TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ**

**MÔN MÔ HÌNH ĐỒ THỊ CÓ XÁC SUẤT**

**PHÂN TÍCH Ý KIẾN NGƯỜI DÙNG**

**BẰNG PHƯƠNG PHÁP HỌC SÂU**

**KẾT HỢP CNN-LSTM**

**(VIETNAMESE SENTIMENT ANALYSIS USING DEEP LEARNING)**

*Người hướng dẫn*: **TS NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **TRƯƠNG TRỌNG VINH - 186005039**

**LÊ ANH KHOA – 186005031**

Lớp **: 10050301**

Khoá  **: 12/2018**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019**

LỜI CẢM ƠN

Tôi xin gửi lời cám ơn chân thành đến thầy Nguyễn Chí Thiện đã hướng dẫn, định hướng cho tôi hoàn thành đề tài. Đồng thời cũng không quên bày tỏ sự biết ơn đến cán bộ giảng viên trường đại học Tôn Đức Thắng đã trang bị cho tôi kiến thức nền tảng cũng như những thắc trong quá trình học tập và nghiên cứu. Cuối cùng xin cảm ơn bạn bè, gia đình và người thân đã ở bên động viên và cùng nhau cố gắng trong suốt quá trình nghiên cứu.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trương Trọng Vinh*

*Lê Anh Khoa*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Trong bài báo lần này chúng tôi sẽ trình bày phương pháp hiện thực giải quyết bài toán phân tích ý kiến người dùng (sentiment analysis) bằng cách áp dụng phương pháp học sâu sữ dụng hybrid model CNN-LSTM và Word Embedding Fasttext.

MỤC LỤC

[**LỜI CẢM ƠN i**](#_Toc26720962)

[**PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii**](#_Toc26720963)

[**TÓM TẮT iv**](#_Toc26720964)

[**MỤC LỤC 1**](#_Toc26720965)

[**DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4**](#_Toc26720966)

[**CHƯƠNG 1 – PHÁT BIỂU BÀI TOÁN 5**](#_Toc26720967)

[1.1 Đặt vấn đề 5](#_Toc26720968)

[1.2 Cách tiếp cận 6](#_Toc26720969)

[1.3. Ý nghĩa khoa học và thực tiễn của bài toán 7](#_Toc26720970)

[**CHƯƠNG 2 –GIẢI QUYẾT BÀI TOÁN 8**](#_Toc26720971)

[2.1 Tổng quát về phương pháp giải quyết bài toán 8](#_Toc26720972)

[2.1.1 Tiễn xử lý 8](#_Toc26720973)

[2.1.2 Mô hình hóa 8](#_Toc26720974)

[2.1.2.1 Word Presentation 9](#_Toc26720975)

[2.2 Đặc trưng của giải pháp đề xuất 12](#_Toc26720976)

[2.3 Cách đánh giá 12](#_Toc26720977)

[**CHƯƠNG 3 - THỰC NGHIỆM 14**](#_Toc26720978)

[3.1. Dữ liệu 14](#_Toc26720979)

[3.2. Công nghệ 14](#_Toc26720980)

[3.2.1. Ngôn ngữ lập trình 14](#_Toc26720981)

[3.2.2. Thư viện 14](#_Toc26720982)

[3.3. Thực nghiệm 14](#_Toc26720983)

[3.4. Đánh giá kết quả 15](#_Toc26720984)

[**CHƯƠNG 4 - KẾT LUẬN 17**](#_Toc26720985)

[4.1 Kết quả đạt được 17](#_Toc26720986)

[4.2. Hướng phát triển 17](#_Toc26720987)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC CHỮ VIẾT TẮT**

CNN Convolutional Neural Network

LSTM Long Short Term Memory

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1: Sơ lượt mô hình 9](#_Toc26720944)

[Hình 2: Kiến trúc mô hình 11](#_Toc26720945)

[Hình 3: Model F1 score 15](#_Toc26720946)

[Hình 4: Model loss value 15](#_Toc26720947)

**DANH MỤC BẢNG**

[Bảng 1: So sánh phân tích ý kiến người dùng giữa phương pháp truyền thống và phương pháp sử dụng machine learning 6](#_Toc26718587)

[Bảng 2: Bảng so sánh Phase B (Aspect-Polarity) F1 score của domain restaurant 16](#_Toc26718588)

CHƯƠNG 1 – PHÁT BIỂU BÀI TOÁN

1.1 Đặt vấn đề

Với sự phát triển của thương mại điện tử việc đánh giá phản hồi của người dùng

là một yêu cầu bắt buộc đối với các nhà cung cấp. Tuy nhiên nếu hàng ngày có hang triệu đánh giá về sản phẩm thì việc phân tích đánh giá là bất khả thi, nên đây là lúc các mô hình học máy thể hiện được sức mạnh của mình.

Và đó là một trong những lý do người ta muốn hiện thực hóa bài toán phân tích sắc thái của bình luận.

Đó không phải là một bài toán mới, tuy nhiên hành vi của con người là khó đoán và ngẫu nhiên chứ không có một quy luật nào cả, đồng thời các vấn đề khác như sự nhập nhằng trong ngôn ngữ khiến cho việc dạy cho máy tính nắm bắt và dự đoán chính xác cực kỳ khó khăn.

Việc sử dụng các mô hình học sâu khiến cho quá trình training và re-training dễ dàng hơn, tuy nhiên vấn đề gặp phải chính là dữ liệu training phải thu thập từ đầu. Quá trình phân loại và gán nhãn cho dữ liệu train cũng làm bằng cách thủ công nên sẽ tiêu tốn rất nhiều thời gian để có được một tập data training hoàn chỉnh.

|  |  |  |
| --- | --- | --- |
|  | Con người | Machine Learning |
| Ưu điểm | Độ chính xác cao, có thể phân loại được cả những thảo luận mang ý mỉa mai, từ lóng, nghĩa bóng,… | Hoàn toàn tự động cho phép xử lý toàn bộ các bình luận một cách nhanh chóng, đồng thời tiết kiệm thời gian và chi phí. |
| Nhược điểm | Đòi hỏi nguồn nhân lực lớn và mất nhiều thời gian.  Khả năng con người có giới hạn nên không thể xử lý quá nhiều bình luận trong một ngày | Ngữ cảnh: một từ có thể mang ý nghĩa tích cực hoặc tiêu cực trong từng hoàn cảnh khác nhau.  Sự nhập nhằng: một câu dù chứa từ khóa tích cực hoặc tiêu cực thì cũng chưa hẳn thể hiện sắc thái của câu. Ngược lại, một câu không chứa từ khóa tích cực hoặc tiêu cực nào lại có thể mang một sắc thái nhất định.  Sự mỉa mai: không thể phân biệt được đâu là từ mang ý nghĩa thông dụng và từ mang ý nghĩa mỉa mai.  Từ lóng, từ địa phương, từ viết tắt, ngôn ngữ chat,…: Với sự phong phú của ngôn ngữ nói chung, và tiếng Việt nói riêng. |

Bảng 1: So sánh phân tích ý kiến người dùng giữa phương pháp truyền thống và phương pháp sử dụng machine learning

1.2 Cách tiếp cận

Bao gồm 3 giai đoạn:

* Thu thập dữ liệu từ các nguồn trên mạng xã hội hoặc các website. Trong bài báo cáo lần này chúng tôi sử dụng bộ dữ liệu Vietnamese Language and Speech Processing - VLSP 2018.
* Tiền xử lý và mô hình hóa dữ liệu.
* Đánh giá mô hình và cải thiện kết quả.

1.3. Ý nghĩa khoa học và thực tiễn của bài toán

Phân loại cảm xúc hay thái độ của người dùng qua bình luận (comment) trên các trang phim, mv ca nhạc, đánh giá về sản phẩm.

Ứng dụng chatbot.

Ứng dụng các mô hình học sâu để giải quyết bài toán Aspect Base Sentiment Analysis cho tập dữ liệu VLSP 2018.

CHƯƠNG 2 –GIẢI QUYẾT BÀI TOÁN

2.1 Tổng quát về phương pháp giải quyết bài toán

2.1.1 Tiễn xử lý

Giảm kích thước từ vựng trong dữ liệu training:

Chúng tôi nhận thấy trong dữ liệu training có rất nhiều dạng biểu diễn cho đơn vị tiền tệ. Vd: 100k, 2m,.. nên chúng tôi quyết định sẽ thay thế chúng bằng "giá".

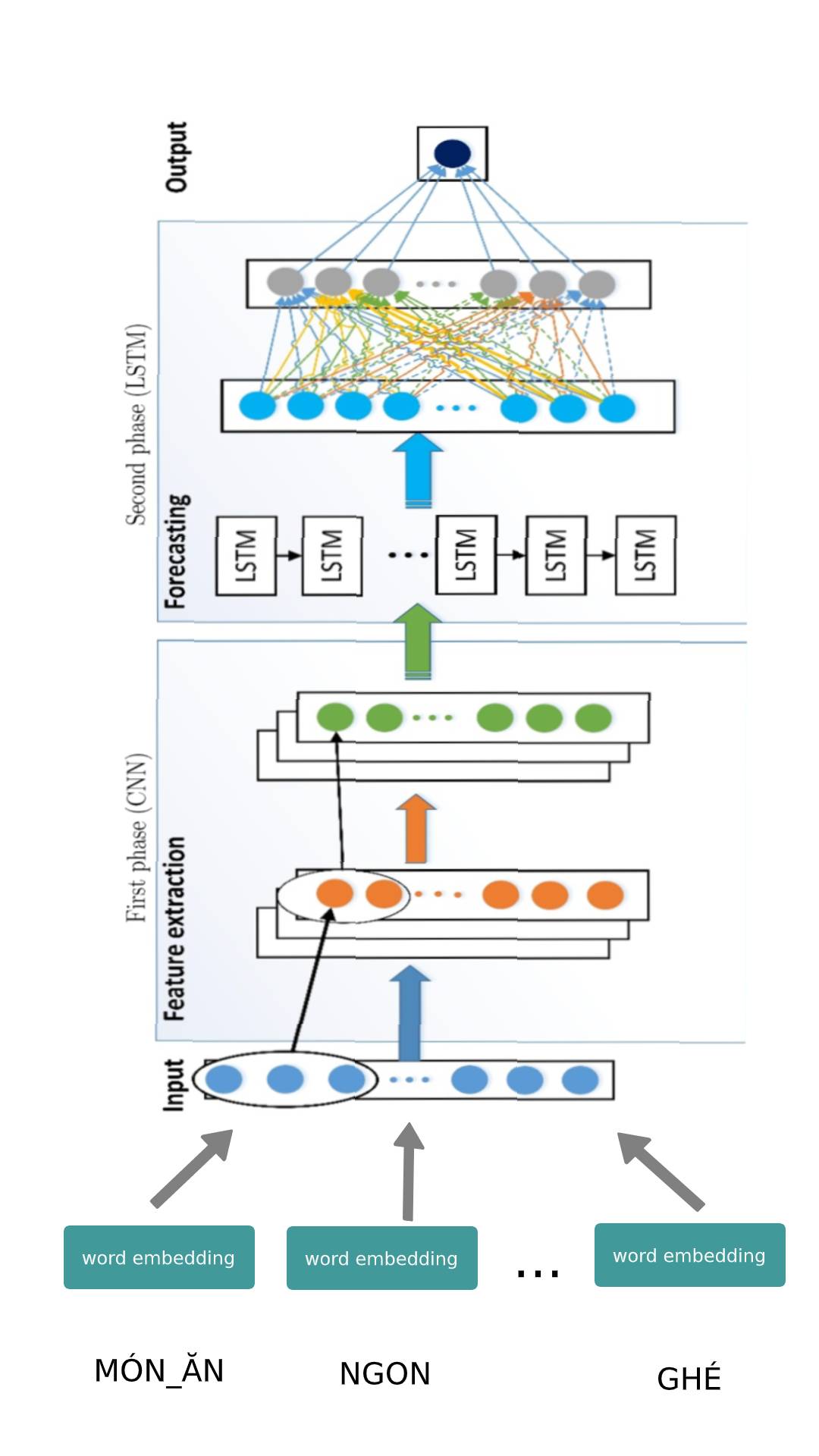
Những kí tự hashtag #abc sẽ được thấy thế bằng "tag"

Chuẩn hóa các đường link website bằng "website".

Do đặc thù bộ dữ liệu là những comment được crawl từ trên internet nên sẽ là văn nói và bị sai chính tả nên ta cần chuẩn hóa chúng. Một số phương pháp chúng tôi sử dụng: chuẩn hóa emoji, chuẩn hóa những từ lập vd: "thíchhhhh" -> "thích",thêm dấu chấm câu vào giữa hai ký tự kề nhau nếu ký tự phía sau in hoa, xóa những ký tự số, sửa lỗi chính tả cho một số từ thông dụng, vd: "k", "ko" -> "không".

Sau khi đã chuẩn hóa dữ liệu chúng tôi sử dụng thư viện pyvi để tokenize câu thành các từ, cuối cùng là sử dụng bộ từ điển loại bỏ đi các stopword.

2.1.2 Mô hình hóa



Hình 1: Sơ lượt mô hình

2.1.2.1 Word Presentation

Chúng tôi sử dụng Word Embedding thay cho One-Hot vector hay tf-idf để biểu diễn dữ liệu đầu vào.

Vì dữ liệu đầu vào là văn nói nên sẽ có nhiều từ sai chính tả hoặc viết tắt, nên chúng tôi sử dụng Fasttext là một phần mở rộng của Word2Vec để biểu diễn dữ liệu đầu vào.

Thay vì training cho đơn vị word, Fasttext sẽ chia text ra làm nhiều đoạn nhỏ được gọi là n-gram cho một từ, ví dụ apple sẽ thành app, ppl, and ple, vector của từ apple sẽ bằng tổng của tất cả cái này. Do vậy, nó xử lý rất tốt cho những trường hợp từ hiếm gặp.

Ví dụ: với mỗi từ trong dữ liệu training sẽ được biểu diễn bằng một vector 300 chiều, độ dài tối đa của một mẩu (sample) là 26 (dựa vào độ dài trung bình của tất cả các sample trong dataset). Vậy để biểu diễn một sample ta cần một ma trận kích thước 26x300.

2.1.2.2 Feature Extraction

CNN: bao gồm các lớp convolutional xếp chồng lên nhau, sữ dụng các hàm kích hoạt phi tuyến tính như Relu hoặc tanh để kích hoạt trọng số ở mỗi node.

Chúng tôi quyết định sử dụng CNN với mục đích rút trích đặc trưng đầu vào và biểu diễn nó dưới dạng dữ liệu có số chiều thấp hơn.

Tuy nhiên, CNN chỉ có thể rút trích được local information còn khả năng biểu diễn context information thì chưa thực sự ấn tượng.

2.1.2.3 Extract Contextual Dependencies

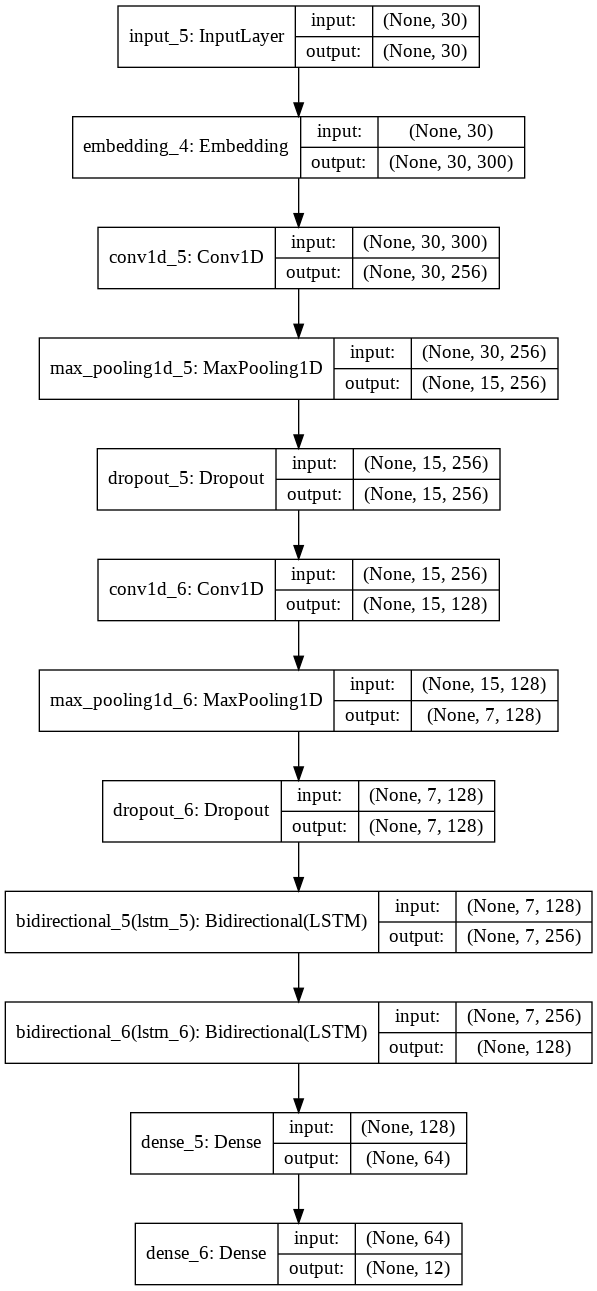
LSTM: Một cái tiến của mạng nơ ron hồi quy, giúp khác phục được vấn đề phụ thuộc xa (long term dependencies).

LSTM có thể biểu diễn context information rất ấn tượng, tuy nhiên thời gian tranning rất lâu.

2.1.2.4 Model Architecture.

Chúng tôi sữ dụng mạng CNN để rút trích local information, và dùng những đặc trưng đã được rút trích của CNN làm input cho LSTM để rút trích Contextual Dependencies information.

Chi tiết của mô hình như sau:



Hình 2: Kiến trúc mô hình

Bao gồm:

* Một tầng Embedding để biểu diễn dữ liệu đâu vào. Tầng Embedding nhận input đầu vào là một ma trận 30x300, với 30 là độ dài trung bình của các mẩu trong tập training.
* Hai tầng Conv1D với số lượng node, kích thước convolutional layer lần lượt là 128-5 và 64-3 theo sau một tầng convolutional là một tầng max pooling, và sữ dụng activation function là relu và dropout của mỗi tầng là 0.5 để giảm overfiting trong quá trình training. Mục tiêu của hai tầng Conv1D là để rút trích các local feature, vậy ma trận 30x300 đầu vào sẽ được biểu diễn lại thành 7x128 giúp cho việc học các đặc trưng ngữ cảnh của hai tầng Bidirectional LSTM được nhanh hơn.
* Kết quả từ tầng từ hai tầng convolutional sẽ là input của hai tầng lstm. Số node của hai tầng đều là 64.
* Một tầng kích hoạt sữ dụng activation function relu.
* Sau cùng là hai tầng fully connected biểu diễn kết quả đầu ra.

Với output của tầng cuối cũng đúng bằng số lượng aspect có thể có trong dữ liệu train. Đối với tập dữ liệu VLSP 2018, domain hotel là 34 và restaurant là 12.

Ví dụ: Ta có tổng số aspect của domain là 5 và được sắp xếp theo thứ tự như sau:

A = {a, b, c, d, e}, và sample có nhãn L = {a-positive, b-negative, c-neutral}

Thì vector biểu diễn sẽ có dạng V = {1, -1, 0.5, -0.5, -0.5}.s

2.2 Đặc trưng của giải pháp đề xuất

Không cần segment dữ liệu đầu vào thành nhiều n-gram như các phương pháp học máy truyền thống.

Các input đầu vào phải có độ dài giống nhau.

Không cần chia thành 2 giai đoạn: aspect detection và sentiment polarity như các phương pháp đã được thực hiện trước đó.

2.3 Cách đánh giá

Đánh giá dựa trên F1 score

A: là tập các giá trị dự đoán (Aspect-Polarity)

B: là tập các giá trị đúng.

CHƯƠNG 3 - THỰC NGHIỆM

3.1. Dữ liệu

Dữ liệu của cuộc thi do VLSP tổ chức , VLSP-2018.

Bao gồm:

|  |  |  |  |
| --- | --- | --- | --- |
| Domain | Dataset | #Review | #Aspect |
| Restaurant | Training  Development  Test | 2961  1290  500 | 9034  3408  2419 |
| Hotel | Training  Development  Test | 3000  2000  600 | 13948  7111  2584 |

3.2. Công nghệ

3.2.1. Ngôn ngữ lập trình

Hệ điều hành Ubuntu, Python 3.6, Google Colaborator

3.2.2. Thư viện

* Numpy
* Pyvi
* Keras
* Pandas
* Matplotlib

3.3. Thực nghiệm

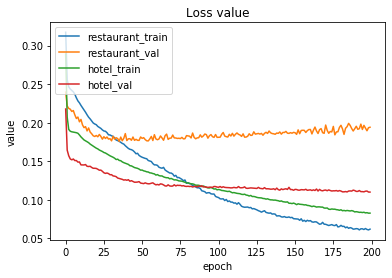
Chúng tôi thực nghiệm trên tập dữ liệu VLSP 2018.

Traning với mức learning-rate là 0.0001, epoch: 200, batch-size: 50 cho mỗi domain.

3.4. Đánh giá kết quả



Hình 3: Model F1 score



Hình 4: Model loss value

So sánh với kết quả đạt được và kết quả được công bố trong bài báo cáo [1] chúng ta có bảng so sánh như sau.

|  |  |  |
| --- | --- | --- |
| Team | Dev | Test |
| SA1 | 0.67 | **0.61** |
| SA2 | 0.59 | 0.58 |
| SA3 | 0.64 | 0.52 |
| **My approach** | **0.7** | **0.61** |

Bảng 2: Bảng so sánh Phase B (Aspect-Polarity) F1 score của domain restaurant

Chúng tôi rút ra các nhận xét như sau:

* F1 score của tập dev đã có sự cái thiện chỉ với 200 epoch.
* F1 score của tập test không có sự thay đổi nguyên nhân là do bias quá cao so với dữ liệu train, độ dài trung bình của một sample trong tập train, dev và test lần lượt được làm tròn là :
  + train: 25.60
  + dev: 23.21
  + test:73.84

Ngoài ra chúng tôi không thực hiện so sánh F1 score domain hotel với bài công bố, vì trong quá trình thực nghiệm thì chúng tôi phát hiện có tổng cộng 34 aspect cho domain này, tuy nhiên trong dữ liệu train chỉ tồn tại tổng 33 aspect.

CHƯƠNG 4 - KẾT LUẬN

4.1 Kết quả đạt được

**Kết quả:**

Áp dụng mô hình hybrid CNN-LSTM để giải quyết bài toàn Aspect Base Sentiment Analysis với tập dữ liệu VLSP 2018.

**Hạn chế:**

Độ chính xác không cao.

Cần phải chia domain cho từng lĩnh vực.

Số chiều của output đầu ra tỉ lệ thuận với số lượng Aspect.

Dữ liệu training ít và bias của tập test sau khi đã pre-processing cao.s

4.2. Hướng phát triển

Tiền xữ lý dữ liệu.

Phát sinh thêm mẩu.

Xây dựng bộ từ điển dự đoán aspect cho các danh từ trong câu. Từ đó đi tìm những tính từ lân cận để dự đoán Polarity tương ứng.

Áp dụng các mô hình state of the art hiện tại để giải quyết bài toán này như BERT.

Biểu diễn dữ liệu input và output theo một phương pháp khác hiệu quả hơn.

**TÀI LIỆU THAM KHẢO**

1. Huyen T M Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, Cuong A Le, VLSP Shared Task: Sentiment Analysis, Journal of Computer Science and Cybernetics, Vol 34, No 4, pp. 283-294, 2018.
2. X. She and D. Zhang, "Text Classification Based on Hybrid CNN-LSTM Hybrid Model," 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2018, pp. 185-189.