ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

Cách tiếp cận hiện đại trong Xử lý ngôn ngữ tự nhiên (055256)

Sentiment Analysis
Áp dụng Học sâu để Phân loại cảm xúc

Giảng viên hướng dẫn: PGS. TS. Quản Thành Thơ

SV thực hiện: Trần Hoàng Duy 1912924

TP. Hồ Chí Minh, Tháng 11 Năm 2022

Mục lục

Danh sách hình vẽ

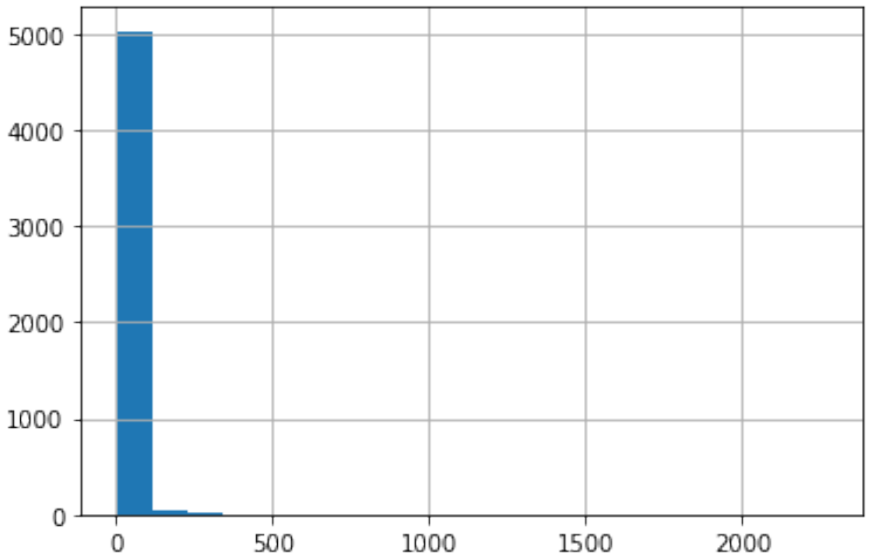
Danh sách bảng

1 Bộ dữ liệu VLSP, tiền xử lý dữ liệu và word embedding

1.1 Bộ dữ liệu VLSP

Một số đặc điểm của bộ dữ liệu VLSP (train):

- Có chiều dài trung bình là 29.35 từ
- Có chiều dài tối đa là 2884 từ
- Có nhiều lỗi chính tả khi nhắc đến tên các hãng sản phẩm công nghệ
- Có nhiều từ lóng, viết tắt
- Có xuất hiện một số emoji và link website



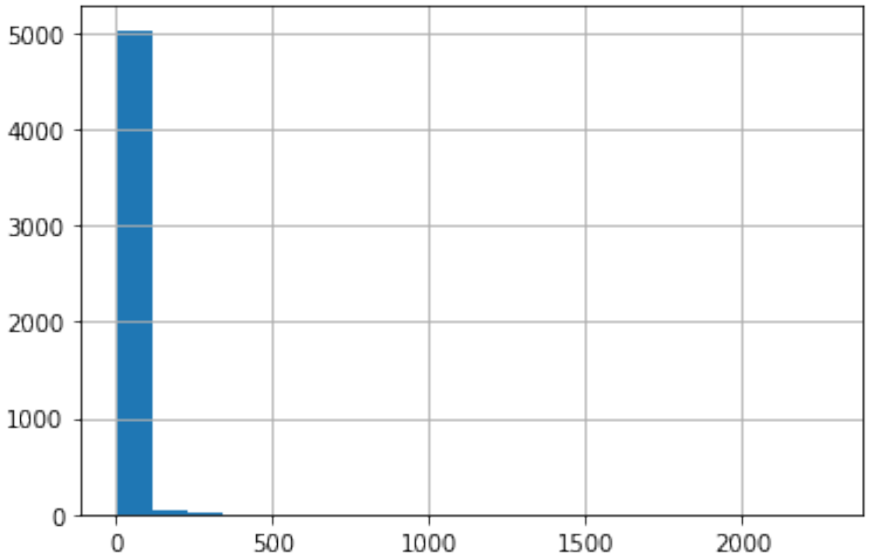
Hình 1: Trực quan hóa về chiều dài của bộ dữ liệu vlsp

1.2 Tiền xử lý dữ liệu VLSP

Để giải quyết các vấn đề đã nêu ở trên, sau đây là các bước tiền xử lý dữ liệu:

- Chuẩn hóa unicode, đưa về dạng chữ viết thường (lowercase)
- Loại bỏ số, các dấu câu (punctuation)
- Loại bỏ emoji, các URL
- Thay thế các tử sai chính tả bằng phiên bản đúng của nó (ví dụ: "android" thay cho "androi", "iphone" thay cho "ifone")

- Thay thế các từ lóng, viết tắt bằng từ đúng của nó ("bình thường" thay cho "bth", "ngon" thay cho "ngol"



Hình 2: Trực quan hóa về chiều dài của bộ dữ liệu vlsp sau tiền xử lý

Kết quả sau khi tiền xử lý, ta thu được bộ dữ liệu có phân bố như sau. Ta sử dụng ViTokenizer của thư viện pyvi (một thư viện nhẹ, nhanh cho tác vụ xử lý ngôn ngữ tiếng Viêt) cho nhiệm vụ tokenize các câu trong bộ dữ liệu thành các token tiếng Việt. Từ điển của tập dữ liệu sau xử lý có tổng cộng 7717 từ.
Tiếp đó ta chuyển từ dạng chuỗi các token thành một mảng các giá trị số, sau đó được padding cho phù hợp với tập dữ liệu. Vì trong các phần sắp tới sẽ có sử dụng LSTM, nên ta lựa chọn cách padding là pre.

1.3 Word Embedding

Dựa vào phân bố chiều dài của tập train sau khi được tiền xử lý, ta có thể thấy đa số dữ liệu nằm ở khoảng dưới 100 từ. Vì vậy ta chọn số MAX_SEQUENCE_LENGTH là 150, vừa phù hợp với đa số trường hợp, vừa tăng tốc cho quá trình huấn luyện.
Trong giai đoạn này, ta sử dụng Word2Vec và một pretrained CBOW để tạo một embedding matrix để truyền vào tầng Embedding.

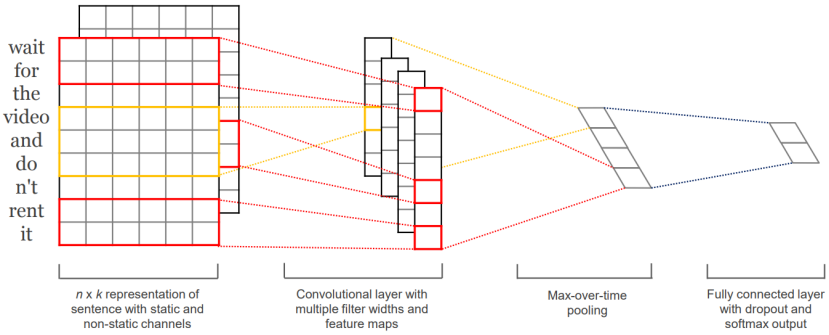
2 Giải quyết bài toán Sentiment Analysis bằng Học sâu

Trong phần sau, ta sẽ áp dụng các kỹ thuật học sâu như CNN, LSTM để giải quyết bài toán

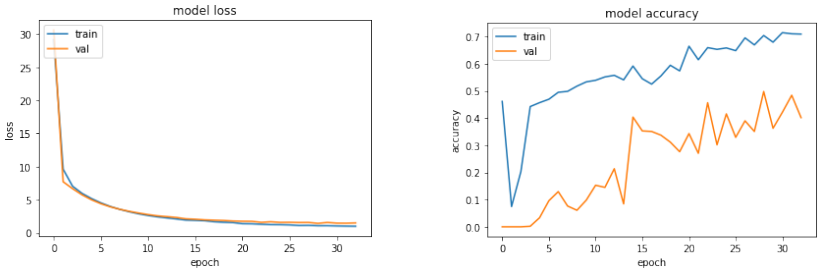
2.1 CNN: Convolutional Neural Network

2.1.1 Mô hình 1: CNN với Convolutional 1D

Mô hình được xây dựng dựa theo mô hình được miêu tả trong bài báo "Convolutional Neural Networks for Sentence Classification". Trong đó, các tầng Convolutional 1D được sử dụng với các filter size khác nhau như trong n-gram. Sau đó kết quả đi qua tầng Max-Pooling 1D và được nối với nhau thông qua tầng Concatenate. Kết quả này được làm phẳng và đi qua tầng Dense (với hàm kích hoạt là softmax) để có được đầu ra là 1 trong 3 trạng thái cảm xúc của câu.



Hình 3: Kiến trúc mô hình phân loại văn bản dựa trên CNN của Yoon Kim



(a) Biểu đồ loss (b) Biểu đồ accuracy

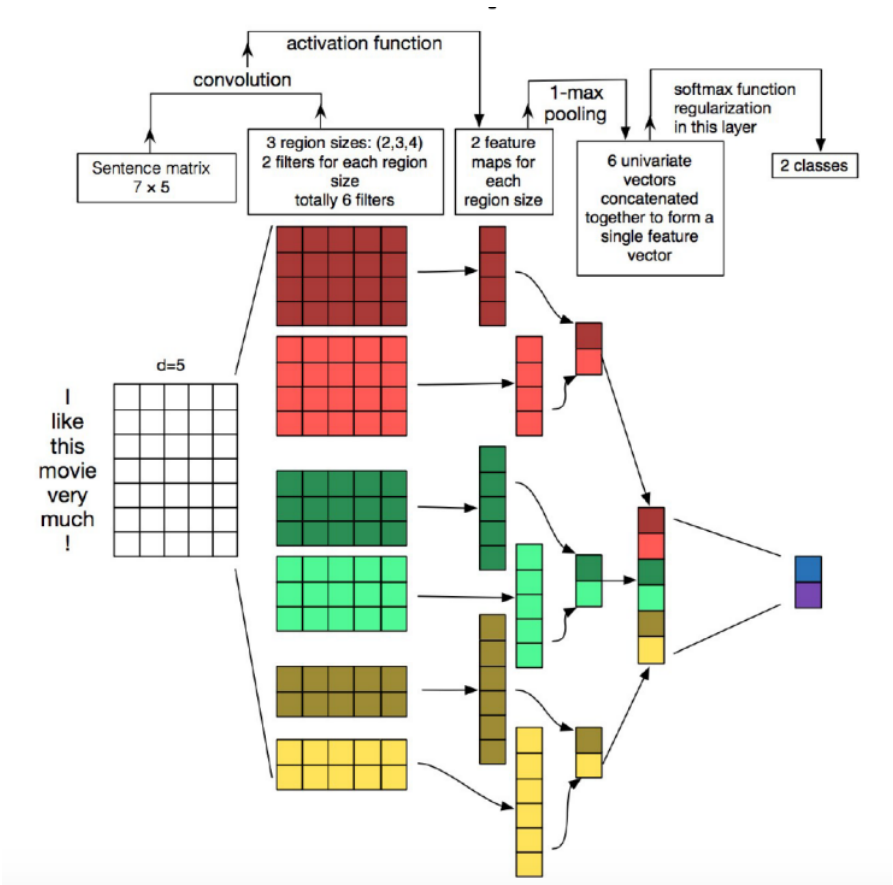
Hình 4: Biểu đồ theo dõi quá trình huấn luyện mô hình 1

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
CNN với Convolution1D	1.387	69.9	70.79	68.1

Bảng 1: Đánh giá mô hình 1 trên tập test

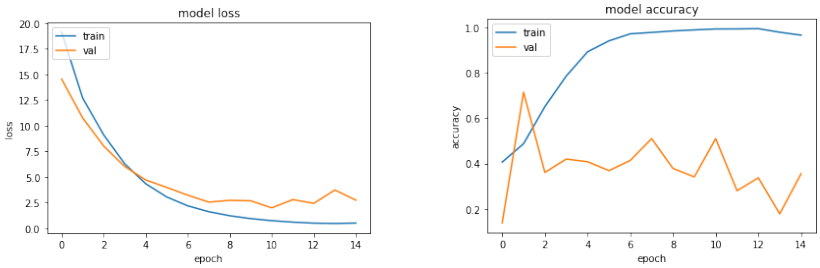
2.1.2 Mô hình 2: CNN với nhiều filter Convolutional1D

Mô hình dưới đây được xây dựng tương tự với mô hình 1, nhưng với mỗi kích thước của filter, ta có 2 tầng Convolutional sử dụng số K đó



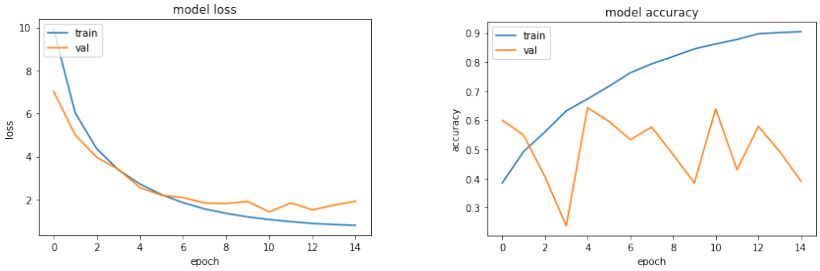
Hình 5: Mô hình phân loại văn bản dựa trên CNN với nhiều filter

Trong quá trình thử nghiệm, MaxPooling1D và AveragePooling1D đã được so sánh. Theo đó, kết quả của mô hình 2.1 (6 tầng MaxPooling) và mô hình 2.2 (3 tầng MaxPooling, 3 tầng AveragePooling) có kết quả cao.



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 6: Biểu đồ theo dõi quá trình huấn luyện mô hình 2.1



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 7: Biểu đồ theo dõi quá trình huấn luyện mô hình 2.2

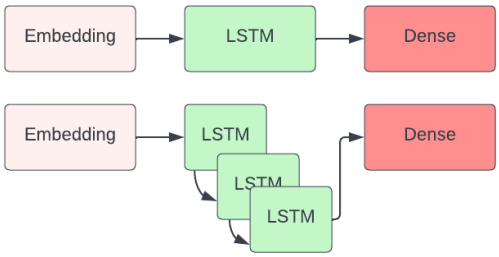
Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
CNN với Convolution1D 6 MaxPool	1.3	66.29	68.29	64.24
CNN với Convolution1D 3 MaxPool, 3 AveragePool	1.257	70.1	71.22	68.1

Bảng 2: Đánh giá mô hình 2.1 và 2.2 trên tập test

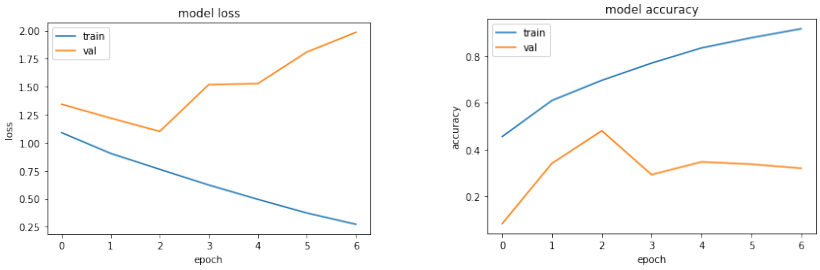
2.2 LSTM: Long Short Term Memory

2.2.1 Mô hình 3: LSTM đơn giản

Trong phần này, 2 mô hình là 3.1 và 3.2 lần lượt sử dụng 1 và 3 tầng LSTM cho việc phân lợp.

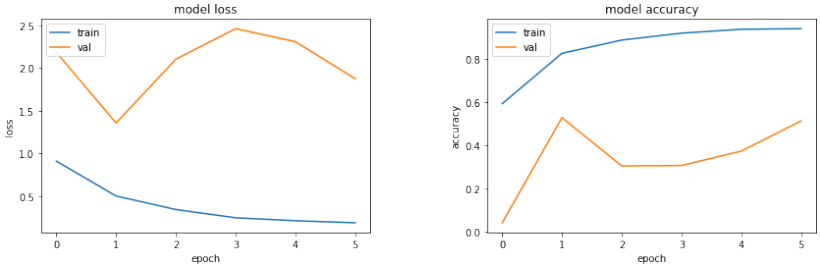


Hình 8: Kiến trúc mô hình phân loại văn bản dựa trên CNN của Yoon Kim



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 9: Biểu đồ theo dõi quá trình huấn luyện mô hình 2.1



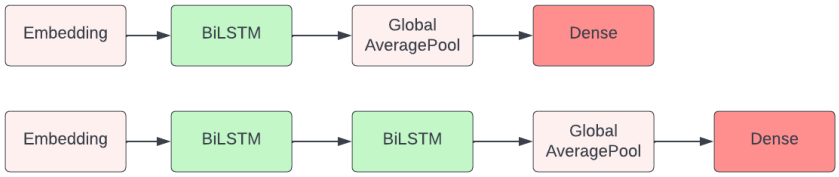
(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 10: Biểu đồ theo dõi quá trình huấn luyện mô hình 2.2

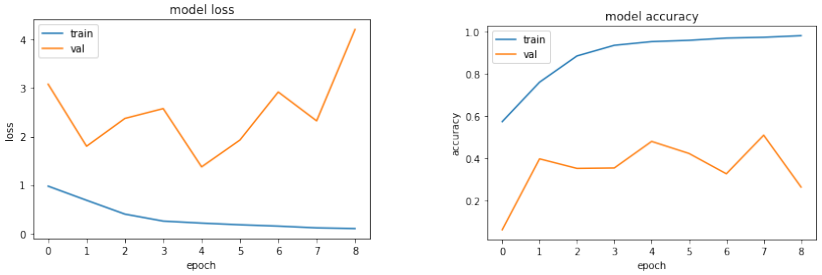
Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
LSTM với 1 tầng	1.31	66.95	67.38	66.84
LSTM với 3 tầng	0.94	68.19	69.04	66.48

Bảng 3: Đánh giá mô hình 3.1 và 3.2 trên tập test

2.2.2 Mô hình 4: Bidirectional LSTM

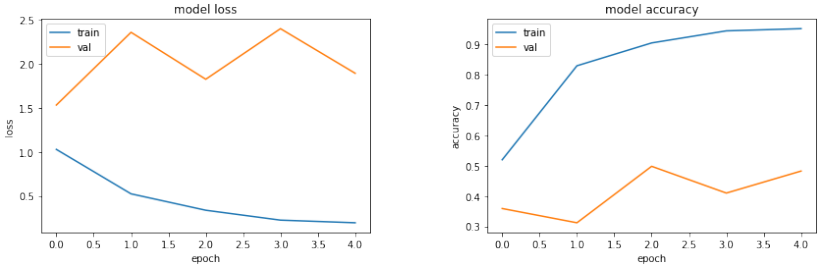


Hình 11: Kiến trúc mô hình phân loại văn bản dựa trên CNN của Yoon Kim



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 12: Biểu đồ theo dõi quá trình huấn luyện mô hình 4.1



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 13: Biểu đồ theo dõi quá trình huấn luyện mô hình 4.2

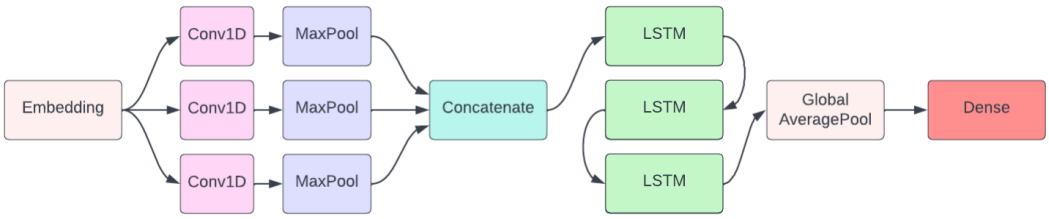
Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
1 tầng BiLSTM, AveragePool	1.05	67.71	68.62	66.86
2 tầng BiLSTM, AveragePool	0.98	68.29	68.74	66.19

Bảng 4: Đánh giá mô hình 4.1 và 4.2 trên tập test

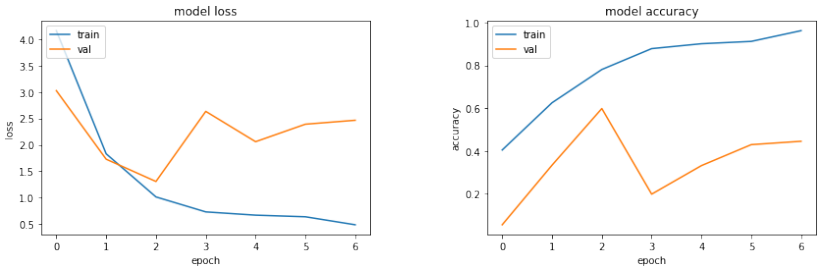
2.3 Kết hợp CNN và LSTM

2.3.1 Mô hình 5: CRNN

Mô hình sau đây được xây dựng dựa trên sự kết hợp của cấu trúc nhiều tầng Conv1D như mô hình 1, 2 và cấu trúc 3 tầng LSTM như của mô hình 4.

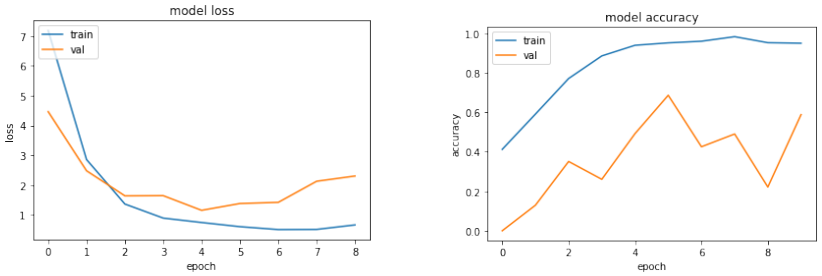


Hình 14: Kiến trúc mô hình CRNN



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 15: Biểu đồ theo dõi quá trình huấn luyện mô hình 5.1



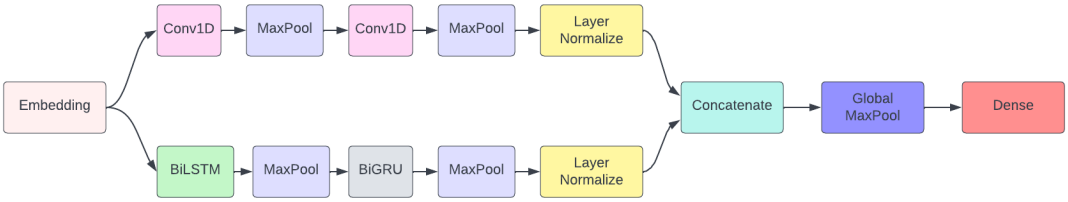
(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 16: Biểu đồ theo dõi quá trình huấn luyện mô hình 5.2

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
CRNN CNN đơn filter, 3 tầng LSTM	1.57	69.52	69.78	68.38
CRNN CNN nhiều filter, 3 tầng LSTM	1.46	69.71	70.01	68.95

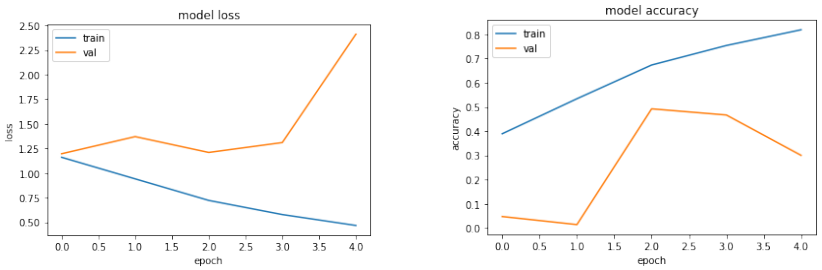
Bảng 5: Đánh giá mô hình 5.1 và 5.2 trên tập test

2.3.2 Mô hình 6: CNN kết hợp BiLSTM



Hình 17: Kiến trúc mô hình kết hợp CNN và BiLSTM

Trong mô hình này, đầu vào sẽ cùng truyền cho CNN và LSTM, sau khi đi qua MaxPool và LayerNormalization, hai kết quả sẽ được nối với nhau và truyền cho các lớp Dense với nhiệm vụ phân lớp



(a) Biểu đồ loss (b) Biểu đồ accuracy

Hình 18: Biểu đồ theo dõi quá trình huấn luyện mô hình 6

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
CNN kết hợp BiLSTM	1.15	69.05	70.43	67.14

Bảng 6: Đánh giá mô hình 6 trên tập test

2.4 Tổng hợp quá trình thử nghiệm

Trong quá trình huấn luyện, hàm mất mát được sử dụng là categorical_crossentropy, optimizer là Adam (với hệ số học là 0.001), tiêu chí đánh giá lần lượt là accuracy, precision và recall.
Mô hình thử nghiệm có callback là ModelCheckpoint với ưu tiên lưu model khi có validation_loss thấp nhất.
Thử nghiệm diễn ra với nhiều cách tiền xử lý dữ liệu khác nhau, các model khác nhau cùng bộ siêu tham số khác nhau. Các model dưới đây là những model cho ra kết quả tốt nhất của mỗi loại.

Model	Loss	Accuracy (%)	Precision (%)	Recall (%)
CNN với Convolution1D	1.387	69.9	70.79	68.1
CNN với Convolution1D 6 MaxPool	1.3	66.29	68.29	64.24
CNN với Convolution1D 3 MaxPool, 3 AveragePool	1.257	70.1	**71.22**	68.1
CNN với Convolution1D 6 MaxPool	1.3	66.29	68.29	64.24
CNN với Convolution1D 3 MaxPool, 3 AveragePool	1.257	**70.1**	**71.22**	68.1
LSTM với 1 tầng	1.31	66.95	67.38	66.84
LSTM với 3 tầng	**0.94**	68.19	69.04	66.48
1 tầng BiLSTM, AveragePool	1.05	67.71	68.62	66.86
2 tầng BiLSTM, AveragePool	0.98	68.29	68.74	66.19
CRNN CNN đơn filter, 3 tầng LSTM	1.57	69.52	69.78	68.38
CRNN CNN nhiều filter, 3 tầng LSTM	1.46	69.71	70.01	**68.95**
CNN kết hợp BiLSTM	1.15	69.05	70.43	67.14

Bảng 7: Tổng hợp các mô hình có kết quả tốt

Danh mục tham khảo

[1] Anand Saran, *Text Classification — CNN with LSTM*, truy cập từ https://anandsarank.medium.com/cnn-with-lstm-for-text-classification-53d18e5f7f5c.

[2] Fushen Yang (2019), *Ensemble sentiment analysis method based on R-CNN and C-RNN with fusion gate*.

[3] Mohammad Sadegh Rasooli (2018), *Cross-lingual sentiment transfer with limited resources*.

[4] Samarth Agrawal, *Sentiment Analysis using LSTM*, truy cập từ https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948.

[5] Quản Thành Thơ, *Slide giáo trình Cách tiếp cận hiện đại trong xử lý ngôn ngữ tự nhiên*.