**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**

**Khoa Công Nghệ Thông Tin**

**ĐỒ ÁN 3**

**Đề tài:**

**Setiment Analysist và Ứng dụng trong phân loại review**

**GVHD:** Th.s Nguyễn Thiên Bảo

**Nhóm 04:**

1. Nguyễn Đoàn Nam Anh 15110005

2. Nguyễn Xuân Tuấn 15110145

***Tp Hồ Chí Minh, Tháng 12-2018***

**Mục Lục**

[I. Đặt vấn đề 4](#_Toc530996226)

[II. Khảo sát phương pháp 4](#_Toc530996227)

[1. Bag of Word 5](#_Toc530996228)

[2. Xây dựng neural networks cho analyze sentiment 6](#_Toc530996229)

[*2.1* *Giải thuật Perceptron Learning Algorithm (PLA)* 6](#_Toc530996230)

[*2.2* *Multi-layer Perceptron* 8](#_Toc530996231)

[2.3 Dự đoán trên Neural Network 9](#_Toc530996232)

[III. XÂY DỰNG NEURON NETWORK CHO BÀI TOÁN SENTIMENT ANALYSIS OF IMDB REVIEWS 11](#_Toc530996233)

[3. TIỀN XỬ LÝ DỮ LIỆU 11](#_Toc530996234)

[3.1 Đọc và làm sạch dữ liệu 11](#_Toc530996235)

[3.2 Xác định một bộ từ vựng 12](#_Toc530996236)

[4. Xây dựng bộ từ vựng 13](#_Toc530996237)

[5. Mô hình Multilayer Perceptron (MLP) cho tập từ vựng 14](#_Toc530996238)

[IV. THỰC NGHIỆM CHƯƠNG TRÌNH 15](#_Toc530996239)

[V. PHÂN CÔNG CÔNG VIỆC 16](#_Toc530996240)

[VI. KẾT LUẬN 16](#_Toc530996241)

[1. NHỮNG ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA CHƯƠNG TRÌNH 16](#_Toc530996242)

[2. HƯỚNG PHÁT TRIỂN CHƯƠNG TRÌNH 16](#_Toc530996243)

Lời Mở Đầu

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning. Từ những thành công của Machine Learning trên thế giới, chúng em quyết định chọn đề tài “Setiment Analysist và ứng dụng trong phân loại review”. Thông qua đó để chúng em hiểu rõ hơn về lĩnh vực Machine Learning.

# Đặt vấn đề

Phân tích quan điểm người dùng ,dưới hình thức ngôn ngữ tự nhiên là một chủ đề thách thức trong Machine Learning .Những từ ngữ tự nhiên có bản chất nhập nhằng ,khó hiểu gây ra không ít khó khăn để mày tính có thể hiểu được .Deep Learning là một phương pháp của học máy cho phép dự đoán đầu ra dựa trên một bộ đầu vào được xử lý thông qua mạng Reuron(Neural networks),áp dụng mô hình Multilayer Perceptron để dự đoán kết quả .Từ những đánh giá về quan điểm của người dùng được chuyển hóa về dạng vector ở đó mỗi từ được gán một trọng số mà từ đó có thể áp dụng phương pháp deep learning đưa ra một đầu ra dự đoán phù hợp. Ở đây chúng ta sẽ áp dụng tập dữ liệu Sentiment Analysis of IMDB Reviews làm tập thử nghiệm để phát triển một hệ thống dự đoán đánh giá phim là yêu thích hoặc không dưới một đánh giá là dạng text.

# Khảo sát phương pháp

Việc đánh giá quan điểm người dùng về một vấn đề nào đó hay cụ thể về việc đánh giá bình một bộ phim ảnh ,sản phẩm ,hay bình luận mạng xã hội ,.. dưới hình thức ngôn ngữ tự nhiên là một loại bài toán sentiment analyze .Các bài toán về ngôn ngữ tự nhiên thông thường gặp khó khăn khi độ dài của các văn bản là khác nhau ,hoặc một số từ khi thêm vào bản khiến nội dung thay đổi hoàn toàn hoặc có nghĩa trái ngược với nội dung ban đầu .

Vì thế thông thường với các dữ liệu thô ban đầu chưa được chỉnh sửa .Cần phải làm sạch dữ liệu khi đó ví dụ loại bỏ những dữ liệu nhiễu ,dấu câu hay những từ xảy ra có tần suất rất thấp ,khi đó các dữ liệu liệu độ dài khác nhau sẽ được đưa về một dạng chuẩn có cùng một chiều sao cho những thông tin (features) ban đầu được giữ lại .

Một trong những mô hình xử lý văn bản có độ dài khác nhau phổ biến được phát triển các mô hình phân tích sentiment analyze này là mô hình Bag of Word .Mô hình chuyển đổi biến đổi thành các vector nơi mà mỗi từ trong đoạn văn bản được gán một trọng số mà những vector này có thể sử dụng trong tính toán mạng nhân tạo Neuron Network

## Bag of Word

Mô hình bag-of-words là một cách để biểu diễn dữ liệu văn bản thành các vector bằng thuật toán học máy , mô tả sự xuất hiện của các từ trong một tài liệu với mỗi phần tử là một số.

Nó có thể được chia thành hai giai đoạn sau:

1. Định nghĩa bộ từ vựng có liên quan dến các từ đã biết

2. Biểu diễn các từ được biết dựa trên bộ từ vựng đã được định nghĩa: Gỉa sử có một đoạn văn bản sau :

***John is more handsome than jenny(1)***

và

***Jenny is taller than John(2)***

Dựa vào trên hai văn bản này ta có một bộ từ vựng gồm 7 từ như sau :

***{John,is,more,handsome,than,Jenny,taller}***

Với mỗi văn bản , ta sẽ tạo ra một vector có chiều bằng 7 ,mỗi phần tử đại diện cho số từ tương ứng xuất hiện trong văn bản đó.

Như vậy ở hai văn bản trên sẽ có hai vector lần lượt là:

(1) [1,1,1,1,1,1,0] (2)[1,1,0,0,1,1,1]

Tuy nhiên Mô hình này có hạn chế là bất kỳ thông tin nào về thứ tự hoặc cấu trúc của các từ trong tài liệu đều bị loại bỏ. Mô hình chỉ quan tâm đến việc các từ đã biết có xuất hiện trong tài liệu hay không, không phải ở đâu trong tài liệu.

Trên thực tế bộ từ vựng có nhiều hơn số từ trên có thể có trăm ngàn hoặc hàng triệu ,vì vậy chiều dài vector sẽ rất dài

Và sẽ có rất nhiều từ không xuất hiện trong bộ từ vựng khi đó các vector thu được có nhiều phần tử bằng 0 .Để lưu trữ hiệu quả ,không nên lưu toàn bộ một vector mà chỉ lưu giá trị các phần tử khác 0 và giá trị tương ứng.

Khi xây dựng được mô hình Bag of word sẽ là một cách để thu được extracted features( biến dữ liệu thô ban đầu thành dữ liệu phù hợp với mục đích sử dụng) được dùng để huấn luyện các thuật toán Classification như neural networks.

## Xây dựng neural networks cho analyze sentiment

Mạng neuron hay nền tảng của deeplearning là một phương pháp được mô phỏng dựa trên hoạt động của hệ thống thần kinh bao gồm số lượng lớn các Neuron được gắn kết để xử lý thông tin , Deep neural networks có hiệu suất cao trên tập dữ liệu ảnh và text data, có thể dễ dàng update mô hình bằng dữ liệu mới thông qua batch propagation. Kiến trúc của mô hình này (số lượng và cấu trúc từng layer) có thể được ứng dụng qua nhiều bài toán khác nhau.Bước đầu để xây dựng mạng neuron cơ bản là giải thuật Perceptron Learning

### *Giải thuật Perceptron Learning Algorithm (PLA)*

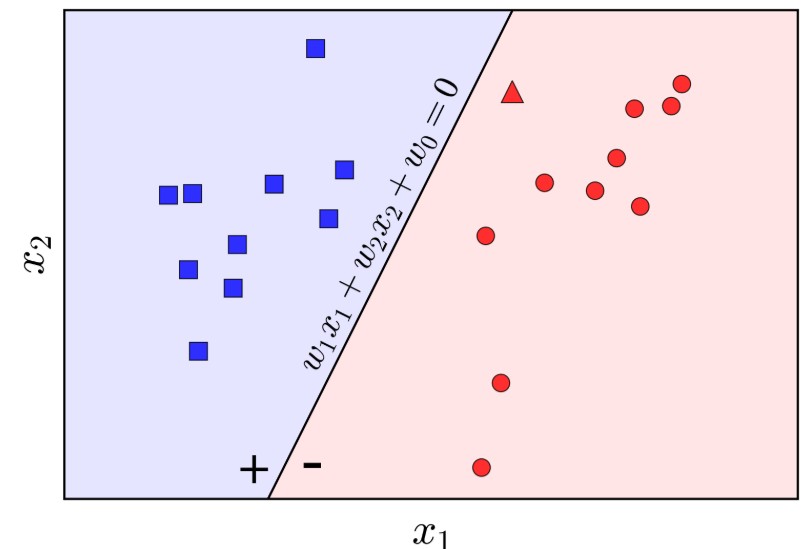
Perceptron là một thuật toán Classification cho trường hợp đơn giản nhất: chỉ có hai class (lớp) (bài toán với chỉ hai class được gọi là binary classification)

Nếu tồn tại một đường phẳng phân chia hai class thì ta gọi hai class đó là linearly separable. Các thuật toán classification tạo ra các boundary là các đường phẳng được gọi chung là Linear Classifier.

Tại một thời điểm boundary là đường phẳng có phương trình:

() = 11 + ⋯ + 𝑑𝑑 + 0

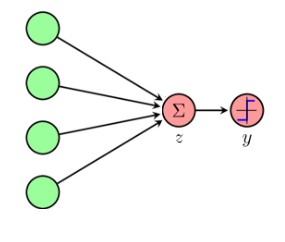
𝑇 = 0



Việc phân loại hai lớp này bằng cách xác định dấu của hàm tổng trên nếu nằm bên trái boundarry thì ta cho là lớp dương hay bên phải thì lớp âm.

Khi đó việc xác định nhãn của phần tử x mới được xác định bằng hàm sgn(𝑇)

Hàm số xác định class của Perceptron label(x)= sgn(𝑇) có thể được mô tả như hình vẽ (được gọi là network) dưới đây



Hình 1.2 Biểu diễn PLA dưới Neuron Network

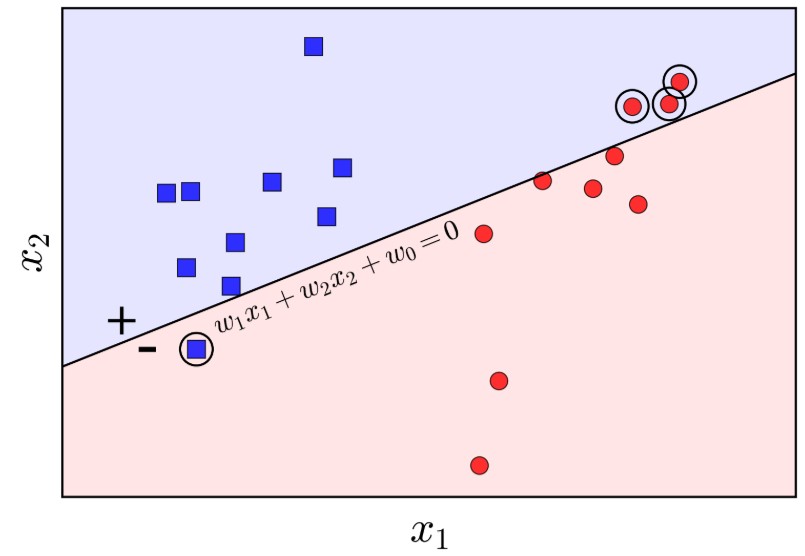
Một mạng neuron bao gồm các thành phần và đặc điểm sau:

* Các Input tương ứng với 1 thuộc tính (attribute) của dữ liệu (patterns).-Đầu ra (output) Là kết quả của mạng Perceptron, đầu ra tương ứng khi nhận vào các đầu vào inputs.

Các trọng số (weights) ứng với từng đầu vào Đây là thành phần quan trọng của mạng Perceptron nói chung và mạng neural nhân tạo nói riêng. Nó thể hiện mức độ quan trọng của từng input trong việc hình thành output.

Quá trình học của mạng Perceptron cũng chính là quá trình thay đổi các trọng số sao cho thu được đầu ra gần với mong đợi nhất.

Khi tính toán từ các input value đã cho, ta có output là Y. Giá trị mong muốn (Desired) là ̂ đã biết trước. Tuy nhiên vì đặc tính Linear Classifier với các tập phân bố ngẫu nhiên PLA không thể classifer toàn bộ các điểm,các điểm không được classifer gọi là misclassified



Hình 1.3 Misclassified trên PLA

Các điểm được khoanh tròn là các điểm bị misclassified (phân lớp lỗi). Điều chúng ta mong muốn là không có điểm nào bị misclassified

Mục đích của việc Learning là làm sao cho delta càng nhỏ càng tốt (Nếu delta =0 là hoàn hảo nhất) bằng cách điều chỉnh trọng số (weights) của các dữ liệu vào.

Nếu xi bị misclassifed, cập nhật w theo công thức: = + 𝑖𝑖

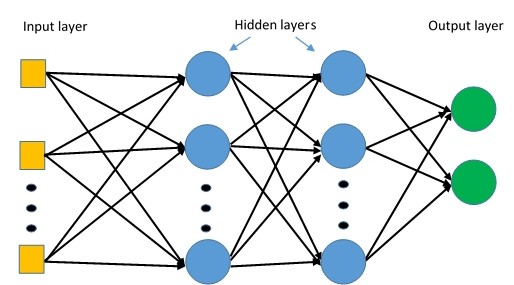
Với 𝑖 = 𝑠(𝑇)

### *Multi-layer Perceptron*

Tập hợp các PLA ta có ta có một mạng neuron có tên là Multi-layer Perceptron MLP),mạng này đóng vai trò giải quyết những các bài toán mà một đơn PLA không giải quyết được bỏi vì tính chất linearly separable của nó ,vì bản chất giải thuật PLA chỉ đưa các đường boundary có dạng tuyến tính ,và vì dữ liệu có tính ngẫu nhiên nên thông thường không thể dùng PLA để classcifer hết được toàn bộ khiến các điểm misclasssified không thể được phân lớp đúng.

Một mạng Multi-layer Perceptron sẽ bao gồm : Input layers và Output layers và

nhiều Hidden layers ở giữa như hình sau :

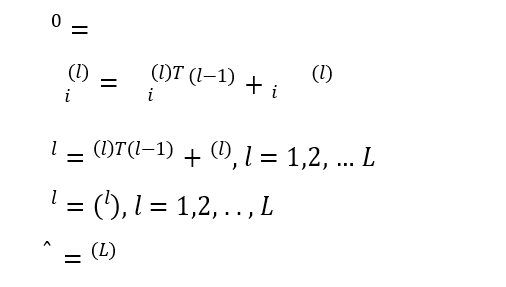


Hình 1.4 Mô hình mạng Multi-layer Perceptron

Hàm sgn(𝑇) được sử dụng bên trên gọi là hàm hàm chuyển đổi (Transfer Function),Hàm tổng của một Neuron cho biết khả năng kích hoạt (Activation) của neuron đó còn gọi là kích hoạt bên trong (internal activation). Các Nueron này có thể sinh ra một output hoặc không trong mạng (nói cách khác rằng có thể output của 1 Neuron có thể được chuyển đến layer tiếp trong mạng Neuron theo hoặc không). Mối quan hệ giữa Internal Activation và kết quả (output) được thể hiện bằng hàm chuyển đổi (Transfer Function).

### *Dự đoán trên Neural Network*

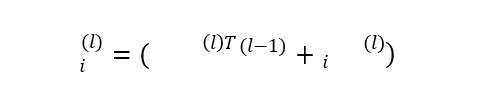
Mạng nơ-ron dự đoán dựa trên forward propagation là các phép nhân ma trận cùng với activation function để thu được kết quả đầu ra.



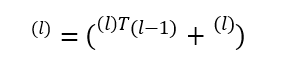
Trong đó,x là input cho dự đoán ̂, 𝑙 là input của layer thứ l, 𝑙 là output của layer thứ l sau khi áp dụng activation function, ̂ là kết quả của hàm dự đoán

Quy tắc cập nhật trọng số cho mạng Multi-layer Perceptron là :

Mỗi output được tính dựa vào công thức:



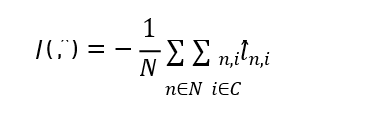
Trong đó f(.) là một (nonlinear) activation function. Ở dạng vector, biểu thức bên trên được viết là:



Qúa trình cập nhật trọng số sao cho sao cho độ lỗi của mô hình đạt được là thấp nhất. Ta gọi hàm độ lỗi của mô hình là loss function. Hàm lỗi thông thường được sử dụng là [cross-entropy loss](https://en.wikipedia.org/wiki/Cross_entropy#Cross-entropy_error_function_and_logistic_regression)

Nếu ta có N dòng dữ liệu huấn luyện, và C nhóm phân lớp ,khi đó loss function

giữa giá trị dự đoán ̂ và y được tính như sau:



Để tối ưu hàm lỗi ,người ta sử dụng phương pháp gradient descent Có hai loại gradient descent, một loại với fixed learning rate được gọi là batch gradient descent, loại còn lại có learning rate thay đổi theo quá trình huấn luyện được gọi là SGD stochastic gradient descent

Ta có thể điều chỉnh tuỳ ý số node ở hidden layer. Tuy nhiên, số node ở hidden layer càng nhiều, nghĩa là số chiều của feature càng lớn thì chi phí tính toán sẽ tăng lên, thời gian hội tụ của hàm gradient descent để tìm được thông số cho mô hình sẽ lâu hơn. Để chọn được số node ở hidden layer là một nghệ thuật và tuỳ theo bài toán cụ thể mà chúng ta đặt ra.

# XÂY DỰNG NEURON NETWORK CHO BÀI TOÁN SENTIMENT ANALYSIS OF IMDB REVIEWS

Data set sử dụng được lấy tại với tập dữ liệu huấn luyện gồm 25000 tập và 25000

tập training ,với mỗi tập là một bình luận về một bộ phim trên IMDB

## TIỀN XỬ LÝ DỮ LIỆU

Trước khi áp dụng phương pháp chuyển đổi vector dữ liệu ban đầu là thô các từ

,dấu câu hoặc số trong văn bản không làm ảnh hưởng để thông tin dữ liệu gọi là dữ liệu nhiễu và cần làm sạch nó trước khi tạo ra bộ từ vựng để tiến hành việc huấn luyện

Chuẩn bị dữ liệu gồm hai quá trình sau:

-Tải và làm sạch dữ liệu

-Xác định bộ từ vựng của các từ phổ biến

### *Đọc và làm sạch dữ liệu*

Có nhiều cách để làm sạch một bộ dữ liệu text ,sau đây là thứ tự một cách phổ biến nhất:

-Xóa bỏ các thẻ html có chứa trong tài liệu

-Tách các từ cách nhau bằng khoảng trắng

-Loại bỏ các dấu câu khỏi các từ

-Loại bỏ các ký tự không phải là bảng chữ cái

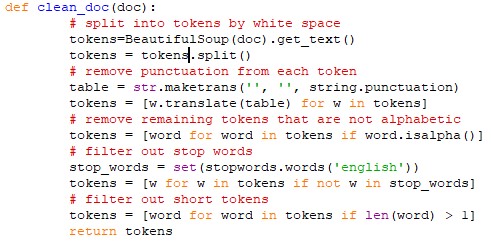
-Loại bỏ các từ có chiều dài <=1 ký tự

### *Xác định một bộ từ vựng*

- Python cung cấp cho người dùng nhiều bộ thư viện áp dụng lên việc xử lý ngôn ngữ tự nhiên ở đây ta sẽ sử dụng một số thư viện để làm sạch dữ liệu đầu vào:

Sử dụng bộ thư viện NLTK với tác dụng làm sạch dữ liệu, xử lý dữ liệu đầu vào cho các thuật toán Machine Learning, ở đây sử dụng để loại bỏ Stopwords (là các từ có tần số xuất hiện nhiều như the, to... các từ này thường mang ít giá trị ý nghĩa và không khác nhau nhiều trong các văn bản khác nhau.)

* Thư viện string import punctuation để loại bỏ tất cả các dấu có có trong văn bản
* Loại bỏ các từ có độ dài <=1 ký tự ,ký tự số không thuộc bảng chữ cái
* Như vậy ta đã có một danh sách từ được làm sạch từ một văn bản
* Ta xây dựng hàm làm sạch dữ liệu như sau :



Môt ví dụ trong dataset (txt\_sentoken2/test/neg/21\_1.txt) sau khi tiến hành làm sạch dữ liệu sẽ có kết quả như sau:

['Bloody', 'Birthday', 'totally', 'rubbish', 'slasher', 'movie', 'beginning', 'end', 'found', 'acting', 'pretty', 'good', 'considering', 'genre', 'movie', 'obvious', 'low', 'budget',

'dont', 'know', 'going', 'cinematography', 'looked', 'ghastly', 'Way', 'oversaturated','Maybe', 'bad', 'transfer', 'DVD', 'maybe', 'always', 'looked', 'like', 'dont', 'know', 'There','really', 'redeeming', 'qualities', 'speak', 'There', 'deaths', 'really', 'gory', 'wouldnt', 'bother','The', 'best', 'thing', 'DVD', 'minute', 'interview', 'producer', 'Max', 'Rosenberg', 'amusing','honest', 'He', 'didnt', 'anything', 'good', 'say', 'director', 'Ed', 'Hunt', 'admits', 'movie', 'failure', 'would', 'like', 'remake', 'believes', 'decent', 'plot', 'However', 'died', 'guess',

'someone', 'else', 'take', 'challenge', 'With', 'way', 'things', 'going', 'last', 'years', 'wouldnt',

'surprise', 'theres', 'least', 'one', 'remake', 'per', 'week', 'cinema', 'day', 'It', 'couldnt', 'worse',

'original', 'suppose', 'couldnt', 'care', 'less', 'whether', 'got', 'remade']

## *Xây dựng bộ từ vựng*

* Sau quá trình làm sạch dữ liệu tiến hành đưa các từ đã được tách rời vào bộ từ vựng chô mô hình Bag of Word ,vì có một số từ rất ít xuất hiện chỉ xảy ra một hoặc hai lần nên để hạn chế số chiều của Bag of Word ta sẽ loại bỏ những từ này.

Số lượng từ sau khi đọc hết 25000 tập huấn luyện ta thu được là :

158918

[('movie', 41208), ('The', 37828), ('film', 37026), ('one', 23190), ('like', 18872), ('good', 13928), ('This', 12733), ('would', 12010), ('time', 11514), ('really', 11274), ('It',

11232), ('story', 11061), ('even', 10880), ('see', 10850), ('much', 9292), ('get', 8994), ('people', 8472), ('bad', 8458), ('great', 8259), ('made', 7889), ('first', 7878), ('well', 7844), ('also', 7720), ('make', 7598), ('films', 7598), ('movies', 7586), ('could', 7566), ('way',7469), ('dont', 7327), ('characters', 7185), ('think', 7080), ('seen', 6486), ('But', 6485), ('character', 6457), ('watch', 6269), ('many', 6231), ('two', 6172), ('plot', 6135), ('acting',6112), ('And', 6105), ('never', 6099), ('Its', 6099), ('little', 5980), ('know', 5955), ('best',5718), ('love', 5701), ('show', 5684), ('life', 5654), ('ever', 5568), ('better', 5443)]

* Đây là 50 từ có số lượng xuất hiện nhiều nhất trong tất cả review
* Để giảm bớt số chiều ta loại bỏ thêm một số từ xuất hiện chỉ một lần duy nhất: Số chiều dữ liệu còn 54085
* Để có thể xây dựng mô hình Bag of Word cho dự đoán cần xác định bộ từ vựng cho trước ở đây là các từ đã được làm sạch từ những bộ dữ liệu review text và được lưu trong bộ từ vựng vocab.txt
* Sử dụng thư viện Tokenizer được cung cấp sẵn bởi Keras chuyển bộ từ vựng trên thành một vector có số chiều bằng số lượng từ có trong từ vựng.
* Như vậy đã có thể áp dụng lên một mô hình Multilayer Perceptron (MLP) với đầu vào lần lượt là các bình luận chuyển hóa thành dạng vector có số chiều là số lượng từ có trong từ vựng và mỗi phần từ của vector được gắn một trọng số tương ứng.

## Mô hình Multilayer Perceptron (MLP) cho tập từ vựng

Mô hình dự đoán sử dụng giải thuật Multilayer Perceptron (MLP) để phân loại tài liệu đã chuyển thành vector là khẳng định hay phủ định

Mô hình này sẽ nhận đầu vào input layer bằng với số lượng từ có trong bộ từ vựng và độ dài của các tài liệu đầu vào

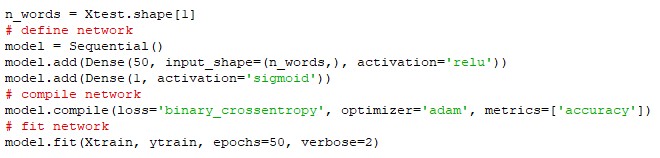
Sử dụng một single hidden layer với 50 neurons và một hàm activation là “relu”

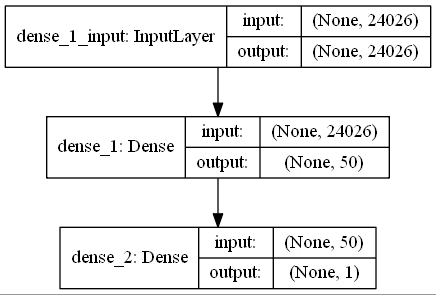
cho tốc độ xử lý nhanh hơn .

Sử dụng hàm lỗi cross-entropy loss và hàm tối ưu stochastic gradient descent

The output layer là một neuron với hàm activation là “ sigmoid” cho dự đoán 0

với nhận xét phủ định và 1 cho nhận xét khẳng định





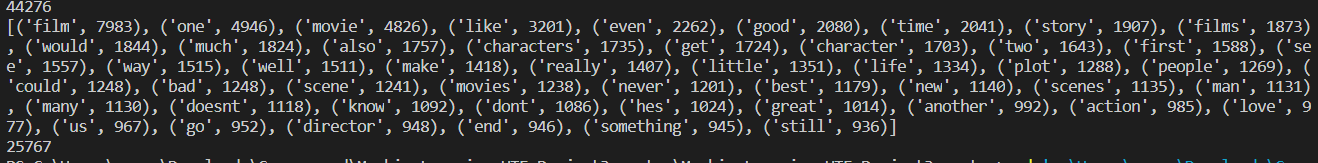
Hình 3.1 Mô hình MLP cho bài toán sentiment analyzis of IMDB reviews

Như vậy đã hoàn thành được một mô hình dự đoán cho cho bài toán Sentiment

Analyze,cụ thể là sentiment analyzis of IMDB reviews .

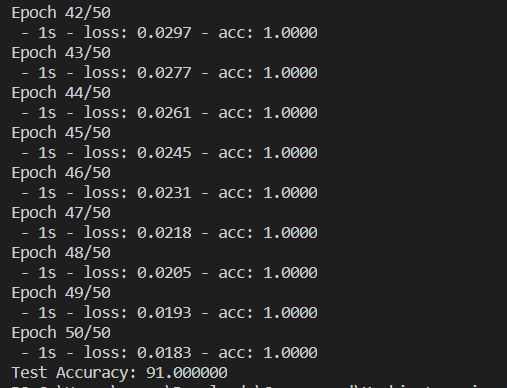
# THỰC NGHIỆM CHƯƠNG TRÌNH

* Xây dựng tập từ vựng từ dataset



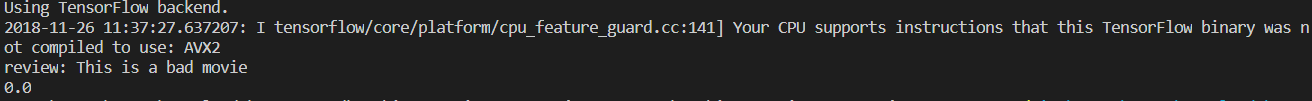
* Training cho model và lưu model lại, kết quả hàm lỗi được tối ưu sau mỗi epoch và độ chính xác được tăng lên

Kết quả hàm lỗi và độ chính xác của mô hình:



* Sau 50 lần Epoch mô hình có kết quả hàm lỗi đạt kết quả tương đối thấp 0.0183

Và có độ chính xác của mô hình là 91%



* Nhập review, chương trình sẽ tự phân loại (1 là good review, 0 là bad review)

# PHÂN CÔNG CÔNG VIỆC

|  |  |
| --- | --- |
| **TÊN SINH VIÊN** | **ĐÁNH GIÁ PHẦN TRĂM ĐÓNG GÓP** |
| NGUYỄN ĐOÀN NAM ANH | 100% |
| NGUYỄN XUÂN TUẤN | 100% |

# KẾT LUẬN

## NHỮNG ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA CHƯƠNG TRÌNH

* Xét về ưu điểm chương trình đã thực hiên đúng yêu cầu của đề tài.
* Xét về nhược điểm thì chương trình chưa nhận dạng đươc tiếng viêt, phần trăm chính xác còn khá thấp.

## HƯỚNG PHÁT TRIỂN CHƯƠNG TRÌNH

* Đối với project này chúng em muốn phát triển lên thành 1 phần mềm có thể review được đa ngôn ngữ.
* Cải tiến mạng neuron ,thay thế cấu trúc mạng nhiều lớp hơn cho tăng khả năng chính xác của mô hình.