hỏi (Question Answering – QA) có khả năng tự động trả lời câu hỏi của con người ở dạng ngôn ngữ tự nhiên bằng cách truy xuất thông tin từ một tập hợp tài liệu. Một hệ thống QA đặc trưng thường bao gồm ba mô đun:
  - Xử lý truy vấn (Query Processing Module) – tiến hành phân loại câu hỏi và mở rộng truy vấn.
  - Xử lý tài liệu (Document Processing Module) – tiến hành truy xuất thông tin để tìm ra tài liệu thích hợp.
  - Xử lý câu trả lời (Answer Processing Module) – trích chọn câu trả lời từ tài liệu đã được truy xuất.
- Tóm tắt văn bản tự động (Automatic Text Summarization) là bài toán thu gọn văn bản đầu vào để cho ra một bản tóm tắt ngắn gọn với những nội dung quan trọng nhất của văn bản gốc. Có hai phương pháp chính trong tóm tắt: phương pháp trích xuất (Extractive) và phương pháp tóm lược ý (Abstractive). Những bản tóm tắt trích xuất được hình thành bằng cách ghép một số câu được lấy nguyên từ văn bản cần thu gọn. Những bản tóm lược ý thường truyền đạt những thông tin chính của đầu vào và có thể sử dụng lại những cụm từ hay mệnh đề trong đó, nhưng nhìn chung được thể hiện ở ngôn ngữ của người tóm tắt.
- Chatbot là việc chương trình máy tính có khả năng trò chuyện (Chat), hỏi đáp với con người qua hình thức hội thoại dưới dạng văn bản (Text). Chatbot thường được sử dụng trong ứng dụng hỗ trợ khách hàng, giúp người dùng tìm kiếm thông tin sản phẩm, hoặc giải đáp thắc mắc, tư vấn sản phẩm, dịch vụ cho khách hàng.

- Kiểm lỗi chính tả tự động là việc sử dụng máy tính để tự động phát hiện các lỗi chính tả trong văn bản (lỗi từ vựng, lỗi ngữ pháp, lỗi ngữ nghĩa) và đưa ra gợi ý cách chỉnh sửa lỗi.



Hình Một số ứng dụng của NLP

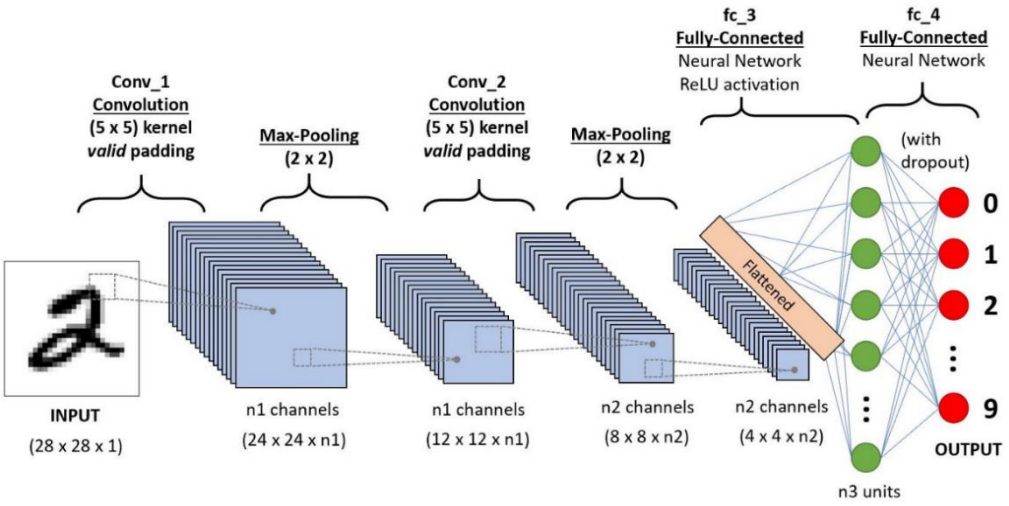
2.2. Dự đoán cảm xúc tiếp cận theo phương pháp Học máy

Phân tích cảm xúc đã được định nghĩa là tính toán nghiên cứu ý kiến, tình cảm và cảm xúc thể hiện trong văn bản. Nói cách khác, khai thác ý kiến là một phương pháp trích xuất ý kiến của người đã tạo ra một tài liệu cụ thể gần đây đã trở thành mối quan tâm nghiên cứu lớn nhất trong mạng xã hội. Tầm quan trọng ngày càng tăng của phân tích tình cảm tăng dần cùng với sự phát triển của phương tiện truyền thông xã hội như đánh giá, thảo luận diễn đàn, và mạng xã hội. Đặc biệt, trong thời đại phát triển kỹ thuật số, chúng ta hiện có một khối lượng dữ liệu lớn được ghi lại dưới dạng văn bản để phân tích.

Học máy là một nhánh của Trí tuệ nhân tạo (Artificial Intelligence – AI). Một lĩnh vực giúp máy tự đồng hiểu được các dữ liệu đưa vào đào tạo mà không cần lập trình cụ thể. Học máy tập trung vào việc phát triển những chương trình máy tính có thể truy cập vào dữ liệu và sử dụng nó để tự học. Sau đó giải quyết những vấn đề như phân loại, phân nhóm, đưa ra dự đoán về một vấn đề nào đó có liên quan đến bộ dữ liệu đầu vào. Học máy có thể được chia thành 4 phần:

- **Học máy có giám sát:** là một nhóm thuật toán dự đoán dữ liệu đầu ra dựa vào các tập dữ liệu đầu vào được gắn nhãn (kết quả đầu ra) trước.Trong quá trình học, thuật toán sẽ dựa vào cấu trúc dữ liệu để thực hiện trích xuất và tính toán. Hai bài toán cơ bản trong học có giám sát là phân loại và hồi quy. Đối với bài toán phân loại thì dữ đoán phân chia loại dữ liệu có cùng thuộc tính, còn đối với hồi quy thì cho ra kết quả là một số thực cụ thể.
- **Học máy không giám sát:** là một nhóm thuật toán sử dụng dữ liệu không có nhãn. Các thuật toán theo cách tiếp cận này hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu. Hay nói cách khác, sử dụng các phương pháp này thiên về việc mô tả tính chất hay đặc tính của dữ liệu. Thuật toán loại này được gọi là học máy không có giám sát vì không giống như Học có giám sát, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Các bài toán cơ bản của học không giám sát: phân nhóm dự trên sữ liên quan của các dữ liệu trong nhóm, tích hợp khai phá một số quy luậ dựa trên nhiều dữ liệu cho trước.
- **Học bán giám sát** là thuật toán kết hợp cả hai thuật toán có giám sát và không giám sát. Áp dụng với một phần tập dữ liệu đã được dán nhãn, phần còn lại thì không được dán nhãn.
- **Học củng cố** là thuật toán giúp hệ thống tự động xác định các hành vi để đạt hiệu quả tối ưu nhất. Trong nghiên cứu này, chúng tôi chọn phương pháp học có giám sát để áp dụng cho bài toán phân loại cảm xúc khách hàng dựa trên bình luận.

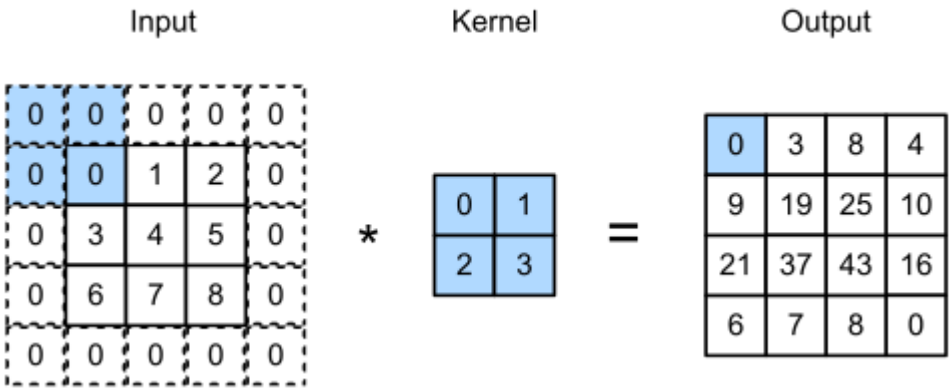
2.3. Mạng thần kinh tích chập



Hình Ví dụ về một mạng thần kinh tích chập CNN

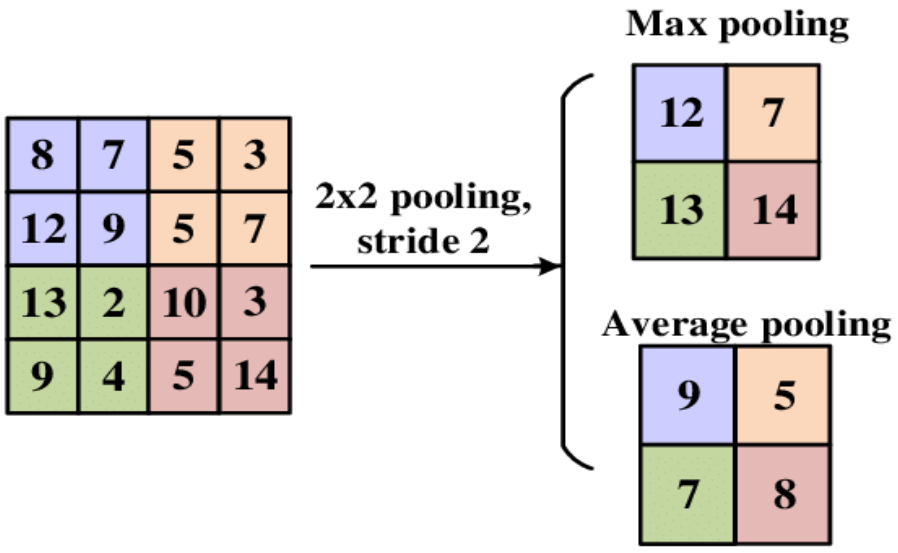
Convolutional Neural Network (CNN) được tạm dịch là: Mạng thần kinh tích chập. Đây được xem là một trong những mô hình của Deep Learning – tập hợp các thuật toán để có mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng nhiều lớp xử lý cấu trúc phức tạp. Hiểu đơn giản, CNN là một lớp của mạng nơ-ron sâu , được áp dụng phổ biến nhất để phân tích hình ảnh trực quan. Hiện tại, chúng ta chưa có định nghĩa một cách chính xác nhất về thuật toán CNN. Mạng CNN được thiết kế với mục đích xử lý dữ liệu thông qua nhiều lớp mảng. Ngoài ra, CNN có thể giúp bạn tạo ra được hệ thống thông minh, phản ứng với độ chính xác khá cao. Convolutional Neural Network có các lớp cơ bản:

- **Convolutional Layer**: Là một hidden layer, gồm một tập các feature maps là các bản scan từ input đầu vào ban đầu. Convolutional Filter hay còn gọi là Kernel, là một ma trận sẽ quét ma trận dữ liệu đầu vào từ trái sang phải, từ trên xuống dưới.



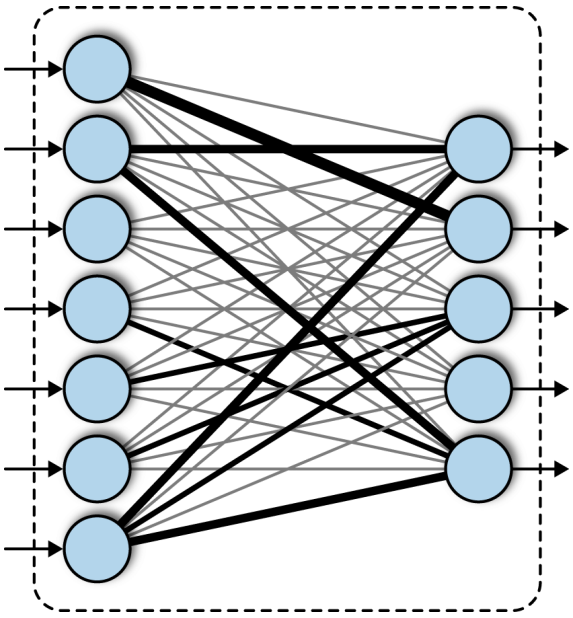
Hình Ví dụ về Convolutional layer

- **Pooling Layer:** Khi đầu vào quá lớn, các lớp Pooling Layer sẽ được dịch chuyển vào giữa những lớp Convolutional Layer nhằm giảm các Parameter. Từ đó giúp mô hình chạy nhanh hơn, tối ưu hơn về mặt thời gian, tránh overfitting. Pooling layer được biết đến với hai loại phổ biến: Max Pooling và Average Pooling. Các Pooling window sẽ trượt trên ma trận dữ liệu đầu vào, sau đó chọn 1 giá trị lớn nhất đối với Max Pooling hay lấy giá trị trung bình làm đại diện đối với Average Pooling.



Hình Minh họa các hoạt động của Max Poolling

- **Fully Connected Layer:** là layer mà mối nút của nó kết nối với tất cả các nút ở layer kế trước nó



Hình Ví dụ về Fully Connected layer

Mạng Convolutional Neural Network là tập hợp nhiều lớp Convolutional chồng lên nhau, sử dụng các hàm Nonlinear Activation và tanh để kích hoạt các trọng số trong các node. Ở mỗi lớp CNN, sau khi được các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho những lớp tiếp theo. Mỗi Layer kết tiếp sẽ là kết quả Convolution từ Layer trước đó nên chúng ta có được các kết nối cục bộ. Thông qua quá trình huấn luyện mạng, các lớp Layer CNN tự động học các giá trị được thể hiện qua các lớp Filter.

2.4. Một số thư viện và thuật toán được sử dụng trong mô hình

2.4.1. Thư viện Numpy

Hình Logo thư viện Numpy Python

NumPy hay Numeric Python là thư viện lõi phục vụ cho khoa học máy tính của Python. Nó cung cấp một đối tượng mảng đa chiều hiệu suất cao và các công cụ để làm việc với các mảng này NumPy chứa các tính năng khác nhau bao gồm những tính năng quan trọng sau:

- Đối tượng mảng đa chiều chiều mạnh mẽ
- Các chức năng broadcasting
- Phép biến đổi Fourier, khả năng số ngẫu nhiên
- Các công cụ để tích hợp mã C / C ++ và Fortran.

Bên cạnh những công dụng khoa học rõ ràng, NumPy cũng có thể được sử dụng như một nơi chứa dữ liệu chung đa chiều hiệu quả. Các kiểu dữ liệu tùy ý có thể được xác định bằng cách sử dụng NumPy, cho phép NumPy tích hợp liền mạch và nhanh chóng với nhiều loại cơ sở dữ liệu.

NumPy là một thư viện Python được viết một phần bằng Python và hầu hết các phần được viết bằng C hoặc C ++. Và nó cũng hỗ trợ các phần mở rộng bằng các ngôn ngữ khác, thường là C ++ và Fortran.

2.4.2. Thư viện pandas

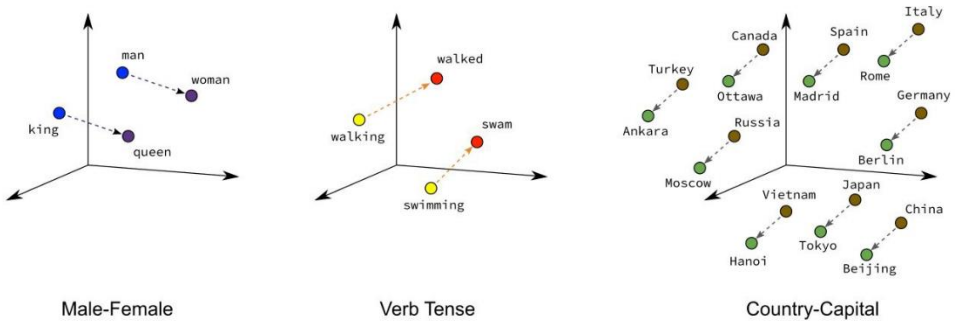
Hình Logo thư viện Pandas Python

Pandas là một thư viện mã nguồn mở được xây dựng dựa trên NumPy, sử dụng thao tác và phân tích dữ liệu, được thiết kế để cho phép bạn làm việc với dữ liệu được gắn nhãn hoặc quan hệ theo cách trực quan hơn:

- Có thể xử lý tập dữ liệu khác nhau về định dạng: chuỗi thời gian, bảng không đồng nhất, ma trận dữ liệu
- Khả năng import dữ liệu từ nhiều nguồn khác nhau như CSV, DB/SQL
- Có thể xử lý vô số phép toán cho tập dữ liệu: subsetting, slicing, filtering, merging, groupBy, re-ordering, and re-shaping,..
- Xử lý dữ liệu mất mát theo ý người dùng mong muốn: bỏ qua hoặc chuyển sang 0
- Xử lý, phân tích dữ liệu tốt như mô hình hoá và thống kê
- Tích hợp tốt với các thư viện khác của python
- Cung cấp hiệu suất tốt

2.4.3. Thư viện gensim

Gensim là một thư viện xử lý ngôn ngữ tự nhiên mã nguồn mở phổ biến. Nó sử dụng các mô hình học thuật hàng đầu để thực hiện các nhiệm vụ phức tạp như xây dựng các vector từ và thực hiện xác định chủ đề và so sánh tài liệu.



Hình Ví dụ về Vector từ được xử lý bằng Gensim

Trên đây là một số ví dụ về véc-tơ từ ở dạng trực quan. Véc-tơ từ được đào tạo từ một kho dữ liệu lớn và là một biểu diễn đa chiều của một từ hoặc dữ liệu. Bạn có thể coi nó như một mảng đa chiều với các đặc trưng có mật độ thưa (nhiều số 0 và một số số 1). Với những vector này, chúng ta có thể thấy mối quan hệ giữa các từ hoặc dữ liệu dựa trên mức độ gần hay xa của chúng và cũng như những so sánh tương tự mà chúng ta tìm thấy. Ví dụ, trong biểu đồ bên trên này, chúng ta có thể thấy rằng phép toán vector king trừ queen xấp xỉ bằng man trừ woman. Hoặc

từ Spain. Thuật toán học sâu được sử dụng để tạo vector từ đã có thể chắt lọc ý nghĩa này dựa trên cách những từ đó được sử dụng trong toàn bộ văn bản.

2.4.4. Thư viện Tensorflow

Tensorflow là một thư viện mã nguồn mở phục vụ cho hoạt động Machine Learning. Nó được xây dựng bởi Google, vì thế chúng ta có thể yên tâm về độ ổn định của nó khi sử dụng.

Hình Logo thư viện Tensorflow Python

Chúng ta đều biết rằng, trong quy trình phát triển một phần mềm bất kỳ đòi hỏi rất nhiều đoạn mã cũng như thuật toán được triển khai. Thuật toán vừa để phân tích, tổng hợp dữ liệu vừa là nền tảng để phần mềm có thể khởi chạy. Tuy nhiên chương trình càng lớn thì khối lượng phép toán càng nhiều. Cách tính toán thủ công không thể đảm bảo hiệu suất như mong muốn được. Vì thế Tensorflow xuất hiện như một chương trình hỗ trợ tính toán bằng cách tiếp cận mạnh mẽ các phép tính và bài toán phức tạp. Nhờ có Tensorflow, người dùng có thể đơn giản hóa toán học thông qua các đồ thị luồng dữ liệu tổng hợp.

Về cơ bản Tensorflow sẽ giúp người dùng tạo ra các biểu đồ luồng dữ liệu hoặc những cấu trúc mô tả. Đây cũng là lý do tại sao Tensorflow được coi như là một framework. Những khung sườn này sẽ hướng dẫn dữ liệu làm cách nào để đi qua một biểu đồ hoặc một series nodes đang được xử lý. Lúc này, mỗi nodes sẽ

đại diện cho một hoạt động toán học cần xử lý. Còn mỗi kết nối hoặc mỗi edge sẽ được coi như một tensor hoặc một mảng dữ liệu đa chiều.

2.4.5. Thư viện Flask

Hình Logo Flask Fameworks

Flask là một web frameworks, nó thuộc loại micro-framework được xây dựng bằng ngôn ngữ lập trình Python. Flask cho phép bạn xây dựng các ứng dụng web từ đơn giản tới phức tạp. Nó có thể xây dựng các api nhỏ, ứng dụng web chẳng hạn như các trang web, blog, trang wiki hoặc một website dựa theo thời gian hay thậm chí là một trang web thương mại. Flask cung cấp cho bạn công cụ, các thư viện và các công nghệ hỗ trợ bạn làm những công việc trên.

Flask là một micro-framework. Điều này có nghĩa Flask là một môi trường độc lập, ít sử dụng các thư viện khác bên ngoài. Do vậy, Flask có ưu điểm là nhẹ, có rất ít lỗi do ít bị phụ thuộc cũng như dễ dàng phát hiện và xử lý các lỗi bảo mật.

2.4.6. Mô hình Word2vec

Mô hình word2vec có 2 phương pháp chính là skip-grams và CBOW như sau:

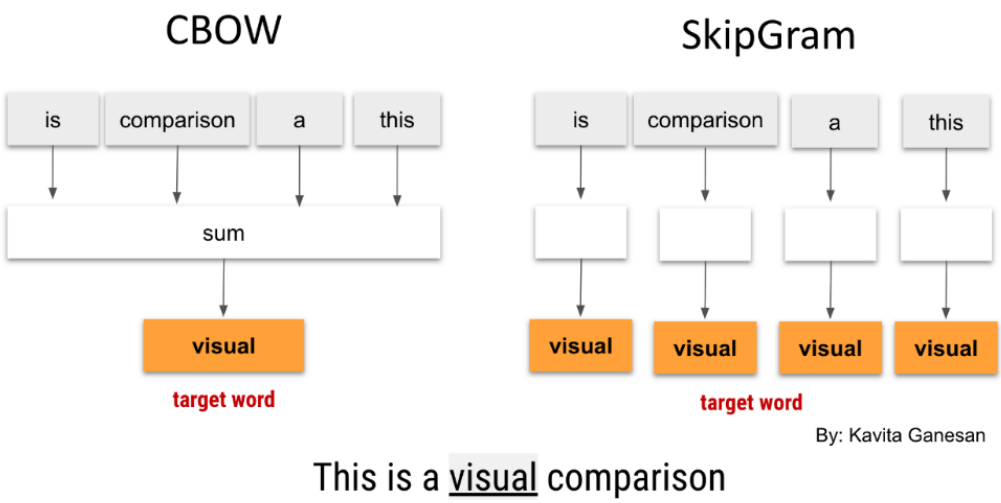
- Skip-grams: Giả sử chúng ta có một câu văn như sau: “Tôi muốn một chiếc cốc màu_xanh đựng hoa quả dầm”. Để thu được một phép nhúng từ tốt hơn chúng ta sẽ lựa chọn ra ngẫu nhiên các từ làm bối cảnh (context). Dựa trên từ bối cảnh,

các từ mục tiêu (target) sẽ được xác định nằm trong phạm vi xung quanh từ bối cảnh. Chẳng hạn ta với việc lựa chọn từ cốc làm bối cảnh nếu lấy từ tiếp theo, từ liền trước, từ cách đó liền trước 2, 3 từ ta sẽ lần lượt thu được các từ mục tiêu như sau:

Bối cảnh	Mục tiêu
cốc	màu_xanh
cốc	chiếc
cốc	một
cốc	muốn

Các nghiên cứu cho thấy từ mục tiêu sẽ được giải thích tốt hơn nếu được học theo các từ bối cảnh. Do đó mô hình skip-grams tìm cách xây dựng một thuật toán học có giám sát có đầu vào là các từ bối cảnh => đầu ra là từ mục tiêu.

- CBOW: Chúng ta nhận thấy rằng mô hình skip-grams sẽ rất tốn chi phí để tính toán vì mẫu số xác xuất là tổng của rất nhiều số mũ cơ số tự nhiên. Để hạn chế chi phí tính toán mô hình CBOW (Continuous Bag of Words) được áp dụng. Về cơ bản thì CBOW là một quá trình ngược lại của skip-grams. Khi đó input của skip-grams sẽ được sử dụng làm output trong CBOW và ngược lại
  - Kiến trúc mạng nơ ron của CBOW sẽ gồm 3 layers:
  - Input layers: Là các từ bối cảnh xung quanh từ mục tiêu.
  - Projection layer: Lấy trung bình véc tơ biểu diễn của toàn bộ các từ input để tạo ra một véc tơ đặc trưng.
  - Output layer: Là một dense layers áp dụng hàm softmax để dự báo xác xuất của từ mục tiêu.



Hình Kiến trúc thuật toán CBOW và SkipGram

Chương 3: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CẢM XÚC ĐÁNH GIÁ

3.1. Thu thập dữ liệu

Tập dữ liệu được sử dụng để huấn luyện mô hình trong đề tài này là tập dữ liệu UIT-ViSFD - Vietnamese Smartphone Feedback Dataset của nhóm tác giả với đại diện là tác giả Luong Luc Phan. Tập dữ liệu bao gồm các đánh giá của khách hàng về điện thoại thông minh được trích xuất từ website thương mại điện tử lớn ở Việt Nam. Các thuộc tính trong bộ dữ liệu:

- comment: Nội dung đánh giá của khách hàng
- n_star: Số sao người mua, người dùng đánh giá cho điện thoại thông minh
- data_time: Thời gian đánh giá được đăng tải
- label: các nhãn của bình luận

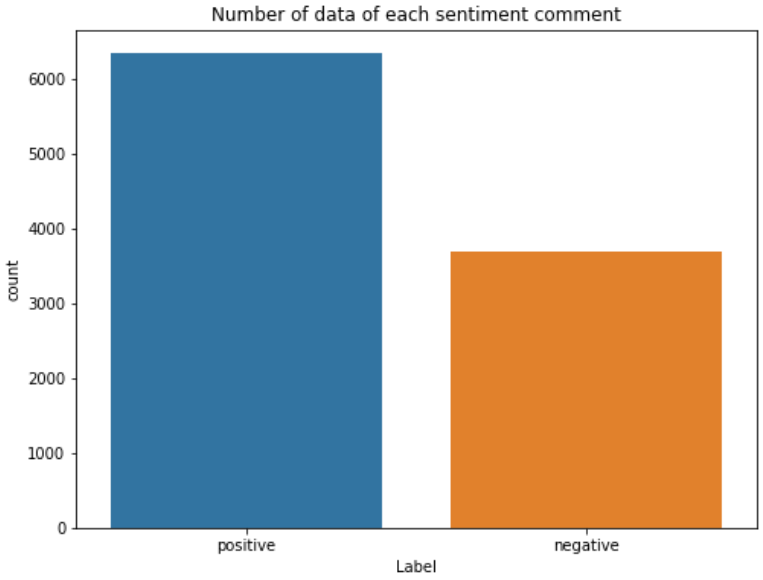
3.2. Tổng quan dữ liệu

Sử dụng 10010 đánh giá của khách hàng trong tập dữ liệu để chia ra làm dữ liệu huấn luyện (traning data) và dữ liệu đánh giá(validation data)

- Training data: [0:9000]
- Validation data: [9000:10010]

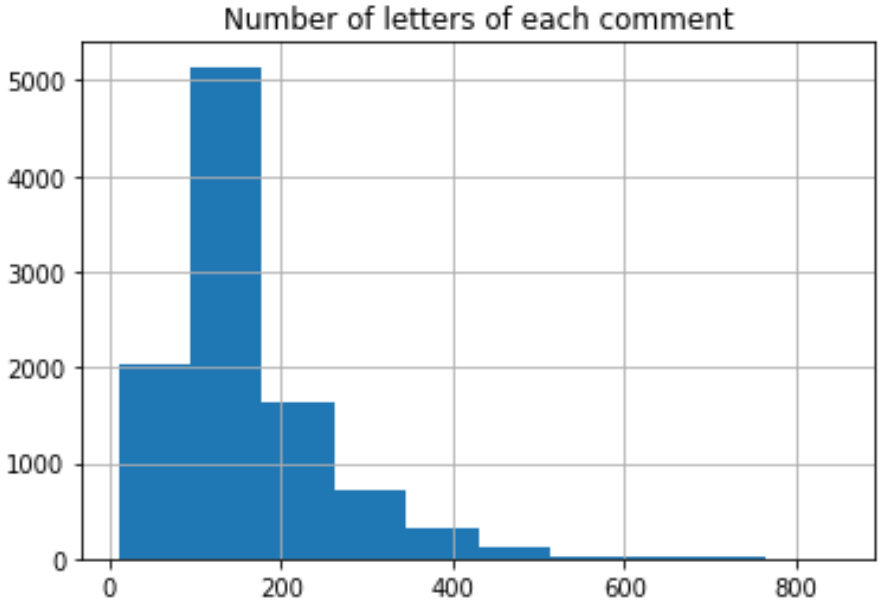
Một số đặc điểm của tập dữ liệu:

- Số lượng đánh giá tiêu cực và tích cực trên tập dữ liệu



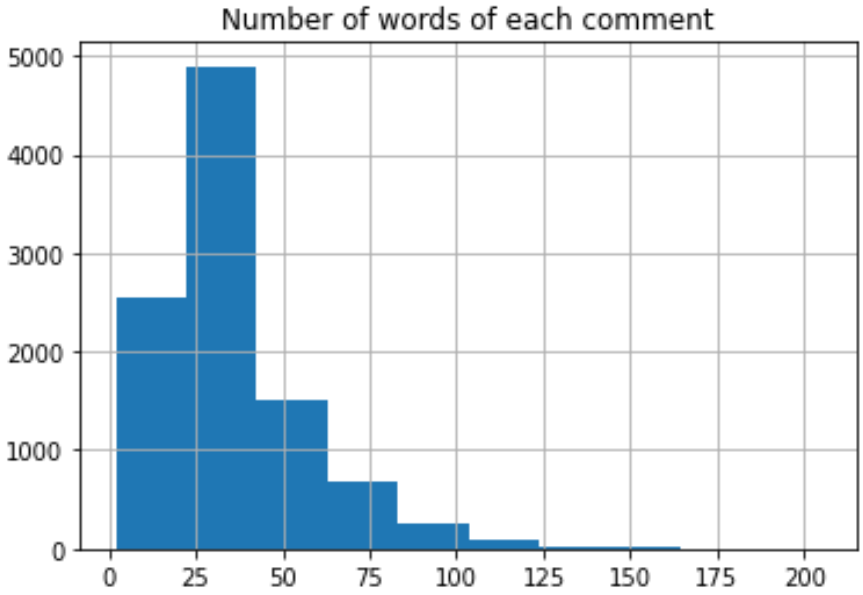
Hình Số lượng đánh giá tích cực và tiêu cực trong tập dữ liệu

- Số kí tự mỗi đánh giá tập trung trong khoảng 400 kí tự trở xuống



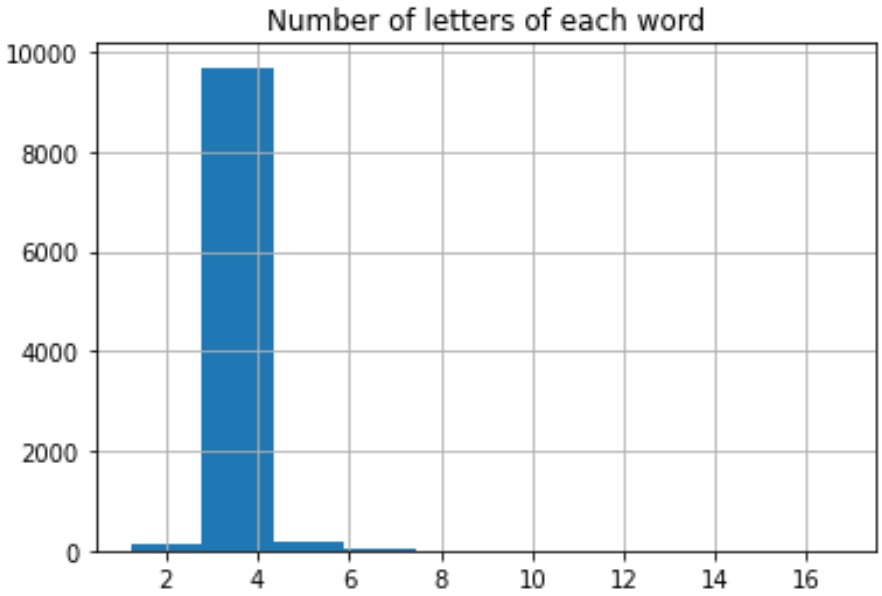
Hình Biểu đồ phân bố số lượng kí tự trong mỗi đánh giá

- Số từ trong mỗi đánh giá dài nhất lên đến 165 từ tập trung trong khoảng dưới 100 từ



Hình Biểu đồ phân bố số lượng từ có trong mỗi đánh giá

- Số kí tự trong mỗi từ dài nhất khoảng 7 kí tự và tập trung ở 4 kí tự



Hình Biểu đồ phân bố số lượng kí tự trong mỗi từ

3.3. Tiền xử lý dữ liệu

Dữ liệu trên tập dữ liệu trên là tập dữ liệu dạng thô, chưa qua xử lý. Do đó dữ liệu có thể bị thiếu, bị rỗng, các đánh giá của khách hàng có thể sai chính tả, quá dài, quá ngắn, chứa các kí tự gây nhiễu như các kí tự đặc biệt, biểu tượng icon, dấu câu, chữ số.

- **Kiểm tra và chỉnh sửa tập dữ liệu:** kiểm tra thống kê những dư liệu bị trống, những dữ liệu sai không mang ý nghĩa, những dữ liệu bị trùng. Sau đó xóa bỏ tất cả những dữ liệu đó để làm sạch tập dữ liệu
- **Xóa các kí tự đặc biệt:** các kí tự đặt biệt không mang nhiều ý nghĩa phân loại, ngược lại có thể làm nhiễu trong quá trình phân tích. Trong bước này sẽ xóa bỏ các kí tự đặc biệt, các biểu tượng icon, các dấu câu, các chữ số là những tác nhân có thể gây nhiễu cho mô hình
- **Chuẩn hóa từ:** trong các câu đánh giá của khách hàng có thể có những kí tự in hoa, in thường khác nhau, những từ được viết tắt hoặc viết sai chính tả hoặc là những teencode. Chẳng hạn như từ "không" sẽ có thể được viết là: kh, ko, k, khong, hong, hk,.. Việc này ảnh hưởng rất nhiều đến kết quản phân tích. Do đó trongm bước này sẽ tiến hành chuẩn hóa chữ hóa thành chữ thường, nối từ mang nghĩa "không" với từ liền kề sau nó (ví dụ: không được => không được)
- **Tách từ:** tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Đây là bước quan trọng trong quá trình xử lý ngôn ngữ tự nhiên đặc biệt là tiếng Việt. Trong bước này, sẽ xác định các từ đơn từ ghép và ghép từ có nghĩa không với từ đứng sau nó. Ví dụ từ "không tốt" máy sẽ hiểu được hai từ "không" và "tốt" và 2 từ này có thể mang 2 nghĩa tiêu cực và tích cực. Nhưng sau khi chuẩn hóa thì từ này sẽ đổi thành "không_tốt " và nó mang ý nghĩa cảm xúc tiêu cực hoàn toàn.

3.4. Gắn nhãn dữ liệu

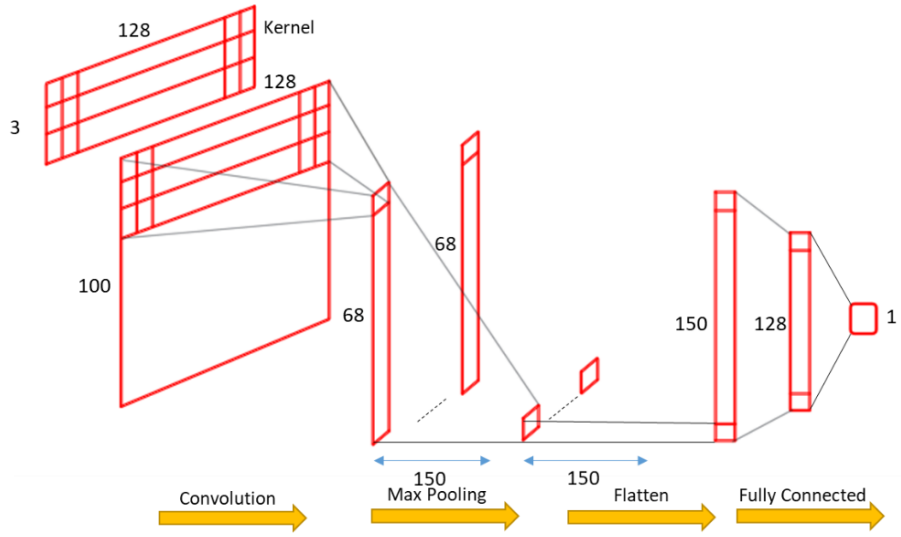
Áp dụng phương pháp phân loại cảm xúc dựa trên điểm số đánh giá (Rating) cụ thể là số sao (n_star) mà khách hàng đánh giá cho sản phẩm để gắn nhãn cho dữ liệu. Tập dữ liệu sẽ được gắn nhãn theo quy tắc sau:

- Star < 4: đánh giá dưới 4 sao sẽ được gắn nhãn là tiêu cực (negative)
- Star >= 4: đánh giá từ 4 sao trở lên sẽ được gắn nhãn là tích cực (positive)

3.5. Vector hóa dữ liệu

- Xây dựng một mô hình vector hóa từ: xây dựng một mô hình Word embedding để biểu diễn tập từ vựng sang dạng ma trận mà vẫn giữ được các đặc tính và ý nghĩa của từ. Công cụ được sử dụng ở bước này là thư viện gensim và mô hình Word2vec. Mô hình Word2vec được sử dụng trong bài toán này là mô hình tự tạo từ dữ liệu training không sử dụng mô hình pretrain.
- Tiến hành vector hóa câu: số lượng đánh giá có độ dài gần 100 chiếm đa số nên sử dụng 100 từ đầu tiên để tao vector hóa đại diện cho bình luận đó. Phương pháp là vector 100 từ đầu của đánh giá thành vector 128 chiều, xếp các vector từ trên xuống tạo thành ma trận có kích thước 100x128.

3.6. Xây dựng mô hình CNN cho bài toán



Hình Mô hình CNN được sử dụng trong trong bài toán

Thiết kế hệ thống CNN với Convolution (150 kernel 3x128) và Max Pooling (68x1)

- **Bước Convolution:**
  - Số lượng kernel được chọn: 150
- **Bước Pooling:**
  - Sử dụng Max Pooling với size (68x1)
- **Bước Fully Connected:**
  - Sử dụng Dense (128, activation = “relu”)

- Sử dụng Dense (128, activation = “sigmoid”) để đưa ra dự đoán

Tóm tắt mô hình CNN:

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_5 (Conv2D)           (None, 98, 1, 150)        57750

 max_pooling2d_5 (MaxPooling  (None, 1, 1, 150)        0
 2D)

 dropout_5 (Dropout)         (None, 1, 1, 150)         0

 flatten_5 (Flatten)         (None, 150)               0

 dense_10 (Dense)            (None, 128)               19328

 dense_11 (Dense)            (None, 1)                 129

=================================================================
Total params: 77,207
Trainable params: 77,207
Non-trainable params: 0
```

Hình Bảng tóm tắt các lớp mô hình CNN

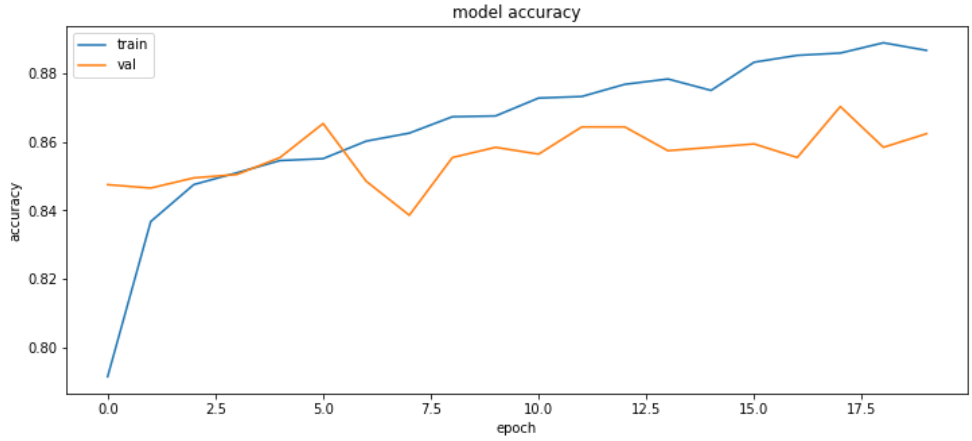
Vì Sigmoid trả về giá trị dự đoán từ 0-1 nên sử dụng hàm round làm tròn:

- Nếu giá trị dự đoán > 0.5 thì đánh giá đó là Tích cực (Positive)
- Nếu giá trị dự đoán <= 0.5 thì đánh giá đó là Tiêu cực (Negative)

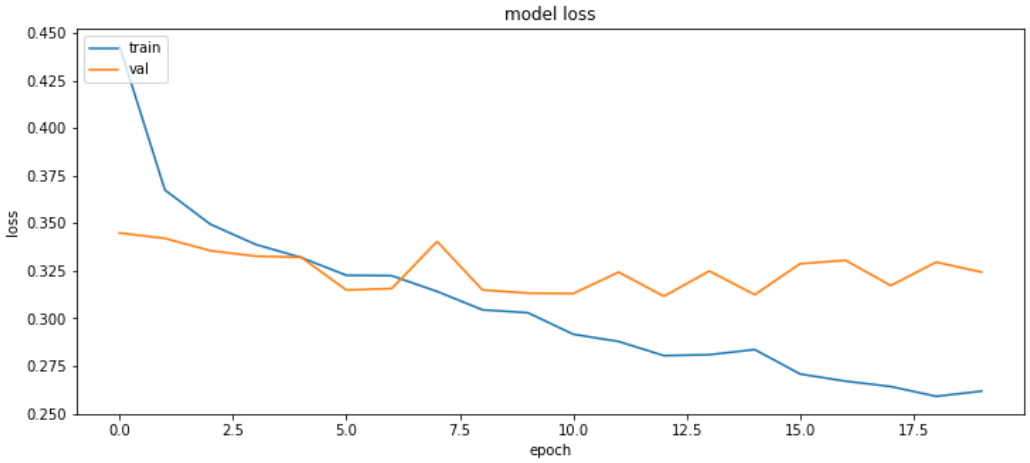
Tiến hành huấn luyện và đánh giá mô hình trên 2 tập dữ liệu Training data và Validation data với: bactch size = 30, epochs = 20. Hàm loss sử dụng binary_crossentropy, optimizer sử dụng adam.

3.7. Kết quả thực nghiệm và đánh giá mô hình

Mô hình đạt độ chính xác trên Training data là 88,67% và trên Validation data là 86,24 %



Hình Độ chính xác của mô hình trên tập Traing và Validation



Hình Giá trị mất mát của mô hình trên tập Training và Validation

Dự đoán một vài đánh giá thô và xem kết quả dự đoán của mô hình:

```
predict('máy này chơi game không được')
```

```
Tiêu cực
0.34
```

```
predict('máy bắt wifi chưa được tốt cho lắm')
```

```
Tiêu cực
0.29
```

Hình Dự đoán 2 đánh giá dễ dự đoán sai

Ta thấy, 2 đánh giá này là đánh giá dễ dự đoán sai vì nó sử dụng từ “không”, “chưa” để phủ định lại từ mang tính chất tốt. Nếu không xử lí nối từ không và từ kế tiếp nó mà đi huấn luyện mô hình thì mô hình sẽ dự đoán sai trong trường hợp này.

```
predict('máy đẹp nhưng pin yếu')
```

```
Tiêu cực
0.14
```

```
predict('pin yếu nhưng máy đẹp')
```

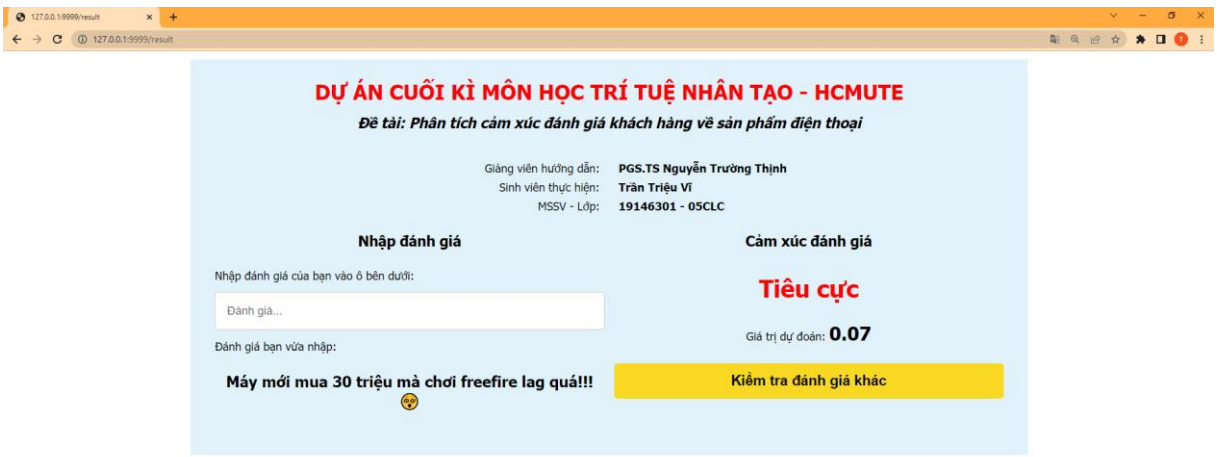
```
Tích cực
0.53
```

Hình Dự đoán 2 đánh giá cùng thuộc tính nhưng khác nghĩa

Hai đánh giá này của 2 khách hàng có các thuộc tính như pin yếu và máy đẹp giống nhau nhưng khi hoán đổi vị trí cho nhau thì sẽ ra hai đánh giá mang ý nghĩa khác nhau. Và mô hình đã làm tốt công việc dữ đoán của nó.

Chương 4: XÂY DỰNG GIAO DIỆN WEBSITE DỰ ĐOÁN ĐÁNH GIÁ KHÁCH HÀNG

❖ Tổng quan về website phân tích đánh giá khách hàng



Hình Giao diện nhập đánh giá của Website

Hình Giao diện xem kết quả phân tích đánh giá của Website

Website được xây dự dựa trên ngôn ngữ HTML và thư viện Flask có sẵn của Python, bao gồm 2 giao diện: giao diện nhập đánh giá và giao diện trả kết quả. Bố cục của website bao gồm các phần:

- Thông tin: tên đồ án, tên đề tài, thông tin giản viên, sinh viên
- Phần nhập đánh giá: bao gồm một ô để người dùng nhập vào đánh giá mà họ muốn phân tích xem là tiêu cực hay tích cực và một nút "Kiểm tra đánh giá" để hoàn thành thao tác nhập đánh giá của người dùng , tiến hành phân tích và chuyển qua giao diện hiển thị kết quả
- Phần cảm xúc đánh giá: hiển thị phân tích đánh giá của người dùng nhập vào là tiêu cực hay tích cực, giá trị mà mô hình dự đoán và một nút "Kiểm tra đánh giá khác" để quay lại giao diện nhập đánh giá.

Chương 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong đề tài này, đã hoàn thành được một giải pháp ứng dụng trong lĩnh vực phân tích ngôn ngữ tự nhiên, cụ thể là phân tích cảm xúc khách hàng dựa trên bình luận được lấy về từ trang web thương mại điện tử lớn tại Việt Nam. Giải pháp được thực hiện trên thuật toán Deep Learning cho độ chính xác trên 88%. Giải quyết được bài toán trong thời kì bùng nổ dữ liệu đó là cung cấp thông tin về những tải nghiệm của khách hàng về sản phẩm cũng như dịch vụ mà cá nhân, doanh nghiệp đang cung cấp hoặc sẽ cung cấp nhằm cải thiện sản phẩn, dịch vụ từ đó đưa ra

những chiến lược kinh doanh phù hợp. Ngoài ra, nghiên cứu này sẽ là tiền đề cho các ứng dụng phân tích dữ liệu, sử dụng giải pháp để tích hợp vào các ứng dụng khác với mục đích khảo sát cảm xúc của khách hàng đối với tất cả các sản phẩm, dịch vụ khác nhau.

Hoàn thành việc xây dựng giao diện website trực quan để người dùng dễ dàng sử dụng phân tích đánh giá mong muốn

5.2. Hướng phát triển

- Thu thập nhiều dữ liệu hơn từ nhiều nguồn khác nhau để có thể phân tích đánh giá khách hàng trên nhiều sản phẩm, dịch vụ
- Sử dụng nhiều thuật toán Machine Learning để tìm ra mô mình tối ưu nhất
- Cải tiến website với giao diện đẹp hơn, và thêm nhiều chức năng như: tải tệp bao gồm hàng loạt đánh giá, phân tích một lúc nhiều đánh giá của khách hàng và cho ra kết quả dưới dạng biểu đồ trực quan giúp người dùng dễ quan sát.
- Xây dựng hệ thống tự động thu thập dữ liệu. Tự động thu thập dữ liệu từ trang web được chỉ đinh, xử lý cơ bản dữ liệu trước khi đưa vào cơ sở dữ liệu.
- Xây dựng ứng dụng trên thiết bị di động đồng bộ với website giúp người dùng thuận tiện trong thao tác và nhanh chóng hơn.

TÀI LIỆU THAM KHẢO

1. Pham Dinh Khanh, Bài 3 - Mô hình Word2Vec, https://phamdinhkhanh.github.io/2019/04/29/ModelWord2Vec.html
2. Nguyễn Văn Hiếu, Flask python là gì? Thư viện flask trong lập trình Python, https://nguyenvanhieu.vn/thu-vien-flask-python-la-gi/
3. Got It Vietnam, Tổng quan thư viện NumPy trong Python, https://vn.got-it.ai/blog/tong-quan-thu-vien-numpy-trong-python
4. Nguyen Van Hoang, Giới thiệu về Pandas (một thư viện phổ biến của Python cho việc phân tích dữ liệu), https://viblo.asia/p/gioi-thieu-ve-pandas-mot-thu-vien-pho-bien-cua-python-cho-viec-phan-tich-du-lieu-aWj53Nnel6m
5. Nguyễn Văn Hiếu, [Khóa học tensorflow] Bài 1 – Tổng quan về thư viện Tensorflow, https://nguyenvanhieu.vn/thu-vien-tensorflow/

6. Luong Luc Phan, UIT-ViSFD, https://github.com/LuongPhan/UIT-ViSFD
7. Nguyễn Hồng Đoàn, Sentiment-Analysis-VLSP2016-BaseML, https://github.com/doanbk/Sentiment-Analysis-VLSP2016-BaseML
8. Trà My, Convolutional Neural Network là gì? Cách chọn tham số cho Convolutional Neural Network chuẩn chỉnh, https://wiki.tino.org/convolutional-neural-network-la-gi/