CHƯƠNG 1: TỔNG QUAN

1.1 Lý do chọn đề tài

Phân tích quan điểm bình luận (SA) là phát hiện và nhận biết các biểu cảm như: thích, yêu, ghét,.. với các mức độ: tích cực, tiêu cực, trung tính về một đối tượng, sự vật sự việc nào đó trong một đoạn văn bản.

SA là một bài toán thuộc lĩnh vực xử lý ngôn ngôn ngữ tự nhiên.

Thông thường trong một đoạn văn bản sẽ bao gồm nhiều khía cạnh mà người ta muốn truyền tải và ứng với mỗi khía cạnh lại mang một màu sắc biểu cảm khác nhau, điều này sẽ ảnh hưởng đến kết quả nhận biết và thất thoát thông tin về cảm xúc. Ví dụ: Sản phẩm A mang đến người dùng một trải nghiệm tuyệt vời với giá cả phải chăng, tuy nhiên tôi không thích cách đóng gói sản phẩm và thời gian giao hàng khá lâu. Chính sự nhập nhằng này là một trong những nguyên nhân làm tăng độ khó của bài toán, khi ta phải nhận biết các khía cạnh tồn tại trong đoạn văn bản và quan điểm bình luận tương ứng, đồng nghĩa với việc thông tin ta nhận được sẽ mang tính chi tiết hơn.

SA được ứng dụng vào các lĩnh vực liên quan đến quản trị và đề xuất chiến lược kinh doanh như: quản trị thương hiệu doanh nghiệp và sản phẩm, phân tích các ý kiến phản hồi từ người dùng,..Ngoài ra, các mạng xã hội hiện nay đóng một vài trò then chốt trong việc lèo lái dư luận, mỗi phút mỗi giây đều có một lượng lớn thông tin chưa được kiểm duyệt tính đúng đắn, xác thực được người dùng đẩy lên các mạng xã hội. Chính điều này đặt ra thách thức lớn cho các doanh nghiệp khi phải liên tục lắng nghe những thông tin trên mạng xã hội với mục đích phát hiện và ngăn chặn sự lan rộng của những thông tin mang tính tiêu cực, bất lợi với chính những doanh nghiệp.

Bài toán phân tích quan điểm bình luận dựa trên khía cạnh (Aspect Based-Sentiment Analysis - ABSA) không phải là một bài toán mới, trên thế giới đã có nhiều công trình nghiên cứu liên quan và đạt được thành tựu đáng kể như [Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence](https://paperswithcode.com/paper/utilizing-bert-for-aspect-based-sentiment) cho tập dữ liệu SentiHood and SemEval-2014 Task 4 datasets. Còn tại Việt Nam thì ta có các công trình nghiên cứu theo phương pháp chia bài toán thành hai giai đoạn: nhận biết khía cạnh và phân tích cảm xúc của khía cạnh đó như Aspect-Based Sentiment Analysis of Vietnamese Texts with Deep Learning (Long bac) hay các nghiên cứu xoay quay bộ dữ liệu VLSP 2018. Các công trình nghiên cứu đã có dành tiếng Việt là không nhiều và không đạt được độ chính xác cao.

Đồng thời cũng chưa có các công bố về các nghiên cứu giải quyết bài toán ABSA cho tiếng Việt bằng áp dụng mô hình ngôn ngữ state of the art như: BERT, RoBERTa,..

1.2 Mục tiêu nghiên cứu

Thử nghiệm tính hiệu quả khi áp dụng mô hình ngôn ngữ BERT và PhoBERT so với các nghiên cứu trước đó cho bài toán ABSA trên tập dữ liệu VLSP 2018.

1.3 Đối tượng và phạm vi nghiên cứu

1.4 Phương pháp thực hiện

1.5 Ý nghĩa thực tiễn

1.6 Kết cấu đề tài

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Phân tích quan điểm bình luận và phân tích quan điểm bình luận dựa trên khía cạnh

2.1.1 Phân tích quan điểm bình luận

Phân tích quan điểm là một kỹ thuật xử lý ngôn ngữ tự nhiên dùng để xác định được dữ liệu thường là dữ liệu dạng văn bản thuộc mức độ cảm xúc nào: tích cực, tiêu cực, trung tính,...

Phân tích quan điểm thường được sử dụng trong doanh nghiệp như một cách để theo dõi các ý kiến phản hồi cũng như thấu hiểu nhu cầu của khách hàng.

Các thông tin liên quan đến sản phẩm, dịch vụ trên internet như các đánh giá, quan điểm cá nhân, nhận xét không những rất có giá trị đối với nhà sản xuất cung cấp dịch vụ và mà còn hữu ích cho chính người dùng mới. Ví dụ: Ngày nay một người muốn mua một cái tai nghe không dây thì trước khi ra cửa hàng, họ sẽ tìm thông tin liên quan đến sản phẩm như các đánh giá khen chê của những người từng được trải nghiệm sản phẩm, mức độ phổ biến, độ nhận diện thương hiệu,... Tương tự các nhà cung cấp sản phẩm sẽ phải lắng nghe những thông tin đó để có thể sửa chữa cải tiến những mặt còn hạn chế, hay cung ứng cho thị trường các sản phẩm phù hợp thị hiếu người dùng.

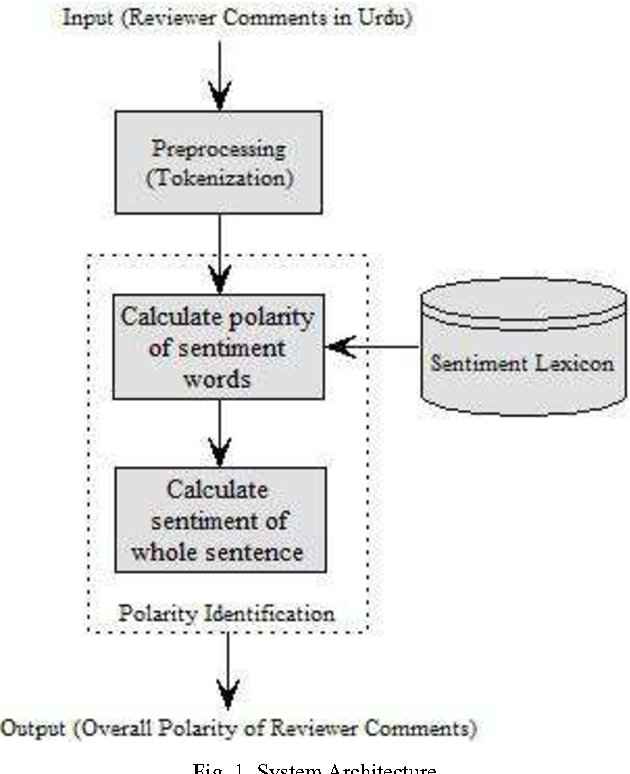
Các phương pháp giải quyết bài toán phân tích quan điểm:

Phương pháp dựa trên từ điển các từ thể hiện mức độ cảm xúc:

Kết quả của việc dự đoán quan điểm sẽ dựa vào kết quả tìm kiếm của các từ riêng lẻ thể hiện cảm xúc, điểm số tương ứng của các từ mang ý nghĩa tích cực, tiêu cực xuất hiện trong các câu, sau đó tổng hợp tất cả để quyết định văn bản sẽ mang quan điểm cảm xúc nào.

Là một trong những phương pháp đơn giản nhất, ít tốn chi phí thực hiện để giải quyết bài toán phân tích quan điểm bình luận. Độ chính xác của phương pháp phụ thuộc vào độ tốt của từ điển cảm xúc.

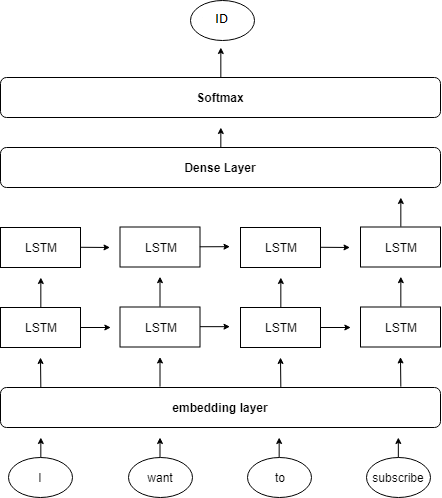
Tuy nhiên phương pháp dựa trên từ điển thể hiện cảm xúc thường không đem lại hiệu quả cũng như độ chính xác thấp. Nguyên nhân là do không thể hiện được mối quan hệ của các từ trong câu, thất thoát các thông tin quan trọng khi bỏ qua việc xét thứ tự của các từ.



Phương pháp Deep Learning Neural Network

Sự phát triển của cơ sở hạ tầng, tốc độ xử lý của phần cứng là một trong những nguyên nhân kéo theo sự phát triển của phương pháp học sâu Deep Learning. Hướng tiếp cận theo phương pháp Deep Learning cũng đạt được nhiều thành tựu đáng kể trong lĩnh vực trí tuệ nhân tạo nói chung và xử lý ngôn ngữ tự nhiên, bài toán phân tích quan điểm bình luận nói riêng.

Một trong số những phương pháp phổ biển để giải quyết bài toán phân tích quan điểm bình luận là kết hợp Word Embeding như Word2Vec, Fastext,... để biểu diễn văn bản đầu vào và mô hình Recurrent Neural Network (RNN) hay một biến thể giải quyết vấn đề phụ thuộc xa mà RNN đang mắc phải là Long Short Term Memory Neural Network (LSTMs).



Phương pháp kết hợp Rule-bases (dựa trên luật) và Corpus-bases (dựa trên ngữ liệu)

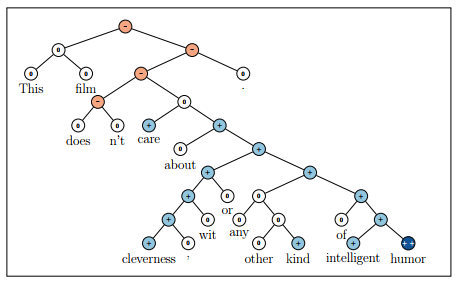
Chung ta có công trình nghiên cứu Recursive Deep Models for Semantic Compositionality

Over a Sentiment Treebank là một ví dụ điển hình cho phương pháp này.

Theo đó, phương pháp chính là sự kết hợp sử dụng mô hình Deep Learning Recursive Neural Network (RNN) và hệ tri thức chuyên gia trong xử lý ngôn ngữ tự nhiên được gọi là Sentiment Treebank.

Sentiment Tree là cây phân tích cú pháp của 1 câu văn, mỗi nút trong cây kèm theo bộ trọng số cảm xúc lần lượt là: rất tiêu cực (very negative), tiêu cực (negative), trung tính (neutral), tích cực (positive) và rất tích cực (very positive). Nhãn toàn cục của nút sẽ được quyết định bằng trọng số của nhãn lớn nhất.

Dựa trên công bố thì độ chính xác của mô hình khi dự đoán cảm xúc cho 1 câu đơn là 85,4%. Tuy nhiên, nhược điểm của phương pháp này ở chổ chỉ xử lý tốt cho dữ liệu đầu vào là một câu đơn.



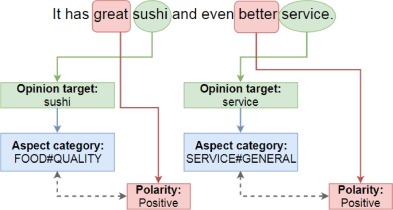
2.1.2 Phân tích quan điểm bình luận dựa trên khía cạnh

Phân tích quan điểm người dùng dựa trên khía cạnh là bài toán có mục tiêu xa hơn so với bài toán phân tích quan điểm người dùng. Mục tiêu của bài toán này chính là nhận biết các khía cạnh khác nhau tồn tại trong đoạn văn bản một cách tự động đồng thời xác định quan điểm cảm xúc ứng với khía cạnh đó.

Ví dụ [1]: Sản phẩm A mang đến người dùng một trải nghiệm tuyệt vời với giá cả phải chăng, tuy nhiên tôi không thích cách đóng gói sản phẩm và thời gian giao hàng khá lâu.

Nếu bài toán phân tích quan điểm thông thường thì kết quả dự đoán của câu trên sẽ là trung tính do trong câu tồn tại cảm xúc tích cực: “trải nghiệm tuyệt vời với giá cả phải chăng” và cảm xúc tiêu cực: “không thích cách đóng gói sản phẩm và thời gian giao hàng khá lâu”.

Tuy nhiên đối với bài toán phân tích quan điểm người dùng dựa trên khía cạnh thì kết quả dự đoán sẽ là một cặp khía cạnh - nhãn cảm xúc. Theo ví dụ trên thì kết quả tương ứng sẽ là: chung - tích cực, giá - tích cực, trải nghiệm giao nhân - tiêu cực. Dễ dàng nhận thấy là kết quả từ bài toán phân tích quan điểm người dùng dựa trên khía cạnh là chi tiết và mang lại giá trị phân tích nhiều hơn so với bài toán phân tích quan điểm thông thường.



Phương pháp kết hợp hai giai đoạn nhận biết khía cạnh (Aspect Extractor) và phân tích quan điểm bình luận (Sentiment Classification)

Bài toán phân tích quan điểm bình luận dựa trên khía cạnh sẽ được chia nhỏ thành hai bài toán con:

Aspect Extractor: dùng để nhận biết các khía cạnh ứng với từng lĩnh vực được đề cập trong đoạn văn bản. Các hướng tiếp cận phổ biến gồm có:

Phương pháp sử dụng tần suất xuất hiện: đoạn văn bản đầu vào sẽ được rút trích ra những danh từ có trong các câu bằng part-of-speech (POS) tagger, sau đó phân tích và quyết định xem danh từ đó sẽ thuộc vào khía cạnh nào dựa vào tần suất xuất hiện. Ví dụ [1]: Sản phẩm A mang đến người dùng một trải nghiệm tuyệt vời với giá cả phải chăng, tuy nhiên tôi không thích cách đóng gói sản phẩm và thời gian giao hàng khá lâu. Danh từ “giá cả” trong câu trên sẽ thường xuyên xuất hiện ứng với khía cạnh “giá” vì vậy sẽ tồn tại khía cạnh “giá” trong câu.

Phương pháp áp dụng học có giám sát Supervise Learning:

Xây dựng một mô hình machine learning model nhận biết có khía cạnh nào được nhắc đến trong đoạn văn bản.

Phương pháp dựa vào Topic Models

Topic modeling là một phương pháp học không giám sát unsupervised learning, giả sử rẳng mỗi đoạn văn bản sẽ bao gồm nhiều chủ đề, với mỗi chủ đề là một phân phối xác suất. Hai mô hình tiêu biểu là pLSA (Probabilistic Latent Semantic Analysis) (Hofmann, 1999)[1] and LDA (Latent Dirichlet allocation) (Blei, Ng and Jordan, 2003; Griffiths and Steyvers, 2003; Steyvers and Griffiths, 2007). Vậy ứng với bài toán phân tích quan điểm người dùng dựa trên khía cạnh thì mỗi chủ đề sẽ tương ứng với một khía cạnh.

Sentiment Classification:

Phương pháp sử dụng từ điển:

Tương tự phương pháp phân tích quan điểm người dùng dựa trên từ điển được mô tả ở mục [].

Phương pháp sử dụng machine learning

Đây là phương pháp được ưa chuộng rất nhiều, có thể kể đến SVMs trong công bố NLP@UIT at VLSP 2018: A SUPERVISED METHOD FOR ASPECT BASED SENTIMENT ANALYSIS. Theo đó, các đặc trưng dùng để huấn luyện sẽ là bi-gram, tri-gram, tất cả danh từ, động từ và tính từ xuất hiện trong câu và số lượng của chúng cũng như các dấu câu.

2.2 BERT

Theo BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

2.3 RoBERTa và PhoBERT

CHƯƠNG 3: THU THẬP VÀ XỬ LÝ DỮ LIỆU

CHƯƠNG 4: THỰC NGHIỆM

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

TÀI LIỆU THAM KHẢO