

# Phân tích cảm xúc bình luận của khách hàng thời gian thực

Tiêu Tự Đạt<sup>1,2</sup>, Hồ Việt Đức<sup>1,2</sup>, Ngô Đình Luân<sup>1,2</sup>, and Ngô Lê Hiếu Kiên<sup>1,2</sup>

<sup>1</sup> Vietnam National University, Ho Chi Minh City, Vietnam

<sup>2</sup> University of Information Technology, Ho Chi Minh City, Vietnam  
{1850589,18520610,18521064,18520952}@gm.uit.edu.vn

**Tóm tắt nội dung** Trước sự phát triển mạnh mẽ về công nghệ thông tin các trang mạng xã hội, dịch vụ mua bán hàng online ngày càng phát triển phổ biến tiêu biểu như twitter, facebook, shopee, tiki. Đối với những trang mạng xã hội đó được nhiều người dùng, tin dùng nên lượng dữ liệu trên các trang khá là lớn và phong phú. Trong nghiên cứu lần này, chúng tôi đánh giá bài toán phân loại với dữ liệu thu thập từ các trang mạng xã hội và sử dụng các model realtime để đánh giá. Đối vs bài toán phân loại, chúng tôi sử dụng dữ liệu chủ yếu thu thập trên twitter và shopee, chủ yếu là các bình luận đánh giá sản phẩm, giới thiệu sản phẩm. Chúng tôi đánh giá các sản phẩm, chất lượng sản phẩm dựa vào sự góp ý của người tiêu dùng.

**Keywords:** sentiment analysis · Real-time · Spark, Selenium

## 1 Mở đầu

Trong chương này, nhóm sẽ giới thiệu tổng quan về bài toán Phân tích cảm xúc thời gian thực và những thành công nhất định khi tiếp cận bài toán này.

### 1.1 Đặt vấn đề

Doanh nghiệp, dịch vụ và tổ chức tiến hành thu thập các thông tin phản hồi của người dùng về các sản phẩm, dịch vụ của họ để đưa ra các hướng đi đúng đắn hơn. Tuy nhiên, với lượng lớn các thông tin phản hồi của người dùng về dịch vụ, sản phẩm nào đó thì người dùng và các doanh nghiệp, tổ chức khó có thể quan tâm được hết. Để giải quyết vấn đề này, họ cần hệ thống có thể phân tích tự động tất cả các phản hồi và tóm tắt lại được tất cả các phản hồi để khách hàng, doanh nghiệp tham khảo và đưa ra các quyết định nhanh chóng.

Hiện nay, những thông tin mà các hệ thống được sử dụng để phân tích các phản hồi của người dùng trên các trang mạng thường chỉ quan tâm đến các thang điểm mà người dùng đánh giá về các sản phẩm, dịch vụ đó. Tuy nhiên, các thang điểm đánh giá phản hồi thì không thể hiện khách quan mức độ hài lòng của người dùng bằng những câu văn, những đoạn bình luận. Do đó, trong nghiên cứu lần này, chúng tôi hiện thực bài toán phân loại với dữ liệu thu thập

từ trang mạng xã hội sử dụng các model realtime để phân tích cảm xúc các bình luận. Đóng gói thành API thông qua những framework có sẵn để các doanh nghiệp có thể ứng dụng sử dụng trong thực tế.

## 1.2 Đối tượng và phạm vi nghiên cứu

### Đối tượng nghiên cứu

Các câu bình luận của người dùng người mua trên sàn thương mại điện tử được khai thác từ những phản hồi của người dùng trên các trang mạng xã hội.

### Phạm vi nghiên cứu

Những bình luận của người dùng trên các trang mạng về các địa điểm ăn uống như nhà hàng, quán ăn, quán bánh,... Phản hồi của các người dùng này sau khi thực hiện phản hồi sản phẩm trên sàn thương mại điện tử. Đối với nghiên cứu này, chúng tôi thực hiện trên bài toán phân tích cảm xúc bình luận/phản hồi của người dùng.

## 1.3 Mục tiêu của nghiên cứu

Trong nghiên cứu này, tiến hành nghiên cứu, tìm hiểu và giải quyết các mục tiêu chính như sau:

- Nghiên cứu về công nghệ dữ liệu lớn để có thể thực hiện crawling các bình luận từ tweeter, shoppe ở đây là spark.
- Thực hiện training với nhiều model để so sánh kết quả mang lại của các model.
- Xây dựng chương trình minh họa cho phép đơn vị kinh doanh thu thập bình luận của khách hàng khi trải nghiệm sử dụng dịch vụ.

### Kết quả nghiên cứu

Với những nội dung được thực thi trong nghiên cứu này, tôi đạt được những kết quả như sau:

- Thành công sử dụng spark để cào lượng lớn dữ liệu có trên tweeter về phục vụ mục đích huấn luyện mô hình
- Đánh giá tính hiệu quả của mô hình dựa trên 2 phương pháp máy học là ClassifierDLApproach vs LogisticRegression trên sparkNLP
- Xây dựng mô phỏng - website, cho phép người dùng tự động trực tiếp thu thập các bình luận từ trang bán hàng của mình (hiện tại chỉ hỗ trợ Shopee). Tiến hành tạo báo cáo tự động, phân tích insights dựa trên các bình luận thu thập được. Các thông tin trong báo cáo gồm, tỉ lệ tích cực/tiêu cực/trung tính từ các bình luận của người dùng. Ngoài ra còn cho phép người dùng nhập một bình luận bất kì để phân tích cảm xúc bình luận đó.

#### 1.4 Phát biểu bài toán

Trong đồ án này, chúng tôi tiến hành tìm hiểu và nghiên cứu bài toán phân tích bình luận/đánh giá của người dùng trên sàn thương mại điện tử (e-com). Bài toán xác định trạng thái cảm xúc, từ câu bình luận của người dùng và các trạng thái cảm xúc được quan tâm đến trong nghiên cứu này là tích cực (positive), trung tính (neutral) và tiêu cực (negative). Nói một cách ngắn gọn, Phân tích cảm xúc thời gian thực là bài toán chính trong đồ án này được phát biểu như sau:

- Đầu vào: Câu bình luận của người dùng về dịch vụ, sản phẩm trong các sàn thương mại điện tử.
- Đầu ra: Cảm xúc tương ứng là tích cực (positive), trung tính (neutral) và tiêu cực (negative).

#### 1.5 Cấu trúc đồ án

Cấu trúc đồ án được chia thành 7 chương, cấu trúc được trình bày như sau:

- Chương 1: Mở đầu. Trình bày lý do chọn nghiên cứu, đối tượng và phạm vi nghiên cứu, mục tiêu cũng như kết quả đạt được.
- Chương 2: Tổng quan. Giới thiệu khái niệm về bài toán phân tích cảm xúc người dùng, phân tích các hướng nghiên cứu đã được thực hiện trong và ngoài nước liên quan đến bài toán này. Trình bày bài toán trong nghiên cứu này tiến hành nghiên cứu và thực hiện.
- Chương 3: Trình bày khái lược về bộ dữ liệu được sử dụng và tiến trình tiền xử lý dữ liệu trước khi huấn luyện mô hình.
- Chương 4: Mô hình phân tích cảm xúc người dùng. Đồng thời, trình bày cơ sở lý thuyết của các phương pháp được sử dụng để tiến hành thử nghiệm.
- Chương 5: Phân tích và kết quả thực nghiệm. Trình bày quá trình cài đặt thử nghiệm, các bảng thông số và phân tích kết quả giữa các thử nghiệm.
- Chương 6: Phân tích ứng dụng, công nghệ thực hiện.
- Chương 7: Kết luận và hướng phát triển. Tổng kết các kết quả quan trọng đã đạt được trong nghiên cứu, những hạn chế chưa được giải quyết và hướng phát triển trong tương lai.

## 2 Tổng quan

Trình bày tổng quát bài toán phân tích cảm xúc người dùng thời gian thực (real-time sentiment analysis - RTSA) là một trong số hai bài toán con của phân tích khía cạnh cảm xúc (aspect-based sentiment analysis - ABSA). Đối với bài toán SA thông thường, từ một câu bình luận của người dùng, kết quả trả về là các trạng thái cảm xúc tương ứng đối với câu đó. Và, đối với bài toán ABSA, phải tiến hành phân tích chi tiết các khía cạnh được đề cập đến trong câu bình luận và cả trạng thái cảm xúc tương ứng cho từng khía cạnh đó.

## 2.1 Tổng quan về phân tích cảm xúc

Trong những năm gần đây, Phân tích cảm xúc (SA) được cộng đồng nghiên cứu thuộc lĩnh vực NLP được đông đảo cộng đồng trong lẫn ngoài nước rất quan tâm. Đây là quá trình xác định và phân loại văn bản thành các cảm xúc khác nhau — ví dụ, cảm xúc tích cực, tiêu cực hoặc trung tính — hoặc cảm xúc — chẳng hạn như vui, buồn, tức giận hoặc ghê tởm — để xác định thái độ của con người đối với chủ thể hoặc thực thể cụ thể.

Phân tích cảm xúc cũng là một trong những công tác quan trọng trong lĩnh vực NLP. Không chỉ quan trọng trong học thuật, nghiên cứu mà còn có ý nghĩa thiết yếu trong các ngành công nghiệp – dịch vụ, cụ thể là việc nhận biết hành vi và thái độ của khách hàng về sản phẩm và dịch vụ mà họ sử dụng.

Với công nghiệp - dịch vụ nói chung, SA được dùng như một công cụ mạnh mẽ để tự động hóa quy trình phân tích và đánh giá ý kiến của người dùng. Các ý kiến người dùng đó thường được thu thập từ trang mạng xã hội, hoặc trang thu thập nhận xét khách hàng về chất lượng cũng như mức độ hài lòng.

## 2.2 Tình hình nghiên cứu

### Tình hình nghiên cứu trên thế giới

Khái niệm phân tích cảm xúc (sentiment analysis) xuất hiện lần đầu tiên trong công trình của Nasukawa và Yi [1]. Khái niệm phân tích ý kiến (opinion mining) xuất hiện lần đầu trong công trình của Dave và Pennock [2]. Tuy nhiên, nghiên cứu được xem là đầu tiên đặt nền móng cho phân tích ý kiến là nghiên cứu của Pang và các cộng sự [3]. Từ đó các nghiên cứu trong bài toán này càng được quan tâm và phát triển.

### Tình hình nghiên cứu trong nước

Bên cạnh những công trình nghiên cứu trên thế giới, bài toán phân tích cảm xúc người cũng thu hút được cộng đồng nghiên cứu trong nước trên đa dạng các miền dữ liệu khác nhau như nhà hàng, khách sạn, điện tử...v.v. Theo như chúng tôi tìm hiểu, công trình nghiên cứu đầu tiên về phân tích ý kiến trên tiếng Việt được thực hiện bởi Kieu và Pham [4]. Trên cấp độ câu văn và xây dựng một hệ thống dựa trên rule-based system, với nền tảng Gate, tiến hành thực nghiệm đánh giá trên bộ ngữ liệu về miền dữ liệu máy tính.

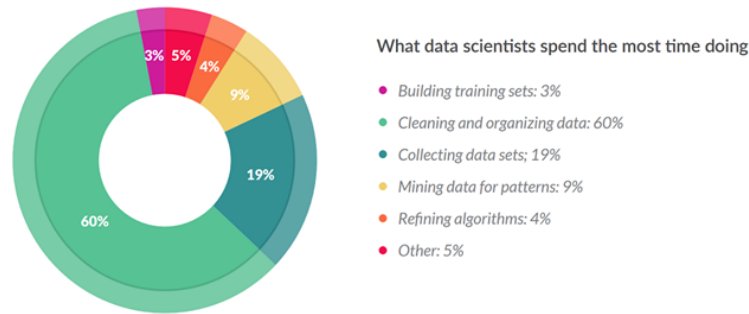
Vu và cộng sự trình bày một nghiên cứu về việc khai thác ý kiến dựa trên khía cạnh đánh giá sản phẩm bằng cách trích xuất các ý kiến rõ ràng hoặc ngụ ý sử dụng quy tắc cú pháp tiếng Việt. Tiếp theo, Le và các cộng sự [5] đề xuất phương pháp học bán giám sát GK-LDA cho việc trích xuất và phân loại các thuật ngữ khía cạnh cho văn bản tiếng Việt.

## 3 Bộ dữ liệu

### 3.1 Tổng quan dữ liệu

Trong chương này, chúng tôi trình bày sơ bộ thông tin về bộ ngữ liệu được sử dụng cho bài toán phân tích ý kiến bình luận của người dùng. Đối với các

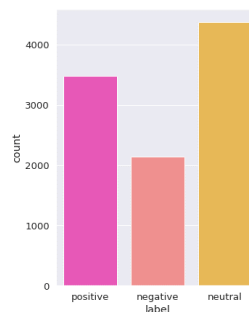
hệ thống dự đoán hiện nay, phần lớn hệ thống được xây dựng từ các thuật toán học có giám sát (supervised learning)[6]. Với thuật toán học có giám sát, chúng ta cần phải có một bộ ngữ liệu mẫu được gán nhãn để tiến hành huấn luyện cho hệ thống. Bộ ngữ liệu mẫu để huấn luyện quyết định chất lượng của hệ thống.



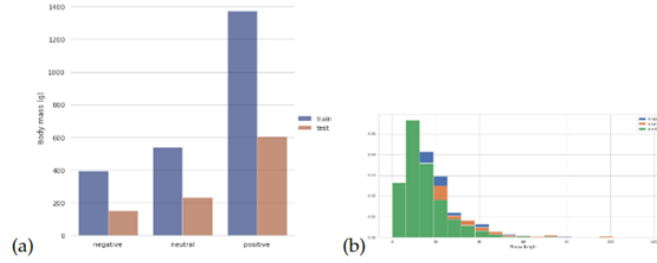
**Hình 1.** Tỷ lệ thời gian dành cho các giai đoạn để xây dựng một hệ thống khoa học ngữ liệu. Nguồn: báo cáo về học ngữ liệu 2016 [6]

Từ hình 1 trích từ báo cáo về khoa học ngữ liệu [6], ta có thể thấy được quá trình xây dựng một hệ thống khoa học ngữ liệu thì gần như toàn bộ thời gian liên quan đến ngữ liệu với 88% tổng thời gian (9% thời gian dành cho công tác phân tích ngữ liệu, 19% thời gian cho việc thu thập ngữ liệu và 60% thời gian cho việc làm sạch và tổ chức lại ngữ liệu). Vì vậy để có thể tập trung vào nghiên cứu và phát triển ứng dụng chúng tôi sử dụng pretrain language model có sẵn trên huggingface.co để gán nhãn dữ liệu sau khi cào từ tweeter. Hình 3 thống kê số lượng nhãn sau khi thực hiện sử dụng pretrain RoBERTa gán nhãn.

Sau khi gán nhãn chúng tôi chia train test để train model và xem xét dữ liệu.



**Hình 2.** Thống kê số lượng nhãn sau khi sử dụng pretrain model.



**Hình 3.** Bộ dữ liệu (a) Phân bố nhãn của tập dữ liệu tương ứng trên tập huấn luyện và tập kiểm tra (b) Tỷ lệ độ dài của câu ứng với mỗi nhãn có trong dữ liệu.

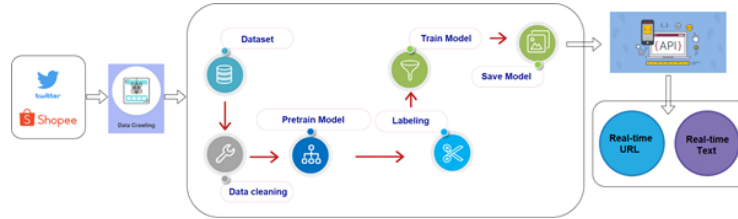
### 3.2 Cào dữ liệu và tiền xử lý dữ liệu

Bộ dữ liệu được trình bày trong đề án của chúng tôi được lấy từ nền tảng tweeter. Bộ dữ liệu được thu thập bằng spark. Đối với dữ liệu thu về chỉ là dữ liệu thô nên đòi hỏi quá trình tiền xử lý để model pretrain có thể gán nhãn tốt hơn và tiện sau này huấn luyện mô hình.

Công tác tiền xử lý, làm sạch ngữ liệu được tiến hành qua một số công tác được chúng tôi trình bày ở bên dưới:

- Đầu tiên, xóa các ký tự như dấu câu, biểu tượng, thẻ bắt đầu bằng icon, hashtag, link URL, hoặc từ không phải là chữ và số ở hai đầu giữa.
- Bỏ các ví dụ rỗng, nhiễu và stopword trong tập dữ liệu.
- Cuối cùng các bình luận ít hơn 10 ký tự cũng được chúng tôi loại bỏ.

Hình 4 dưới đây là tổng quát về hệ thống phân loại thời gian thực cho thương mại điện tử cụ thể là shopee.



**Hình 4.** Tổng quát về hệ thống phân loại cảm xúc thời gian thực của chúng tôi.

## 4 Huấn luyện mô hình

Chúng tôi sử dụng Google colab để huấn luyện. Sử dụng Spark NLP và Spark ML để huấn luyện mô hình.

Trong bài báo này 2 mô hình chúng tôi sử dụng là LogisticRegression và ClassifierDLApproach. Dùng 2 mô hình này với các mô hình Embeddings khác nhau để tìm ra mô hình kết hợp phù hợp nhất với bài toán.

Kết hợp 2 mô hình máy học là LogisticRegression và ClassifierDLApproach với các mô hình Embeddings mà thư viện Spark NLP hỗ trợ tạo ra các mô hình kết hợp. Với các mô hình Embeddings và 2 mô hình máy học trên, chúng tôi đã huấn luyện được 11 mô hình kết hợp gồm:

- ClassifierDLApproach + Bert embeddins
- ClassifierDLApproach + Glove embeddings
- ClassifierDLApproach + Universal Sentence Embeddings
- ClassifierDLApproach + ELMO Embeddings
- Classifier DL + Glove + Basic text processing
- LogisticRegression + Bert Embeddings
- LogisticRegression + Glove embeddings
- LogisticRegression + Universal Sentence Embeddings
- LogisticRegression + ELMO Embeddings
- LogisticRegression + CountVectorizer
- LogisticRegression + TFIDF

Cài đặt mô hình:

Đối với các mô hình kết hợp với ClassifierDLApproach: Batch\_Size = 64, Epoch = 20, Learning\_Rate = 5e - 3, Drop\_Out = 0.5.

Đối với các mô hình kết hợp với LogisticRegression: maxIter = 20, reg-Param = 0.3, elasticNetParam = 0.

Kết quả từng mô hình sẽ được chúng tôi tổng kết ở phần 5.

## 5 Kết quả thực nghiệm

Để đánh giá mô hình chúng tôi chia bộ dữ liệu với tỉ lệ là 8:2 trong đó 80% dữ liệu train và 20% dữ liệu test.

**Bảng 1.** Kết quả các mô hình kết hợp của mô hình ClassifierDLApproach với các mô hình kết hợp của ClassifierDLApproach mô hình ClassifierDLApproach + ELMO Embeddings đạt kết quả cao nhất với F1\_score 0.78.

Mô hình kết hợp	Avg Precision	Avg Recall	Avg F1_score
ClassifierDLApproach + Bert embeddins	0.35	0.32	0.25
ClassifierDLApproach + Glove embeddings	0.74	0.76	0.75
ClassifierDLApproach + Universal Sentence Embeddings	0.75	0.76	0.76
ClassifierDLApproach + ELMO Embeddings	<b>0.78</b>	<b>0.79</b>	<b>0.78</b>
Classifier DL + Glove + Basic text processing	0.71	0.76	0.73

**Bảng 2.** Bảng 2: Kết quả các mô hình kết hợp của mô hình LogisticRegression với các mô hình kết hợp của ClassifierDLApproach mô hình LogisticRegression + CountVectorizer có kết quả cao nhất với F1\_score 0.79 nên chúng tôi chọn mô hình này làm mô hình để dự đoán Real-time.

Mô hình kết hợp	Avg Precision	Avg Recall	Avg F1_score
LogisticRegression + Bert Embeddings + Bert embeddins	0.68	0.67	0.67
LogisticRegression + Glove embeddings + Glove embeddings	0.69	0.61	0.63
LogisticRegression + Universal Sentence Embeddings	0.72	0.68	0.69
LogisticRegression + ELMO Embeddings	0.73	0.69	0.70
LogisticRegression + CountVectorizer	<b>0.82</b>	<b>0.78</b>	<b>0.79</b>
LogisticRegression + TFIDF	0.81	0.77	0.79

## 6 Ứng dụng

### 6.1 Đặc tả phần mềm

#### Objective

**Vision.** Ứng dụng đặt mục tiêu nằm trong top nhà cung cấp các giải pháp tự động phân tích trải nghiệm của các nhóm khách hàng và xây dựng lợi thế cạnh tranh.

**Goal.** (a) Tất cả vì trải nghiệm khách hàng.(b) Cải tiến liên tục. (c) Tập trung thực thi. (d) Đơn giản hóa quy trình. (simplicity is the ultimate sophistication).

Nitiatives

- Tiến hành tự động thu thập dữ liệu.
- Tiến hành tự động gán nhãn những bình luận thu thập được
- Người dùng nhập từng đoạn text để phân tích bình luận **persona(s)**.

#### Thiết kế

##### Nhập đoạn text

Ảnh 5 là giao diện Ứng dụng cho phép người dùng nhập 1 đoạn text bất kỳ vào để phân tích cảm xúc.

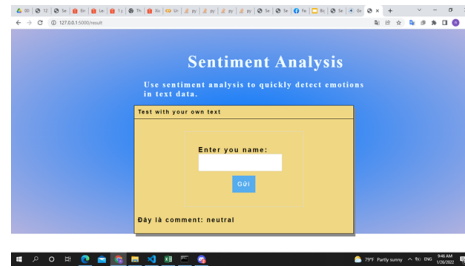
*Giao diện ứng dụng cho phép nhập link shoppe để phân loại cảm xúc từng bình luận có trong đó*

Ứng dụng cho phép người dùng nhập đường dẫn sản phẩm trên sàn thương mại điện tử shoppe để phân loại bình luận. Giao diện ứng dụng ở hình 6.

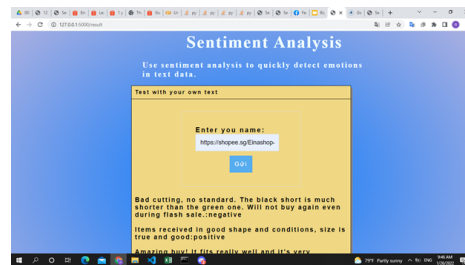
### 6.2 Công Nghệ

Flask là một framework python mã nguồn mở được sử dụng để xây dựng các ứng dụng web trực quan hóa dữ liệu tương tác. Nó được phát triển và được phát hành vào giữa năm 2017.





**Hình 5.** Giao diện ứng dụng cho phép nhập một đoạn text bất kì để phân loại cảm xúc.



**Hình 6.** Nhập url.

Selenium WebDriver là một trong những công cụ phổ biến nhất khi nói đến tự động hóa giao diện người dùng Web. Khung Selenium có thể được sử dụng với nhiều ngôn ngữ lập trình như Python, Java, C , v.v.

Apache Spark đang là sự chú ý của nhiều doanh nghiệp trong việc ứng dụng công nghệ vào phân tích và xử lý dữ liệu nhanh chóng.

## 7 Kết luận và hướng phát triển

### 7.1 Hạn chế

- Về ứng dụng, hiện vẫn chưa thiết lập trên hệ thống server có GPU, khiến việc trải nghiệm của người dùng đa phần chưa mượt mà và cần bổ sung thêm nhiều tính năng bổ ích.
- Chưa có sự so sánh khi sử dụng mô hình các pretrain model cho web demo.

### 7.2 Hướng phát triển trong tương lai

Ngôn ngữ vốn rất phức tạp và quá trình định lượng, thẩm định chất lượng cảm xúc từ văn bản cũng không hề dễ dàng. Ngay cả việc giao tiếp thường nhật giữa người với nhau thì chúng ta cũng khó thể nào biết được cảm xúc của đối phương dành cho mình ra sao. Có chăng điều khác biệt nằm ở máy móc vì chúng ta sẽ cần phải hạn chế yếu tố định tính trong giao tiếp và chuyển đổi các đặc

tính ấy thành các điểm số, thang đo nhằm mang tính định lượng cho cảm xúc con người. Từ những hạn chế được nêu ở trên, nhóm có những đề xuất như sau nhằm phát triển đồ án:

- Sử dụng phương pháp này để thực hiện tiếp cận trên nhiều bộ dữ liệu tiếng Việt với đa dạng văn cảnh khác nhau.
- Mở rộng bài toán cũng là điều rất quan trọng. Có thể nói, việc phân tích tổng quan cảm xúc người dùng biểu hiện trong câu nói là chưa đủ, mà chúng ta cần phải là chi tiết hơn rất nhiều. Và bài toán mà chúng tôi rất muốn tiếp tục phát triển và mở rộng là phân tích khía cạnh/cảm xúc người dùng, đồng thời, cũng là chức năng muốn được thực thi trên trang web hiện có của mình. Lúc bấy giờ, cá nhân người bán hàng sẽ hiểu rõ và sâu hơn về đối tượng khách hàng của mình.

## Tài liệu

1. Tetsuya Nasukawa and Jeonghee Yi. “Sentiment analysis: Capturing favorability using natural language processing”. In: Proceedings of the 2nd international conference on Knowledge capture. 2003, pp. 70–77.
2. Kushal Dave, Steve Lawrence, and David M Pennock. “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”. In: Proceedings of the 12th international conference on World Wide Web. 2003, pp. 519–528.
3. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Thumbs up? Sentiment classification using machine learning techniques”. In: arXiv preprint cs/0205070 (2002).
4. Binh Thanh Kieu and Son Bao Pham. “Sentiment analysis for Vietnamese”. In: 2010 Second International Conference on Knowledge and Systems Engineering. IEEE. 2010, pp. 152–157.
5. Vojtech Franc and Václav Hlaváč. “Multi-class support vector machine”. In: Object recognition supported by user interaction for service robots. Vol. 2. IEEE. 2002, pp. 236–239.
6. Binh Thanh Kieu and Son Bao Pham. “Sentiment analysis for Vietnamese”. In: 2010 Second International Conference on Knowledge and Systems Engineering. IEEE. 2010, pp. 152–157.