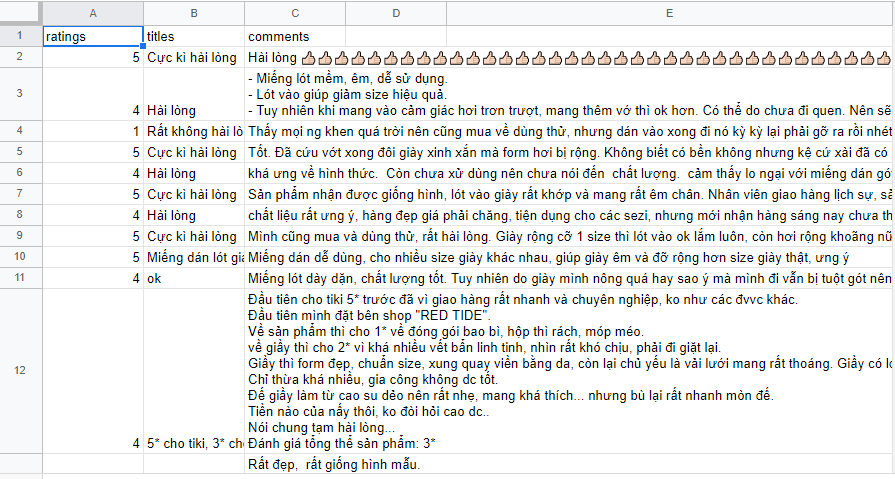
*Nguyễn Trường Sơn*

**1. Xử lí dữ liệu**

Dữ liệu được thu thập bằng cách clone từ các web thương mại điện tử như Tiki, Lazada,… hay tham khảo từ các nguồn như ngân hàng và chứng khoán.

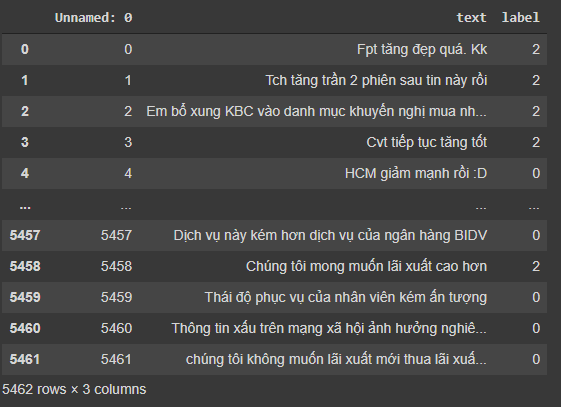


Hình 1: Dữ liệu thu được từ Tiki.

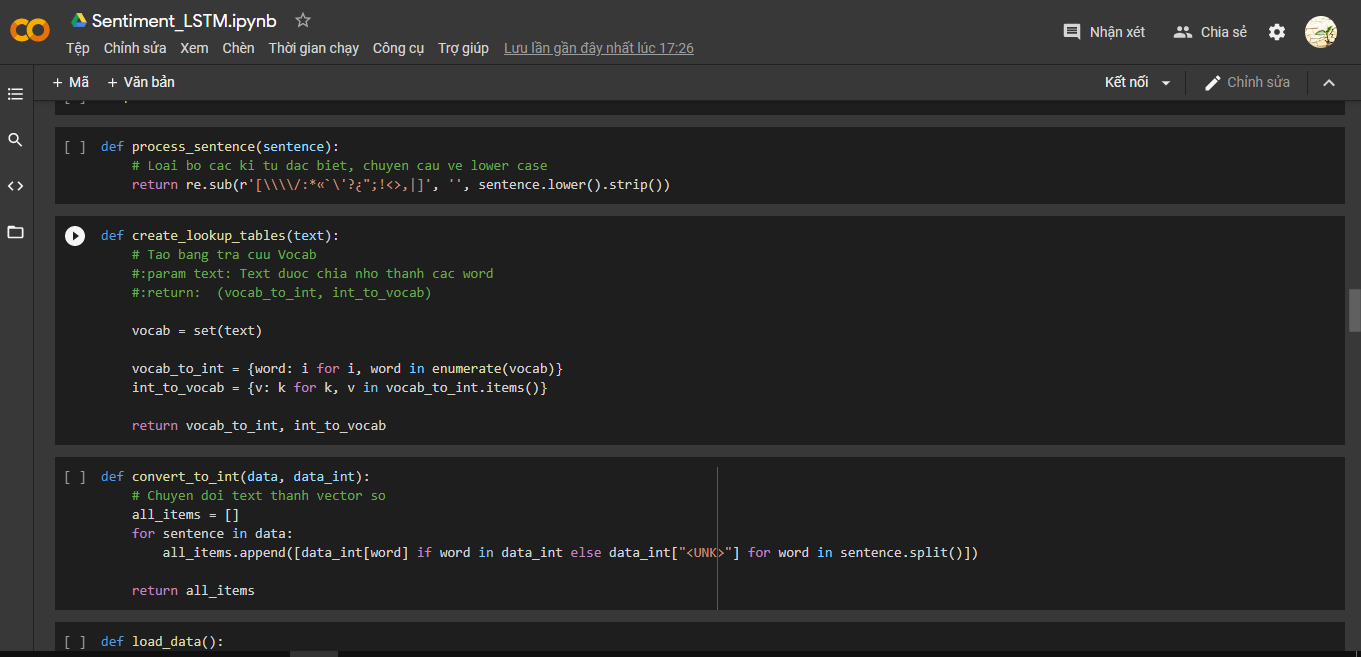
Sau khi thu thập được dữ liệu, dữ liệu sẽ qua xử lí bằng nhiều cách như: loại bỏ các dữ liệu lỗi, outlier, xử lí imbalanced.



Kết quả thu được như sau:



Tiếp theo là sẽ xử lí để phù hợp với dữ liệu input của model. Bằng cách loại bỏ các kí tự đặc biệt, chuyển các câu về dạng Vectorize.



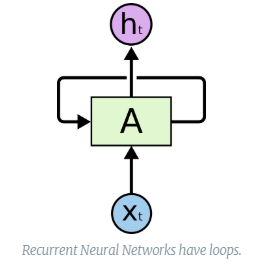
**2. LSTM**

*2.1. RNN*

Con người không bắt đầu suy nghĩ của họ từ đầu tại tất cả các thời điểm. Cũng như bạn đang đọc bài viết này, bạn hiểu mỗi chữ ở đây dựa vào từ bạn đã hiểu các chữ trước đó chứ không phải là đọc tới đâu ném hết đi tới đó, rồi lại bắt đầu suy nghĩ lại từ đầu tới chữ bạn đang đọc. Tức là tư duy đã có một bộ nhớ để lưu lại những gì diễn ra trước đó.

Tuy nhiên các mô hình mạng nơ-ron truyền thống thì không thể làm được việc đó, đó có thể coi là một khuyết điểm chính của mạng nơ-ron truyền thống. Ví dụ, bạn muốn phân loại các bối cảnh xảy ra ở tất cả các thời điểm trong một bộ phim, thì đúng là không rõ làm thế nào để có thể hiểu được một tình huống trong phim mà lại phụ thuộc vào các tình huống trước đó nếusử dụng các mạng nơ-ron truyền thống.

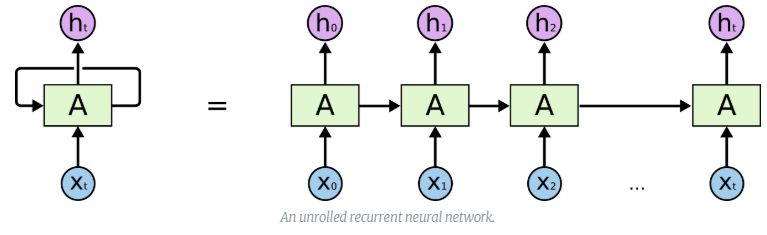
Mạng nơ-ron hồi quy (Recurrent Neural Network) sinh ra để giải quyết vấn đề đó. Mạng này chứa các vòng lặp bên trong cho phép thông tin có thể lưu lại được.



Hình 2: Mạng RNN.

Hình vẽ trên mô tả một đoạn của mạng nơ-ron hồi quy *A* với đầu vào là *xt*​ và đầu ra là *ht*​. Một vòng lặp cho phép thông tin có thể được truyền từ bước này qua bước này qua bước khác của mạng nơ-ron.

Các vòng lặp này khiến cho mạng nơ-ron hồi quy trông có vẻ khó hiểu. Tuy nhiên, nếu bạn để ý một chút thì nó không khác mấy so với các mạng nơ-ron thuần. Một mạng nơ-ron hồi quy có thể được coi là nhiều bản sao chép của cùng một mạng, trong đó mỗi đầu ra của mạng này là đầu vào của một mạng sao chép khác.



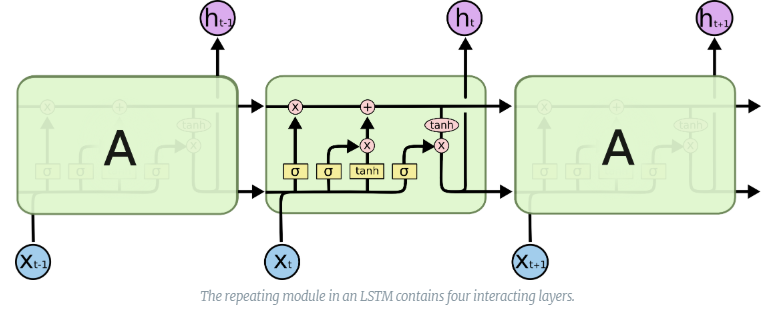
Trong vài năm gần đây, việc ứng dụng RNN đã đưa ra được nhiều kết quả không thể tin nổi trong nhiều lĩnh vực: nhận dạng giọng nói, mô hình hóa ngôn ngữ, dịch máy, mô tả ảnh,…

Đằng sau sự thành công này chính là sự đóng góp của LSTM. LSTM là một dạng đặc biệt của mạng nơ-ron hồi quy, với nhiều bài toán thì nó tốt hơn mạng hồi quy thuần. Hầu hết các kết quả thú vị thu được từ mạng RNN là được sử dụng với LSTM.

*2.2. Mạng LSTM*

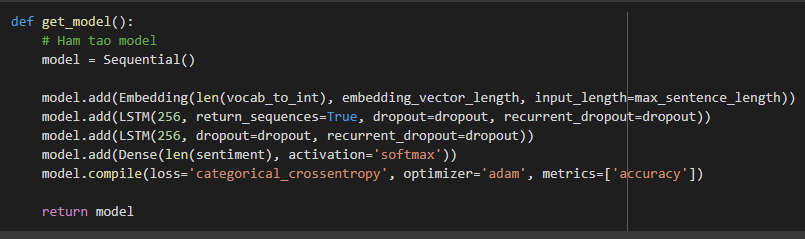
*Mạng* bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi [Hochreiter & Schmidhuber (1997)](http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

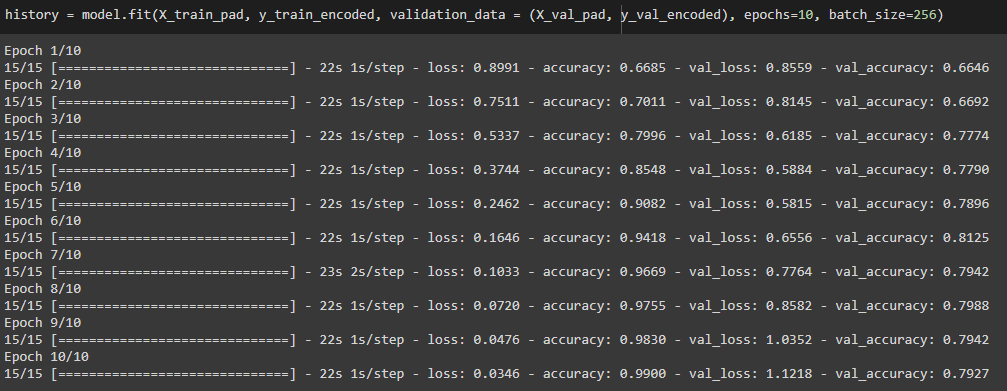


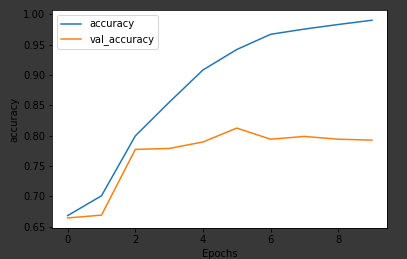
*2.3. Implement và tiến hành train.*

Giờ chúng ta sẽ dung model LSTM để tiến hành phân tích cảm xúc qua văn bản( Sentiment Analysis).



Bây giờ, chúng ta sẽ tiến hành train. Sau 10 epoch, kết quả thu được như sau:





Hình 4: Kết quả training.

***Nhận xét***:

Model đã bị overfiting. Model đã học rất tốt trên tập train, nhưng lại có performance không tốt trên tập test.