ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

VÕ HỒNG THIÊN
LÊ TUYẾT MAI

KHÓA LUẬN TỐT NGHIỆP

NGHIÊN CỨU TẦM ẢNH HƯỞNG CỦA TÁCH TỪ TRÊN CÁC BÀI TOÁN NHẬN DẠNG CHUỖI TIẾNG VIỆT

THE INFLUENCE WORD SEGMENTATION ON VIETNAMESE SPAN DETECTION

CỬ NHÂN NGÀNH CÔNG NGHỆ THÔNG TIN
ĐỊNH HƯỚNG NHẬT BẢN

TP. HỒ CHÍ MINH, 2022

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

Võ Hồng Thiên – 18521432

Lê Tuyết Mai – 18521080

KHÓA LUẬN TỐT NGHIỆP

NGHIÊN CỨU TẦM ẢNH HƯỞNG CỦA TÁCH TỪ TRÊN CÁC BÀI TOÁN NHẬN DẠNG CHUỖI TIẾNG VIỆT

THE INFLUENCE WORD SEGMENTATION ON VIETNAMESE SPAN DETECTION

CỬ NHÂN NGÀNH CÔNG NGHỆ THÔNG TIN
ĐỊNH HƯỚNG NHẬT BẢN

GIẢNG VIÊN HƯỚNG DẪN
ThS. NGUYỄN VĂN KIỆT
TS. NGUYỄN LƯU THUỲ NGÂN

TP. HỒ CHÍ MINH, 2022

LỜI CẢM ƠN

Đầu tiên, chúng tôi xin gửi lời cảm ơn chân thành đến tập thể quý thầy cô Trường Đại học Công nghệ Thông tin – Đại học Quốc Gia TP.HCM và quý thầy cô khoa Khoa học và Kỹ thuật thông tin đã giúp cho chúng tôi có những kiến thức cơ bản làm nền tảng để thực hiện nghiên cứu này.

Nhóm chúng tôi xin gửi lời cảm ơn chân thành tới ThS. Nguyễn Văn Kiệt và TS. Nguyễn Lưu Thùy Ngân đã đồng hành và theo sát nhóm chúng tôi để hướng dẫn, quan tâm, lo lắng và chỉnh sửa để có được khóa luận tốt nghiệp tốt nhất. Thầy và cô là hai người truyền nguồn cảm hứng, kiến thức để nhóm có đủ nhiệt huyết để thực hiện khóa luận tới cuối cùng.

Nhóm chúng tôi cũng gửi lời cảm ơn tới các anh, chị và các bạn trong nhóm NLP@UIT đã hỗ trợ chia sẻ kinh nghiệm cùng các góp ý quý giá cho nhóm để nhóm có kết quả chỉn chu nhất.

Tiếp theo, chúng tôi muốn cảm ơn tới thầy cô đã truyền đạt các kiến thức quý báu từ khi bước chân vào nhà trường, kiến thức chúng tôi tích lũy được từ quý thầy cô đã giúp ích cho chúng tôi thực hiện khóa luận rất nhiều.

Cuối cùng, chúng tôi xin cảm ơn đến gia đình và bạn bè đã động viên, khuyến khích và truyền năng lượng tích cực cho nhóm để hoàn thành khóa luận.

Xin chân thành cảm ơn quý Thầy/Cô.

Mục lục

Danh mục hình

Danh mục bảng

Danh mục từ viết tắt

ABSA	**A**spect **B**ase **S**entiment **A**nalysis
AI	**A**rtificial **I**ntelligence
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CRF	**C**onditional **R**andom **F**ield
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
MRC	**M**achine **R**eading **C**omprehension
NLP	**N**atural **L**anguage **P**rocessing
OOV	**O**ut **O**f **V**ocabulary
SOTA	**S**tate **O**f **T**he **A**rt
TP	**T**rue **P**ositive
TN	**T**rue **N**egative

TÓM TẮT KHOÁ LUẬN

Khi xử lý các bài toán xử lý ngôn ngữ tự nhiên, chúng tôi nhận ra tầm quan trọng của việc tách từ trong việc tăng hiệu suất đào tạo các mô hình. Để có cái nhìn tổng thể và khách quan về việc sử dụng phương pháp tách từ như thế nào cho bài toán nhận dạng chuỗi trên tiếng Việt, trong khoá luận này chúng tôi sẽ tiến hành so sánh sự tác động của phương pháp tách từ tiếng Việt dựa trên so sánh thực nghiệm hai bài toán đại diện cho nhận diện đa chuỗi là bài toán nhận diện cảm xúc theo khía cạnh (ABSA) - nhận diện nhiều thực thể (multi span detection) và đại diện cho nhận diện đơn chuỗi là bài toán đọc hiểu tự động (MRC) - nhận diện một thực thể (single span detection). Đối với bài toán nhận diện cảm xúc theo khía cạnh, chúng tôi tiến hành so sánh thực nghiệm trên bộ dữ liệu tiếng Việt đã được công bố là UIT-ViSD4SA [27]. Đối với bài toán đọc hiểu tự động, chúng tôi tiến hành so sánh thực nghiệm trên bộ dữ liệu tiếng Việt đã được công bố là UIT-ViQuAD 1.0 [17]. Chúng tôi tiến hành so sánh thực nghiệm trên các mô hình SOTA hiện nay như là: PhoBERT [16], XLM-RoBERTa [4]. Đối với bài toán nhận diện cảm xúc theo khía cạnh, chúng tôi tiến hành so sánh thêm với một loại nhúng từ được đào tạo trước (pretrain word embedding) dành cho tiếng Việt là PhoW2V [15] kết hợp với mô hình BiLSTM-CRF [11].

Chương 1. TỔNG QUAN

1.1 Đặt vấn đề

Xử lý ngôn ngữ tự nhiên (Natural Languge Processing - NLP) là một nhánh của Trí tuệ nhân tạo (Artificial Intelligence - AI) cung cấp cho máy tính khả năng hiểu ngôn ngữ viết và nói của con người. Dễ dàng kể đến một số ứng dụng của NLP trong kiểm tra chính tả, tự động điền từ hay câu, phát hiện thư rác, trợ lý ảo trên điện thoại và ô tô. Tuy nhiên, ít ai biết rằng máy móc hoạt động với các con số chứ không phải chữ cái. Vì vậy, để làm việc với một lượng lớn dữ liệu văn bản có sẵn, việc tiền xử lý văn bản (text preprocessing) đóng vai trò hết sức quan trọng. Bản thân tiền xử lý văn bản bao gồm nhiều giai đoạn, và một trong số những kỹ thuật rất được quan tâm đối với tiếng Việt đó là tách từ (word segmentation). Vậy tại sao nó lại được quan tâm? Bởi vì các ngôn ngữ của các quốc gia khác nhau mang những đặc điểm khác nhau.

1.1.1 Đặc điểm ngôn ngữ

Ngôn ngữ trên thế giới được chia làm ba loại chính:

Ngôn ngữ hoà kết (Flexional) là nhóm ngôn ngữ có sự phân tách ngữ nghĩa các từ trong câu theo khoảng cách trắng. Khi ngắt các từ trong câu bằng khoảng thì mỗi từ đều mang ý nghĩa và không có quá nhiều từ ghép gây nhập nhằng ngữ nghĩa. Các ngôn ngữ thuộc nhóm ngôn ngữ hoà kết như Đức, Latin, Hi Lạp, Anh, Nga,...

Ngôn ngữ chắp dính (Agglutinate) là nhóm ngôn ngữ được xây dựng bằng cách ghép các kí tự riêng lẻ với nhau. Đặc điểm của nhóm ngôn ngữ này là không

có khoảng cách trắng giữa các từ. Các từ được xây dựng từ một hay nhiều kí tự. Các ngôn ngữ thuộc nhóm ngôn ngữ chắp dính như Thổ Nhĩ Kỳ, Mông Cổ, Nhật Bản, Triều Tiên,...

Ngôn ngữ đơn lập (Isolate) là nhóm ngôn ngữ có sự kết hợp giữa hai ngôn ngữ trên. Ngôn ngữ đơn lập vẫn được ghép bởi các từ và có khoảng trắng ngăn cách các từ trong câu. Tuy nhiên, ngôn ngữ này có một đặc điểm là không phải các từ đơn nào cũng có ý nghĩa và khi ghép các từ đơn không có nghĩa lại thành một từ có nghĩa gọi là từ láy, hoặc ghép các từ đơn có nghĩa lại thành một từ có nghĩa khác gọi là từ ghép. Chính vì đặc điểm này mà ngôn ngữ trong nhóm ngôn ngữ đơn lập rất dễ gây nhập nhằng ý nghĩa nếu các từ không được tách một cách chuẩn xác theo ngữ nghĩa và hoàn cảnh. Điều này tạo nên một thách thức lớn đối với các bài toán xử lý ngôn ngữ tự nhiên được xây dựng cho ngôn ngữ này. Các ngôn ngữ thuộc nhóm ngôn ngữ đơn lập như **Việt Nam**, Indonesia, Hán,...

Vì đặc điểm ngôn ngữ khác nhau nên các kỹ thuật tiền xử lý đối với các ngôn ngữ khác nhau cũng khác nhau đặc biệt là kỹ thuật tách từ. Dưới đây chúng tôi sẽ trình bày các kỹ thuật tách từ phổ biến hiện nay.

1.1.2 Kỹ thuật tách từ

Kỹ thuật tách từ (word segmentation) là một trong những giai đoạn quan trọng trong quá trình tiền xử lý văn bản. Cho dù đang làm việc với các kỹ thuật xử lý ngôn ngữ truyền thống hay các kỹ thuật học sâu nâng cao thì vẫn không thể bỏ qua giai đoạn này. Nói một cách đơn giản, word segmentation là quá trình tách một cụm từ, câu, đoạn văn, một hoặc nhiều tài liệu văn bản thành các đơn vị nhỏ hơn. Mỗi đơn vị nhỏ hơn này được gọi là token.

Đầu ra khi thực hiện các kỹ thuật tách từ là các token. Có thể coi token là các khối xây dựng của NLP và tất cả các mô hình NLP đều xử lý văn bản thô ở cấp độ các token. Chúng được sử dụng để tạo từ vựng trong một kho ngữ liệu. Từ

vựng này sau đó được chuyển đổi thành số và giúp chúng ta xây dựng mô hình. Token có thể là một từ (word), một từ phụ (sub-word) hoặc thậm chí là một ký tự (character). Các thuật toán khác nhau tuân theo các quy trình khác nhau trong việc thực hiện mã hóa.

Đối với ngôn ngữ hoà kết điển hình như tiếng Anh thì word segmentation tương đồng với tokenization. Tuy nhiên đối với những ngôn ngữ mà ranh giới của từ không được đánh dấu bằng dấu cách thì word segmentation bao gồm: ghép các từ trong một câu thành những từ ghép có ý nghĩa và kết hợp với tokenization.

Có ba kỹ thuật word segmentation phổ biến, các kỹ thuật này hoạt động khác nhau và có những ưu điểm và nhược điểm riêng sẽ được phân tích cụ thể bên dưới.

1.1.2.1 Word-based tokenization algorithm (Thuật toán mã hóa dựa trên từ)

Đây là kỹ thuật được sử dụng phổ biến trong phân tích văn bản. Nó chia một đoạn văn bản thành các từ (ví dụ tiếng Việt) hoặc âm tiết (ví dụ tiếng Anh) dựa trên dấu phân cách. Dấu phân cách hay được sử dụng chính là dấu cách trắng. Tuy nhiên, cũng có thể tách văn bản không theo dấu phân cách. Ví dụ kỹ thuật trong tiếng Việt vì một từ trong tiếng Việt có thể chứa 2 hoặc 3 âm tiết được nối với nhau bởi dấu gạch nối. Ví dụ: ["Let", "us", "learn", "tokenization."].

1.1.2.2 Subword-based tokenization algorithm (Thuật toán mã hóa dựa trên từ phụ)

Đây là một giải pháp nằm giữa mã hóa dựa trên từ và ký tự. Ý tưởng chính là giải quyết đồng thời các vấn đề của mã hóa dựa trên từ (kích thước từ vựng rất lớn, có nhiều tokens OOV) và mã hóa dựa trên ký tự (chuỗi rất dài và token riêng lẻ rất khó xử lý đối với ngữ nghĩa từ).

Hạn chế của kỹ thuật này là nó dẫn đến một kho ngữ liệu khổng lồ và một lượng từ vựng lớn, khiến mô hình cồng kềnh hơn và đòi hỏi nhiều tài nguyên tính toán hơn. Bên cạnh đó, một hạn chế nữa là liên quan đến các từ sai chính tả. Nếu kho ngữ liệu có từ "knowledge" viết sai chính tả thành "knowldge", mô hình sẽ gán token OOV cho từ sau đó. Do đó, để giải quyết tất cả những vấn đề này, các nhà nghiên cứu đã đưa ra kỹ thuật mã hóa dựa trên ký tự.

Hầu hết các mô hình tiếng Anh đều sử dụng các dạng thuật toán của mã hóa từ phụ, trong đó, phổ biến là WordPeces được sử dụng bởi BERT và DistilBERT, Unigram của XLNet và ALBERT, và Bye-Pair Encoding của GPT-2 và RoBERTa.

Mã hóa dựa trên từ khóa phụ cho phép mô hình có kích thước từ vựng phù hợp và cũng có thể học các biểu diễn độc lập theo ngữ cảnh có ý nghĩa. Mô hình thậm chí có thể xử lý một từ mà nó chưa từng thấy trước đây vì sự phân tách có thể dẫn đến các từ phụ đã biết. Ví dụ: ["Let", "us", "learn", "token", "ization."].

1.1.2.3 Character-based tokenization algorithm (Thuật toán mã hóa dựa trên ký tự)

Mã hóa dựa trên ký tự chia văn bản thô thành các ký tự riêng lẻ. Logic đằng sau mã hóa này là một ngôn ngữ có nhiều từ khác nhau nhưng có một số ký tự cố định. Điều này dẫn đến một lượng từ vựng rất nhỏ. Ví dụ tiếng Anh có 256 ký tự khác nhau (chữ cái, số, ký tự đặc biệt) trong khi chứa gần 170,000 từ trong vốn từ vựng. Do đó, mã hóa dựa trên ký tự sẽ sử dụng ít token hơn so với mã hóa dựa trên từ. Thuật toán sẽ chia câu thành các ký tự, ở đây là từng chữ cái một.

Một trong những lợi thế chính của mã hóa dựa trên ký tự là sẽ không có hoặc rất ít từ không xác định hoặc các từ không có trong bộ từ vựng (Out Of Vocabulary - OOV). Do đó, nó có thể biểu diễn các từ chưa biết (những từ không được nhìn thấy trong quá trình huấn luyện) bằng cách biểu diễn cho mỗi ký tự.

Một ưu điểm khác là các từ sai chính tả có thể được viết đúng chính tả lại, thay vì có thể đánh dấu chúng là mã thông báo OOV và làm mất thông tin.

Loại mã hóa này khá đơn giản và có thể làm giảm độ phức tạp của bộ nhớ và thời gian. Tuy nhiên, một ký tự thường không mang bất kỳ ý nghĩa hoặc thông tin nào như một từ. Ngoài ra, tuy kỹ thuật này giúp giảm kích thước từ vựng nhưng lại làm tăng độ dài chuỗi trong mã hóa dựa trên ký tự. Mỗi từ được chia thành từng ký tự và do đó, chuỗi mã hóa dài hơn nhiều so với văn bản thô ban đầu. Ví dụ: ["L", "e", "t", "u", "s",...].

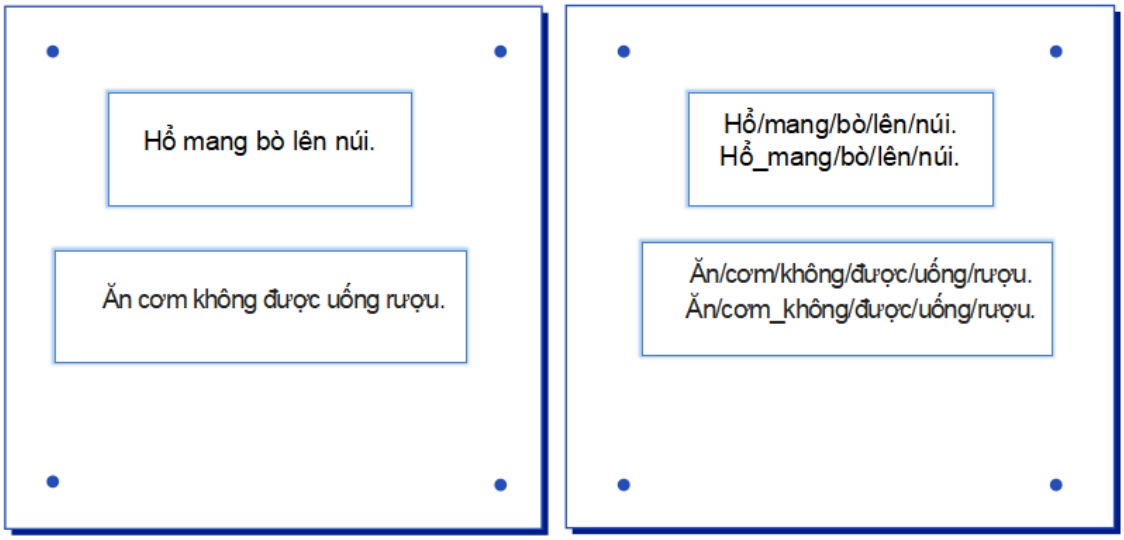
Thực tế, các mô hình NLP sử dụng các phương pháp tách từ phù hợp theo từng ngôn ngữ. Tuỳ thuộc vào từng bài toán, mà cùng một văn bản có thể được xử lý dưới các loại token khác nhau. Mỗi token thường có tính duy nhất và được biểu diễn bằng một ID, các ID này là một cách mã hoá hay cách định danh token trên không gian số.

Chính vì lý do đó tách từ được xem là bước xử lý quan trọng đối với các hệ thống Xử Lý Ngôn Ngữ Tự Nhiên, đặc biệt là đối với các ngôn ngữ thuộc vùng Đông Á theo loại hình ngôn ngữ đơn lập, ví dụ: tiếng Trung Quốc và đặc biệt là tiếng Việt. Với các ngôn ngữ thuộc loại hình này, ranh giới từ không chỉ đơn giản là những khoảng trắng như trong các ngôn ngữ thuộc loại hình hòa kết như tiếng Anh, mà phải có sự liên hệ chặt chẽ giữa các tiếng với nhau, một từ có thể cấu tạo bởi một hoặc nhiều tiếng. Vì vậy đối với các ngôn ngữ thuộc vùng Đông Á, vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ.

Hiện nay, có rất nhiều thư viện của Python như là NLTK, spaCy, Keras, Gensim hỗ trợ việc xử lý tách từ. Đặc biệt hơn đối với tiếng Việt là VnCoreNLP. Đây là thư viện đang được xếp hạng khá cao về độ chính xác mà nó đem lại.

Đối với tiếng Anh hoặc các ngôn ngữ không phải là ngôn ngữ đơn lập thì từ là một nhóm các kí tự có nghĩa được tách biệt bằng khoảng trắng trong câu. Vì thế mà kỹ thuật tách từ trở nên rất đơn giản. Tuy nhiên trong tiếng Việt dấu cách được dùng để phân tách các âm tiết (tiếng) chứ không phải các từ. Mang đặc

trưng là từ Tiếng Việt biến đổi hình thái, ranh giới từ không được xác định mặc nhiên bằng khoảng trắng. Cho nên có trường hợp một câu có thể có nhiều ngữ nghĩa khác nhau tuỳ vào kỹ thuật tách từ như thế nào, gây nhập nhằng về ngữ nghĩa của câu.



HÌNH 1.1: Mô tả mối liên hệ giữa tách từ và ngữ nghĩa.

Như hình 1.1, các ví dụ đã cho chúng ta thấy được sự tác động mạnh mẽ của ý nghĩa câu khi tách từ khác nhau. Ở ví dụ đầu tiên "Hổ mang bò lên núi.". Từ hổ mang nếu được tách theo tiếng sẽ có thể hiểu theo ý nghĩa con hổ kết hợp với động từ mang (trong mang vác)

Để tiến hành thử nghiệm, chúng tôi sử dụng thư viện VnCoreNLP [30].

Trên đây, chúng tôi đã giới thiệu về ý nghĩa của việc tách từ trong xử lý ngôn ngữ Tiếng Việt, để có cái nhìn chính xác và khách quan hơn về vấn đề tách từ đối với các bài toán NLP, chúng tôi tiến hành so sánh dựa trên thực nghiệm đối với bài toán nhận diện chuỗi (span detection).

1.2 Đối tượng và phạm vi nghiên cứu

1.2.1 Đối tượng

- Bài toán nhận diện đơn thực thể (single span detection) - bài toán đọc hiểu máy.
- Bài toán nhận diện đa thực thể (multi span detection) - bài toán nhận diện cảm xúc theo khía cạnh.

1.2.2 Phạm vi

- Bài toán phân tích cảm xúc khía cạnh: các bình luận trên mạng xã hội về lĩnh vực công nghệ.
- Bài toán đọc hiểu máy: Các bài báo trên Wikipedia.

1.3 Giới thiệu bài toán

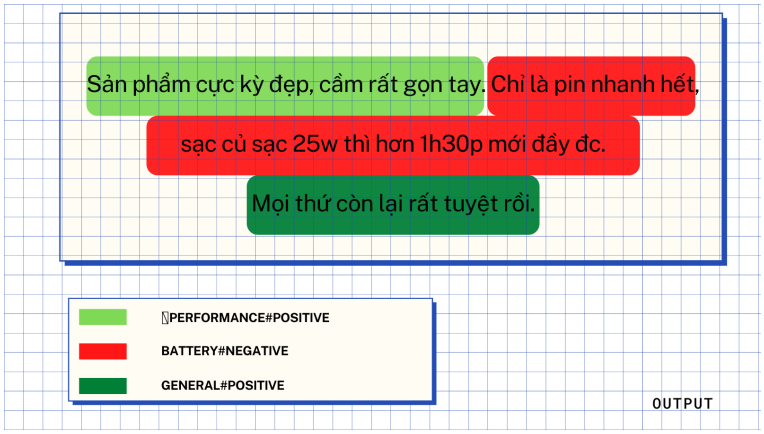
Các bài toán nhận diện chuỗi được chia làm hai bài toán nhỏ hơn. Đó là nhận diện chuỗi đơn (single span detection) và nhận diện chuỗi đa (multi span detection). Trong mỗi bài toán, chúng tôi chọn một bài toán đại diện làm đối tượng thử nghiệm của chúng tôi. Đối với nhận diện chuỗi đơn, chúng tôi chọn bài toán đọc hiểu tự động. Và đối với nhận diện chuỗi đa, chúng tôi chọn bài toán nhận diện cảm xúc dựa trên khía cạnh.

1.3.1 Bài toán nhận diện cảm xúc theo khía cạnh

Trong những năm gần đây, ý kiến và đánh giá phản hồi của khách hàng ngày một được chú trọng và quan tâm nhiều hơn. Cùng với việc gia tăng nhanh chóng của lượng dữ liệu đánh giá sản phẩm trên internet. Việc thu thập các đánh

giá và xây dựng các phân tích trên lượng dữ liệu thu thập để cải thiện dịch vụ khách hàng đang dần trở thành xu hướng. Vì thế mà bài toán phân tích cảm xúc trở nên được chú trọng hơn hẳn. Tuy nhiên, việc thực hiện phân tích cảm xúc, phản hồi của khách hàng chỉ theo phương diện cảm xúc tích cực hay tiêu cực mà không thể chỉ rõ ra được điểm tích cực hay tiêu cực cụ thể ở đâu để cải thiện cũng là một bài toán thách thức đối với các doanh nghiệp. Vì thế mà xu hướng hiện nay đang hướng tới bài toán phân tích cảm xúc theo từng khía cạnh. Nó giải quyết được vấn đề phân tích cảm xúc phản hồi của khách hàng không những thế còn đưa ra cảm xúc theo từng khía cạnh cụ thể. Hơn nữa, bài toán này còn trích xuất được thông tin mang yếu tố cảm xúc của khách hàng theo từng khía cạnh.

HÌNH 1.2: Mô tả đầu vào bài toán phân tích cảm xúc theo khía cạnh.



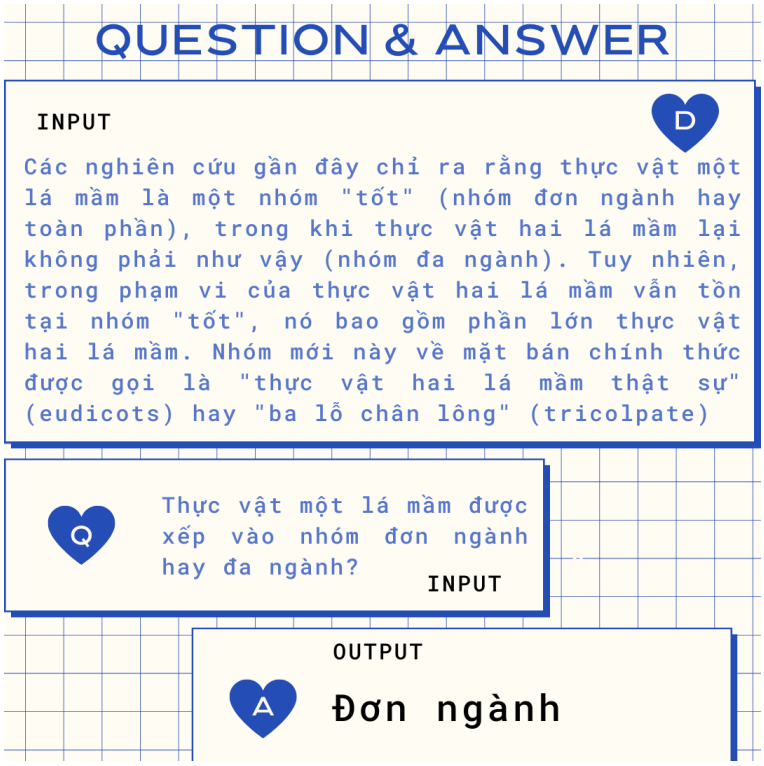
HÌNH 1.3: Mô tả đầu ra bài toán phân tích cảm xúc theo khía cạnh.

1.3.2 Bài toán hệ thống hỏi đáp

Đọc hiểu tự động (Machine Reading Comprehension - MRC) được biết đến là một trong những lĩnh vực nghiên cứu chính trong xử lý ngôn ngữ tự nhiên (NLP). MRC là nhiệm vụ làm cho máy tính đọc, hiểu văn bản và trả lời các câu hỏi liên quan đến văn bản đó, đọc hiểu là một thách thức lớn đối với máy tính. MRC đã không còn xa lạ trong những năm gần đây và đạt được nhiều kết quả khá ấn tượng. Việc trả lời một câu hỏi dựa vào nội dung của bài văn cho sẵn đôi khi không hề dễ đối với chúng ta, vì vậy để máy tính trả lời được thì đó cũng là thách thức lớn. Thử thách đặt ra là làm cho một hệ thống máy tính hiểu ngôn ngữ mà con người sử dụng, từ đó có sự tương tác với con người thông qua các ngữ cảnh cụ thể, đây chính là nhiệm vụ mà các máy đọc hiểu văn bản sẽ đảm nhiệm.

1.4 Mục tiêu

Khi thực hiện nghiên cứu các bài toán xử lý ngôn ngữ tự nhiên gần đây, chúng tôi nhận thấy các bài toán về nhận dạng chuỗi đang rất được quan tâm bởi tính ứng dụng cao mà nó đem lại, giải quyết được rất nhiều vấn đề. Tuy



HÌNH 1.4: Mô tả đầu vào đầu ra của bài toán đọc hiểu tự động.

nhiên, độ chính xác của các bài toán này vẫn chưa cao 45,70% (trên bộ dữ liệu UIT-ViSD4SA) [27] đối với bài toán nhận diện cảm xúc theo khía cạnh. Vì thế mà chúng tôi mong muốn đề xuất phương pháp để tăng hiệu suất các bài toán nhận dạng chuỗi. Ngoài các kỹ thuật tinh chỉnh (fine-turning) hay đề xuất các mô hình mới thì chúng tôi lại quan tâm đến kỹ thuật tiền xử lý dữ liệu (preprocessing) bởi chúng tôi tin rằng nó cũng có thể khiến hiệu suất cải thiện đáng kể nếu dữ liệu được tiền xử lý một cách hiệu quả.

Kỹ thuật tiền xử lý dữ liệu gồm có rất nhiều nhưng chúng tôi đặc biệt quan tâm đến kỹ thuật word-segmentation bởi nó mang đặc trưng của ngôn ngữ và hơn hết là đối với tiếng Việt. Vì thế mà chúng tôi quyết định thực hiện nghiên cứu tầm ảnh hưởng của nó đối với các bài toán nhận dạng chuỗi.

Mục đích của khóa luận tốt nghiệp này là nghiên cứu và đưa ra so sánh kết

luận cho việc sử dụng phương pháp tách từ cho văn bản tiếng Việt. Để làm được điều này, chúng tôi tập trung ba mục tiêu chính:

- Đầu tiên, là xây dựng thử nghiệm trên hai bài toán nhận dạng chuỗi là nhận diện cảm xúc theo khía cạnh (đại diện cho bài toán nhận dạng đa chuỗi) và bài toán đọc hiểu máy (đại diện cho bài toán nhận dạng đơn chuỗi).
- Thứ hai, là xây dựng thử nghiệm hai bài toán liệt kê trên với hai phương pháp word segmentation là theo âm tiết và theo từ.
- Thứ ba, phân tích, so sánh và đưa ra kết luận cho hai phương pháp word segmentation nêu trên đối với cả hai bài toán nhận dạng chuỗi.
- Ngoài ra, chúng tôi còn tiến hành so sánh trên bài toán nhận diện đoạn mang ý nghĩa cảm xúc và nhận diện đoạn mang ý nghĩa khía cạnh.

1.5 Cấu trúc khóa luận

Chương 1: Tổng quan

Trong chương này, chúng tôi giới thiệu tổng quan về khái niệm tách từ, bài toán phân tích cảm xúc khía cạnh và bài toán đọc hiểu tự động. Tầm quan trọng của việc tách từ trong các tác vụ nhận diện chuỗi, khả năng ứng dụng của đề tài mà chúng tôi thực hiện cùng với đó là giới thiệu lý do chúng tôi chọn đề tài này để nghiên cứu.

Chương 2: Các công trình nghiên cứu liên quan

Trong chương này, chúng tôi trình bày một số công trình nghiên cứu liên quan trên thế giới và trong nước đến hai bài toán nhận diện chuỗi mà chúng tôi sẽ tiến hành thử nghiệm trong nghiên cứu này. Đó là bài toán nhận diện cảm xúc theo khía cạnh và bài toán hệ thống đọc hiểu máy.

Chương 3: Cơ sở lý thuyết, thử nghiệm

Trong chương này, chúng tôi giới thiệu về một số đặc điểm của bộ dữ liệu được sử dụng trong khóa luận, các mô hình sử dụng, các mô hình liên quan và phương pháp đánh giá mô hình. Cùng với đó là quy trình thử nghiệm các mô hình để so sánh.

Chương 4: Kết quả

Trong chương này, chúng tôi sẽ trình bày phân tích kết quả đạt được từ hai bài toán và các lỗi gặp phải.

Chương 5: Kết luận và hướng phát triển

Trong chương này, chúng tôi sẽ chỉ ra sự đóng góp, các vấn đề khó khăn gặp phải và hướng phát triển của khóa luận

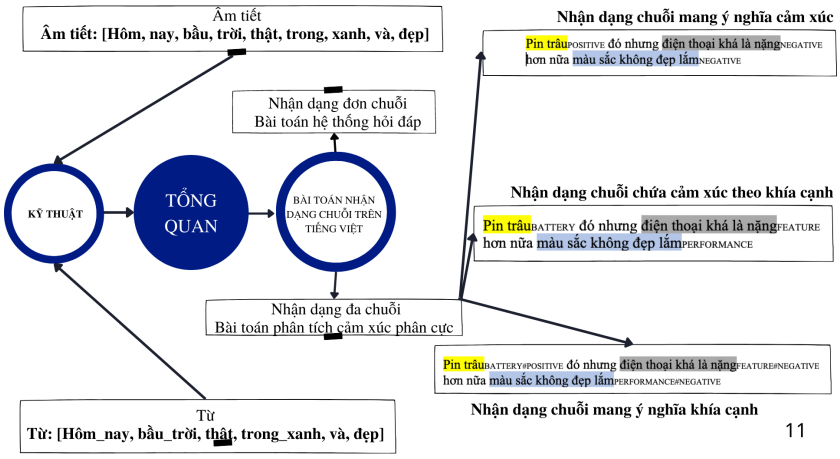
1.6 Tính ứng dụng của đề tài

Trong khoá luận này, chúng tôi sẽ làm rõ liệu tách từ theo từ hay tách từ theo âm tiết sẽ mang lại kết quả tốt hơn cho các bài toán nhận diện chuỗi. Cụ thể hơn, đó là hai bài toán đọc hiểu tự động và phân tích cảm xúc dựa trên khía cạnh. Chúng tôi thực hiện đề tài này với mục đích:

- Tạo tiền đề cơ sở cho các nghiên cứu để lựa chọn phương pháp tách từ phù hợp với mô hình thực hiện hơn. Tránh việc phải thử nghiệm nhiều lần, lãng phí thời gian và chi phí.
- Góp phần nâng cao độ chính xác dựa trên sự lựa chọn phương pháp tách từ phù hợp với mô hình đang thực hiện.

1.7 Kết luận

Tóm lại, trong khoá luận này, chúng tôi thực hiện nghiên cứu kỹ thuật tách từ nào sẽ phù hợp cho bài toán nhận dạng chuỗi trên tiếng Việt. Tổng quan khoá luận được thể hiện như hình 4.1.



HÌNH 1.5: Tổng quan khoá luận.

Chương 2. CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Trong thời gian gần đây, việc phát triển chóng mặt của các trang mạng xã hội cũng như các trang thương mại điện tử giúp cho các doanh nghiệp có thể tìm hiểu được nhiều hơn thông tin đánh giá của khách hàng dành cho sản phẩm của mình. Tuy nhiên, cùng với lợi ích trên thì các doanh nghiệp cũng đang phải đối mặt với "cơn lũ" dữ liệu. Việc khai thác được triệt để những dữ liệu này là điểm mấu chốt để các doanh nghiệp có thể triển khai nhanh chóng các quyết định, cải thiện sản phẩm của mình giúp cho doanh thu cao hơn khi đưa sản phẩm ra thị trường. Vấn đề được đặt ra là có quá nhiều thông tin để xử lý không thể sử dụng nguồn lực con người để đọc và phân tích đánh giá. Tuy nhiên hầu hết các dữ liệu đánh giá sản phẩm của người dùng đều là dữ liệu phi cấu trúc.

Để giải quyết bài toán trên, những năm trước đây, người ta sẽ sử dụng các bài toán nhận diện cảm xúc (sử dụng phương pháp phân lớp theo câu - Text Classification). Tuy nhiên phương pháp này cũng mang đến rất nhiều vấn đề cần quan tâm như việc nó sẽ dễ nhận diện sai khi bình luận của khách hàng mang nhiều ý nghĩa khác nhau.

Ví dụ đối với câu bình luận: "Máy xài khá ổn đấy. Tuy nhiên màn hình của nó có vẻ hơi nhỏ nhưng được cái pin khá trâu.", nếu là bài toán phân tích cảm xúc sẽ rất khó để có thể đánh giá bình luận trên là tích cực hay tiêu cực.

Vì thế mà những năm gần đây các bài toán nhận dạng chuỗi trở nên được chú ý hơn hẳn bởi độ tiện dụng của nó. Bài toán nhận dạng chuỗi gồm hai loại chính là: Nhận dạng đơn chuỗi (Single Span Detection) và Nhận dạng đa chuỗi (Mutil

Span Detection). Bài toán nhận dạng đơn chuỗi được ứng dụng nhiều trong thực tế như các hệ thống hỏi đáp, đọc hiểu tự động,... Bài toán nhận dạng đa chuỗi cũng được quan tâm rất nhiều trong những năm gần đây, có thể kể đến bài toán nhận diện đoạn văn bản mang ý nghĩa cảm xúc khía cạnh giải quyết được các vấn đề bất cập được kể trên.

Ví dụ đối với ví dụ đề cập ở trên, bài toán có thể cho kết quả (Máy xài khá ổn-GENERAL#NEUTRAL, màn hình của nó có vẻ hơi nhỏ-SCREEN#NEGATIVE, pin khá trâu-BATTERY#POSITIVE). Có thể thấy nó khắc phục được nhược điểm nhập nhằng, hơn thế nữa còn có thể trích xuất đoạn văn bản mang ý nghĩa cảm xúc.

Hai bài toán được đề cập ở trên tuy rằng ứng dụng rất tốt trong các ứng dụng thực tế nhưng hiện tại hiệu suất của nó trên các bộ dữ liệu tiếng Việt vẫn chưa được đánh giá cao. Bài toán Nhận diện cảm xúc theo khía cạnh đạt 45,70% [27]. Vì thế mà chúng tôi nghiên cứu tầm ảnh hưởng của tách từ đối với hai bài toán này với mục tiêu đưa ra được một phương pháp tách từ hiệu quả giúp cải thiện hiệu suất cho bài toán nhận dạng chuỗi. Cụ thể là bài toán phân tích cảm xúc khía cạnh và bài toán đọc hiểu máy

2.1 Công trình trên thế giới

2.1.1 Bài toán phân tích cảm xúc dựa trên khía cạnh

Phân tích cảm xúc dựa trên khía cạnh (ABSA) bao gồm ba nhiệm vụ cơ bản: trích xuất cụm từ khía cạnh, trích xuất cụm từ ý kiến và phân loại tình cảm ở cấp độ khía cạnh.

Phân tích cảm xúc dựa trên khía cạnh đã được nghiên cứu và giới thiệu bởi Hu và các cộng sự vào năm 2004 [10], về các khía cạnh đánh giá sản phẩm áp dụng một bộ quy tắc dựa trên các quan sát các số liệu thống kê. Bài toán Phân

tích khía cạnh cảm xúc bắt đầu được biết đến từ cuộc thi SemEval 2015 task 12 và SemEval 2016. Trong đánh giá ngữ nghĩa SemEval2014 [22], phân tích cảm xúc dựa trên khía cạnh lần đầu tiên được giới thiệu như một nhiệm vụ chung đối với các bài đánh giá tiếng Anh ở cấp độ câu cho hai lĩnh vực là nhà hàng và khách sạn.

Trong nhiệm vụ 12 của SemEval2015 (SE-ABSA15) [20], phân tích cảm xúc dựa trên khía cạnh ở cấp độ văn bản, có ba tác vụ dành cho người tham gia:

- Nhận diện khía cạnh.
- Biểu hiện mục tiêu ý kiến (Opinion Target Expression-OTE).
- Nhận diện cảm xúc.

Ở tác vụ 1 và tác vụ 3, hệ thống phải có khả năng trích xuất các danh mục khía cạnh và xác định cảm xúc đối với từng khía cạnh trong các câu hoặc bài đánh giá. Trong SemEval2015 Task 12 tập dữ liệu đã được xây dựng dựa trên SemEval2014 Task 14 (SE-ABSA14). Trong SE-ABSA15 mô tả danh mục khía cạnh của nó như một loại thực thể được kết hợp với một loại thuộc tính (ví dụ:Food#Style.)

Trong task 5 của SemEval2016 [21]: Phân tích cảm xúc dựa trên khía cạnh. Bài báo này mô tả nhiệm vụ được chia sẻ của SemEval 2016 về phân tích cảm xúc dựa trên khía cạnh (ABSA), sự tiếp nối của các nhiệm vụ tương ứng của năm 2014 và 2015. Trong năm thứ ba, nhiệm vụ đã cung cấp 19 tập dữ liệu đào tạo và 20 tập dữ liệu thử nghiệm cho 7 lĩnh vực (Khách sạn, Điện tử gia dụng, Viễn thông, Bảo tàng,...) và 8 ngôn ngữ khác (tiếng Hà Lan, tiếng Pháp, tiếng Nga, tiếng Tây Ban Nha, tiếng Thổ Nhĩ Kỳ và tiếng Ả Rập,...) cũng như một thủ tục đánh giá chung. Từ các bộ dữ liệu này, 25 cho cấp độ câu và 14 cho ABSA cấp văn bản; cái sau được giới thiệu lần đầu tiên dưới dạng nhiệm vụ con trong SemEval. Nhiệm vụ thu hút 245 bài dự thi từ 29 đội.

Với sự gia tăng của nội dung do người dùng cung cấp, với sự quan tâm đến cảm xúc và ý kiến phản hồi, lượng dữ liệu khai thác được từ văn bản đã tăng

lên nhanh chóng, cả trong học thuật và kinh doanh. Tuy nhiên, phần lớn các phương pháp tiếp cận hiện tại cố gắng phát hiện tính tích cực tổng thể của một câu, đoạn văn hoặc khoảng văn bản, bất kể các thực thể được đề cập (ví dụ: máy tính xách tay, pin, màn hình) và các thuộc tính của chúng (ví dụ: giá cả, thiết kế, chất lượng). Nhiệm vụ phân tích cảm xúc dựa trên các khía cạnh của SemEval-2015 (SE-ABSA15) là sự tiếp nối của Nhiệm vụ 4 của SemEval-2014 (SE-ABSA14). Trong phân tích tình cảm dựa trên khía cạnh, mục đích là xác định các khía cạnh của các thực thể và tình cảm được thể hiện cho từng khía cạnh. Mục tiêu cuối cùng là có thể tạo ra các bản tóm tắt liệt kê tất cả các khía cạnh và cực tổng thể của chúng.

Phân tích cảm xúc dựa trên khía cạnh nhận được nhiều sự quan tâm. Chứng kiến sự phát triển nhanh chóng của nghiên cứu trong Xử lý ngôn ngữ tự nhiên các công việc ban đầu chỉ tập trung vào việc giải quyết một trong những nhiệm vụ phụ này một cách riêng lẻ. Một số công việc gần đây tập trung vào việc giải quyết sự kết hợp của hai nhiệm vụ phụ, ví dụ: trích xuất các thuật ngữ khía cạnh cùng với các phân cực tình cảm hoặc trích xuất các thuật ngữ khía cạnh và ý kiến một cách khôn ngoan. Gần đây hơn, nhiệm vụ trích xuất bộ ba đã được đề xuất, tức là trích xuất bộ ba (thuật ngữ khía cạnh, thuật ngữ quan điểm, phân cực cảm xúc) từ một câu. Tuy nhiên, các cách tiếp cận trước đây không giải quyết được tất cả các nhiệm vụ con trong một khuôn khổ thống nhất. Năm 2021, Mao và các cộng sự [14] đã giới thiệu nhiệm vụ trích xuất bộ ba.

2.1.2 Bài toán đọc hiểu tự động

Khái niệm đọc hiểu máy hay đọc hiểu tự động (MRC) xuất hiện từ rất sớm từ những năm 1970. Nhưng đến năm 2015, các nhà nghiên cứu DeepMind [9] đã đề xuất một giải pháp mới để tạo dữ liệu có giám sát với quy mô lớn để học các mô hình MRC. Đánh dấu sự hồi sinh của bài toán MRC, với sự xuất hiện hàng loạt

của các bộ dữ liệu lớn dựa trên giải pháp mới. Đầu tiên là Bộ dữ liệu CNN [8].

Họ cũng đề xuất một mô hình Neural Network-mô hình LSTM dựa trên Attention-based có tên là Attentive Reader [8]. Và chứng minh rằng nó vượt trội hơn các phương pháp tiếp cận NLP trước đó bằng một sự cách biệt rất lớn. Trong các thí nghiệm, Attentive Reader đạt độ chính xác 63,8% trong khi các hệ thống NLP trước đó thu được tối đa 50,9% trên bộ dữ liệu CNN.

Tiếp đó, năm 2016 Chen và các cộng sự [2] đã chỉ ra rằng một mô hình mạng neural được thiết kế đơn giản và cẩn thận có thể nâng hiệu suất lên 72,4% trên tập dữ liệu CNN. Một sự phát triển vượt bậc với 8,6% so với trước đó. Và cũng trong báo cáo này, Chen và các cộng sự cũng đã chứng minh rằng các mô hình mạng neural có khả năng nhận dạng các từ vựng và cụm từ phù hợp tốt hơn so với các feature-based classifier. Tuy nhiên, mặc dù tập dữ liệu CNN cung cấp một con đường đầy hứa hẹn để đào tạo các mô hình thống kê hiệu quả nhưng Chen và các cộng sự đã kết luận rằng tập dữ liệu dường như bị nhiễu do phương pháp tạo dữ liệu gây ra lỗi tham chiếu và hạn chế để có thể thúc đẩy tiến trình hơn nữa.

Để giải quyết những hạn chế của tập dữ liệu CNN, Rajpurkar và cộng sự (2016) [23] đã thu thập một tập dữ liệu mới có tên là THE STANFORD QUESTION ANSWERING DATASET (SQuAD) [23]. Bộ dữ liệu chứa 107.785 cặp câu hỏi-câu trả lời dựa trên 536 bài báo Wikipedia. Các câu hỏi do cộng đồng đặt ra và câu trả lời cho mỗi câu hỏi là một khoảng văn bản từ đoạn đọc tương ứng.

SQuAD là bộ dữ liệu đọc hiểu quy mô lớn đầu tiên với các câu hỏi tự nhiên. Nhờ chất lượng cao và đánh giá tự động đáng tin cậy, bộ dữ liệu này đã thu hút sự quan tâm to lớn trong cộng đồng NLP và trở thành tiêu chuẩn trung tâm trong lĩnh vực này. Nó lần lượt truyền cảm hứng cho một loạt các mô hình MRC mới (Wang và Jiang, 2017 [32], Seo và các cộng sự, 2017 [26], Chen và các cộng sự, 2017 [3], Wang và các cộng sự, 2017 [33], Yu và các cộng sự, 2018 [34]) và sự tiến bộ diễn ra nhanh chóng, tính đến tháng 10 năm 2018, hệ thống hoạt động tốt nhất đã đạt được điểm F1 là 91,8% (Delvin và các cộng sự, 2018) [6]. Kết quả đã vượt quá hiệu

suất ước tính của con người là 91,2%, trong khi một feature-based classifier được xây dựng vào đầu năm 2016 chỉ thu được tỷ lệ F1 là 51,0%.

Cuộc khảo sát toàn diện về các khía cạnh khác nhau của các hệ thống MRC, bao gồm các phương pháp, cấu trúc, đầu vào, đầu ra và những nghiên cứu mới của Baradaran và cộng sự 2020 [1] đã một lần nữa khẳng định sự phát triển của MRC. Theo báo cáo từ năm 2016 đến năm 2020, đã có 241 bài báo về MRC chất lượng được đưa ra. Báo cáo cũng chứng minh rằng trọng tâm của nghiên cứu đã thay đổi trong những năm gần đây từ việc trích xuất câu trả lời sang tạo câu trả lời, từ đọc đơn tài liệu sang đọc đa tài liệu và từ việc học từ đầu đến sử dụng các pretrained.

Năm 2021, Rust và các cộng sự [25] đã công bố báo cáo về phân tích về phân đoạn tách từ với đối tượng là các mô hình pretrained đơn ngữ (nhóm tác giả không sử dụng mô hình pretrained tiếng việt trong báo cáo này) và mô hình pretrained đa ngữ. Báo cáo chỉ ra rằng, đối với các tác vụ trên một ngôn ngữ duy nhất, các mô hình pretrained đơn ngữ cho kết quả tốt hơn mô hình pretrained đa ngữ. Tuy nhiên, đối với ngôn ngữ cùng thuộc nhóm ngôn ngữ đơn lập với tiếng việt như Indonesia thì kết quả lại đi ngược lại. Mô hình đa ngôn ngữ cho kết quả cao hơn với tất cả bài toán trừ bài toán nhận diện cảm xúc. Đó là một trong những động lực để chúng tôi lựa chọn đề tài này.

2.2 Công trình trong nước

2.2.1 Bài toán phân tích tình cảm dựa trên khía cạnh

Bài toán Phân tích cảm xúc dựa trên khía cạnh đã được nghiên cứu rộng rãi trên nhiều thứ tiếng khác nhau. Năm 2021, Dang và các cộng sự đã giới thiệu một kiến trúc Mạng thần kinh hợp pháp [5] để phát hiện khía cạnh cho tiếng Việt đạt đạt được điểm F1 là 80,40% cho miền nhà hàng và 69,25% cho miền khách sạn

trên tập dữ liệu của thử thách VLSP 2018 [13].

Năm 2019, Nguyen và các cộng sự đã giới thiệu một tập dữ liệu cho các nghiên cứu về hai nhiệm vụ phụ: phát hiện khía cạnh và phát hiện phân cực [18]. Bộ dữ liệu bao gồm 7.828 đánh giá về nhà hàng. Kết quả đạt được điểm F1 là 87,13% cho phát hiện khía cạnh và điểm F1 là 59,20% cho phát hiện phân cực.

Năm 2021, Thanh và các cộng sự [27] đã giới thiệu bộ dữ liệu UIT-ViSD4SA tại hội nghị PACLIC35 cho bài toán nhận diện cảm xúc theo khía cạnh. Tác vụ này kết hợp được cả ba bài toán phụ của nó là trích xuất khía cạnh, trích xuất mang ý nghĩa cảm xúc và phân cực cảm xúc theo khía cạnh.

2.2.2 Bài toán đọc hiểu máy

Bên cạnh các công trình nước ngoài đang diễn ra sôi nổi thì các đề tài về đọc hiểu tự động cũng đang thu hút nhiều sự chú ý ở trong nước.

Do và cộng sự đã giới thiệu một bộ dữ liệu mới UIT-ViWikiQA [7], bộ dữ liệu đầu tiên để đánh giá khả năng đọc hiểu của máy dựa trên trích xuất câu bằng tiếng Việt. Bộ dữ liệu bao gồm: 23.074 câu trả lời dựa trên 5.109 đoạn của 174 bài viết trên Wikipedia tiếng Việt. Kết quả tốt nhất với Exact Match (EM) là 85,97% và F1-score là 88,77% trên mô hình XLM-R (phiên bản Large).

Năm 2021 ở thử thách VLSP 2021 với nhiệm vụ MRC, bộ dữ liệu UIT-ViQuAD 2.0 [29] do Nguyen và cộng sự giới thiệu đã được chọn để tham dự. Bộ dữ liệu vượt bậc so với các bộ dữ liệu cho tiếng Việt về nhiệm vụ MRC hiện hành bằng cách thêm vào những câu hỏi không thể trả lời. Từ đó giúp máy tiếp cận gần hơn với các cách thức xử lý giống với con người. Bộ dữ liệu bao gồm hơn 35.000 câu hỏi-câu trả lời. Với 23.000 câu hỏi trong UIT-ViQuAD 1.0 [17] kết hợp hơn 12.000 câu hỏi không thể trả lời được được tạo ra bởi cộng đồng. Đạt kết quả 77,24% với độ đo F1-score và 67,43% với độ đo (EM) trên tập private test.

2.3 Kết luận

Lĩnh vực xử lý ngôn ngữ tự nhiên nói chung và các bài toán nhận diện chuỗi tự động trong văn bản từ trước đến nay vẫn nhận được rất nhiều sự quan tâm trên thế giới và cả trong nước. Tuy nhiên, sau khi nghiên cứu các công trình liên quan, chúng tôi nhận thấy có hai điểm hạn chế:

- Thứ nhất, các công trình nghiên cứu về các bài toán nhận diện chuỗi chỉ mới được nghiên cứu và quan tâm những năm gần đây, đặc biệt phải kể đến là bài toán nhận diện cảm xúc theo khía cạnh kết hợp được cả ba tác vụ là trích xuất cụm từ khía cạnh, trích xuất cụm từ ý kiến và phân loại tình cảm theo khía cạnh. Các bài toán trước chỉ tập trung nhiều vào phân tích cảm xúc theo khía cạnh là chủ yếu.

- Thứ hai, chúng tôi nhận thấy rằng các công trình nghiên cứu về nhận diện chuỗi hiện tại chủ yếu tập trung trên tiếng Anh. Trong khi đó các công trình nghiên cứu trên tiếng Việt (mang đặc trưng khác với tiếng Anh) lại không có nhiều và chưa được đầu tư. Chúng tôi quyết định thực hiện nghiên cứu tầm ảnh hưởng của tách từ trên các bài toán nhận dạng chuỗi tiếng việt để thúc đẩy nghiên cứu trên các bài toán nhận dạng chuỗi và đăc biệt là đối với tiếng Việt.

Dựa trên các nghiên cứu trước đó và các hạn chế được nêu ra ở trên, mục tiêu đề tài của chúng tôi là tìm đáp án cho câu hỏi "Phương pháp word-segmentation nào sẽ phù hợp với bài toán nhận dạng chuỗi cho tiếng Việt ở thời điểm hiện tại?". Nghiên cứu của chúng tôi sẽ giúp thúc đẩy sự phát triển trong việc cải thiện hiệu suất đối với các bài toán nhận dạng chuỗi đối với tiếng Việt.

Chương 3. CƠ SỞ LÝ THUYẾT, THỬ NGHIỆM

3.1 Dữ liệu sử dụng trong thử nghiệm

3.1.1 Bộ dữ liệu UIT-ViSD4SA

Trong bài báo cáo này, chúng tôi sử dụng bộ dữ liệu tiếng Việt Vietnamese feedback dataset toward span detection aspect category sentiment analysis (UIT - ViSD4SA) [27] cho bài toán ABSA.

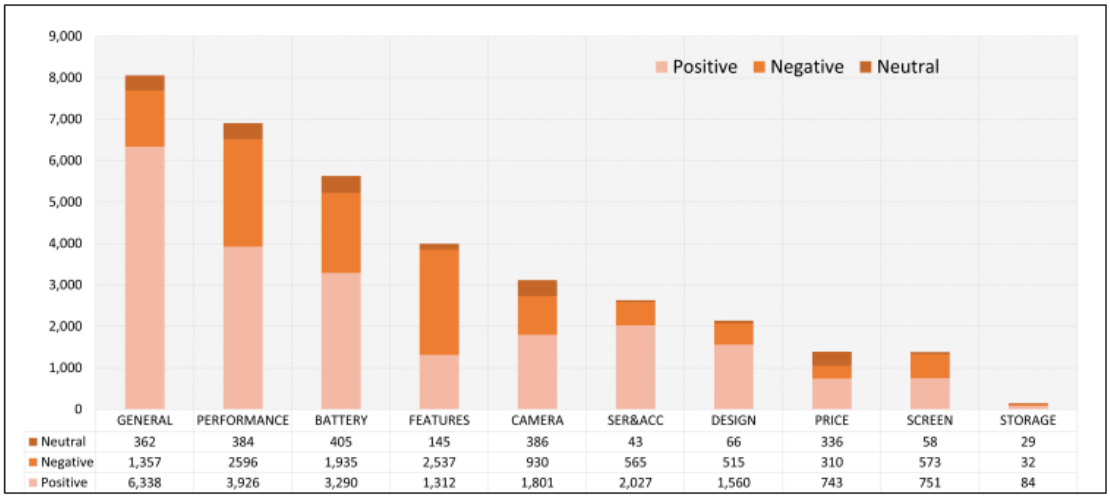
Bộ dữ liệu UIT-ViSD4SA được phát triển dựa trên bộ dữ liệu UIT-ViSFD được đề xuất bởi Phan và các cộng sự, 2021 [19]. Cốt lõi của bộ dữ liệu được hình thành dự trên sự kết hợp giữa những định nghĩa và quy tắc gán nhãn mới với hướng dẫn gán nhãn của bộ dữ liệu gốc.

Bộ dữ liệu bao gồm 35.396 chú thích được gán dựa trên 10 khía cạnh (Bảng 3.1) và 3 phương diện cảm xúc (Positive, Negative, Neutral) trên 11.122 lời nhận xét, lời bình luận về các sản phẩm công nghệ trên các trên mạng xã hội cho việc đánh giá.

Khía cạnh	Định nghĩa
SCREEN	Người dùng bình luận về chất lượng màn hình, kích cỡ, màu sắc hoặc công nghệ hiển thị.
CAMERA	Các ý kiến đề cập đến chất lượng của máy ảnh, độ rung, độ trễ, tiêu điểm hoặc màu hình ảnh.
FEATURES	Người dùng đề cập đến các tính năng, cảm biến vân tay, kết nối WiFi, chạm hoặc phát hiện khuôn mặt của điện thoại.
BATTERY	Các ý kiến mô tả dung lượng pin hoặc chất lượng pin.
PERFORMANCE	Các đánh giá mô tả khả năng ram, bộ xử lý, hiệu suất sử dụng hoặc độ trơn tru của điện thoại.
STORAGE	Các ý kiến đề cập đến dung lượng lưu trữ, khả năng mở rộng công suất thông qua thẻ nhớ.
DESIGN	Các đánh giá đề cập đến phong cách, thiết kế hoặc vỏ.
PRICE	Các ý kiến trình bày về giá của điện thoại.
GENERAL	Các đánh giá chung về điện thoại.
SER&ACC	Các ý kiến đề cập đến dịch vụ bán hàng, bảo hành hoặc đánh giá các phụ kiện của điện thoại.

BẢNG 3.1: Danh sách các khía cạnh và định nghĩa [19].

Bộ dữ liệu được chia ngẫu nhiên thành ba tập dữ liệu: train, dev, test với tỉ lệ 7:1:2. Sự phân phối của các khía cạnh kết hợp với cảm xúc được thể hiện ở hình 3.1 và bảng 3.2:



HÌNH 3.1: Sự phân phối nhãn trong bộ dữ liệu UIT-ViSD4SA [27].

Tập	Số bình luận	Trung bình cộng số đoạn có ý nghĩa trên một bình luận	Trung bình độ dài đoạn	Tích cực	Tiêu cực	Trung tính	Tổng cộng
Train	7,784	3.2	32.6	15,356	7,793	1,560	35,396
Dev	1,113	3.1	32.4	2,110	1,144	241	
Test	2,225	3.2	32.5	4,266	2,269	413	

BẢNG 3.2: Số liệu thống kê tổng quan của bộ dữ liệu UIT-ViSD4SA [17].

Ví dụ minh họa bình luận được gán nhãn trong bộ dữ liệu UIT-ViSD4SA được thể hiện ở hình 3.2.

	Gold labels	Aspect prediction	Polarity prediction	Aspect#polarity prediction
1	tôi cảm thấy, **loa có tiếng gì đó phát ra**FEATURES#NEGATIVE, mặc dù k chạm vào điện thoại.còn lại **in trâu**BATTERY#POSITIVE, **màn nét**SCREEN#POSITIVE, **chơi game ấm**PERFORMANCE#NEGATIVE, nhưng **loa dè**FEATURES#NEGATIVE *i feel,* ***there're some sound from the speaker****, even though I don't touch the phone.the rest is* ***battery last long****,* ***the screen is clear****,* ***play game phone is warm****, but* ***noisy speaker***	màn nétSCREEN chơi game ấmPERFORMANCE loa dèFEATURES	màn nét, chơi game ấm, nhưng loa dèNEUTRALX	in trâuBATTERY#POSITIVE màn nétSCREEN#POSITIVE chơi game ấmPERFORMANCE#POSITIVEX loa dèFEATURES#NEGATIVE
2	Sử dụng hơn 3 tháng thấy **máy rất tốt**GENERAL#POSITIVE , **dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầy**BATTERY#POSITIVE. **Rất thích dark mode**PERFORMANCE#POSITIVE. *Using more than 3 months find that* ***the device is really goo****,* ***using till 2 days to need to charge , take 2-3 hours to full****.* ***Really like the dark mode****.*	máy rất tốtGENERAL dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầyBATTERY Rất thích dark mode CAMERAX	máy rất tốtPOSITIVEL dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầyPOSITIVE Rất thích dark mode POSITIVE	máy rất tốtGENERAL#POSITIVE dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầyBATTERY#POSITIVE Rất thích dark mode FEATURES#POSITIVEX

HÌNH 3.2: Bình luận được gán nhãn theo khía cạnh cảm xúc trong bộ dữ liệu UIT-ViSD4SA [17].

3.1.2 Bộ dữ liệu UIT-ViQuAD 1.0

Bộ dữ liệu tiếng Việt UIT-ViQuAD 1.0 [17] bao gồm hơn 23.000 cặp câu hỏi - câu trả lời do con người tạo ra dựa trên 5.109 đoạn văn của 174 bài viết tiếng Việt từ Wikipedia. Bộ dữ liệu phù hợp để đánh giá các mô hình MRC trên tiếng Việt.

Các bài viết tiếng Việt từ trang web Wikipedia được chọn bằng cách: sử dụng Project Nayuki's Wikipedia's internal PageRanks để có được một bộ 5.000 bài viết hàng đầu của Việt Nam, từ đó chọn ngẫu nhiên các bài viết để tạo dữ liệu.

Cấu trúc mỗi dữ liệu bao gồm:

Đoạn văn: Mỗi đoạn văn là một đoạn tương ứng với một đoạn văn trong một bài viết. Hình ảnh, mục lục và bảng được loại trừ. Các đoạn văn ngắn hơn 300 ký tự hoặc chứa nhiều ký tự đặc biệt sẽ bị loại bỏ.

Câu hỏi: Mỗi câu hỏi được con người đặt ra dựa trên đoạn văn.

Câu trả lời: Câu trả lời cho câu hỏi là một khoảng thuộc đoạn văn.

Mô tả điển hình cho bài toán được thể hiện ở hình 3.3.

Passage: Nước biển có độ mặn không đồng đều trên toàn thế giới mặc dù phần lớn có độ mặn nằm trong khoảng từ **3,1%** tới 3,8%. Khi sự pha trộn với nước ngọt đổ ra từ các con sông hay gần các sông băng đang tan chảy thì nước biển nhạt hơn một cách đáng kể. Nước biển nhạt nhất có tại **vịnh Phần Lan**, một phần của biển Baltic.
(**English**: Seawater has uneven salinity throughout the world although most salinity ranges from **3.1%** to 3.8%. When the mix with freshwater pouring from rivers or near glaciers is melting, the seawater is significantly lighter. The lightest seawater is found in the **Gulf of Finland**, a part of the Baltic Sea.)

Question: Độ mặn thấp nhất của nước biển là bao nhiêu? (**English**: What is the lowest salinity of seawater?)
Answer: **3.1%** (**English**: 3.1%)

Question: Nước biển ở đâu có hàm lượng muối thấp nhất? (**English**: Where is the lowest salt content?)
Answer: **Vịnh Phần Lan**. (**English**: Gulf of Finland.)

HÌNH 3.3: Một dữ liệu trong bộ dữ liệu UIT-ViQuAD 1.0 [17].

Các thông tin tổng quát của bộ dữ liệu UIT-ViQuAD 1.0 được thể hiện chi tiết qua bảng 3.3.

	Train	**Dev**	**Test**	**All**
Number of articles	138	18	18	174
Number of pasages	4.101	515	493	5.109
Number of questions	18.579	2.285	2.210	**23.074**
Average passage length	153,9	147,9	155,0	153,4
Average question length	12,2	11,9	12,2	12,2
Average answer length	8,1	8,4	8,9	8,2
Vocabulary size	36.174	9.184	9.792	41.773

BẢNG 3.3: Tổng quan bộ dữ liệu UIT-ViQuAD 1.0 [17].

Số liệu thống kê về bộ dữ liệu UIT-ViQuAD 1.0 theo ba loại độ dài bao gồm độ dài câu hỏi (bảng 3.4), độ dài câu trả lời (bảng 3.4) và độ dài đoạn văn (bảng 3.5)

Đọ dài	**Câu hỏi**				**Câu trả lời**			
	Train	Dev	Test	All	Train	Dev	Test	All
1-5	1,03	1,44	0,95	1,06	**54,12**	**50,63**	**52,26**	**53,60**
6-10	35,99	38,38	33,21	35,96	19,95	22,14	19,10	20,08
11-15	**44,97**	**44,29**	**49,05**	**45,29**	10,86	10,81	10,81	10,85
16-20	15,01	13,61	14,07	14,78	6,28	7,48	6,83	6,45
>20	3,00	2,28	2,71	2,90	8,80	8,93	11,00	9,02

BẢNG 3.4: Thống kê độ dài câu hỏi và câu trả lời trên bộ dữ liệu UIT-ViQuAD 1.0 [17].

Length	**Passage**			
	Train	**Dev**	**Test**	**All**
<101	11,41	10,10	11,16	11,25
101-150	**47,50**	**53,59**	**45,44**	**47,92**
151-200	24,99	23,69	28,60	25,21
201-250	9,41	8,93	9,94	9,41
251-300	4,02	2,52	1,83	3,66
>300	2,66	1,17	3,04	2,54

BẢNG 3.5: Thống kê độ dài đoạn văn trên bộ dữ liệu UIT-ViQuAD 1.0 [17].

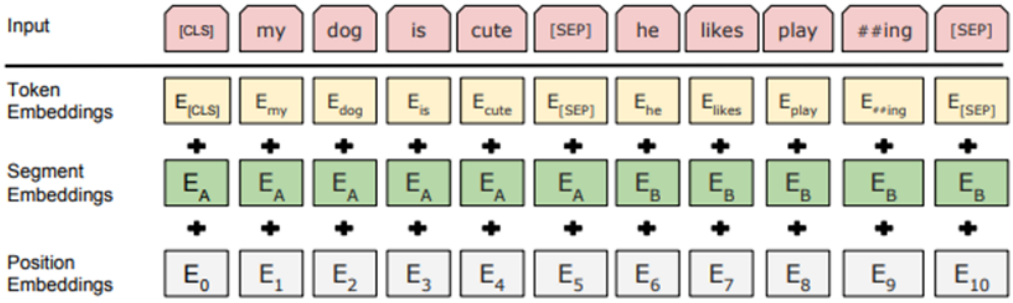
3.2 Mô hình

3.2.1 Mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) là một kiến trúc rất nổi tiếng trong lĩnh vực NLP được đề xuất bởi Devlin và cộng sự vào năm 2018 [34]. Nhiệm vụ của BERT là biểu diễn không giám sát các mối quan hệ giữa các từ trong ngữ cảnh hai chiều thông qua các vector từ. Tính tới thời điểm hiện tại BERT vẫn được đánh giá là một trong những mô hình tốt nhất trong việc biểu diễn từ, và mô hình tiên tiến này đạt được hiệu suất tốt trên các bộ dữ liệu như SQuAD [23].

BERT có kiến trúc đa tầng, bao gồm nhiều lớp Bidirectional Transformer Encoder. Các lớp này có nhiệm vụ mã hóa và giải mã các từ, vị trí và quan hệ giữa

các từ trong câu. Để biểu diễn các đầu vào, BERT sử dụng WordPiece embedding để biểu diễn các từ và các ký tự đặc biệt để đánh dấu các vị trí trong câu, cụ thể Hình 3.4.



HÌNH 3.4: Biểu diễn từ đầu vào mô hình Bert [34].

Với mỗi input đầu vào, dữ liệu sẽ được xử lý thông qua 3 quá trình:

- Token Embedding: Đầu vào được đánh dấu theo quy ước thêm ký tự [CLS] làm ký tự bắt đầu, thêm [SEP] ở vị trí phân tách giữa các câu và cuối câu. Các từ sau khi được embedding thông qua WordPiece, sẽ được tách thành các thành phần ví dụ: playing => (play, ##ing)
- Segment Embedding: Các token sẽ được đánh dấu thuộc thành phần nào trong câu
- Position Embedding: Cuối cùng, tất cả chúng được đánh dấu thứ tự phục vụ cho quy trình tiếp theo.

Mỗi mô hình BERT thực hiện 2 nhiệm vụ chính, bao gồm:

- Masked LM: che giấu đi một số token đầu vào một cách ngẫu nhiên và sau đó chỉ dự đoán các token được giấu đi đó.

- Next Sentence Prediction: dự đoán thứ tự giữa 2 câu bất kỳ trong văn bản, xem xét mối quan hệ giữa 2 câu để biết chúng có phải là những câu liên tiếp nhau hay không. Nhiệm vụ này phục vụ cho mục đích tìm kiếm và hiểu nội dung kết hợp từ nhiều câu.

3.2.2 Mô hình RoBERTa

Mô hình được Liu và cộng sự công bố vào năm 2019 [12] nhằm tối ưu hóa việc đào tạo kiến trúc BERT để mất ít thời gian hơn trong quá trình pre-training.

Về kiến trúc, được thiết kế tương tự như BERT, đều sử dụng kiến trúc mã hóa (encoder) của Transformer với hàm kích hoạt GELU làm kiến trúc chính, sử dụng kiến trúc self-attention hai chiều để hiểu bối cảnh của một từ. Nhưng đối với RoBERTa, có những điểm thay đổi so với BERT:

- Loại bỏ mục tiêu Next Sentence Prediction. Các tác giả đã thử nghiệm với việc thêm hay loại bỏ Next Sentence Prediction vào các phiên bản khác nhau và kết luận rằng việc loại bỏ Next Sentence Prediction sẽ cải thiện một chút hiệu suất tác vụ.

- Đào tạo với batch size lớn hơn và sequences dài hơn: Ban đầu BERT được đào tạo cho 1 triệu step với batch size là 256 sequences. Đối với RoBERTa, các tác giả đã đào tạo mô hình với 125 step với batch size là 2.000 sequences và 31.000 steps với batch size 8.000 sequences. Điều này có hai lợi thế, batch size lớn cải thiện sự khó hiểu trên masked language modeling objective và cũng như độ chính xác của tác vụ cuối cùng (end-task). Batch size lớn cũng dễ dàng thực hiện song song thông qua đào tạo song song phân tán hơn.

- Tự động thay đổi masking pattern: trong cấu trúc của BERT, việc tạo mask chỉ diễn ra một lần duy nhất ở bước tiền xử lý dữ liệu. Làm cho một mask

trở thành mask duy nhất và ảnh hưởng đến khả năng mô hình. Để tránh việc đó nhóm tác giả đã sử dụng tự động thay đổi masking pattern.

3.2.3 Mô hình PhoBert

PhoBERT [16] hiện nay đang là mô hình đào tạo trước được xem là đi đầu xu hướng (State Of The Art-SOTA) đối với các mô hình được đào tạo trước trên tiếng Việt. Nó được huấn luyện đơn ngữ, chỉ huấn luyện dành riêng cho tiếng Việt. Việc huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa của Facebook được Facebook giới thiệu giữa năm 2019 là một cải tiến so với BERT trước đây.

Tương tự như BERT, PhoBERT cũng có hai phiên bản là PhoBERT base với mười hai transformers block và PhoBERT large với hai mươi bốn transformers block.

PhoBERT được huấn luyện trên khoảng 20GB dữ liệu bao gồm 1GB trên tập dữ liệu trên trang Wikipedia dành cho tiếng Việt và 19GB còn lại lấy từ tập dữ liệu báo tiếng Việt (Vietnamese News). Đây là một lượng dữ liệu đủ lớn để huấn luyện một mô hình về đào tạo trước.

Vì PhoBERT dựa trên RoBERTa nên chỉ sử dụng tác vụ Masked Language Model để train, bỏ đi tác vụ Next Sentence Prediction.

Ngoài ra, PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder.

Mô hình PhoBERT có kiến trúc tương tự như RoBERTa mà RoBERTa lại được phát triển từ BERT vì thế chúng tôi sẽ đề cập đến BERT và RoBERTa trong nội dung này.

3.2.4 Mô hình XLM-R

Trái với PhoBERT được đề cập ở phần trên, XLM-RoBERTa (XLM-R) [4] được đề xuất nhằm hướng tới một mô hình đạt hiệu suất cao trong hiểu biết đa

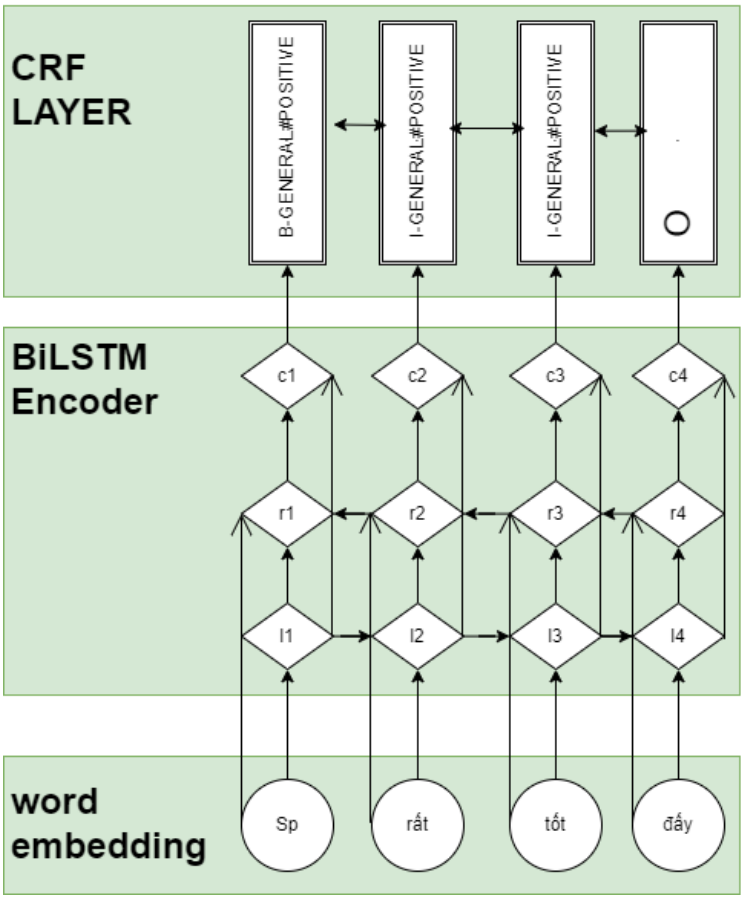
ngôn ngữ. XLM-R được đề xuất bởi Conneau và cộng sự vào 2019. XLM-RoBERTa là phiên bản đa ngôn ngữ của RoBERTa. Nó được đào tạo trước trên 2.5TB dữ liệu CommonCrawl đã lọc trên 100 ngôn ngữ.

XLM-R sử dụng các kỹ thuật đào tạo tự giám sát để đạt được hiệu suất cao trong hiểu biết đa ngôn ngữ, một nhiệm vụ trong đó là mô hình được đào tạo bằng một ngôn ngữ và sau đó được sử dụng với các ngôn ngữ khác mà không cần bổ sung dữ liệu đào tạo. Về kiến trúc, XLM-R được tạo nên từ XLM và RoBERTa [12]. XLM-R hoạt động rất tốt trên cả các ngôn ngữ có nguồn tài nguyên thấp như tiếng Việt.

RoBERTa là một mô hình transformer được đào tạo trước trên một tập dữ liệu lớn theo cách tự giám sát. Điều này có nghĩa là nó chỉ được đào tạo trước trên các văn bản thô, không có con người ghi nhãn chúng theo bất kỳ cách nào (đó là lý do tại sao nó có thể sử dụng nhiều dữ liệu có sẵn công khai) với một quy trình tự động để tạo đầu vào và nhãn từ các văn bản đó. Chính xác hơn, nó đã được đào tạo trước với mục tiêu mô hình hóa ngôn ngữ được che giấu (MLM). Lấy một câu, mô hình che dấu ngẫu nhiên 15% số từ trong đầu vào, sau đó chạy toàn bộ câu được che thông qua mô hình và phải dự đoán các từ đã che. Điều này khác với các mạng neural truyền thống (RNN) thường thấy các từ lần lượt. Nó cho phép mô hình học cách biểu diễn hai chiều của câu. Bằng cách này, mô hình học cách biểu diễn bên trong của 100 ngôn ngữ mà sau đó có thể được sử dụng để trích xuất các tính năng hữu ích cho các tác vụ hạ lưu: ví dụ: nếu bạn có một tập dữ liệu gồm các câu được gắn nhãn, bạn có thể đào tạo một bộ phân loại tiêu chuẩn bằng cách sử dụng các tính năng do XLM-RoBERTa làm đầu vào.

3.2.5 Mô hình BiLSTM-CRF

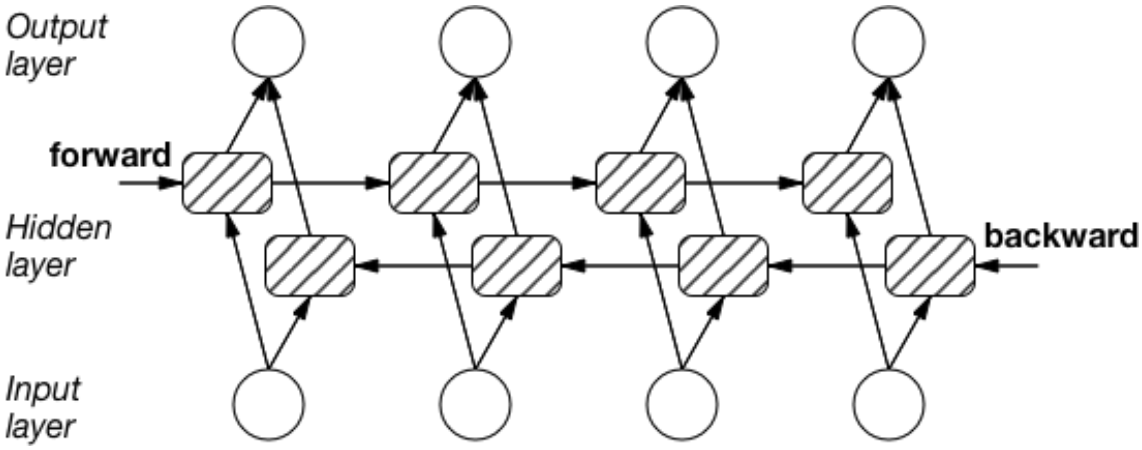
Kiến trúc mô hình BiLSTM-CRF cho bài toán nhận diện cảm xúc theo khía cạnh được mô tả như hình 3.5.



HÌNH 3.5: Mô tả mô hình BiLSTM-CRF cho bài toán ABSA.

Word Embedding (nhúng từ): Lớp này sẽ chuyển đổi mỗi từ thành một vectơ với kích thước cố định. Trong khóa luận này chúng tôi sử dụng nhúng từ được đào tạo trước là PhoW2V với hai phiên bản 100 chiều và 300 chiều.

Bộ mã hóa BiLSTM (Bidirectional Long Short-Term Memory): Bộ mã hóa này chứa hai lớp LSTM (Long Short-Term Memory) và có chức năng tìm hiểu thông tin từ ngữ cảnh.

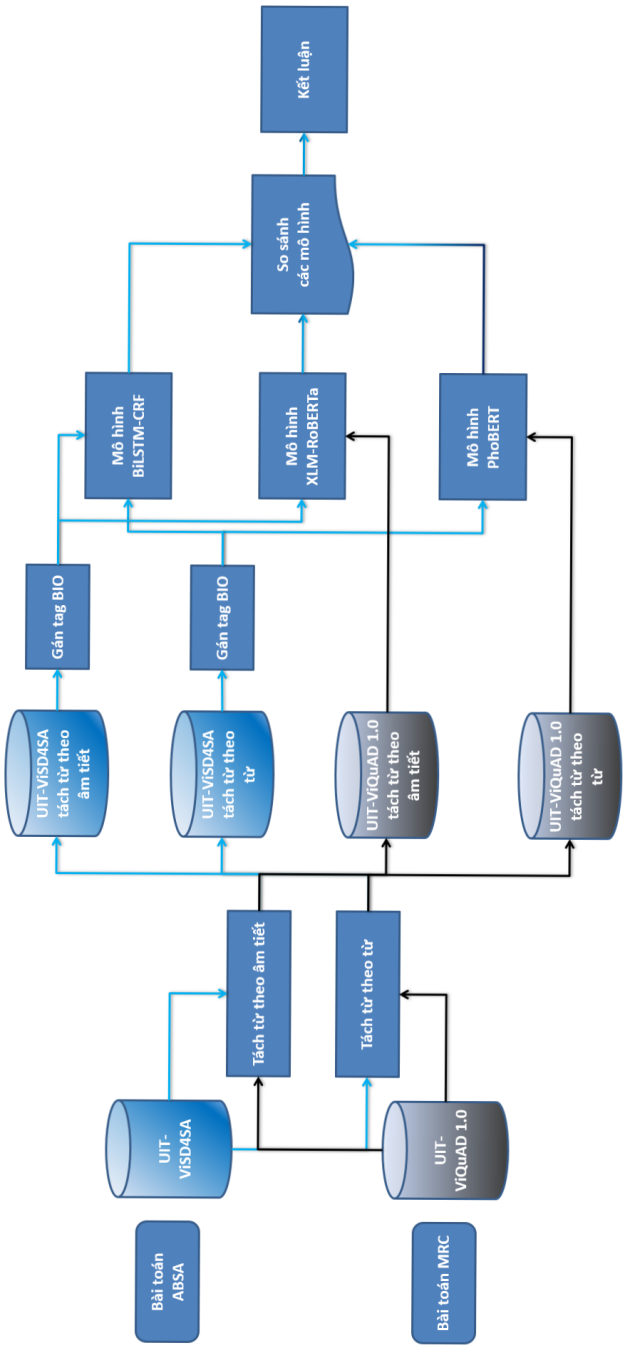


HÌNH 3.6: Mô hình BiLSTM.

Conditional Random Field (CRF) [31] là một mô hình xác suất cho các bài toán dự đoán có cấu trúc và đã được áp dụng rất thành công trong rất nhiều lĩnh vực như thị giác máy tính, xử lý ngôn ngữ tự nhiên, sinh-tin học,... Trong mô hình CRF, các nút (node) chứa dữ liệu đầu vào và các node chứa dữ liệu đầu ra được kết nối trực tiếp với nhau, đối nghịch với kiến trúc của LSTM hoặc BiLSTM trong đó các đầu vào và đầu ra được kết nối gián tiếp qua các ô nhớ (memory cell). Lớp này có thể thêm một số hạn chế vào các nhãn dự đoán cuối cùng để đảm bảo chúng hợp lệ. Những hạn chế này có thể được học bởi lớp CRF tự động từ tập dữ liệu đào tạo trong quá trình đào tạo.

3.3 Quy trình thực hiện

Quy trình thực hiện được mô tả chi tiết như sơ đồ 3.7.



HÌNH 3.7: Quy trình thực hiện.

3.4 Phương pháp đánh giá

3.4.1 Bài toán MRC

Để đánh giá hiệu suất của mô hình, đối với bài toán MRC chúng tôi sử dụng hai độ đo là Exact Match (EM) và F1-Score (F1) dựa trên các đánh giá trên bộ SQuAD [23] hay chi tiết các độ đo được trình bày cụ thể trong nghiên cứu của Zeng và cộng sự [35].

3.4.1.1 Exact Match (EM)

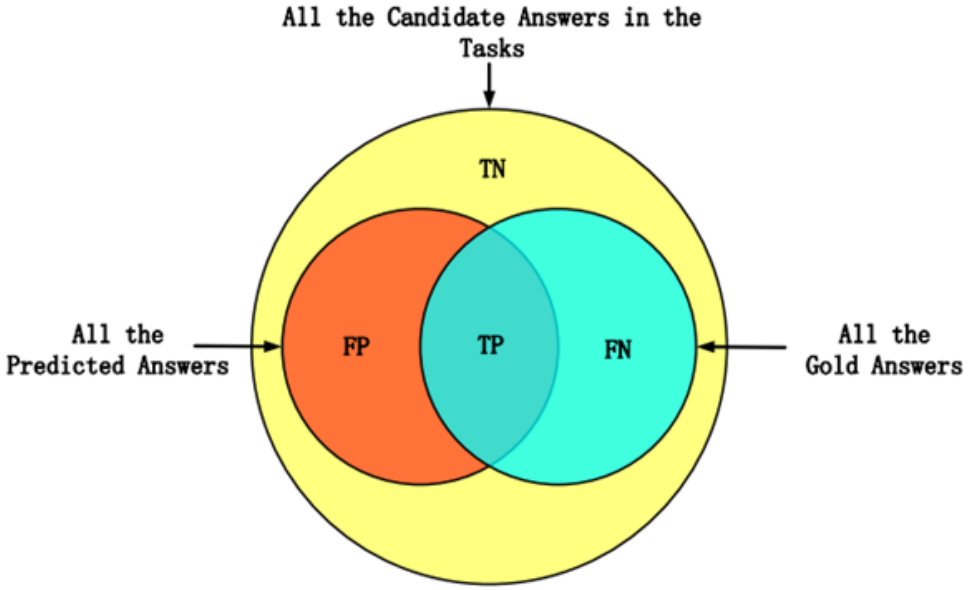
Nếu câu trả lời đúng cho câu hỏi là một câu hoặc một cụm từ, có thể một số từ trong câu trả lời do hệ thống tạo ra là câu trả lời đúng và các từ khác không phải là câu trả lời đúng. Trong trường hợp này, Exact Match đại diện cho tỷ lệ phần trăm câu hỏi mà câu trả lời do hệ thống tạo ra hoàn toàn khớp với câu trả lời đúng, có nghĩa là mọi từ đều giống nhau. Kết hợp chính xác thường được viết tắt là EM. Ví dụ: nếu một tác vụ MRC chứa N câu hỏi, mỗi câu hỏi tương ứng với một câu trả lời đúng, câu trả lời có thể là một từ, cụm từ hoặc câu và số câu hỏi mà hệ thống trả lời đúng là M. Exact Match sau đó có thể được tính như sau:

$$\text{EM} = \frac{M}{N}$$

3.4.1.2 F1-score

Số liệu này đo lường sự chồng chéo giữa dự đoán và câu trả lời đúng. Theo phương pháp đánh giá trong SQuAD, coi câu trả lời dự đoán và câu trả lời đúng là túi token, trong khi bỏ qua tất cả các dấu chấm câu và các từ bài viết như "a" và "an" hoặc "the". F1-score được tính thông qua hai độ đo khác là Precision và Recall, chi tiết của các độ đo chúng tôi sẽ trình bày trong phần bên dưới:

- Precision: Thể hiện phần trăm trùng lặp giữa từ trong câu trả lời thật và câu trả lời dự đoán. Để có được Precision ở cấp độ từ, trước tiên chúng ta cần hiểu ý nghĩa của true positive (TP), false positive (FP), true negative (TN), and false negative (FN) như trong Hình 3.8



HÌNH 3.8: Biểu diễn trùng lặp từ giữa câu trả lời đúng và trả lời được dự đoán.

Hình 3.8 mô tả:

- TP thể hiện các từ giống nhau giữa câu trả lời dự đoán và câu trả lời đúng.
- FP thể hiện các từ không có trong câu trả lời đúng nhưng trong câu trả lời dự đoán.
- FN thể hiện các từ không có trong câu trả lời dự đoán mà là câu trả lời đúng.

Precision ở cấp độ từ cho một câu hỏi được tính như sau:

$$\text{Precision} = \frac{\text{Num}(TP)}{\text{Num}(TP) + \text{Num}(FP)}$$

Trong đó Num (TP) là số lượng từ TP, Num (FP) là số lượng từ FP.

- Recall: Đại diện cho tỷ lệ phần trăm từ trong câu trả lời đúng đã được dự đoán chính xác trong một câu hỏi.

$$\text{Recall } = \frac{\text{Num}(TP)}{\text{Num}(TP) + \text{Num}(FN)}$$

Trong đó Num (TP) là số lượng từ TP, Num (FN) là số lượng từ FN. F1-score được tính như sau:

$$F1 = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

3.4.2 Bài toán ABSA

Để đánh giá hiệu suất của mô hình, đối với bài toán ABSA chúng tôi sử dụng độ đo marco F1-score dựa trên các đánh giá trên bộ UIT-ViSD4SA [27].

3.4.2.1 Macro F1-score

Sự cân bằng không đồng đều về các nhãn/lớp của một bộ dữ liệu là một vấn đề lớn đối với các bài toán nhiều nhãn/lớp. Dẫn đến sự ảnh hưởng xấu giữa những nhãn/lớp có số lượng lớn với các nhãn/lớp có số lượng ít.

Marco F1-score là một phiên bản của F1-score được sử dụng để đánh giá chất lượng của các vấn đề với nhiều nhãn nhị phân hoặc nhiều lớp. Giải quyết vấn đề sự ảnh hưởng xấu của phân bố nhãn không đồng đều gây ra.

Marco F1-score được tính bằng trung bình F1 của các nhãn/lớp có trong bộ dữ liệu với trọng lượng của mỗi F1 của nhãn/lớp là như nhau. Công thức tính Marco F1-score như sau:

$$\text{Macro-F } = \frac{1}{N} \sum_{i=0}^{N} F_i$$

Trong đó:

- i là chỉ số nhãn/lớp của bộ dữ liệu
- N là tổng số nhãn/lớp của bộ dữ liệu

Đối với hai thành phần cốt lõi của F1-score là Precision và Recall. Công thức tính phiên bản Marco như sau

$$\text{Macro-P } = \frac{1}{N}\sum_{i=0}^{N} P_i$$

$$\text{Macro-R } = \frac{1}{N}\sum_{i=0}^{N} R_i$$

Trong đó:

- i là chỉ số nhãn/lớp của bộ dữ liệu
- N là tổng số nhãn/lớp của bộ dữ liệu

3.5 Thông số cài đặt mô hình

Khi thực hiện các thử nghiệm trong khóa luận này, chúng tôi sử dụng GPU được cung cấp bởi Google (phiên bản sử dụng Google Colab Pro).

3.5.1 Bài toán ABSA

Để xây dựng bài toán nhận diện cảm xúc theo khía cạnh, chúng tôi tiến hành xử lý dữ liệu theo định dạng IOB (Inside Outside Beginning).

3.5.1.1 Định dạng IOB

IOB là định dạng gắn thẻ được Ramshaw và Marcus [24] giới thiệu năm 1995. Trong bài toán nhận dạng cảm xúc theo khía cạnh, chúng tôi sử dụng định dạng IOB để xử lý dữ liệu nhãn. Mô tả dữ liệu nhãn được mô tả ở phần phụ lục A .

3.5.1.2 Tham số mô hình PhoBERT và mô hình XLM-R

Khi thực hiện xây dựng mô hình PhoBERT cho phiên bản tách theo từ và mô hình XLM-R cho phiên bản tách theo tiếng, chúng tôi sử dụng T-NER [28] là công cụ được viết bằng Python sử dụng cho tác vụ tinh chỉnh mô hình ngôn ngữ nhận dạng thực thể được triển khai trên Pytorch. Các tham số được sử dụng để tinh chỉnh được mô tả như bảng sau.

Mô hình PhoBERT&XLM-R	Tham số
random_seed	48
learning_rate	1e-5
total_step	2.000
warmup_step	10
batch_size	16
max_seq_length	300

BẢNG 3.6: Tham số mô hình PhoBERT và XLM-R cho bài toán ABSA.

3.5.1.3 Tham số mô hình BiLSTM-CRF

Mô hình BiLSTM-CRF	Tham số
seed	48
threads	1.200
min_freq	120
bucket	32
batch_size	5.000
embed_dropout	0.33
epoch	30
patient	10

BẢNG 3.7: Tham số mô hình BiLSTM-CRF.

3.5.2 Bài toán MRC

3.5.2.1 Mô hình PhoBERT và mô hình XLM-R

Đối với bài toán MRC, khi thực hiện xây dựng mô hình PhoBERT cho phiên bản tách theo từ và mô hình XLM-R cho phiên bản tách theo tiếng, các tham số tinh chỉnh của chúng tôi như bảng 3.8.

PhoBERT & XLM-R	Tham số
batch_size	8
learning_rate	2e-5
max_seq_lenght	256
doc_stride	81
max_query_length	81

BẢNG 3.8: Tham số mô hình PhoBERT và XLM-R cho bài toán MRC.

3.6 Kết luận

Trong chương này chúng tôi đã thực hiện phân tích bộ dữ liệu sử dụng trong khóa luận là UIT-ViSD4SA cho bài toán nhận diện cảm xúc theo khía cạnh

và UIT-ViQuAD 1.0 cho bài toán hệ thống hỏi đáp. Chúng tôi cũng đưa ra cơ sở lý thuyết mà chúng tôi đã nghiên cứu để thực hiện khóa luận này, gồm có các mô hình đào tạo trước không chỉ phù hợp với vấn đề mà chúng tôi quan tâm mà nó còn đang được đánh giá là mô hình SOTA hiện nay. Bên cạnh đó chúng tôi cũng đưa ra cơ sở lý thuyết của mô hình BiLSTM-CRF, mô hình chúng tôi sử dụng kết hợp với nhúng từ được đào tạo trước để có thể đưa ra kết luận khách quan hơn về tác động của các kỹ thuật tách từ. Trong chương này, chúng tôi cũng mô tả quy trình thực hiện một cách tổng quan và cũng đưa ra các thông số chi tiết được sử dụng để xâ dựng mô hình.Ngoài ra, các độ đo được sử dụng để đánh giá kết quả mô hình cũng được mô tả chi tiết ở chương này.

Chương 4. KẾT QUẢ

4.1 Kết quả tổng quan

Trong phần này, chúng tôi sẽ trình bày tổng quan các kết quả thử nghiệm với các mô hình PhoBERT và XLM-RoBerta dựa trên hai phương pháp tách từ (tách từ theo từ và tách từ theo âm tiết) đối với hai bài toán MRC và ABSA. Ngoài ra chúng tôi còn đưa ra kết quả thử nghiệm cho mô hình BiLSTM-CRF kết hợp với pretrain word embedding PhoW2V cho bài toán ABSA.

4.1.1 Bài toán ABSA

(Chú thích: R: Recall, P: Precision, F: F1-score.)

Với bài toán ABSA, bảng số liệu dưới đây thể hiện kết quả tổng quan của hai tập dữ liệu (tập phát triển-Dev Set và tập kiểm tra-Test Set) trên ba bài toán:bài toán nhận diện cảm xúc dựa trên khía cạnh (Aspect_Polarity), bài toán nhận diện khía cạnh (Aspect) và bài toán nhận diện cảm xúc (Polarity) theo ba độ đo recall, precision và F1-score.

(%)			Dev Set			Test Set		
			P(macro)	R(macro)	F(macro)	P(macro)	R(macro)	F(macro)
Aspect_Polarity	Syllable	XLM-R	42,65	54,28	47,77	42,65	53,93	47,63
		BiLSTM-CRF(PhoW2V)	64,70	63,75	64,07	62,70	62,90	62,72
	Word	PhoBERT	44,07	50,51	47,07	44,33	51,99	47,85
		BiLSTM-CRF(PhoW2V)	60,86	59,13	59,85	61,56	58,86	60,05
Aspect	Syllable	XLM-R	44,32	59,17	50,68	42,52	57,60	48,93
		BiLSTM-CRF(PhoW2V)	63,70	61,91	62,67	64,46	62,85	63,57
	Word	PhoBERT	45,69	58,48	51,30	44,86	58,58	50,81
		BiLSTM-CRF(PhoW2V)	63,78	60,46	61,94	65,66	63,11	64,29
Polarity	Syllable	XLM-R	42,37	56,46	48,41	41,56	55,52	47,54
		BiLSTM-CRF(PhoW2V)	57,03	55,05	56,02	56,84	54,81	55,80
	Word	PhoBERT	42,29	54,63	47,67	42,04	54,88	47,61
		BiLSTM-CRF(PhoW2V)	43,84	44,21	43,94	44,17	44,14	44,10

BẢNG 4.1: Kết quả tổng quan bài toán nhận diện cảm xúc theo khía cạnh.

4.1.2 Bài toán MRC

Đối với bài toán MRC, Bảng số liệu dưới đây thể hiện kết quả tổng quan của hai tập dữ liệu (tập phát triển-Dev Set và tập kiểm tra-Test Set) trên bài toán đọc hiểu tự động theo hai độ đo EM và F1-score.

Mô hình	**Dev Set**		**Test Set**	
	EM	**F1_score**	**EM**	**F1_score**
PhoBert_Large	66,85	85,50	63,32	83,42
PhoBert_Base	64,17	82,39	61,10	80,78
XLM-R_Large	**71,12**	**87,99**	**69,07**	**86,80**
XLM-R_Base	64,04	81,96	61,09	81,05

BẢNG 4.2: Kết quả tổng quan bài toán MRC.

4.2 Phân tích kết quả chi tiết

4.2.1 Bài toán ABSA

Trong phần này, chúng tôi sẽ tiến hành phân tích kết quả theo từng bài toán: nhận diện cảm xúc theo khía cạnh, nhận diện khía cạnh, nhận diện cảm xúc. Các kết quả dưới đây đều được đánh giá trên tập kiểm tra-Test Set.

Đối với các mô hình pre-train gồm có (XLM-R với kỹ thuật tách từ theo âm tiết và mô hình PhoBERT với kỹ thuật tách từ theo từ) thì trong cả ba bài toán là nhận diện cảm xúc theo khía cạnh, nhận diện khía cạnh và nhận diện cảm xúc mô hình PhoBERT đều cho kết quả tốt hơn. **Kỹ thuật tách theo từ đem lại kết quả tốt hơn cho cả ba bài toán là nhận diện cảm xúc theo khía cạnh, nhận diện khía cạnh và nhận diện cảm xúc đối với mô hình pretrain(PhoBERT).**

Tuy nhiên đối với mô hình BiLSTM-CRF, hiệu suất mô hình có sự khác biệt khi sử dụng các kỹ thuật tách từ khác nhau.

Mô hình BiLSTM-CRF kết hợp với pretrain word embedding PhoW2V có sự khác biệt rất lớn giữa các bài toán khác nhau. Đối với bài toán nhận diện khía cạnh, kỹ thuật tách theo từ cho kết quả tốt hơn. Tuy nhiên đối với bài toán nhận diện cảm xúc theo khía cạnh, theo kết quả thử nghiệm của chúng tôi, bài toán lại đạt kết quả tốt hơn khi tách từ theo âm tiết. Đối với bài toán nhận diện cảm xúc,

Sau đây là các phân tích chi tiết cũng như các kết qủa thử nghiệm theo từng bài toán.

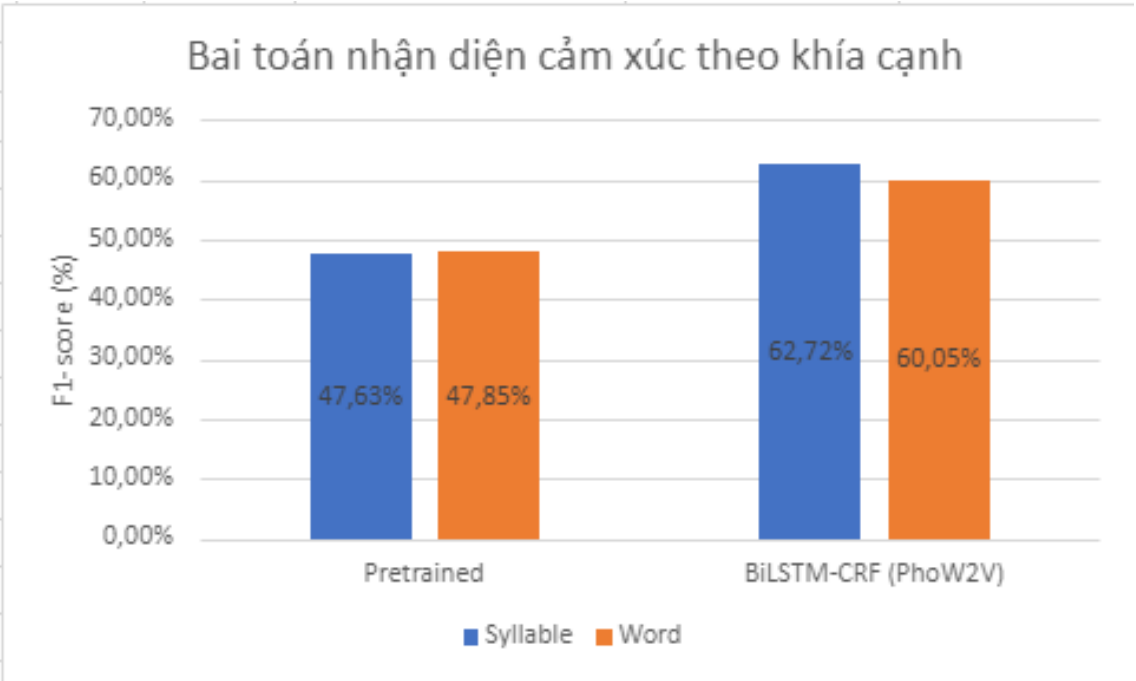
4.2.1.1 Bài toán nhận diện cảm xúc theo khía cạnh

Để hiểu rõ hơn các mô hình Nhận diện cảm xúc theo khía cạnh bị ảnh hưởng bởi các kỹ thuật tách từ như thế nào, chúng tôi tiến hành phân tích các mô hình trên ba độ đo precision, recall và f1-score.

Aspect_Polarity (%)		P(macro)	R(macro)	F(macro)
Syllable	XLM-R	42,65	53,93	47,63
	BiLSTM-CRF(PhoW2V)	62,70	62,90	62,72
Word	PhoBERT	44,33	51,99	47,85
	BiLSTM-CRF(PhoW2V)	61,56	58,86	60,05

BẢNG 4.3: Kết quả tổng quan bài toán nhận diện cảm xúc theo khía cạnh.

Mô hình nhận diện cảm xúc theo khía cạnh cho kết quả tích cực khi được áp dụng kỹ thuật tách theo từ, chúng ta có thể thấy hiệu suất mô hình đạt 62,72% với độ đo f1-score(macro), vượt trội hơn hẳn các mô hình còn lại. Thể hiện ở Hình4.1.



HÌNH 4.1: Phân tích hiệu suất mô hình trên bài toán nhận diện cảm xúc theo khía cạnh.

Biểu đồ cho thấy các mô hình BiLSTM-CRF kết hợp với các pretrain embedding cho kết quả tốt hơn so với các mô hình pre-train. Mô hình pre-train trong hình 4.1 là mô hình XLM-R đối với kỹ thuật tách từ theo âm tiết và mô hình PhoBERT đối với kỹ thuật tách từ theo từ.

Bài toán nhận diện cảm xúc theo khía cạnh trên đối với các mô hình pre-train thì kỹ thuật tách từ theo từ-tách từ theo tiếng Việt cho kết quả tốt hơn. Tuy nhiên sự chênh lệch giữa các mô hình không quá lớn. Mô hình XLM-R với kỹ thuật tách từ theo âm tiết đạt 47.63% đối với độ đo f1-score(macro), mô hình PhoBERT với kỹ thuật tách từ theo từ đạt 47.85% với cùng độ đo. Sự chênh lệch chỉ đạt 0.22%.

Tuy nhiên đối với mô hình BiLSTM-CRF, hiệu suất mô hình được cải thiện rất rõ rệt. Hiệu suất đạt cao nhất 62,72% khi kết hợp với pretrain word embedding PhoW2V phiên bản 100 chiều.

Khi thử nghiệm bài toán này, Kim và các cộng sự đã đề xuất nhiều phương pháp kết hợp với kỹ thuật tách từ theo âm tiết. Kết quả các mô hình được thể hiện ở bảng 4.4

System	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
Aspect-polarity (syllable)	64,87	54,55	57,98	48,77	34,27	37,64
Aspect-polarity (syllable + char)	59,51	57,56	58,52	43,66	37,53	39,30
Aspect-polarity (syllable + char + XLM-R-base)	60,71	61,62	61,16	46,18	43,42	44,37
Aspect-polarity (syllable + char + XLM-R-large)	61,78	62,99	**62,38**	46,84	45,46	**45,70**

BẢNG 4.4: Kết quả bài toán nhận diện cảm xúc theo khía cạnh được công bố từ Kim và các cộng sự [27].

Aspect_Polarity(%)	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
syllable	89,85	88,79	89,32	64,46	62,85	63,57
word	90,46	89,66	90,06	65,66	63,11	64,29

BẢNG 4.5: Kết quả bài toán nhận diện cảm xúc theo khía cạnh mô hình BiLSTM-CRF.

4.2.1.2 Bài toán nhận diện khía cạnh

Mô hình nhận diện cảm xúc theo khía cạnh cho kết quả tích cực khi được áp dụng kỹ thuật tách theo từ, chúng ta có thể thấy hiệu suất mô hình đạt 64,29% với độ đo f1-score(macro), vượt trội hơn hẳn các mô hình còn lại. Thể hiện ở Hình4.2.

Aspect (%)		P(macro)	R(macro)	F(macro)
Syllable	XLM-R	42,52	57,60	48,93
	BiLSTM-CRF(PhoW2V)	64,46	62,85	63,57
Word	PhoBERT	44,86	58,58	50,81
	BiLSTM-CRF(PhoW2V)	65,66	63,11	64,29

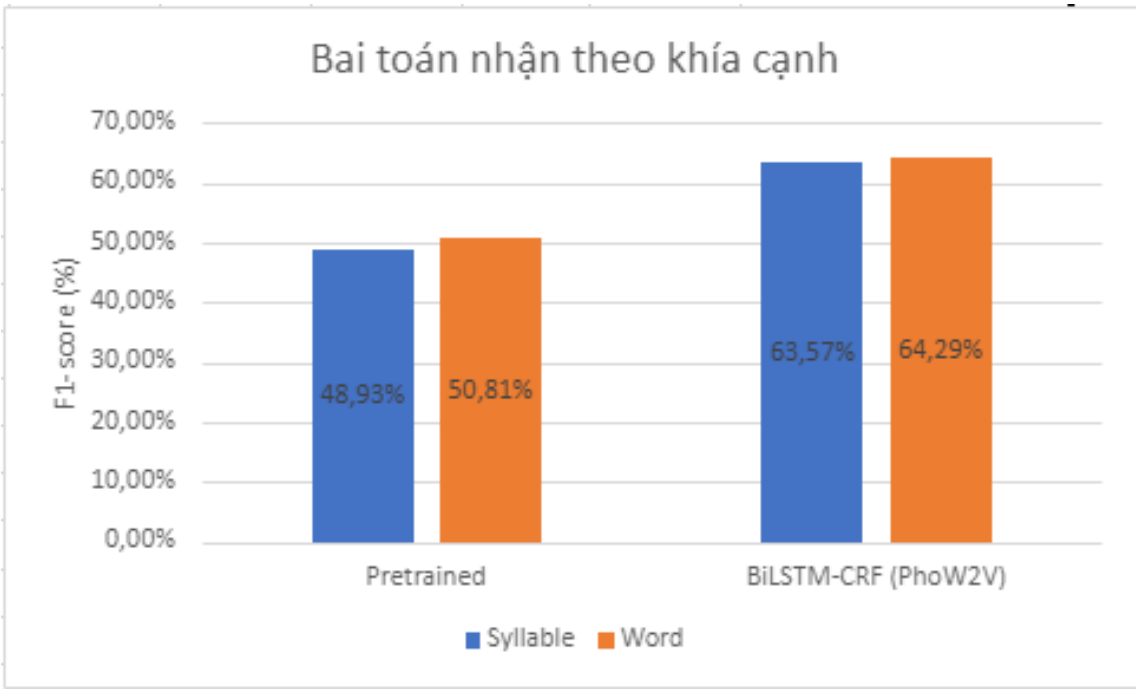
BẢNG 4.6: Kết quả tổng quan bài toán nhận diện khía cạnh.

Bài toán nhận diện khía cạnh trên đối với các mô hình pre-train thì kỹ thuật tách từ theo từ-tách từ theo tiếng Việt cho kết quả tốt hơn. Tuy nhiên sự chênh lệch giữa các mô hình không quá lớn. Mô hình XLM-R với kỹ thuật tách từ theo âm tiết đạt 48.93% đối với độ đo f1-score(macro), mô hình PhoBERT với kỹ thuật tách từ theo từ đạt 50.81% với cùng độ đo. Sự chênh lệch chỉ đạt 1.88%.

Tuy nhiên đối với mô hình BiLSTM-CRF, hiệu suất mô hình được cải thiện rất rõ rệt. Hiệu suất đạt cao nhất 64,29% khi kết hợp với nhúng từ phiên bản từ PhoW2V phiên bản 100 chiều.

Khi thử nghiệm bài toán này, Kim và các cộng sự đã đề xuất nhiều phương pháp kết hợp với kỹ thuật tách từ theo âm tiết. Kết quả các mô hình được thể hiện

ở bảng 4.7



HÌNH 4.2: Phân tích hiệu suất mô hình trên bài toán nhận diện khía cạnh.

System	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
Aspect (syllable)	64,55	60,86	62,65	62,76	57,28	59,74
Aspect(syllable + char)	63,78	62,11	62,93	61,64	58,91	60,21
Aspect(syllable + char + XLM-R-base)	65,63	65,15	65,39	62,88	61,62	62,17
Aspect(syllable + char + XLM-R-large)	64,96	66,85	**65,89**	62,00	63,56	**62,76**

BẢNG 4.7: Kết quả bài toán nhận diện khía cạnh được công bố từ Kim và các cộng sự [27].

Aspect(%)	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
syllable	88,66	88,64	88,65	62,70	62,90	62,72
word	89,33	86,76	88.02	61,56	58,86	60,05

BẢNG 4.8: Kết quả bài toán nhận diện khía cạnh mô hình BiLSTM-CRF.

4.2.1.3 Bài toán nhận diện cảm xúc

Polarity (%)		P(macro)	R(macro)	F(macro)
Syllable	XLM-R	41,56	55,52	47,54
	BiLSTM-CRF(PhoW2V)	56,84	54,81	55,80
Word	PhoBERT	42,04	54,88	47,61
	BiLSTM-CRF(PhoW2V)	44,17	44,14	44,10

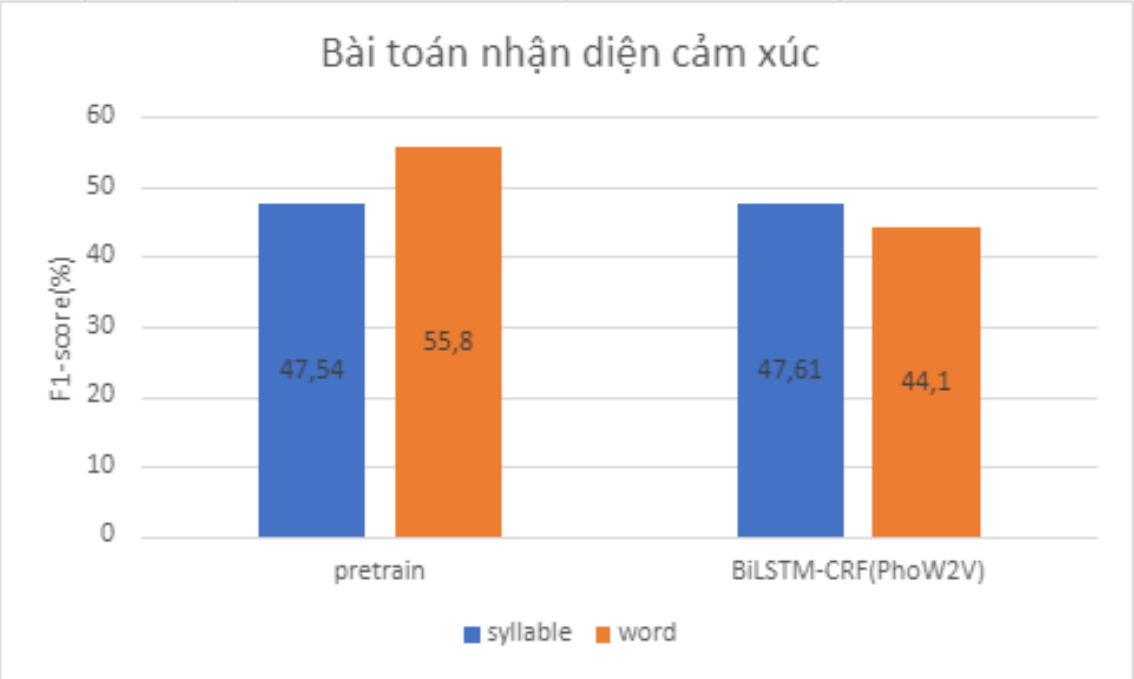
BẢNG 4.9: Kết quả tổng quan bài toán nhận diện cảm xúc.

Tuy nhiên, đối với bài toán nhận diện cảm xúc, kết quả thử nghiệm lại cho thấy ảnh hưởng đáng ngạc nhiên của mô hình khi sử dụng kỹ thuật tách từ theo âm tiết vớimô hình BiLSTM-CRF kết hợp với nhúng từ được đào tạo trước PhoW2V. Kết quả đem lại 55,80% cho mô hình này.

Đối với bài toán này, các mô hình pre-train cũng không tạo sự ảnh hưởng khác biệt rõ rệt đối với kỹ thuật tách từ. Tuy nhiên, đối với kỹ thuật tách từ theo từ vẫn cho kết quả tốt hơn 7%. Mô hình pre-train XLM-R với kỹ thuật tách từ theo âm tiết đạt 47.54% và mô hình PhoBERT với kỹ thuật tách từ theo từ đạt kết quả 47.61%

4.2.2 Bài toán MRC

Với bài toán MRC, các mô hình pretrained (XLM-RoBerta với kỹ thuật tách từ theo âm tiết và mô hình PhoBERT với kỹ thuật tách từ theo từ) được sử dụng thì



HÌNH 4.3: Phân tích hiệu suất mô hình trên bài toán nhận diện cảm xúc.

System	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
Polarity (syllable)	52.36	50.10	51.20	46.71	38.37	41.05
Polarity(syllable + char)	52.12	51.00	51.55	44.44	38.79	40.68
Polarity(syllable + char + XLM-R-base)	54.88	55.91	55.39	46.87	46.39	46.57
Polarity(syllable + char + XLM-R-large)	56.89	59.78	**58.30**	49.00	50.60	**49.77**

BẢNG 4.10: Kết quả bài toán nhận diện khía cạnh được công bố từ Kim và các cộng sự [27].

mô hình XLM-RoBerta cho kết quả tốt nhất vơi độ đo EM là 61.09% và F1-score là 86.80% trên tập test với phiên bản XLM-R-large. Lơn hơn phiên bản PhoBERT-large là 5.75% với độ đo EM và 3.38% với độ đo F1-score.

Kết quả này là tương đồng với kết quả của Rust và cộng sự đưa ra đối với một loại ngôn ngữ đơn lập giống tiếng việt đó là ngôn ngữ indonesia.

Polarity(%)	P (micro)	R (micro)	F (micro)	P (macro)	R (macro)	F (macro)
syllable	96,97	94,85	95,90	56,84	54,81	55,80
word	89,50	91,72	90,60	44,17	44,14	44,10

BẢNG 4.11: Kết quả bài toán nhận diện cảm xúc mô hình BiLSTM-CRF.

Sự kết hợp từ theo đặc điểm ngôn ngữ cụ thể (ở đây là tiếng việt), giúp các từ mới tạo ra mang ý nghĩa rõ ràng hơn. Tuy nhiên khi kết hợp nhiều từ mới lại với nhau sẽ dẫn đến những trường hợp gây sự sai lệch trong đọc hiểu của mô hình. Đối với bài toán MRC các sai lệch đó được tạo nên từ: sai trọng tâm câu hỏi, các vấn đề về từ đồng nghĩa, đặc điểm trả lời của ngôn ngữ,...

Đó là những nguyên nhân dẫn đến mô hình đa ngữ, cụ thể là XLM-RoBerta mang lại kết quả cao hơn cho bài toán MRC trên tiếng Việt này.

4.3 Phân tích lỗi

Để có được những hiểu biết sâu sắc hơn, chúng tôi tiến hành phân tích kết quả đạt được đối với từng bài toán bằng cách:

- Lựa chọn ngẫu nhiên 100 dữ liệu trong tập test.
- Tiến hành dự đoán kết quả.
- So sánh kết quả dự đoán với kết quả thực tế.

4.3.1 Bài toán MRC

Đối với bài toán MRC, chúng tôi so sánh câu trả lời dự đoán với bốn câu trả lời thực tế. Nếu câu trả lời dự đoán khớp hoàn toàn với ít nhất một trong bốn câu trả lời thực tế, chúng tôi sẽ bỏ qua. Ngược lại, chúng tôi sẽ kiểm tra với mô hình còn lại. Với cùng dữ liệu đầu vào, liệu rằng kết quả dự đoán có sự khác biệt hay

không? Từ đó, rút ra các đặc điểm của từng mô hình và giải quyết vấn đề đặt ra ở bài báo cáo này.

Sau khi thực hiện phân tích, chúng tôi nhận thấy đối với cả hai mô hình PhoBERT và XLM-Roberta đều có tồn tại ba vấn đề. Chúng tôi đã phân tích và đưa ra những kết quả như sau:

Các câu trả lời nằm ở vị trí mắc nối (Giữa câu, cuối câu, chú thích,...) có dấu câu nằm liền kề thì các dấu câu được tính là một ký tự trong câu dự đoán (Bảng 4.12). Sau đó chúng tôi thực hiện loại bỏ các dấu câu được lấy dư và kết quả cho thấy, các dấu câu được lấy dư không ảnh hưởng đến kết quả của mô hình.

Đoạn văn	Chu kỳ tự quay của Trái Đất xét từ các định tinh, được IERS gọi là ngày định tinh, dài 86.164,098903691 giây thời gian Mặt Trời trung bình (UT1) hay 23h 56m 4,098903691s. Chu kì Trái Đất tự quay xét theo tuế sai hay chuyển động của xuân phân trung bình, bị đặt tên sai là năm thiên văn, dài 86.164,09053083288 giây Mặt Trời trung bình (UT1) hay 23h 56m 4,09053083288s. Vì thế ngày thiên văn ngắn hơn ngày định tinh khoảng **8,4 ms**. Độ dài của ngày Mặt Trời trung bình tính theo giây hệ SI có sẵn tại IERS cho các giai đoạn từ 1623-2005. và 1962-2005.
Câu hỏi	Độ dài của một ngày tính theo thiên văn và tính theo định tinh có khoảng chênh lệnh là bao nhiêu ms?
Câu trả lời	**8,4 ms**
Câu trả lời dự đoán	**8,4 mili giây,**

BẢNG 4.12: Phân tích bài toán MRC 1.

Nhận diện chính xác từ đồng nghĩa là một trong những vấn đề lớn của bài toán MRC. Và đặc biệt hơn là đối với ngôn ngữ đơn lập như tiếng Việt. Cả hai mô hình PhoBERT và XLM-R đều gặp vấn đề với từ đồng nghĩa. Một ví dụ về câu trả lời dự đoán bị sai do chưa hiểu đúng từ đồng nghĩa (Bảng 4.13). "Bị mất quyền kiểm soát khi nào" tương đương với "bị chiếm vào ngày nào" nhưng cả hai mô hình đều hiểu sai và đưa ra kết quả dự đoán sai.

Đoạn văn	Trong Chiến tranh thế giới thứ hai, Lục quân Đế quốc Nhật Bản chiếm Kuala Lumpur vào ngày 11 tháng 1 năm 1942. Người Nhật chiếm đóng thành phố cho đến ngày 15 tháng 8 năm 1945, khi tổng tư lệnh của Đệ thất phương diện quân Nhật Bản tại Singapore và Malaysia là Seishirō Itagaki đầu hàng chính phủ Anh Quốc. Năm 1957, Liên hiệp bang Malaya (Federation of Malaya) giành được độc lập khỏi sự thống trị của người Anh. Kuala Lumpur vẫn là thủ đô khi Malaysia thành lập vào ngày 16 tháng 9 năm 1963.
Câu hỏi	Thủ đô của Malaysia bị mất quyền kiểm soát vào tay Nhật Bản khi nào?
Câu trả lời	ngày 11 tháng 1 năm 1942
XLM-R dự đoán	ngày 15 tháng 8 năm 1945,
PhoBERT dự đoán	ngày 15 tháng 8 năm 1945, khi tổng_tư_lệnh của Đệ thất phương_diện_quân Nhật_Bản tại Singapore và Malaysia là Seishirō_Itagaki đầu_hàng chính_phủ Anh Quốc.

BẢNG 4.13: Ví dụ trả lời bị sai lệnh do từ đồng nghĩa.

Các câu hỏi yêu cầu sự suy luận từ những dữ liệu không hoàn thiện và sự chồng lấp các khái niệm để đưa ra câu trả lời đang là yếu tố gây ảnh hưởng lớn nhất đến cả hai mô hình. Bảng 4.14 thể hiện một ví dụ về dự đoán sai đối với một câu hỏi yêu cầu sự suy luận để trả lời. Khi được hỏi "Sở giao dịch chứng khoán Malaysia có trụ sở ở đâu?", cả hai mô hình PhoBERT và XLM-R đều đưa ra câu trả lời dự đoán là "đặt tại thành phố". Nhưng thành phố là một định nghĩa khái quát về quy mô dân cư, không phải một địa chỉ hay khu vực rõ ràng. Và trước đó, đoạn văn đang nói về thành phố Kuala Lumpur, nên câu trả lời đúng phải là "Kuala Lumpur".

Ngoài các vấn đề chung của cả hai mô hình, mô hình PhoBERT còn gặp phải một vấn đề liên quan đến đặc điểm ngôn ngữ tiếng việt, đó là cách trả lời câu hỏi. Với đặc trưng tiếng việt, khi được hỏi về một vấn đề, câu trả lời sẽ thường bắt đầu bằng một từ nối (Ví dụ Bảng 4.15). Khi so sánh với câu trả lời thì "bắt_nguồn từ" đang bị thừa, và kéo theo đó là kết quả của độ đo EM giảm.

Ngoài việc câu trả lời bắt đầu bằng một từ nối thì vấn đề câu trả lời sẽ bao

Đoạn văn	Kuala Lumpur là trung tâm kinh tế, thương mại, tài chính, bảo hiểm, bất động sản, truyền thông và nghệ thuật của quốc gia. Kuala Lumpur là một thành phố toàn cầu hạng alpha, và là thành phố toàn cầu duy nhất tại Malaysia. Phát triển cơ sở hạ tầng tại các khu vực xung quanh như sân bay quốc tế Kuala Lumpur tại Sepang, sự hình thành Hành lang đa phương tiện siêu cấp và sự mở rộng cảng Klang củng cố hơn nữa tầm quan trọng về kinh tế của thành phố. Bursa Malaysia hay Sở giao dịch chứng khoán Malaysia đặt tại thành phố, tạo thành một trong các hoạt động kinh tế cốt lõi của thành phố.
Câu hỏi	Sở giao dịch chứng khoán Malaysia có trụ sở nằm ở đâu?
Câu trả lời	Kuala Lumpur
Câu trả lời dự đoán	đặt tại thành phố,

BẢNG 4.14: Ví dụ cho câu hỏi yêu cầu sự suy luận bị dự đoán sai.

gồm phần đáp án và phần liên quan (Bảng 4.16) cũng đang giảm độ chính xác của mô hình. Đáp án là "ba dân tộc chính" nhưng theo đặc trưng ngôn ngữ thì đáp án của câu hỏi "ba dân tộc chính đó là gì?" được kèm theo.

Vấn đề đặc điểm ngôn ngữ mà mô hình PhoBERT gặp phải đang giải thích cho kết quả tại sao mô hình đơn ngữ lại cho độ chính xác thấp hơn mô hình đa ngữ trong nhiệm vụ đơn ngữ.

Đoạn văn	Thuật ngữ Deutschland trong tiếng Đức, ban đầu là diutisciu land ("các vùng người Đức") có nguồn gốc từ deutsch (tương tự dutch), bắt nguồn từ tiếng Thượng Đức Cổ diutisc "dân", ban đầu được sử dụng để phân biệt ngôn ngữ của thường dân khỏi tiếng Latinh và các hậu duệ của nó. Đến lượt mình, nó lại bắt nguồn từ tiếng German nguyên thủy *iudiskaz "dân", từ *eudō, bắt nguồn từ tiếng Ấn-Âu nguyên thủy *tewtéh- "người", từ "Teuton" cũng bắt nguồn từ đó. Từ Germany trong tiếng Anh bắt nguồn từ Germania trong tiếng Latinh- là từ được sử dụng sau khi Julius Caesar chọn nó để chỉ các dân tộc phía đông sông Rhine.
Câu hỏi	Nguồn gốc của từ Germany trong tiếng anh là từ đâu?
Câu trả lời	Germania trong tiếng Latinh
XLM-R dự đoán	Germania trong tiếng Latinh-
PhoBERT dự đoán	bắt_nguồn từ Germania trong tiếng Latinh

BẢNG 4.15: Ví dụ trả lời theo đặc điểm ngôn ngữ 1.

Đoạn văn	Kuala_Lumpur có thành_phần cư_dân hỗn_tạp, gồm ba dân_tộc chính của Malaysia : người Mã_Lai, người Hoa và người Ấn, song thành_phố cũng tồn_tại những sự kết_hợp văn_hoá như người Âu-Á, ngoài_ra còn có người Kadazan, người Iban và các sắc_tộc bản_địa từ Đông_Malaysia và Malaysia bán_đảo. Theo điều_tra dân_số năm 2010 do Cơ_quan Thống_kê tiến_hành, tỷ_lệ cư_dân Bumiputera tại Kuala_Lumpur là khoảng 44,2%, trong khi người Hoa chiếm 43,2% và người Ấn chiếm 10,3%. Có một hiện_tượng đáng chú_ý là sự hiện_diện ngày_càng tăng của cư_dân ngoại_quốc tại Kuala_Lumpur, hiện họ chiếm khoảng 9% dân_số thành_phố.
Câu hỏi	Malaysia gồm bao_nhiêu dân_tộc chính?
Câu trả lời	ba dân_tộc chính
XLM-R dự đoán	ba dân tộc chính
PhoBERT dự đoán	ba dân_tộc chính của Malaysia : người Mã_Lai, người Hoa và người Ấn,

BẢNG 4.16: Ví dụ trả lời theo đặc điểm ngôn ngữ 2.

4.3.2 Bài toán ABSA

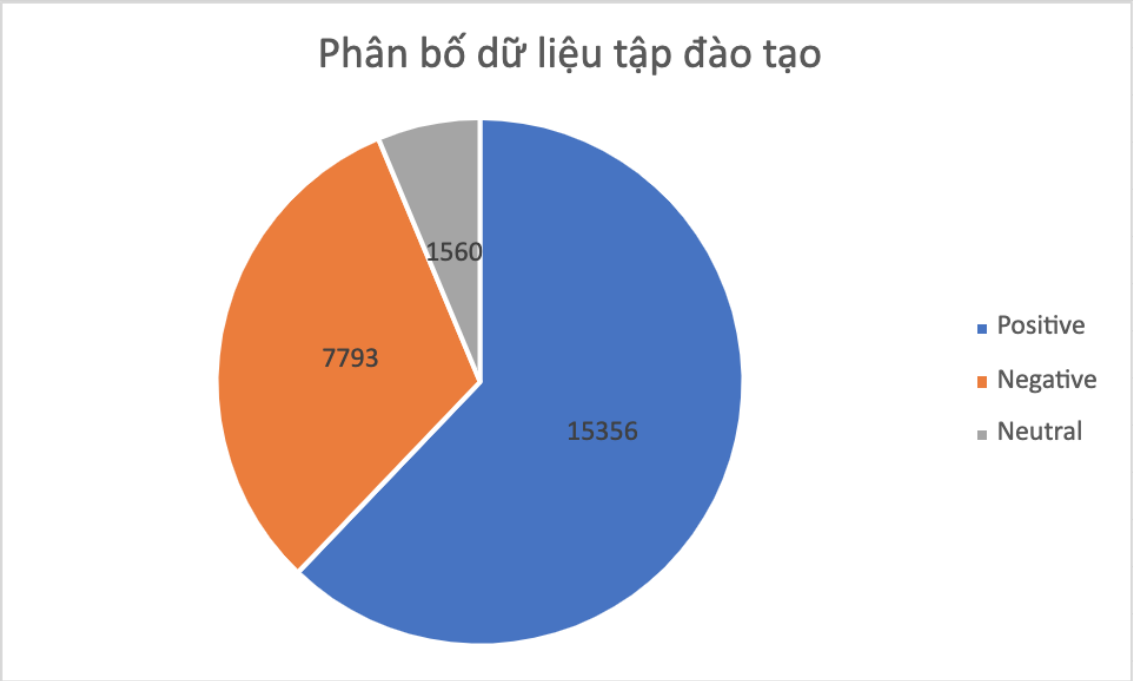
HÌNH 4.4: Hình minh hoạ bình luận mạng xã hội.

Việc thu thập bình luận người dùng từ các diễn đàn, các trang thương mại điện tử hay các trang phân phối tiêu dùng làm dữ liệu đào tạo giúp cho dữ liệu đào tạo được sát với thực tế thì việc đó cũng đem lại rất nhiều thách thức. Sau đây chúng tôi sẽ liệt kê một số lỗi dữ liệu phổ biến mà chúng tôi gặp phải khi đào tạo các mô hình để có thể làm cơ sở cho các phát triển trong nghiên cứu bài toán này. Hình 4.4 là ví dụ điển hình của một số lỗi thường gặp như sai chính tả, sai dấu câu,... mà chúng tôi sẽ phân tích sau đây.

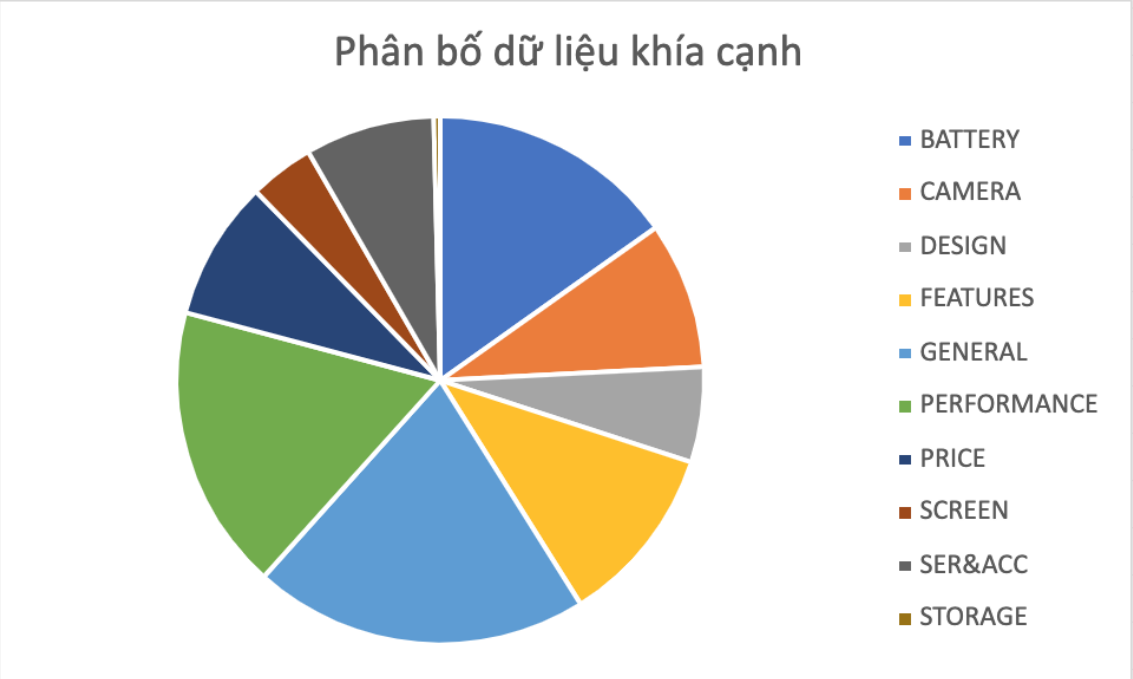
Sự phân bố không đồng đều giữa các nhãn

Do sự phân bố dữ liệu các nhãn không đều (mô tả ở hình 3.2) nên đối với các khía cạnh STORAGE#POSTIVE, STORAGE#NEGATIVE, STORAGE#NEUTRAL chưa nhận diện tốt. Đây có lẽ là vấn đề rất thường gặp phải khi đào tạo các mô hình. Vấn đề mất cân bằng nhãn thường rất gặp phải do số lượng đánh giá của người dùng về các vấn đề rất khác biệt nhau. Đối với thiết bị công nghệ nói chung và các thiết bị điện thoại nói riêng, việc đánh giá của người dùng thường mang tính chung chung nhiều hơn là hướng đến một khía cạnh cụ thể. Hơn thế nữa có rất nhiều khía cạnh được quan tâm hơn so với các khía cạnh khác. Chẳng hạn như số lượng dữ liệu mà chúng tôi thu thập được "Điện thoại khá ổn", "điện thoại khá tốt", "xài ổn đấy" nó mang tính chất là nhận xét chung. Vì thế mà số lượng dữ liệu mang nhận xét chung khá cao và tạo ra độ chênh lệch lớn hơn so với các nhãn dữ liệu khác. Ngoài ra, với những người sử dụng phổ thông thì việc họ quan tâm đến sản phẩm điện thoại sẽ quan tâm nhiều hơn đến thiết kế, lượng pin. Vì thế mà dữ liệu ở hai nhãn này vẫn thuộc ở mức cao do lượng bình luận thu thập ở người dùng phổ thông sẽ nhiều hơn các chuyên gia. Chỉ các chuyên gia hay những người có niềm đam mê với công nghệ họ mới xem xét, phân tích kĩ đến tính năng, chất lượng camera nên số lượng dữ liệu ở hai nhãn này ít hơn hẳn so với ba nhãn liệt kê trên. Tuy nhiên, điều đáng chú ý là các vấn đề như màn hình hay lưu trữ nhận được rất ít sự phản hồi vì thế cũng gây khó khăn cho vấn đề nhận diện các khía cạnh này.

Các bình luận mang yếu tố tích cực cũng chênh lệch nhiều so với những bình luận tiêu cực và bình luận trung tính. Tuy nhiên số lượng bình luận trung tính rất thấp và gây khó khăn cho việc nhận diện nhãn trung tính. **Các nhãn STORAGE và NEUTRAL khó để nhận diện.**

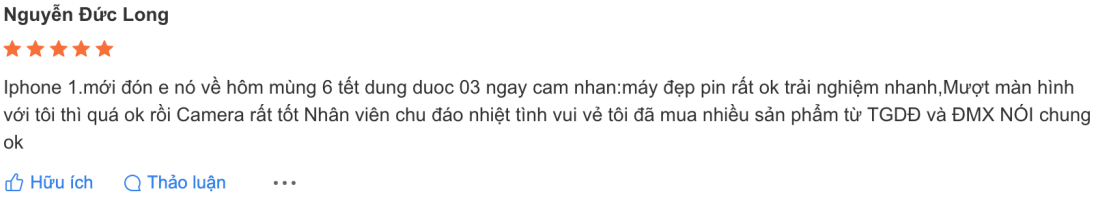


HÌNH 4.5: Phân bố dữ liệu trên tập đào tạo.



HÌNH 4.6: Phân bố nhãn khía cạnh trên tổng ba tập.

Cách đặt dấu câu sai



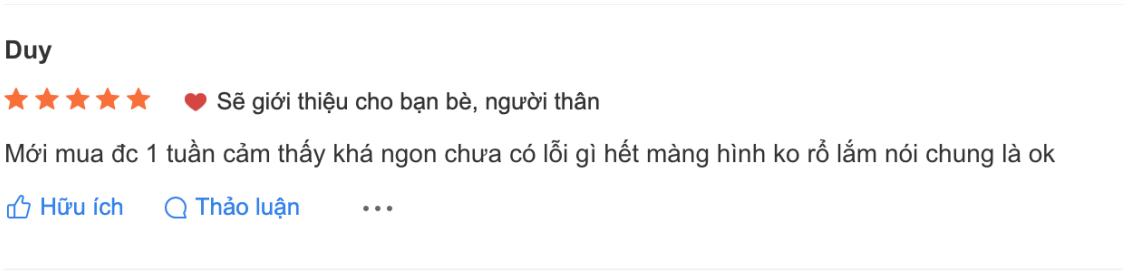
HÌNH 4.7: Hình minh hoạ bình luận 1.

HÌNH 4.8: Hình minh hoạ bình luận 2.

Vì dữ liệu được thu thập ngẫu nhiên từ các trang mạng xã hội hay các diễn đàn mạng về thiết bị công nghệ cho nên không tránh khỏi việc người dùng sử dụng các kí tự lạ, dùng từ vùng miền, viết sai chính tả, viết tắt, dùng từ ám hiệu. Những điều này đem đến một thách thức rất lớn để đào tạo mô hình. Lỗi sai này ảnh hưởng nhiều đến các kỹ thuật tách từ. Ví dụ như ở hình 4.7, bình luận của người dùng là "Iphone 1 .mới.." thay vì "Iphone 1. mới...". Khi tách từ nó sẽ là "Iphone/1/.mới/..." thay vì đúng sẽ là "Iphone/1/./mới/...". Điều này sẽ tạo thêm từ vựng mới và cũng gây khó khăn cho các mô hình pretrain cũng như các pretrain word embedding khi tiến hành xây dựng mô hình nhận diện.

Sai chính tả, viết tắt

Việc sử dụng các từ viết tắt, kí tự teencode hay việc viết sai chính tả cũng tạo nên sự gia tăng từ vựng và cũng đem lại khó khăn cho các mô hình pretrain cũng như các pretrain word embedding. Như hình 4.9, các từ như "dc", "màng hình", "ko" được thấy khá nhiều trong các bình luận người dùng. Điều này tạo



HÌNH 4.9: Phân bố dữ liệu trên tập đào tạo.

nên thách thức không ít cho việc đào tạo mô hình. Ngoài ra các định dạng viết tắt không theo bất kỳ một quy tắc nào như hình 4.4, từ "NVtgdd" có thể được viết theo nhiều cách như "NVTGDD", "nhân viên TGDD", "nv TGDD", cùng một ý nghĩa nhưng số lượng cách viết khác nhau tạo ra một bộ từ vựng rất hơn, hơn thế nữa nó cũng rất khó để nhận diện trong các pretrain.

Số lượng câu có mono span rất ít, không đáng kể

	Train	Dev	Test
Number of Comments	7,786	1,112	2,224
Number of Tokens	283,460	39,023	80,787
Number of Aspects	23,597	3,371	6,742
Average number of aspects per sentence	3.3	3.2	3.3
Average length per sentence	36.4	35.1	36.3

HÌNH 4.10: Hình minh hoạ tổng quan dữ liệu [19].

Số lượng bình luận chỉ nhận diện được một đoạn rất hiếm, trung bình là khoảng 3-4 đoạn mang ý nghĩa cho một câu bình luận vì thế mà bài toán này đem trong mình nhiều thách thức hơn.

Định dạng sai vị trí trong bộ dữ liệu

Ngaoài các yếu tố khách quan do sự ảnh hưởng của người dùng, thì còn một vấn đề khiến độ chính xác bị ảnh hưởng là do nhận diện đoạn mang ý nghĩa trong quá trình gán nhãn dữ liệu. Như hình 4.11, vị trí bắt đầu và kết thúc đoạn mang ý nghĩa rất quan trọng. Ví dụ [0, 11] của đoạn bình luận sẽ là "Pin nhanh hế" làm cho việc xây dựng mô hình khó khăn không chỉ bởi tạo từ vựng mới mà còn gây khó khăn trong việc đánh giá mô hình.

4.4 Kết luận

Chúng tôi đã thử nghiệm và đưa ra các kết luận về sự ảnh hưởng của các kỹ thuật tách từ đối với bài hai bài toán nhận diện chuỗi là hệ thống hỏi đáp và nhận diện cảm xúc. Đồng thời, phân tích các lỗi mà hai bài toán đang gặp phải và giải thích nguyên nhân của từng lỗi.

Bình luận	"Pin nhanh hết vân tay cũng chưa đk nhạy Nhân vien tgdd tư vấn nhiệt tình dễ thương 5 sao về phong cách bán hàng.
Nhãn sai	[0, 13, 'BATTERRY#NEGATIVE'] [14, 39, 'FETURES#NEGATIVE'] [40, 110, 'SER&ACC#POSITIVE']
Nhãn sai	[0, 11, 'BATTERRY#NEGATIVE'] [14, 39, 'FETURES#NEGATIVE'] [40, 113, 'SER&ACC#POSITIVE']
Nhãn đúng	[0, 13, 'BATTERRY#NEGATIVE'] [14, 39, 'FETURES#NEGATIVE'] [40, 113, 'SER&ACC#POSITIVE']

HÌNH 4.11: Hình minh hoạ lỗi sai toạ độ đoạn.

Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Đóng góp

Với đề tài này, chúng tôi hi vọng có thể tạo tiền đề cho các nghiên cứu phát triển các bài toán nhận diện chuỗi sau này. Sau đây là một số đóng góp chính của chúng tôi khi thực hiện đề tài này. Trong báo cáo khóa luận tốt nghiệp này, chúng tôi đã đạt được một số kết quả nhất định:

5.1.1 Bài toán ABSA

Đối với riêng bài toán nhận diện cảm xúc theo khía cạnh. Chúng tôi đã tiến hành thử nghiệm và đạt một số kết quả như sau:

- Cải thiện được hiệu suất mô hình so với các công bố trước đó trên cùng bộ dữ liệu.
- Xây dựng mô hình nhận diện cảm xúc theo khía cạnh, nhận diện cảm xúc, nhận diện khía cạnh sử dụng BiLSTM-CRF, PhoBERT, XLM-R.
- Xây dựng mô hình nhận diện cảm xúc theo khía cạnh đạt hiệu suất tốt nhất là 62,72% với độ đo f1-score(macro).
- Xây dựng mô hình nhận diện khía cạnh đạt hiệu suất tốt nhất là 64,29% với độ đo f1-score(macro).

- Xây dựng mô hình nhận diện cảm xúc đạt hiệu suất tốt nhất là 55,80% với độ đo f1-score(macro).

- Trả lời các câu hỏi mục tiêu đã đề ra như sau:

  - Kỹ thuật **tách theo từ** đạt hiệu suất cao hơn trên mô hình pretrain. Mô hình PhoBERT đạt hiệu suất cao hơn so với XLM-R cho cả 3 bài toán. Tuy nhiên các mô hình pretrain lại không đem lại hiệu suất cao bằng mô hình BiLSTM-CRF.
  - Đối với **bài toán nhận diện cảm xúc theo khía cạnh**, việc kết hợp **kỹ thuật tách theo âm tiết** với mô hình BiLSTM-CRF sử dụng PhoW2V cho kết quả tốt nhất trong thử nghiệm.
  - Đối với **bài toán nhận diện khía cạnh** việc kết hợp **kỹ thuật tách theo từ** với mô hình BiLSTM-CRF sử dụng PhoW2V cũng cho kết quả tốt nhất trong thử nghiệm.
  - Đối với **bài toán nhận diện cảm xúc** vẫn là mô hình BiLSTM-CRF sử dụng PhoW2V nhưng ở bài toán này **kỹ thuật tách theo âm tiết** lại chiếm ưu thế hơn.

5.1.2 Bài toán MRC

- Xây dựng hệ thống đọc hiểu máy xây dựng hai mô hình PhoBERT và XLM-RoBerta.

- Thực hiện thử nghiệm trên bốn phiên bản: PhoBERT-base, PhoBERT-large, XLM-RoBerta-base, XLM-RoBerta-large.

- Đối với bài toán MRC tách từ theo âm tiết cho kết quả tốt nhất trong thử nghiệm.

Kết quả nhóm chúng tôi đưa ra là phù hợp với kết quả của Rust và cộng sự [25] công bố. Đối với ngôn ngữ đơn lập, tùy vào bài toán mà tách từ theo âm tiết hay tách từ theo tiếng sẽ mang lại kết quả tốt hơn.

5.2 Khó khăn

Đối với bài toán Nhận diện cảm xúc theo khía cạnh.

- Miền dữ liệu còn hạn chế.
- Hiệu suất mô hình nhận diện cảm xúc theo khía cạnh chưa cao.
- Bộ dữ liệu bất đồng bộ giữa các nhãn.
- Chưa xử lý tốt các vấn đề đã nêu ở mục 3.6.

Đối với bài toán đọc hiểu tự động.

- Chưa xử lý tốt các từ đồng nghĩa.
- Chưa xử lý tốt các câu suy luận khó.

5.3 Hướng phát triển

Một số hướng phát triển cho đồ án này bao gồm:

- Cải thiện hiệu suất mô hình.
- Thử nghiệm mô hình trên nhiều miền dữ liệu khác.

5.4 Kết luận

Hình 5.1 mô tả trực quan hơn câu trả lời cho câu hỏi mục tiêu mà chúng tôi đặt ra ban đầu là "Tầm ảnh hưởng của kỹ thuật tách từ trên các bài toán nhận dạng chuỗi tiếng Việt. Cột mô hình thể hiện mô hình phù hợp và cho kết quả cao nhất so với các mô hình chúng tôi thử nghiệm đề cập ở trên . Cột âm tiết và từ cho biết sự kỹ thuật tách từ nào phù hợp với bài toán hơn.

Bài toán	MÔ HÌNH	ÂM TIẾT	TỪ
Nhận diện cảm xúc theo khía cạnh	BiLSTM-CRF (PhoW2V 300 chiều)	✓	
Nhận diện khía cạnh	BiLSTM-CRF (PhoW2V 300 chiều)		✓
Nhận diện cảm xúc	BiLSTM-CRF (PhoW2V 100 chiều)	✓	
Hệ thống hỏi đáp	XLM-R	✓	

HÌNH 5.1: Bảng kết luận.

Tài liệu tham khảo

[1] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *CoRR*, vol. abs/2001.01582, 2020. arXiv: 2001.01582. [Online]. Available: http://arxiv.org/abs/2001.01582.

[2] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/Daily Mail reading comprehension task," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 2358–2367. DOI: 10.18653/v1/P16-1223. [Online]. Available: https://aclanthology.org/P16-1223.

[3] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. DOI: 10.18653/v1/P17-1171. [Online]. Available: https://aclanthology.org/P17-1171.

[4] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, *Unsupervised cross-lingual representation learning at scale*, Aug. 2020. [Online]. Available: https://arxiv.org/pdf/1911.02116.

[5] T. Dang, V. Duc, K. Nguyen, and N. Nguyen, "Deep learning for aspect detection on vietnamese reviews," Nov. 2018, pp. 104–109. DOI: 10.1109/NICS.2018.8606857.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[7] P. N. Do, N. D. Nguyen, T. V. Huynh, K. V. Nguyen, A. G. Nguyen, and N. L. Nguyen, "Sentence extraction-based machine reading comprehension for vietnamese," *CoRR*, vol. abs/2105.09043, 2021. arXiv: 2105.09043. [Online]. Available: https://arxiv.org/abs/2105.09043.

[8] K. M. Hermann, T. Kociský, E. Grefenstette, *et al.*, "Teaching machines to read and comprehend," in *NIPS*, 2015, pp. 1693–1701. [Online]. Available: http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

[9] K. M. Hermann, T. Kočiský, E. Grefenstette, *et al.*, *Teaching machines to read and comprehend*, Nov. 2015, pp. 1693–1701. [Online]. Available: https://arxiv.org/pdf/1506.03340.

[10] M. Hu and B. Liu, "Mining opinion features in customer reviews," Jul. 2004.

[11] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[12] Y. Liu, M. Ott, N. Goyal, *et al.*, *Roberta: A robustly optimized bert pretraining approach*, Jul. 2019. [Online]. Available: https://arxiv.org/pdf/1907.11692.

[13] L. Mai and B. Le, "Aspect-based sentiment analysis of vietnamese texts with deep learning," in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds., Cham: Springer International Publishing, 2018, pp. 149–158, ISBN: 978-3-319-75417-8.

[14] Y. Mao, Y. Shen, C. Yu, and L. Cai, *A joint training dual-mrc framework for aspect based sentiment analysis*, 15, May 2021, pp. 13 543–13 551. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17597.

[15] A. T. Nguyen, M. H. Dao, and D. Q. Nguyen, "A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4079–4085.

[16] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042.

[17] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen, "A Vietnamese dataset for evaluating machine reading comprehension," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2595–2605. DOI: 10.18653/v1/2020.coling-main.233. [Online]. Available: https://aclanthology.org/2020.coling-main.233.

[18] M.-H. Nguyen, T. M. Nguyen, D. Van Thin, and N. L.-T. Nguyen, "A corpus for aspect-based sentiment analysis in vietnamese," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019, pp. 1–5. DOI: 10.1109/KSE.2019.8919448.

[19] L. L. Phan, P. H. Pham, K. T. Nguyen, *et al.*, "SA2SL: from aspect-based sentiment analysis to social listening system for business intelligence," *CoRR*, vol. abs/2105.15079, 2021. arXiv: 2105.15079. [Online]. Available: https://arxiv.org/abs/2105.15079.

[20] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, *SemEval-2015 task 12: Aspect based sentiment analysis*, Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 486–495.

DOI: 10.18653/v1/S15-2082. [Online]. Available: https://aclanthology.org/S15-2082.

[21] M. Pontiki, D. Galanis, H. Papageorgiou, *et al.*, *SemEval-2016 task 5: Aspect based sentiment analysis*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 19–30. DOI: 10.18653/v1/S16-1002. [Online]. Available: https://aclanthology.org/S16-1002.

[22] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, *SemEval-2014 task 4: Aspect based sentiment analysis*, Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. DOI: 10.3115/v1/S14-2004. [Online]. Available: https://aclanthology.org/S14-2004.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016. arXiv: 1606.05250. [Online]. Available: http://arxiv.org/abs/1606.05250.

[24] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*, Springer, 1999, pp. 157–176.

[25] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How good is your tokenizer? on the monolingual performance of multilingual language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021, Online, August 1-6, 2021*, 2021, pp. 3118–3135. [Online]. Available: https://arxiv.org/abs/2012.15613.

[26] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *CoRR*, vol. abs/1611.01603, 2016. arXiv: 1611.01603. [Online]. Available: http://arxiv.org/abs/1611.01603.

[27] K. N. T. Thanh, S. H. Khai, P. P. Huynh, L. P. Luc, D.-V. Nguyen, and K. N. Van, "Span detection for Vietnamese aspect-based sentiment analysis," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China: Association for Computational Lingustics, Nov. 2021, pp. 318–328. [Online]. Available: https://aclanthology.org/2021.paclic-1.34.

[28] A. Ushio and J. Camacho-Collados, "T-NER: An all-round python library for transformer-based named entity recognition," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Online: Association for Computational Linguistics, Apr. 2021, pp. 53–62. [Online]. Available: https://www.aclweb.org/anthology/2021.eacl-demos.7.

[29] K. Van Nguyen, S. Q. Tran, L. T. Nguyen, T. Van Huynh, S. T. Luu, and N. L.-T. Nguyen, *Vlsp 2021 - vimrc challenge: Vietnamese machine reading comprehension*, 2022. DOI: 10.48550/ARXIV.2203.11400. [Online]. Available: https://arxiv.org/abs/2203.11400.

[30] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "Vncorenlp: A vietnamese natural language processing toolkit," *arXiv preprint arXiv:1801.01331*, 2018.

[31] H. M. Wallach, "Conditional random fields: An introduction," *Technical Reports (CIS)*, p. 22, 2004.

[32] S. Wang and J. Jiang, "Machine comprehension using match-lstm and answer pointer," *CoRR*, vol. abs/1608.07905, 2016. arXiv: 1608.07905. [Online]. Available: http://arxiv.org/abs/1608.07905.

[33] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated self-matching networks for reading comprehension and question answering," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 189–198. DOI: 10.18653/v1/P17-1018. [Online]. Available: https://aclanthology.org/P17-1018.

[34] A. W. Yu, D. Dohan, M. Luong, *et al.*, "Qanet: Combining local convolution with global self-attention for reading comprehension," *CoRR*, vol. abs/1804.09541, 2018. arXiv: 1804.09541. [Online]. Available: http://arxiv.org/abs/1804.09541.

[35] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, "A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets," *arXiv preprint arXiv:2006.11880*, 2020.

Phụ lục A.

A.1 Nhãn bài toán nhận diện cảm xúc theo khía cạnh.

A.2 Nhãn bài toán nhận diện khía cạnh

A.3 Nhãn bài toán nhận diện cảm xúc

Nhãn bài toán ABSA	B-STORAGE#POSITIVE
B-SCREEN#POSITIVE	I-STORAGE#POSITIVE
I-SCREEN#POSITIVE	B-STORAGE#NEGATIVE
B-SCREEN#NEGATIVE	I-STORAGE#NEGATIVE
I-SCREEN#NEGATIVE	B-STORAGE#NEUTRAL
B-SCREEN#NEUTRAL	I-STORAGE#NEUTRAL
I-SCREEN#NEUTRAL	B-DESIGN#POSITIVE
B-CAMERA#POSITIVE	I-DESIGN#POSITIVE
I-CAMERA#POSITIVE	B-DESIGN#NEGATIVE
B-CAMERA#NEGATIVE	I-DESIGN#NEGATIVE
I-CAMERA#NEGATIVE	B-DESIGN#NEUTRAL
B-CAMERA#NEUTRAL	I-DESIGN#NEUTRAL
I-CAMERA#NEUTRAL	B-PRICE#POSITIVE
B-FEATURES#POSITIVE	I-PRICE#POSITIVE
I-FEATURES#POSITIVE	B-PRICE#NEGATIVE
B-FEATURES#NEGATIVE	I-PRICE#NEGATIVE
I-FEATURES#NEGATIVE	B-PRICE#NEUTRAL
B-FEATURES#NEUTRAL	I-PRICE#NEUTRAL
I-FEATURES#NEUTRAL	B-GENERAL#POSITIVE
B-BATTERY#POSITIVE	I-GENERAL#POSITIVE
I-BATTERY#POSITIVE	B-GENERAL#NEGATIVE
B-BATTERY#NEGATIVE	I-GENERAL#NEGATIVE
I-BATTERY#NEGATIVE	B-GENERAL#NEUTRAL
B-BATTERY#NEUTRAL	I-GENERAL#NEUTRAL
I-BATTERY#NEUTRAL	B-SER&ACC#POSITIVE
B-PERFORMANCE#POSITIVE	I-SER&ACC#POSITIVE
I-PERFORMANCE#POSITIVE	B-SER&ACC#NEGATIVE
B-PERFORMANCE#NEGATIVE	I-SER&ACC#NEGATIVE
I-PERFORMANCE#NEGATIVE	B-SER&ACC#NEUTRAL
B-PERFORMANCE#NEUTRAL	I-SER&ACC#NEUTRAL
I-PERFORMANCE#NEUTRAL	O

BẢNG A.1: Nhãn bài toán nhận diện cảm xúc theo khía cạnh.

Nhãn bài toán nhận diện khía cạnh
B-SCREEN
I-SCREEN
B-CAMERA
I-CAMERA
B-FEATURES
I-FEATURES
B-BATTERY
I-BATTERY
B-PERFORMANCE
I-PERFORMANCE
B-STORAGE
I-STORAGE
B-DESIGN
I-DESIGN
B-PRICE
I-PRICE
B-GENERAL
I-GENERAL
B-SER&ACC
I-SER&ACC

BẢNG A.2: Nhãn bài toán nhận diện khía cạnh.

Nhãn bài toán nhận diên cảm xúc
B-POSITIVE
I-POSITIVE
B-NEGATIVE
I-NEGATIVE
B-NEUTRAL
I-NEUTRAL

BẢNG A.3: Nhãn bài toán nhận diện cảm xúc.