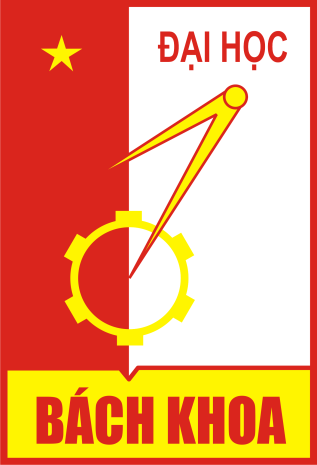


**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

VIỆN CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO BÀI TẬP LỚN**

HỌC PHẦN KHAI PHÁ WEB

ĐỀ TÀI: PHÂN TÍCH CẢM XÚC BÌNH LUẬN

TIẾNG VIỆT

*Giảng viên hướng dẫn:* **TS. NGUYỄN KIÊM HIẾU**

*Nhóm sinh viên thực hiện:* **Trần Lương Nguyên – 20152732**

**Nguyễn Văn Công Linh – 20152223**

**Lê Trịnh Thành – 20153380**

**Phạm Vũ Tiến – 20153789**

Hà Nội, 12/2019

**MỤC LỤC**

[PHẦN 1. GIỚI THIỆU BÀI TOÁN 3](#_Toc27576572)

[1.1. Lý do lựa chọn bài toán 3](#_Toc27576573)

[1.2. Mô tả bài toán 3](#_Toc27576574)

[PHẦN 2. QUÁ TRÌNH THỰC HIỆN 4](#_Toc27576575)

[2.1. Thu thập và phân tích đặc điểm của dữ liệu 4](#_Toc27576576)

[2.2. Tiền xử lý dữ liệu 5](#_Toc27576577)

[2.3. Vector hoá dữ liệu 5](#_Toc27576578)

[2.4. Xây dựng và huấn luyện mô hình 6](#_Toc27576579)

[2.4.1. Lý thuyết mô hình Convolutional Neural Network (CNN) 6](#_Toc27576580)

[2.4.2. Xây dựng mô hình CNN cho bài toán phân tích bình luận: 8](#_Toc27576581)

[PHẦN 3. KẾT QUẢ THỰC NGHIỆM 9](#_Toc27576582)

[3.1. So sánh kết quả accuracy và loss khi thay đổi các tham số 9](#_Toc27576583)

[3.2. Đánh giá trên tập test 12](#_Toc27576584)

[PHẦN 4. KẾT LUẬN 13](#_Toc27576585)

[PHẦN 5. TÀI LIỆU THAM KHẢO 14](#_Toc27576586)

**DANH MỤC HÌNH VẼ**

[Hình 1. Một vài bình luận đã được gán nhãn 4](#_Toc27384686)

[Hình 2. Kết quả phân tích nhãn 4](#_Toc27384687)

[Hình 3. Kết quả phân tích độ dài bình luận 5](#_Toc27384688)

[Hình 4. Ví dụ tiền xử lý dữ liệu 5](#_Toc27384689)

[Hình 5. Convolutional layer 6](#_Toc27384690)

[Hình 6. Pooling layer 7](#_Toc27384691)

[Hình 7. Fully Connected Layer 7](#_Toc27384692)

[Hình 8. Thiết kế CNN với Convolution (150 kernel 3x128) và MaxPooling (68x1) 8](#_Toc27384693)

# **PHẦN 1. GIỚI THIỆU BÀI TOÁN**

## **1.1. Lý do lựa chọn bài toán**

Phân tích phản hồi khách hàng (hay được gọi là Sentiment Analysis) là một bài toán thuộc lĩnh vực xử lý ngôn ngữ tự nhiên dưới dạng văn bản.

Tại sao chúng ta lại cần phân tích phản hồi khách hàng?

Ngày nay, với sự bùng nổ của các trang thương mại điện tử, có quá nhiều các sản phẩm, nhãn hàng được đăng bán trên internet. Bạn là một khách hàng, bạn luôn muốn tìm một sản phẩm tốt từ một thương hiệu uy tín. Làm sao để bạn có thể biết các sản phẩm bạn đang xem có thực sự tốt, chất lượng. Dựa trên các bình luận của những khách hàng đã mua và sử dụng, bạn có thể dựa trên đó để quyết định xem mua hay không nên mua. Đứng về phía các nhà cung cấp sản phẩm, dựa trên những thông tin khách hàng, họ có thể đưa ra được những cải tiến phù hợp, lắng nghe, phân tích những thứ khách hàng thực sự cần.

Do đó phân tích phản hồi của khách hàng là bài toán rất quan trọng trong thực tế, cần có giải pháp tốt để phân tích và khai thác hiệu quả từ các phản hồi đó.

## **1.2. Mô tả bài toán**

**Input:** Tập các bình luận về điện thoại di động trên các trang thương mại điện tử (Lazada, Shopee, Thegioididong).

**Output:** Mô hình dự đoán cảm xúc bình luận.

Bài toán đặt ra: Cần tìm cách thu thập dữ liệu (phản hồi của khách hàng) qua hoạt động mua bán online mặt hàng điện thoại di động và các loại phụ kiện điện thoại. Phân tích, xử lý dữ liệu, trích xuất đặc trưng của bình luận đã có, qua đó xây dựng mô hình dự đoán sắc thái cảm xúc của bình luận mới.

# **PHẦN 2. QUÁ TRÌNH THỰC HIỆN**

## **2.1. Thu thập và phân tích đặc điểm của dữ liệu**

– Sử dụng spider, scapy, splash để crawl dữ liệu trên các trang thương mại điện tử (Lazada, Shopee, Thegioididong). Thu được 12772 điểm dữ liệu dạng JSON.

– Đọc từng bình luận và gán nhãn cho từng điểm dữ liệu vừa thu thập:

+ Nhãn **-1**: bình luận tiêu cực.

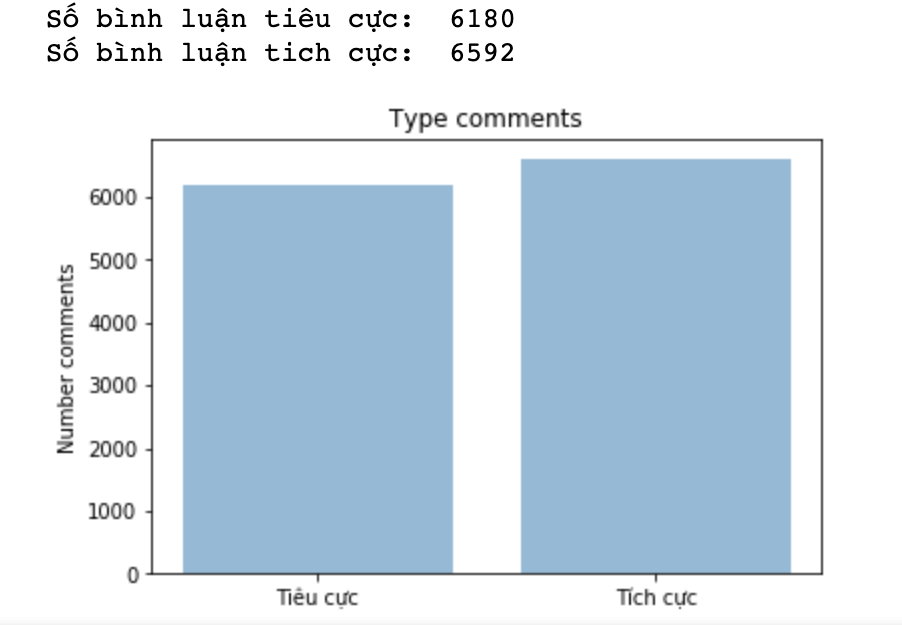
+ Nhãn **1**: bình luận tích cực.

A screenshot of a cell phone

Description automatically generated

Hình 1. Một vài bình luận đã được gán nhãn

– Phân tích số lượng các bình luận tích cực và tiêu cực:



Hình 2. Kết quả phân tích nhãn

– Phân tích độ dài bình luận:

A screenshot of a cell phone

Description automatically generated

Hình 3. Kết quả phân tích độ dài bình luận

Nhận xét: Độ dài phổ biến của các bình luận thu thập được là *70 từ/1 bình luận.*

## **2.2. Tiền xử lý dữ liệu**

– Chuẩn hoá về chữ thường

– Tách từ

– Loại bỏ số, dấu câu và các ký tự đặc biệt

A screenshot of a cell phone

Description automatically generated

Hình 4. Ví dụ tiền xử lý dữ liệu

## **2.3. Vector hoá dữ liệu**

**Bước 1:** Xây dựng mô hình vector hoá từ:

Xây dựng mô hình word embedding – mô hình chuyển một tập từ vựng về một không gian vector mà vẫn giữ được đặc trưng ý nghĩa của từ.

*Input:* 12772 câu bình luận đã tiền xử lý ở bước trước.

*Công cụ:* thư viện gensim.

*Ouput:* mô hình word2vec:

Kích thước từ điển: hơn 9600 từ

Chức năng: biến đổi một từ có trong bộ từ điển thành một vector 128 chiều.

**Bước 2:** Tiến hành vector hoá câu:

Vì số lượng bình luận có độ dài nhỏ hơn 70 chiếm đa số, nên chỉ sử dụng 70 từ đầu tiên trong các bình luận để tạo vector đại diện cho bình luận đó.

Cụ thể như sau:

*Input:* Bình luận đã được tiền xử lý.

*Phương pháp:*

Vector hoá 70 từ đầu của bình luận thành vector 128 chiều.

Xếp các vector từ từ trên xuống tạo thành ma trận có kích thước 70x128.

*Output:* ma trận kích thước 70x128 biểu diễn cho bình luận.

Sau bước này, ta có 12772 ma trận, mỗi ma trận có kích thước 70x128, đây sẽ là đầu vào của mô hình trong bước sau.

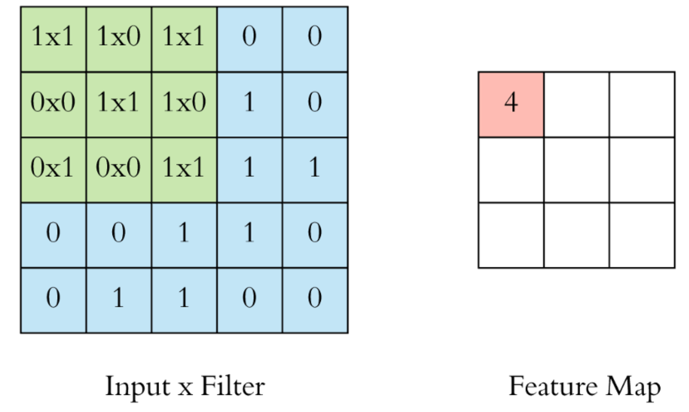
## **2.4. Xây dựng và huấn luyện mô hình**

### **2.4.1. Lý thuyết mô hình Convolutional Neural Network (CNN)**

Convolutional Neural Network là một mạng nơron tích chập. Các khái niệm quan trọng trong CNN: Convolutional Layer, Pooling Layer, Fully Connected Layer.

**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.

A close up of a keyboard

Description automatically generated

Hình 5. Convolutional layer (Nguồn:https://viblo.asia/p/ung-dung-convolutional-neural-network-trong-bai-toan-phan-loai-anh-4dbZNg8ylYM )

**Pooling Layer:**

Tầng pooling layer có mục đích làm giảm hyperparameter cần tính toán, từ đó giúp mô hình chạy nhanh hơn, tối ưu về mặt thời gian, tránh overfitting. Có 2 loại Pooling thường dung là MaxPooling và AveragePooling. Cơ chế pooling tương tự với tầng Convolutional Layer, một pooling window sẽ trượt trên ma trận dữ liệu đầu vào, sau đó chọn 1 giá trị lớn nhất đối với MaxPooling hay lấy giá trị trung bình làm đại diện đối với AveragePooling.

A close up of a clock

Description automatically generated

Hình 6. Pooling layer (Nguồn:https://viblo.asia/p/ung-dung-convolutional-neural-network-trong-bai-toan-phan-loai-anh-4dbZNg8ylYM )

**Fully Connected Layer:**

Fully Connected Layer là layer mà mỗi nút của nó kết nối với tất cả các nút của layer kế trước.

**A close up of a logo

Description automatically generated**

Hình 7. Fully Connected Layer (Nguồn: https://machinelearningcoban.com)

### **2.4.2. Xây dựng mô hình CNN cho bài toán phân tích bình luận:**

A picture containing text

Description automatically generated

Hình 8. Thiết kế CNN với Convolution (150 kernel 3x128) và MaxPooling (68x1)

***Chia dữ liệu với 12772 điểm dữ liệu:***

Tập train: [0:10000]

Tập validate: [10000:11500]

Tập test: [11500:12772]

***Bước Convolution:***

Thử nghiệm với số lượng kernel thay đổi: 1, 100, 150

Thử nghiệm với kích thước kernel thay đổi: 2x128, 3x128, 5x128

***Bước Pooling:***

Thử nghiệm Max Pooling với size = 69x1 (với kernel 2x128)

Thử nghiệm Max Pooling với size = 68x1 (với kernel 3x128)

Thử nghiệm Max Pooling với size = 66x1 (với kernel 5x128)

# **PHẦN 3. KẾT QUẢ THỰC NGHIỆM**

## **3.1. So sánh kết quả accuracy và loss khi thay đổi các tham số**

*Thay đổi kích thước kernel:*

A close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

*Thay đổi số lượng kernel:*

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

Theo dõi thay đổi của acc và loss của tập train và test khi số epoch tăng dần.

A close up of a map

Description automatically generated

## **3.2. Đánh giá trên tập test**

|  |  |  |
| --- | --- | --- |
| **Model** | **F1 micro** | **F1 macro** |
| CNN (150 Kernel có size = (3,128)) | 0.901 | 0.820 |
| SVM | 0.9053 | 0.9051 |

# **PHẦN 4. KẾT LUẬN**

Bài toán phân tích quan điểm nói riêng cũng như các bài toán Xử lý ngôn ngữ tự nhiên nói chung, có rất nhiều mô hình deep learning, các thuật toán machine learning có thể áp dụng và thu được kết quả khá tốt.

Mô hình Convolutional Neural Network được ứng dụng rất phổ biển trong xử lý hình ảnh và khi áp dụng vào bài toán phân tích quan điểm cũng đem lại kết quả khá tốt. Điều đó cho thấy sức mạnh và sự phát triển của các mô hình deep learning khi mà sức mạnh tính toán của máy tính ngày càng tăng, cùng với đó là dữ liệu trong thực tế ngày càng tăng lên nhiều lần.

# **PHẦN 5. TÀI LIỆU THAM KHẢO**

1. Convolution Neural Network - <https://nttuan8.com/>

2. Phân tích phản hồi khách hàng - <https://viblo.asia/p/phan-tich-phan-hoi-khach-hang-hieu-qua-voi-machine-learningvietnamese-sentiment-analysis-Eb85opXOK2G>

3. Convolution Neural Network - <http://cs231n.github.io/convolutional-networks/>

4. Word2vec - <https://viblo.asia/p/so-luoc-word2vec-6J3ZgBoAKmB>