**Shape, square

Description automatically generated**

**TRƯỜNG ĐẠI HỌC QUY NHƠN**

KHOA CÔNG NGHỆ THÔNG TIN

-----🙞🙜🕮🙞🙜-----

A blue and white logo

Description automatically generated with low confidence

**BÁO CÁO BÀI TẬP LỚN**

**HỌC PHẦN: XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**CHỦ ĐỀ: MÔ HÌNH NGÔN NGỮ- PHÁT HIỆN TIN GIẢ**

***Giảng viên hướng dẫn:*** TS. LÊ QUANG HÙNG

***Lớp học phần:*** 222105029601

***Sinh viên thực hiện:*** TRƯƠNG QUANG HƯNG

***Mã sinh viên:*** 4351050117

***Khoa:*** CÔNG NGHỆ THÔNG TIN

***Ngành và khóa:*** CNTT K43A

***Bình Định, Tháng 5 năm 2023***

**MỤC LỤC**

[**I.** **GIỚI THIỆU** 3](#_Toc135038083)

[**II.** **PHÁT HIỆN TIN GIẢ TỰ ĐỘNG** 3](#_Toc135038084)

[**1.** **Dựa trên các kỹ thuật học máy truyền thống** 3](#_Toc135038085)

[1.1. Phương pháp 3](#_Toc135038086)

[1.1.2. Ưu điểm 7](#_Toc135038087)

[1.1.3. Hạn chế 7](#_Toc135038088)

[**2.** **Dựa trên các kỹ thuật học sâu** 7](#_Toc135038089)

[2.1. Phương pháp 7](#_Toc135038090)

[2.2. Ưu điểm 11](#_Toc135038091)

[2.3. Hạn chế 12](#_Toc135038092)

[**III.** **MÔ HÌNH ĐƯỢC SỬ DỤNG** 12](#_Toc135038093)

[**IV.** **KẾT QUẢ** 16](#_Toc135038094)

[**V.** **KẾT LUẬN** 18](#_Toc135038095)

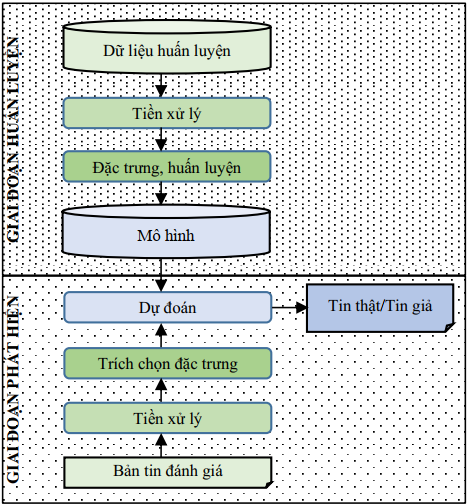
1. **GIỚI THIỆU**

Trong những năm gần đây, trí tuệ nhân tạo (AI) đã mang lại những thay đổi quan trọng trong lĩnh vực công nghệ thông tin và kiến trúc, chẳng hạn như sử dụng và phát triển hệ thống giao thông thông minh, trợ lý cá nhân ảo, phẫu thuật bằng robot và có thể tác động lớn nhất đến cuộc sống của chúng ta là ngôn ngữ tự nhiên.

Ngày nay, nhờ có internet, chất lượng và số lượng tin tức tăng lên hàng ngày, đồng thời cách người tiêu dùng tiếp cận và quản lý thông tin trực tuyến hàng ngày cũng không ngừng thay đổi. Những người trẻ tuổi, đặc biệt là những người thiếu kinh nghiệm, sử dụng các nền tảng truyền thông xã hội, ứng dụng di động hoặc các trang web đơn giản và năng động để trích xuất thông tin cần thiết một cách nhanh chóng và dễ dàng, nhiều lần mà không cần phân biệt. Sự đa dạng của tin tức trực tuyến có thể làm tăng sự tham gia vào các cuộc bầu cử dân chủ, tạo cơ hội cho mọi người tham gia hoặc thậm chí thay đổi quan điểm và tâm lý. Tuy nhiên, các công nghệ và tính năng mới có thể được sử dụng thông qua các nền tảng truyền thông xã hội để lan truyền tin giả trên quy mô lớn, tạo ra thông tin được cá nhân hóa và trở nên hiệu quả hơn trong các chiến dịch thông tin sai lệch. Do đó, cần có những nguồn thông tin đáng tin cậy và đáng tin cậy để công chúng không trở thành con mồi cho ý định của những kẻ muốn thao túng thực tế. Một số nhà nghiên cứu cũng cho rằng các chính trị gia dân túy sử dụng tin giả để làm suy yếu chính quyền.

1. **PHÁT HIỆN TIN GIẢ TỰ ĐỘNG**
2. **Dựa trên các kỹ thuật học máy truyền thống**
   1. Phương pháp

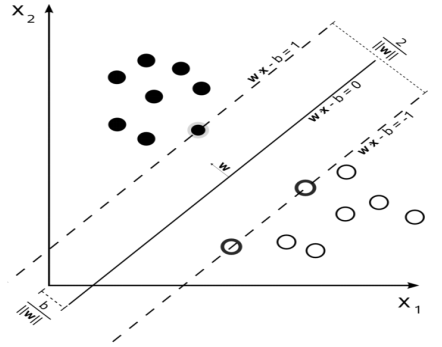
Đa số các nghiên cứu theo hướng này đều sử dụng phương pháp học máy giám sát (Supervised Learning) hoặc học máy bán giám sát (Semi-Supervised Learning) để  
huấn luyện tạo mô hình nhằm mục đích phân loại tập dữ liệu và dự đoán.  
Mô hình tổng quát của hướng tiếp cận này như Hình 1.



***Hình 1.*** *Mô hình học máy để phát hiện tin giả*

Bước đầu tiên trong mô hình này là giai đoạn thu thập tập dữ liệu để xây dựng cơ sở dữ liệu huấn luyện. Trong cơ sở dữ liệu này bao gồm các bản tin đã được gán nhãn là tin giả hoặc tin thật. Trong trường hợp học máy giám sát, tất cả các dữ liệu dùng để huấn luyện đều phải được gán nhãn, trong trường hợp học bán giám sát thì bao gồm cả dữ liệu đã gán nhãn và chưa gán nhãn. Giai đoạn tiền xử lý cho phép sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để làm sạch dữ liệu, loại bỏ các thông tin không có ích và biểu diễn lại dữ liệu. Giai đoạn trích chọn đặc trưng cho phép trích lọc những đặc trưng ngôn ngữ cần thiết phục vụ cho việc phân loại, nhận dạng nội dung. Trên cơ sở các đặc trưng đã trích xuất, thực hiện việc huấn luyện theo các thuật toán lựa chọn để xây dựng mô hình đặc trưng. Mô hình này sẽ được sử dụng cho giai đoạn dự đoán một bản tin là tin giả  
hay tin thật. Giai đoạn dự đoán có chức năng đối sánh các đặc trưng  
của bản tin cần đánh giá với mô hình đặc trưng đã tạo ra trong giai đoạn huấn luyện để quyết định xem bản tin đó là tin giả hay tin thật. Có nhiều thuật toán được sử dụng để huấn luyện và dự đoán trong học máy như:

1. **Naïve Bayes**: Thuật toán này hoạt động dựa trên tiếp cận xác suất và định lý Bayes. Nói một cách đơn giản, Naïve Bayes giả định rằng, một thuộc tính trong danh mục này không liên quan gì đến các thuộc tính khác. Ví dụ, trái cây sẽ được phân loại là táo khi có màu đỏ, hình xoáy và đường kính gần 8 cm. Bất kể các thuộc tính này phụ thuộc vào nhau hay các thuộc tính khác. Thuật toán này thường được sử dụng để phân loại văn bản.  
   Xác suất Naïve Bayes:  
   𝑃(𝑐|𝑥) = 𝑃(𝑥|𝑐)𝑃(𝑐)  
   𝑃(𝑥)  
   𝑃(𝑐|𝑋) = 𝑃(𝑥1|𝑐) × 𝑃(𝑥2|𝑐) × . . .× 𝑃(𝑥𝑛|𝑐) × 𝑃(𝑐)  
   - P(c|x) là xác suất của c khi biết x. Trong đó, c là các  
   lớp (nhãn) và x là tập các thuộc tính (đặc trưng).  
   - P(c) là xác suất của lớp c.  
   - P(x|c) là xác suất của x nếu biết c.  
   - P(x) là xác suất của x.  
   **2) Decision Tree**: Cây quyết định là một công cụ quan trọng hoạt động dựa trên cấu trúc giống như biểu đồ luồng được sử dụng chủ yếu cho các bài toán phân loại. Mỗi nút bên trong của cây quyết định chỉ định một điều kiện hoặc một "kiểm tra" trên một thuộc tính và việc phân nhánh được thực hiện trên cơ sở các điều kiện và kết quả kiểm tra. Cuối cùng, nút lá mang nhãn lớp thu được sau khi tính toán tất cả các thuộc tính. Khoảng cách từ gốc đến lá thể hiện quy luật phân loại. Chúng rất quan trọng trong việc tạo ra các biến và tính năng mới hữu ích cho việc khám phá dữ liệu và dự đoán biến mục tiêu khá hiệu quả.  
   **3) Support Vector Machine (SVM)**: Đây là thuật toán hỗ trợ phân loại rất phổ biến và hiệu quả, có thể áp dụng trong học có giám sát hoặc bán giám sát. Mục đích của SVM là phân loại dữ liệu thành hai lớp khác nhau, trong trường hợp này là lớp các tin giả và lớp các tin thật. Với một bộ các mẫu huấn luyện thuộc hai thể loại cho trước, thuật toán huấn luyện SVM xây dựng một mô hình SVM để phân loại các mẫu khác vào một trong hai thể loại đó. Thuật toán SVM chia hai lớp dữ liệu bằng một siêu mặt phẳng d–1 chiều khi số chiều của dữ liệu huấn luyện là d. Trong đó, w.x–b=0 là siêu mặt phẳng thể hiện sự phân tách dữ liệu.



***Hình 2.*** *Mô hình phân lớp bằng SVM*

**4) K-Nearest Neighbors (KNN)**: Một thuật toán đơn giản được sử dụng cho cả nhiệm vụ phân loại và hồi quy. KNN là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có. Ý tưởng của thuật toán KNN cho rằng, những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm dữ liệu x, y có k thuộc tính có thể dựa trên các khoảng cách:  
- Euclidean: √∑k i=1(xi - yi)2;  
- Manhattan: ∑k i=1|xi - yi|;  
- Minkowski: (∑k i=1(|xi - yi|)q)1/q

### 1.1.2. Ưu điểm

Các phương pháp phát hiện tin giả dựa trên học máy để phân tích nội dung có một số ưu điểm nhất định như:  
- Dễ dàng triển khai vì các giải thuật học máy đã được ứng dụng rất nhiều và đã có những cải tiến, hoàn thiện nhất định để đảm bảo chất lượng tốt và thời gian phân tích ngắn.  
- Cho kết quả khá tốt trong trường hợp có một bộ dữ liệu chất lượng và được cập nhật thường xuyên. Việc phát hiện tin giả chỉ đơn thuần là thực hiện phân loại nhị phân vào một trong hai nhóm là tin giả hoặc tin thật.  
- Đây là hướng tiếp cận phát hiện tin giả dựa trên nội dung nên có thể kết hợp với các hướng tiếp cận khác như phân tích về lan truyền tin, mức độ trích dẫn, phân tích hình ảnh... để nâng cao hơn độ chính xác của phát hiện tin giả.

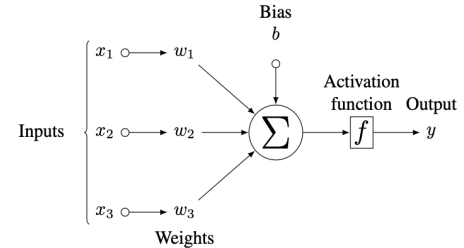
### 1.1.3. Hạn chế

Tuy nhiên, phương pháp học máy có một số hạn chế cần phải tiếp tục nghiên cứu, khắc phục bao gồm:  
- Phải liên tục cập nhật dữ liệu huấn luyện để điều chỉnh mô hình vì trong thực tế dữ liệu tin tức thay đổi hàng ngày,  
hàng giờ, thậm chí hàng giây.  
- Phương pháp này chỉ mới dừng ở phân tích nội dung mà chưa tính đến các yếu tố khác như đặc điểm lan truyền  
tin, hình ảnh,...  
- Phụ thuộc vào lĩnh vực thông tin như chính trị, dịch bệnh, quảng cáo,... nên cần phân loại dữ liệu trước để tăng tốc độ xử lý và độ chính xác.

1. **Dựa trên các kỹ thuật học sâu**
   1. Phương pháp

Các mạng nơ-ron chuyển tiếp sâu được gọi là mạng nơ-ron truyền thẳng hoặc các perceptron nhiều lớp là các mô hình cơ bản của học sâu.

Mục tiêu của mạng nơ-ron chuyển tiếp là làm gần đúng hàm f ∗. Ví dụ, y=f∗(x) ánh xạ đầu vào x (input) thành đầu ra y (output). Mạng nơ-ron chuyển tiếp (forward neural network) định nghĩa một ánh xạ y=f(x; b) và tìm giá trị của các tham số b, dẫn đến giá trị xấp xỉ tốt nhất của hàm f. Mô hình tổng quát của mạng nơ-ron được biểu diễn như Hình 3.



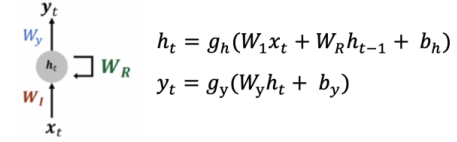
***Hình 3.*** *Mô hình mạng nơ-ron*

Mô hình cơ bản của một tế bào nơ-ron được gọi là tế bào cảm thụ. Perceptron nhận tín hiệu đầu vào x= (x1, x2,..., xn + 1) thông qua các lớp chuyển tiếp để tạo ra véc-tơ w=(w1, w2, ..., wn + 1). Đầu ra Perceptron được cho dưới dạng tích vô hướng của trọng số và véc-tơ, được biến đổi  
bởi hàm kích hoạt: 𝑜𝑢𝑡𝑝𝑢𝑡 = 𝑓(𝑤. 𝑥) = 𝑓(∑𝑛 𝑖=+11 𝑤𝑖𝑥𝑖) Dựa trên mô hình tổng quát này, người ta có thể đề xuất các thuật toán học sâu khác nhau hoạt động tương tự như các thuật toán học máy. Tuy nhiên, có một sự khác biệt chính đó là các thuật toán học sâu có các lớp diễn giải dữ liệu khác nhau. Mạng nơron nhân tạo đề cập đến mạng của các thuật toán như vậy (gọi chung là Perceptron) [17].  
*1) Mạng nơ-ron tích chập*Mạng nơ-ron tích chập (Convolutional neural networks - CNN) là mạng nơ-ron đặc biệt được sử dụng để xử lý dữ liệu. Những dữ liệu này được biểu diễn chính dưới dạng ma trận. Dữ liệu trong trường hợp phát hiện tin giả là tập hợp m văn bản, mỗi văn bản được xử lý và biểu diễn dưới dạng một véc-tơ n chiều. Như vậy, dữ liệu vào sẽ là một ma trận m⊆n. Các mạng tích chập nằm trong các mạng nơron đơn giản sử dụng tích chập của nhiều dữ liệu số có thể có trong một trong các lớp của chúng.

Khái niệm tích chập trong toán học được định nghĩa là:

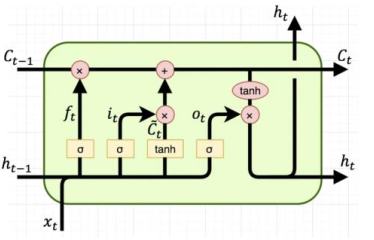


Các lớp tích chập (convolutional layer) có các tham số K (Kernel) đã được học để tự điều chỉnh và lấy ra những thông tin chính xác nhất mà không cần chọn các đặc trưng.  
*2) Mạng nơ-ron hồi quy*Trong mô hình mạng nơ-ron thông thường, chúng ta coi input là các dữ liệu độc lập, không có mối liên hệ với nhau. Tuy nhiên, trong ngôn ngữ tự nhiên thì mối liên hệ giữa các từ và ngữ cảnh đóng một vai trò quan trọng, quyết định ý nghĩa của câu văn. Do đó việc áp dụng mô hình mạng nơ- ron thông thường vào các bài toán xử lý ngôn ngữ tự nhiên thường không đạt kết quả mong muốn.  
Để khắc phục nhược điểm này, chúng ta sử dụng mô hình mạng nơ-ron hồi quy (Recurrent Neural Network - RNN). RNN coi dữ liệu đầu vào là một chuỗi liên tục và có thứ tự (Sequence), nối tiếp nhau theo thứ tự thời gian. Ví dụ như một đoạn văn bản có thể được coi là một chuỗi các từ (words) hoặc là một chuỗi các ký tự (character). Tại thời điểm t, với dữ liệu đầu vào xt ta có kết quả output là yt. Tuy nhiên, khác với mạng nơ-ron thường, yt lại được sử dụng là input để tính kết quả output cho thời điểm (t+1). Điều này cho phép RNN có thể lưu trữ và truyền thông tin đến thời điểm tiếp theo. Mô hình hoạt động của RNN có thể được mô tả trong Hình 4.



***Hình 4.*** *Mô tả cách xử lý của mạng RNN*

Thông thường hàm kích hoạt gh được sử dụng là *tanh*còn gy có thể là hàm *sigmoid* hoặc *softmax* tùy thuộc vào  
từng bài toán cụ thể.  
*3) Mạng bộ nhớ ngắn-dài hạn LSTM*Về mặt lý thuyết thì RNN có thể xử lý và lưu trữ thông tin của một chuỗi dữ liệu với độ dài bất kỳ. Tuy nhiên, trong thực tế thì RNN chỉ tỏ ra hiệu quả với chuỗi dữ liệu có độ dài không quá lớn (short-term memory). Nguyên nhân của vấn đề này là do vấn đề suy giảm gradient  
(gradient được sử dụng để cập nhật giá trị của ma trận trọng số trong RNN và nó có giá trị nhỏ dần theo từng lớp khi thực hiện lan truyền). Khi gradient trở nên rất nhỏ (có giá trị gần bằng 0) thì giá trị của ma trận trọng số sẽ không được cập nhật thêm và do đó mạng Neuron sẽ dừng việc học tại lớp này. Đây cũng chính là lý do khiến cho RNN không thể lưu trữ thông tin của các bước thời gian trước đó trong một chuỗi dữ liệu có độ dài lớn. LSTM (Long Short Term Memory) là một mạng cải tiến của RNN nhằm giải quyết vấn đề ghi nhớ lại giá trị các lớp trước đó. Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó, mỗi nút mạng đã có thể ghi nhớ được mà không cần bất kì can thiệp  
nào. Chi tiết về cách thức xử lý tại một nút mạng của LSTM được mô tả như Hình 5.



***Hình 5.*** *Mô tả một nút mạng trong LSTM*

Trong đó, *ft, it, ot* tương ứng với forget gate (cổng quên), input gate (cổng vào) và output gate (cổng ra).  
- Cổng vào: Bước này sẽ quyết định xem thông tin nào cần bỏ đi từ trạng thái tế bào. Nó sẽ lấy đầu vào là *ht-1* và *xt* rồi đưa ra kết quả là một số trong khoảng [0, 1] cho mỗi số trong trạng thái tế bào *Ct-1*. Nếu là 1 nó sẽ lưu trữ thông tin lại cho sau này, còn 0 sẽ xoá toàn bộ thông tin.  
Hàm ft được tính như sau:  
ft= 𝜎(Wf ∗ xt + Wf ∗ ht-1 + bf)  
- Cổng vào: Bước này quyết định xem thông tin mới nào ta sẽ lưu vào trạng thái tế bào. Việc này gồm 2 phần: Đầu tiên là sử dụng một tầng *sigmoid* được gọi là “tầng cổng vào” để quyết định giá trị nào ta sẽ cập nhập; Tiếp theo là một tầng *tanh* tạo ra một véc-tơ cho giá trị mới 𝐶̃𝑡 nhằm thêm vào cho trạng thái. Sau đó, mạng sẽ kết hợp 2 giá trị đó lại để tạo ra một cập nhập cho trạng thái.  
𝑖t = 𝜎(Wi ∗ xt + W𝑖 ∗ ht-1 + bi)  
𝐶̃𝑡 = tanh(Wc ∗ xt + W𝑐 ∗ ht-1 + b𝑐)  
Ct = 𝑓t ∗ Ct-1 + i𝑡 ∗ 𝐶̃𝑡)  
- Cổng ra: Giá trị đầu ra sẽ dựa vào trạng thái tế bào, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, ta chạy một tầng *sigmoid* để quyết định phần nào của trạng thái tế bào ta muốn xuất ra. Sau đó, cung cấp trạng thái tế bảo qua một hàm *tanh* để có giá trị trong khoảng [-1,1] và nhân nó với đầu ra của cổng *sigmoid* để được giá trị đầu ra mong muốn.  
𝑜t = 𝜎(W0 ∗ xt + W𝑜 ∗ ht-1 + b0)  
ℎt = ot ∗ xt + tanh(𝐶t)

* 1. Ưu điểm

Học sâu là một dạng đặc biệt của học máy nên có tất cả ưu điểm của học máy như đã trình bày ở Mục 1.1.2. Tuy nhiên, khi ứng dụng các kỹ thuật học sâu để phát hiện tin giả, ta có thể áp dụng các kỹ thuật xử lý ở nhiều tầng tương ứng với các lớp mạng thay vì chỉ xử lý tuyến tính như các kỹ thuật học máy. Hơn nữa, các kỹ thuật học sâu cho phép xử lý khối lượng dữ liệu rất lớn và vì vậy rất phù hợp với phát hiện tin giả trong bối cảnh dữ liệu gia tăng nhanh chóng hiện nay.

* 1. Hạn chế

Với kết quả nghiên cứu đến lúc này, việc ứng dụng học sâu trong phát hiện tin giả vẫn gặp phải các vấn đề như đối với các phương pháp học máy đã trình bày ở Mục 1.1.3.  
Hơn nữa, việc tự phát triển và cài đặt một hệ thống dựa trên học sâu là khá phức tạp và tốn kém. Đa số các thử nghiệm hiện nay trên các mạng học sâu được đều sử dụng các bộ thư viện sẵn có của Google (Keras, TensorFlow), Facebook (Pytorch, Caffe), Microsoft (CNTK, Gluon),... Cuối cùng, vấn đề làm thế nào để thiết kế một mạng học sâu xử lý tích hợp các thông tin dựa trên nội dung, phương thức lan truyền, văn phong,... là vẫn còn ở phía trước.

1. **MÔ HÌNH ĐƯỢC SỬ DỤNG**

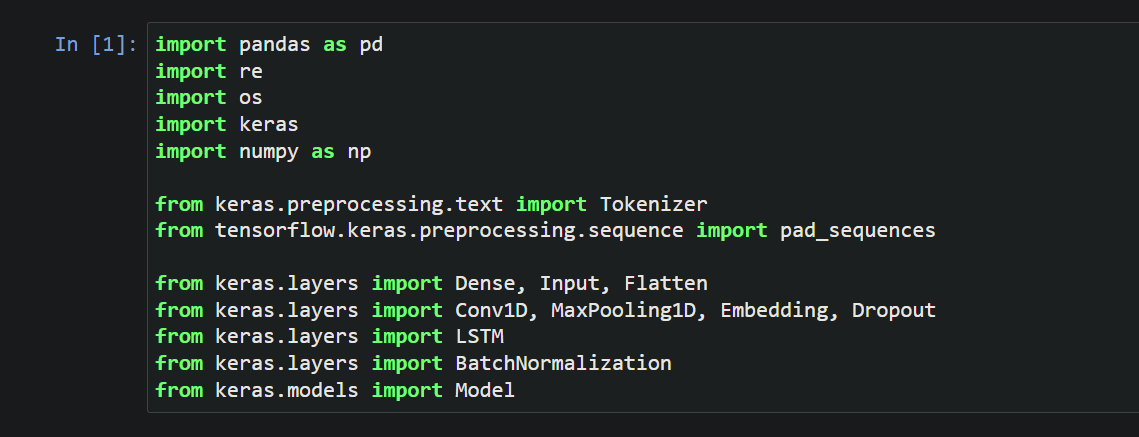
Mạng nơ-ron tích chập (Convolutional neural networks - CNN) là mô hình được em sử dụng để thực hiện bài báo cáo này.

-Bộ dữ liệu:

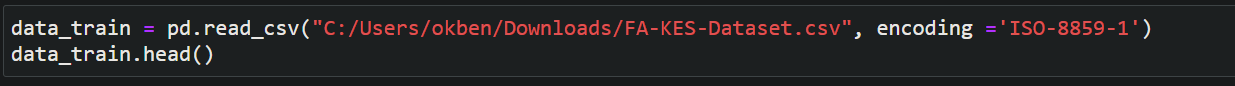




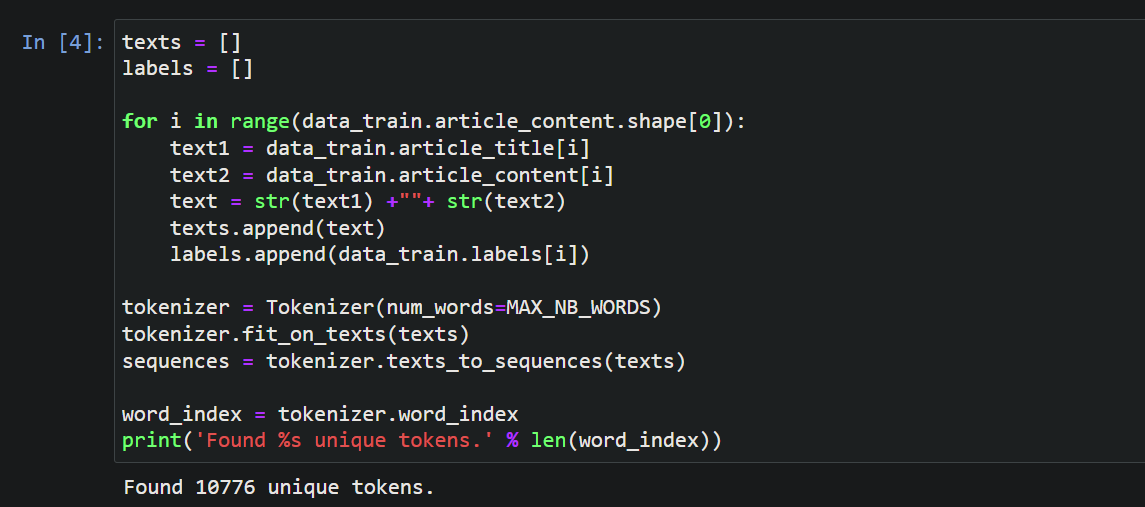
-Nhập pack cần thiết:

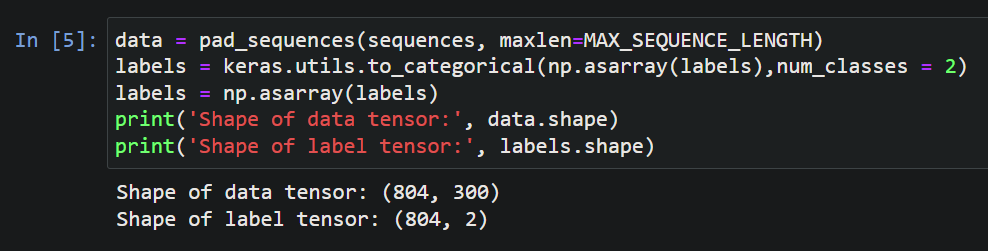


-Đọc bộ dữ liệu:

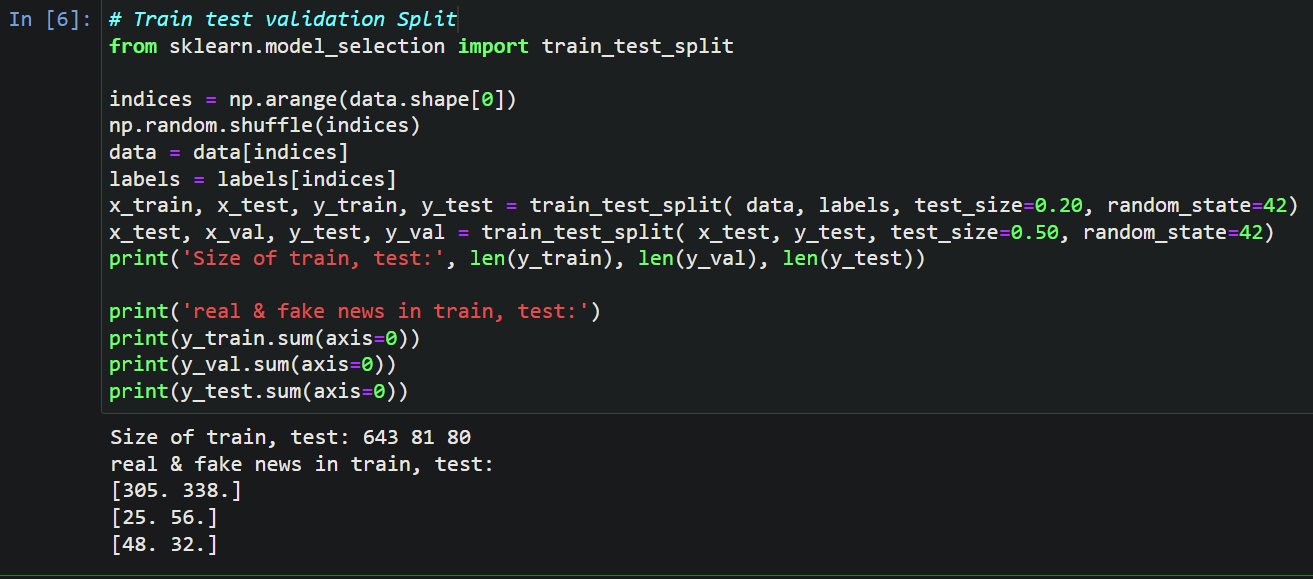


-Tiền xử lý dữ liệu:

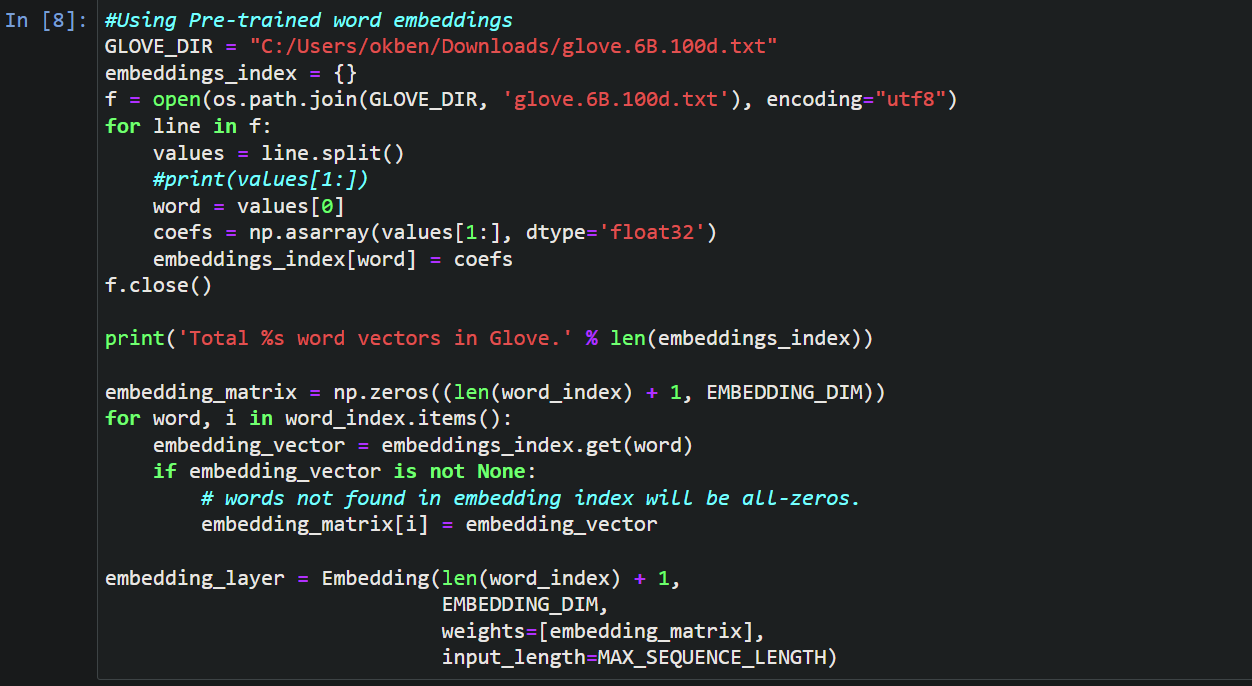




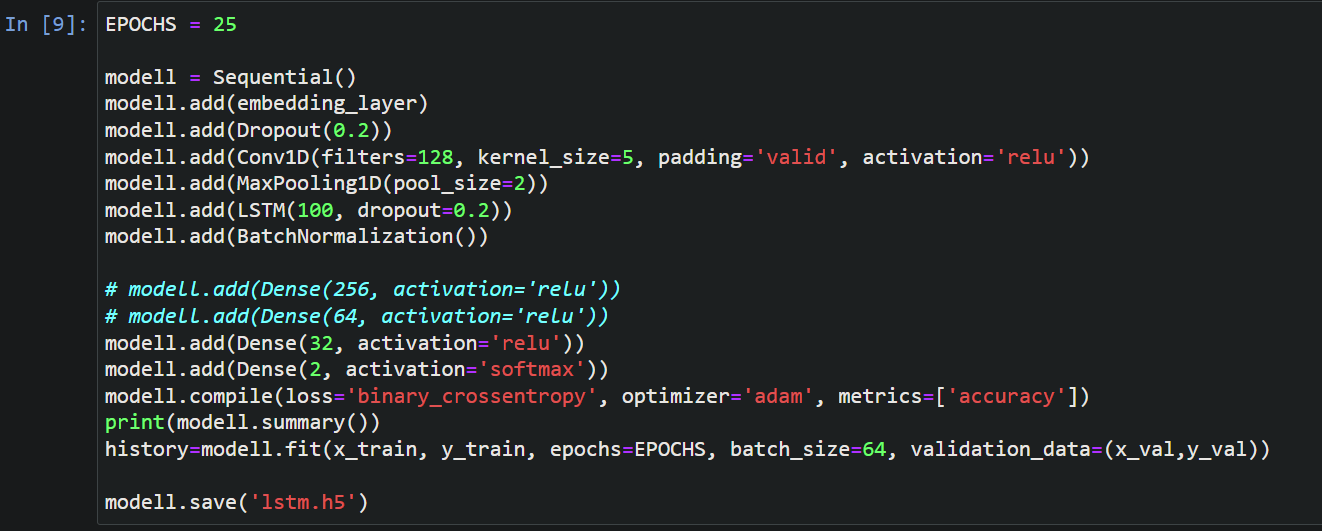
-Train test validation Split:



-Nhúng từ:



-Huấn luyện mô hình CNN &LSTM



Sau khi sử dụng mô hình này em thấy nó các điểm mạnh và điểm yếu như sau.

Điểm mạnh của CNN trong phát hiện tin giả là:

- Xử lý dữ liệu lớn: CNN có thể xử lý số lượng lớn dữ liệu và học được các đặc trưng phức tạp trong dữ liệu, làm cho nó phù hợp cho các ứng dụng phát hiện tin giả trong các tập dữ liệu lớn.

- Tính linh hoạt: CNN có thể được đào tạo trên nhiều loại dữ liệu, chẳng hạn như văn bản, hình ảnh và âm thanh, làm cho nó phù hợp cho phát hiện tin giả trong nhiều loại nội dung trực tuyến.

- Tính chính xác cao: CNN có thể cung cấp kết quả phân loại chính xác và đáng tin cậy với độ chính xác tốt trên các tập dữ liệu đa dạng.

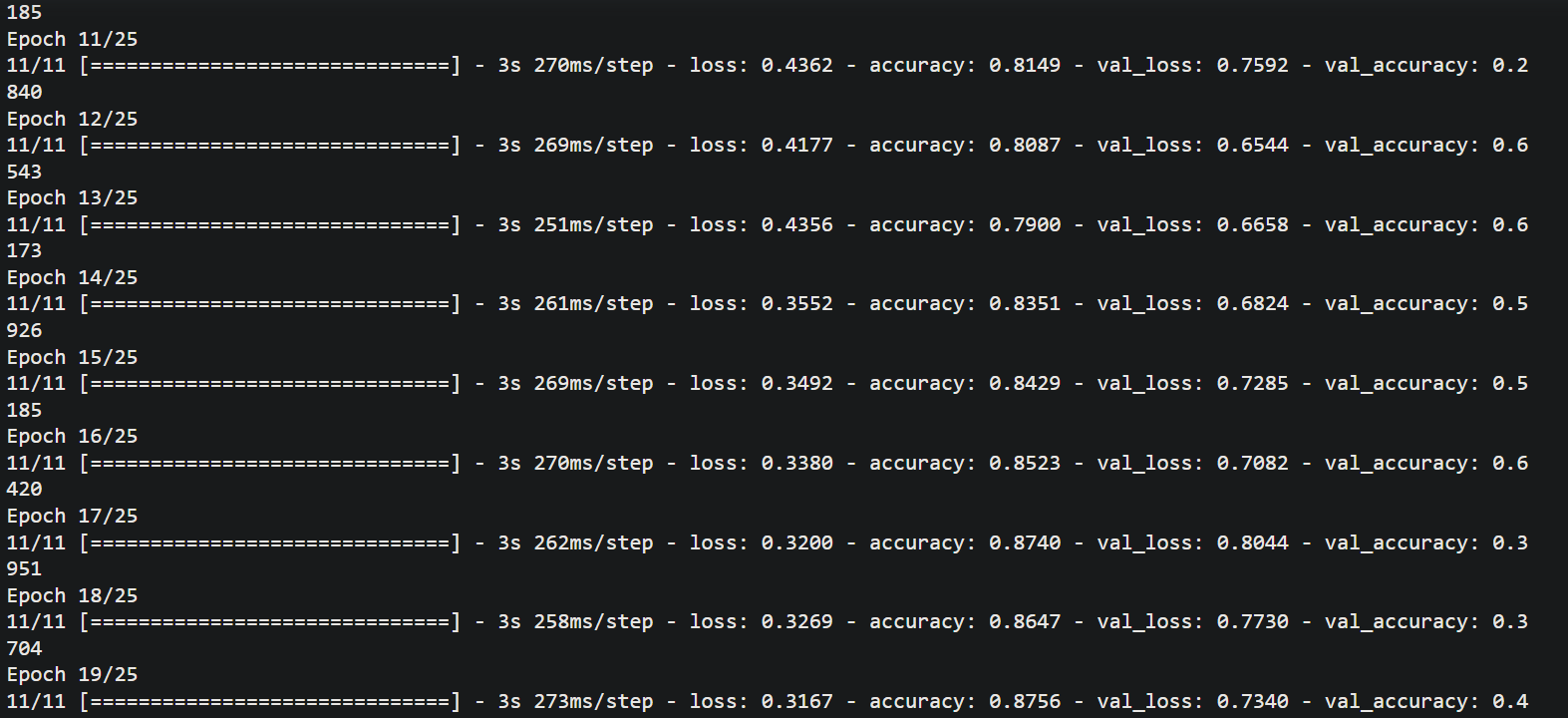
Tuy nhiên, CNN cũng có một số điểm yếu trong phát hiện tin giả, bao gồm:

- Phụ thuộc vào dữ liệu đào tạo: CNN cần được đào tạo trên các tập dữ liệu lớn và đa dạng để đảm bảo tính chính xác và hiệu quả của quá trình phát hiện tin giả. Tuy nhiên, nếu tập dữ liệu đào tạo không đại diện cho các trường hợp khác nhau của tin giả và tin chính thống, mô hình CNN có thể không phát hiện được các trường hợp tin giả mới và khác biệt.

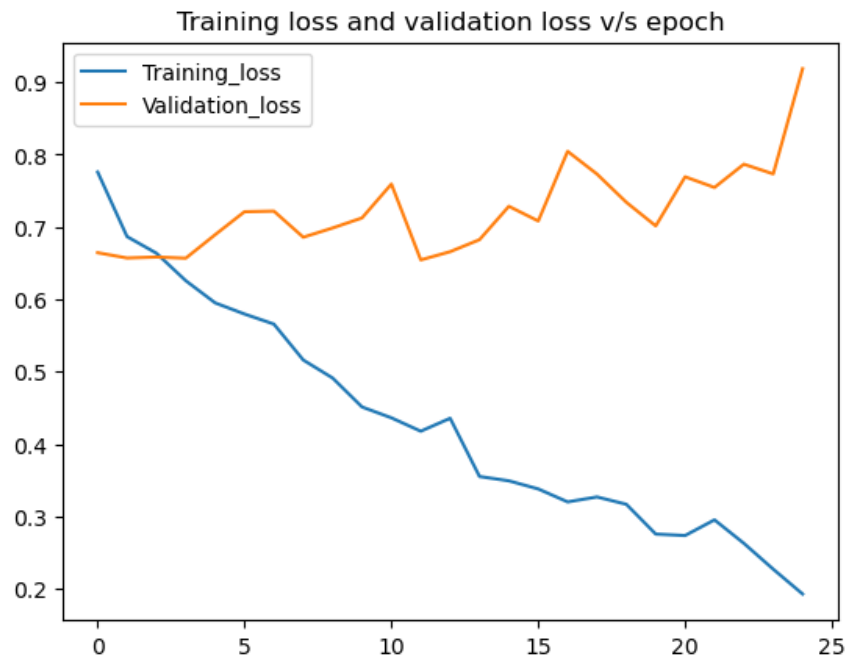
- Không thể xử lý tin giả phức tạp: CNN có thể phát hiện được các đặc trưng cơ bản của tin giả, nhưng nó không thể xử lý được các trường hợp tin giả phức tạp và tinh vi. Để đảm bảo tính chính xác, CNN cần được kết hợp với các kỹ thuật và phương pháp khác để phát hiện các trường hợp tin giả phức tạp hơn.

- Cần nhiều tài nguyên tính toán: CNN cần nhiều tài nguyên tính toán để đào tạo và phân loại dữ liệu, do đó có thể tốn nhiều thời gian và tiền bạc để triển khai trong các ứng dụng thực tế.

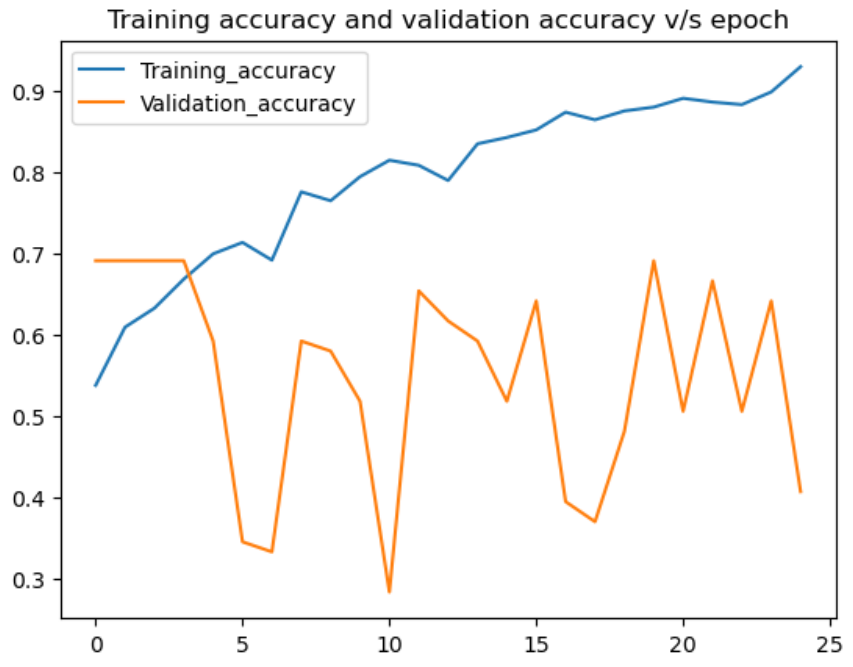
1. **KẾT QUẢ**



***Hình 6.*** *Độ chính xác của mô hình CNN*

**

***Hình 7.*** ***Training loss and validation loss v/s epoch***



***Hình 8.*** ***Training accuracy and validation accuracy v/s epoch***

**Link github demo:** [**https://github.com/kdv4/Fake\_News\_Detection**](https://github.com/kdv4/Fake_News_Detection)

1. **KẾT LUẬN**

CNN là một công nghệ tiên tiến được sử dụng để phát hiện tin giả. Tuy nhiên, nó cần được kết hợp với các kỹ thuật và phương pháp khác để đảm bảo tính chính xác và hiệu quả của quá trình phát hiện tin giả trong các tập dữ liệu đa dạng.

Thông qua quá trình học và nghiên cứu về đề tài này, em đã tiếp thu được nhiều kiến thức về môn như hiểu được một số mô hình xử lý ngôn ngữ tự nhiên. Ngoài ra còn học được thêm cách sử dụng Jupyter, sử dụng hệ điều hành Linux, thêm được 1 số kinh ngiệm viết code và sửa code.

**TÀI LIỆU THAM KHẢO**

[1] <https://www.kaggle.com/datasets/mohamadalhasan/a-fake-news-dataset-around-the-syrian-war>

[2] <https://www.kaggle.com/c/fake-news/data>

[3] <https://www.kaggle.com/code/sahrul/fake-news-detection-using-cnn>

[4] <https://www.mdpi.com/2078-2489/13/3/151>

[5]<https://www.researchgate.net/publication/348580170_Identifying_Fake_News_on_Social_Networks_Based_on_Natural_Language_Processing_Trends_and_Challengs>