APPLICATION OF NATURAL LANGUAGE PROCESSING TECHNIQUES TO ANALYZE TELECOMMUNICATION SERVICE DEMANDS THROUGH SOCIAL MEDIA COMMENTS

Hoang Phuoc Loc[1*], Pham The An[2], Nguyen Thi Tan Dien[3], Le Trung Hieu[2], Huynh Thi Kim Ngan[1]
[1]Quang Tri Teacher Training College, [2]VNPT Quang Tri Branch, [3]Thuan Primary School, Huong Hoa, Quang Tri

ARTICLE INFO

Received: 23/01/2025
Revised: 14/03/2025
Published: 21/03/2025

KEYWORDS

Natural language processing
Needs analysis
Sentiment analysis
Social network
Text classification

ABSTRACT

Analyzing customer needs through social media is a crucial approach to capturing customer feedback on services or products. This process enables companies to develop strategies for improving product offerings, thereby enhancing service quality and business performance. In this study, we collected comment data from the VNPT fanpage, labeled and processed it, and created an experimental dataset comprising over 5,000 sentences. A customer needs analysis model leveraging natural language processing techniques was proposed, based on Facebook's FastText classification method. Additionally, experiments were conducted using other machine learning methods, including Naive Bayes and Support Vector Machine. The experimental results on the constructed dataset revealed that the proposed model utilizing FastText outperformed others, achieving an accuracy rate exceeding 90%. These findings establish a foundation for future research on expanding datasets in this domain and extending customer sentiment analysis to support corporate business strategies effectively.

ỨNG DỤNG CÁC KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN TRONG PHÂN TÍCH NHU CẦU SỬ DỤNG DỊCH VỤ VIỄN THÔNG TỪ CÁC BÌNH LUẬN TRÊN MẠNG XÃ HỘI

Hoàng Phước Lộc[1*], Phạm Thế An[2], Nguyễn Thị Tân Diện[3], Lê Trung Hiếu[2], Huỳnh Thị Kim Ngân[1]
[1]Trường Cao đẳng Sư phạm Quảng Trị, [2]VNPT Chi nhánh Quảng Trị, [3]Trường Tiểu học Thuận, Hướng Hóa, Quảng Trị

THÔNG TIN BÀI BÁO

Ngày nhận bài: 23/01/2025
Ngày hoàn thiện: 14/03/2025
Ngày đăng: 21/03/2025

TỪ KHÓA

Xử lý ngôn ngữ tự nhiên
Phân tích nhu cầu
Phân tích cảm xúc
Mạng xã hội
Phân loại văn bản

TÓM TẮT

Phân tích nhu cầu khách hàng qua mạng xã hội là một trong những kênh quan trọng để nắm bắt được ý kiến phản hồi của khách hàng về dịch vụ hoặc sản phẩm được cung cấp. Từ đó giúp các công ty có chiến lược điều chỉnh sản phẩm nhằm nâng cao chất lượng dịch vụ và hiệu quả kinh doanh. Trong nghiên cứu này, chúng tôi thu thập dữ liệu bình luận từ fanpage của VNPT, sau đó gán nhãn, huấn luyện và tạo tập dữ liệu thực nghiệm (datasets) hơn 5.000 câu. Một mô hình phân tích nhu cầu khách hàng sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên được đề xuất dựa trên phương pháp phân loại FastText của Facebook. Nghiên cứu này cũng tiến hành thực nghiệm sử dụng các phương pháp máy học khác là NaiveBayes và Support Vector Machine. Kết quả thực nghiệm đánh giá mô hình trên datasets đã xây dựng cho thấy mô hình đề xuất sử dụng FastText cho kết quả tốt hơn với độ chính xác trên 90%. Kết quả nghiên cứu này cũng là cơ sở cho các nghiên cứu tiếp theo về mở rộng xây dựng datasets cho lĩnh vực nghiên cứu này và mở rộng bài toán phân tích cảm xúc khách hàng nhằm phục vụ chiến lược kinh doanh của công ty.

DOI: https://doi.org/10.34238/tnu-jst.11945

[*] Corresponding author. *Email: loc_hp@qtttc.edu.vn*

1. Giới thiệu

Mạng xã hội là nơi mang mọi người đến với nhau để trò chuyện, chia sẻ ý tưởng, sở thích và kết nối với nhau qua phương tiện truyền thông xã hội [1]. Không chỉ vậy, mạng xã hội còn là "mảnh đất vàng" cho kinh doanh online, hỗ trợ tìm kiếm khách hàng, tương tác, quảng cáo, xây dựng thương hiệu doanh nghiệp hoặc xây dựng thương hiệu cá nhân. Mạng xã hội có bản chất của hoạt động xã hội, ở đây, người sử dụng có thể đưa ra những bình luận, nhận xét và đánh giá của mình một cách vô tư về các sản phẩm hay dịch vụ họ đã và đang sử dụng. Do đó, thông tin từ mạng xã hội rất có ý nghĩa cho các công ty, nhà cung cấp,... người đã tạo sản phẩm, dịch vụ nếu họ có chiến lược thu thập thông tin phản hồi của người dùng hợp lý để từ đó điều chỉnh sản phẩm, dịch vụ của mình nhằm đáp ứng nhu cầu người dùng.

Trí tuệ nhân tạo (AI) nói chung và công nghệ xử lý ngôn ngữ tự nhiên nói riêng đang trở thành một phần cốt lõi của ngành công nghệ để giúp các doanh nghiệp phân tích nhằm đưa ra các quyết định kinh doanh đúng đắn; hạn chế các sai lầm do phán đoán chủ quan nhằm tạo ra các sản phẩm và dịch vụ sáng tạo, đáp ứng nhu cầu người sử dụng, góp phần gia tăng doanh số kinh doanh của doanh nghiệp [2]. Tuy nhiên, việc sử dụng AI và các kỹ thuật xử lý ngôn ngữ tự nhiên để phân tích nhu cầu người sử dụng các dịch vụ vẫn còn khiêm tốn. Hầu hết các doanh nghiệp chỉ tập trung vào nghiên cứu lĩnh vực kinh doanh hẹp của họ và mỗi ngành lại có đặc điểm và đặc trưng sản phẩm riêng.

Chúng tôi đã sử dụng cổng tìm kiếm về cơ sở dữ liệu công bố khoa học và công nghệ Việt Nam (https://db0.vista.gov.vn/) để tìm kiếm các công bố của Việt Nam liên quan đến chủ đề nghiên cứu về phân tích nhu cầu sử dụng dịch vụ viễn thông từ các bình luận trên mạng xã hội của Vinaphone. Kết quả cho thấy chưa có công bố về lĩnh vực hẹp này.

Hơn nữa, các kỹ thuật được sử dụng để giải quyết bài toán phân tích cảm xúc - phân tích nhu cầu đang được đầu tư nghiên cứu rất mạnh. Kết quả đánh giá tổng quan chỉ ra rằng, phương pháp FastText của Facebook [3] là một tiếp cận mã nguồn mở khá mới và có tiềm năng sử dụng để giải quyết bài toán đang đặt ra. FastText được thiết kế để nhanh chóng huấn luyện và dự đoán, không yêu cầu tài nguyên tính toán cao, nên phù hợp cho các ứng dụng cần xử lý nhanh và triển khai thực tế trên các hệ thống có tài nguyên hạn chế. Trong khi đó, các giải pháp khác như BERT/PhoBERT, LSTM không có được những đặc điểm này. Do đó, áp dụng công nghệ này để giải quyết bài toán phân tích nhu cầu sử dụng dịch vụ viễn thông từ các bình luận trên mạng xã hội của Vinaphone là một lĩnh vực có tính ứng dụng cao, rất cần để đầu tư nghiên cứu đúng mức.

Thực vậy, AI và xử lý ngôn ngữ tự nhiên là một hướng nghiên cứu đang phát triển và có nhiều ứng dụng quan trọng [4]. Tuy nhiên, lĩnh vực này vẫn đang tồn tại những vấn đề hóc búa mà máy tính khó có thể thay thế hoàn toàn con người. Trong những bài toán đang được đặt ra cho các nhà nghiên cứu, có bài toán phân tích cảm xúc và phân tích nhu cầu.

Theo Tang và cộng sự [5], phân tích cảm xúc bao gồm hai dạng phân lớp: phân lớp quan điểm nhị phân và phân lớp quan điểm đa lớp. Cho một tập văn bản cần đánh giá D = $\{d_1, d_2, \ldots, d_n\}$, trong đó d_i là văn bản con thứ i, i = 1... n và một tập đánh giá được xác định trước C = {tích cực (positive), tiêu cực (negative)}. Phân lớp quan điểm nhị phân là phân loại mỗi tài liệu $d_i \subset D$ vào một trong hai lớp: tích cực và tiêu cực. Nếu d_i thuộc lớp tích cực có nghĩa là tài liệu d_i thể hiện quan điểm tích cực. Ngược lại, d_i thuộc tiêu cực có nghĩa tài liệu d_i thể hiện quan điểm tiêu cực. Phân lớp quan điểm đa lớp, kí hiệu C*, thiết lập tập C* = {tích cực mạnh (strong positive), tích cực (positive), trung lập (neutral), tiêu cực (negative), tiêu cực mạnh (negative strong)} và phân loại mỗi $d_i \subset D$ vào một trong các lớp trong C*. Có khá nhiều cách tiếp cận cho bài toán phân tích cảm xúc. Phân tích cảm xúc có thể dựa vào cụm từ thể hiện quan điểm thông qua phương pháp phân tích và gán nhãn từ loại được đề xuất bởi Turney [6]. Phương pháp này được thực hiên theo 2 bước. Bước 1 trích chọn ra các cụm từ chứa tính từ hay trạng từ. Bước 2 xác định xu hướng quan điểm của cụm từ thu được dựa trên độ đo PMI (pointwise mutual information) theo công thức:

$$PMI(\ term\ _1,\ term\ _2) = log_2\left(\frac{Pr(\ term\ _1 \wedge term\ _2)}{Pr(\ term\ _1) Pr(\ term\ _2)}\right) \quad (1)$$

Trong đó:

Pr($term_1 \wedge term_2$): xác suất đồng xuất hiện của từ $term_1$ và từ $term_2$. Pr($term_1$), Pr($term_2$): xác suất mà $term_1$, $term_2$ xuất hiện khi thống kê chúng riêng rẽ.

Phân tích cảm xúc cũng có thể được thực hiện dựa vào phương pháp phân lớp văn bản bằng các kỹ thuật máy học như Bayesian, SVM (Support vector machine), KNN (k-nearest neighbor),… Cách tiếp cận này được Pang và Le [7] áp dụng để đánh giá người xem phim thành hai lớp tích cực và tiêu cực cho kết quả thực nghiệm tốt với độ chính xác 81%.

Phân tích cảm xúc cũng có thể dựa vào hàm tính điểm số được đưa ra bởi Dave và cộng sự [8] thông qua hai bước. Bước 1 tính điểm các từ trong văn bản của tập dữ liệu theo biểu thức (2):

$$score(t_i) = \frac{Pr(t_i|C) - Pr(t_i|C')}{Pr(t_i|C) + Pr(t_i|C')} \quad (2)$$

Trong đó:

t_i là từ cần được tính điểm.

C là một lớp quan điểm; C' là lớp phần bù của C hoặc (not C).

Pr(t|C): xác suất t xuất hiện ở lớp C, được tính bằng số lần xuất hiện của t trong lớp C. Điểm số được chuẩn hóa trong khoảng [-1, 1].

Bước 2, một văn bản mới $d_i = t_1 \ldots t_n$ sẽ được phân lớp theo công thức (3):

$$class(d_i) = \begin{cases} C & eval\ (d_i) > 0 \\ C' & otherwise \end{cases} \quad (3)$$

với $eval(d_i) = \sum_j\ Score\ (t_i)$

Phân tích cảm xúc dựa trên phương pháp máy học đang thu hút nhiều nhà nghiên cứu quan tâm. Điển hình như các nghiên cứu nền tảng của Tang và cộng sự [5], Pang và Lee [7], Taboada [9], Beineke và cộng sự [10], Matsumoto và cộng sự [11]. Các kết quả thực nghiệm từ những phương pháp tiếp cận này đã chứng tỏ có độ tin cậy khá cao.

Bên cạnh các phương pháp được sử dụng phân loại văn bản bằng các kỹ thuật máy học như Bayesian, SVM hay KNN,…, phương pháp FastText của Facebook [3] là một tiếp cận mã nguồn mở khá mới được dùng để phân loại văn bản. Tuy nhiên, sử dụng tiếp cận FastText để phân loại văn bản tiếng Việt vẫn chưa được nghiên cứu một cách thấu đáo. Đặc biệt, ứng dụng FastText vào bài toán phân tích nhu cầu sử dụng dịch vụ VNPT thông qua bình luận bằng tiếng Việt trên mạng xã hội qua trang Fanpage của Vinaphone về các dịch vụ Internet, di động và truyền hình số cần được đầu tư nghiên cứu. Dựa vào tiếp cận này, chúng tôi đề xuất giải pháp ở nội dung tiếp theo để giải quyết bài toán được đặt ra. Kết quả thực nghiệm chứng tỏ được giải pháp đề xuất mang lại hiệu quả tốt và được chỉ ra ở phần thực nghiệm.

2. Phương pháp nghiên cứu

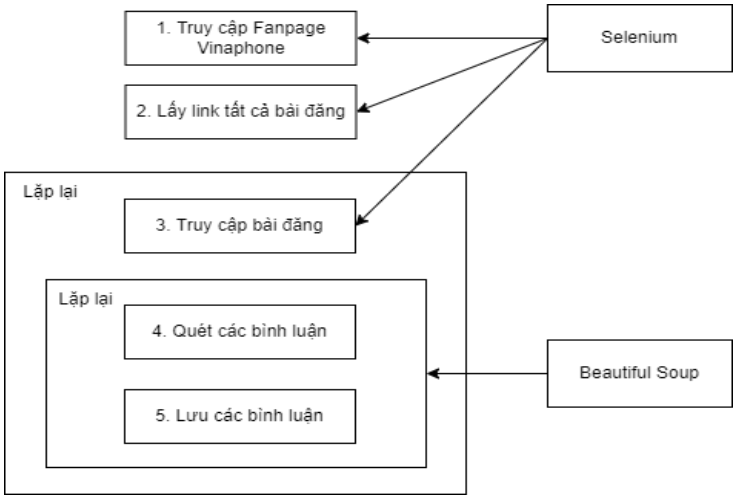
Phân tích cảm xúc (Sentiment analysis) để khai thác quan điểm là một qui trình nghiên cứu rất phức tạp, được nghiên cứu trên nhiều khía cạnh khác nhau. Ở Việt Nam, khai phá quan điểm trên mạng xã hội được coi là một lĩnh vực mới, nhận được nhiều sự quan tâm trong những năm gần đây và chỉ mới đạt được một số kết quả bước đầu. Cụ thể, kết quả tìm kiếm trên cơ sở dữ liệu công bố khoa học và công nghệ Việt Nam (https://db0.vista.gov.vn/) cho thấy các công bố chủ yếu là khám phá về công nghệ phân tích cảm xúc, các công bố về khai phá quan điểm để giải quyết các bài toán chuyên ngành còn rất khiêm tốn. Khai thác quan điểm có vai trò rất quan trọng, giúp các công ty, tổ chức hay cá nhân biết được ý kiến, quan điểm của một bộ phận người quan tâm về vấn đề của mình đang triển khai. Xuất phát từ nhu cầu này, chúng tôi đề xuất mô hình phân tích và đánh giá các bình luận trên mạng xã hội Facebook, tại fanpage của Vinaphone nhằm phân loại các bình luận của khách hàng theo nhu cầu sử dụng các dịch vụ về Internet, di động và truyền hình số. Từ đó, chúng tôi phân tích và đưa ra các chiến lược bán hàng, chính sách chăm sóc khách hàng phù hợp với từng đối tượng khách hàng. Mô hình hệ thống đề xuất ở Hình 1 được mô tả qua bốn bước như sau:

Bước 1: Thu thập bình luận tại fanpage của VNPT Vinaphone (https://www.facebook.com/vinaphonefan).
Bước 2: Tiền xử lý dữ liệu
Bước 3: Huấn luyện và phân lớp câu bình luận
Bước 4: Thử nghiệm và đánh giá kết quả

Hình 1. *Mô hình khai thác nhu cầu của các bình luận trên mạng xã hội*

2.1. Thu thập bình luận

Module thực hiện hai quá trình thu thập dữ liệu và rút trích bình luận được mô tả ở Hình 2. Hình 2 thể hiện 5 bước của quá trình truy cập, thu thập, quét và tiến hành lọc các bình luận trên fanpage Vinaphone. Kết quả thu thập được hơn 5.000 câu bình luận tạo cơ sở dữ liệu cho quá trình nghiên cứu.



Hình 2. *Các bước thu thập bình luận*

Khởi đầu, chúng tôi sử dụng thư viện Beautiful Soup và Selenium trên ngôn ngữ Python để crawler (crawler là phần mềm có khả năng tự động lấy dữ liệu như ảnh, text,… trên WWW) dữ liệu HTML trên website. Mã HTML được phân tích cấu trúc DOM3, theo các luật quy định sẵn, crawler sẽ xác định vùng dữ liệu cần bóc tách: liên kết tương tự, thông tin bình luận cần thu thập. Các liên kết được chọn lọc và lưu trữ trong một hàng đợi URL. Để rút trích đúng mục tiêu bài viết, các liên kết hoặc tiêu đề bài viết được lọc lại theo từ khóa ứng với sản phẩm cần thu thập. Quá trình này được lặp lại cho tới khi không còn liên kết nào trong hàng đợi hoặc đủ số lượng cần thiết.

2.2. Tiền xử lý dữ liệu

Dữ liệu sau khi rút trích được tiền xử lý để có được một tập dữ liệu rõ ràng, không trùng lặp, loại bỏ các liên kết, trích dẫn trong bình luận. Module tiền xử lý này rất quan trọng, bởi lẽ làm giảm nhiễu và sự nhập nhằng cho chương trình, cũng như quá trình thực thi, thực nghiệm chương trình.

- Xóa các biểu tượng cảm xúc, kí tự đặc biệt: Trong phạm vi bài báo này, việc xử lý các ký tự đặc biệt và các biểu tượng cảm xúc chưa mang ý nghĩa phân loại, mặt khác sẽ gây nhiễu trong quá trình phân tích.
- Chuyển dạng từ rõ nghĩa: Người sử dụng thường có thói quen viết tắt, viết các ký hiệu thay

cho từ rõ nghĩa. Chẳng hạn từ "Có thể dùng 4g ko" ("ko" có nghĩa là không), vs (với)… hay dữ liệu không đồng bộ, không chuẩn hóa. Việc này sẽ ảnh hưởng gây nhiễu kết quả phân tích.

- Xóa dòng dữ liệu: Tập dữ liệu thu về sẽ có nhiều dữ liệu bị trống, dữ liệu trống không có ý nghĩa trong quá trình phân tích, gây tốn bộ nhớ lưu trữ.

- Tiến hành gắn nhãn vào tạo file dataset:

Từ tập dữ liệu được xử lý và tách ra từ và câu, chúng tôi gán nhãn để tạo các tập đặc trưng theo các loại nhu cầu để phục vụ cho việc phân loại và gán nhãn của câu. Chúng tôi phát triển một module dựa trên kỹ thuật NaiveBayes từ thư viện Scikit-learn mã nguồn mở để tiến hành phân loại và gán nhãn cho các câu với độ chính xác 85%.

Tập dữ liệu sau khi thu thập và xử lý sẽ được biên tập để chuẩn bị cho quá trình huấn luyện. Dữ liệu được biên tập dưới dạng:

__label__INTERNET Tôi cần lắp đặt Internet tại Đông Hà, nhờ tư vấn. __label__MOBILE Xin hỏi gói cước di động có 3g giá rẻ. *__label__MY_TV MyTV có sử dụng được cho 3 tivi không?*

__label__MOBILE Sim VNPT như nào mới đăng kí đc gói cước đó

__label__MOBILE Ad kiểm tra giúp mình số thuê bao 0912 307 880 đăng ký được gói nào bên trên.

__label__MOBILE Cho mình hỏi 0911200234 dk 4g sao ak

__label__INTERNET Tư vấn giúp mình gói Home TV với.

2.3. Tách từ tiếng Việt

Tách từ có thể nói là giai đoạn quan trọng nhất, ảnh hưởng đến kết quả của mô hình xử lý. Bước này có nhiệm vụ xác định các từ có trong văn bản, kết quả của nó là một tập các từ riêng biệt. Hiện tại có một số công cụ hỗ trợ cho tách từ tiếng Việt như: Mô hình tách từ bằng WFST [12]; công cụ JvnTextPro tách từ [13]; bộ công cụ tách từ vnTokenizer [14].

Nghiên cứu này sử dụng tiếp cận thư viện Underthesea của tác giả Vũ Anh [15], là một bộ Toolkit mã nguồn mở hoàn chỉnh, để tích hợp sử dụng trong mô hình nghiên cứu đề xuất.

2.4. Huấn luyện và phân lớp

Phản hồi sau khi được thu thập sẽ phân thành các lớp khác nhau để phục vụ việc thống kê, tạo báo cáo. Phân lớp văn bản là quá trình gán nhãn các văn bản ngôn ngữ tự nhiên một cách tự động vào một hoặc nhiều lớp cho trước, "nhóm" các đối tượng "giống" nhau vào "một lớp" dựa trên các đặc trưng dữ liệu của chúng. Hệ thống đánh giá phân lớp các bình luận rút trích được thành 3 nhóm: "Nhu cầu dịch vụ Internet", "Nhu cầu dịch vụ di động" và "Nhu cầu dịch vụ MyTV" tương ứng là: "INTERNET", "MOBILE", "MYTV".

Trong nghiên cứu này, chúng tôi phân lớp nhu cầu dựa trên các tiếp cận SVM, NaiveBayes và công cụ FastText của Facebook trên tập dữ liệu hơn 5.000 câu bình luận để phân tích và đánh giá giải thuật tối ưu để sử dụng cho mô hình đề xuất. Quá trình huấn luyện và phân lớp được thực hiện theo mô hình được đề xuất ở Hình 3.

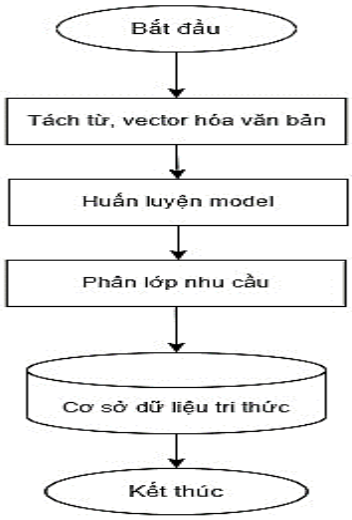
Để thực hiện sơ đồ ở Hình 3, chúng tôi phát triển các module tương ứng sau:

Module 1 - tách từ: Với module này, thư viện Underthesea của tác giả Vũ Anh [15] được sử dụng để tách từ. Tiếp theo tiến hành vector hóa văn bản và trích xuất đặc trưng, chúng tôi dùng Bag-of-words và sử dụng Pineline để chuẩn bị dữ liệu mô hình huấn luyện.

Module 2 - huấn luyện mô hình: Trong module này, phương pháp xử lý ngôn ngữ tự nhiên dựa trên công cụ FastText được sử dụng để huấn luyện mô hình trên dữ liệu từ điển đã thu thập. Chúng tôi cũng sử dụng các mô hình Support Vector Machines và NaiveBayes để cài đặt phân loại nhu cầu nhằm so sánh và đánh giá kết quả đạt được.

Chúng tôi lựa chọn sử dụng FastText để giải quyết bài toán đặt ra vì FastText được thiết kế để nhanh chóng huấn luyện và dự đoán, ngay cả trên các tập dữ liệu lớn. Không yêu cầu tài nguyên tính toán cao, nên phù hợp cho các ứng dụng cần xử lý nhanh và triển khai thực tế trên các hệ

thống có tài nguyên hạn chế. Kỹ thuật sử dụng các vector từ giúp phân loại văn bản đơn giản và hiệu quả, phù hợp với bài toán đặt ra. Trong khi đó, BERT/PhoBERT yêu cầu GPU và tài nguyên mạnh để huấn luyện và suy luận do kích thước mô hình lớn. Thời gian dự đoán thường lâu hơn, đặc biệt khi xử lý lượng lớn văn bản. LSTM huấn luyện lâu hơn do tính chất tuần tự của mạng hồi tiếp.



Hình 3. *Sơ đồ huấn luyện cơ sở tri thức*

Facebook FastText [16] là một thư viện mã nguồn mở được dùng cho biểu diễn từ hay nhúng từ (word embeddings) và phân lớp văn bản. FastText được tạo bởi phòng thí nghiệm trí tuệ nhân tạo của Facebook và hiện tại đang được nghiên cứu để ứng dụng. Mô hình trong đó cho phép tạo các thuật toán máy học có giám sát và không có giám sát để có được biểu diễn vector cho các từ. FastText làm việc tương tự như Word2Vec nhưng huấn luyện bắt đầu ở mức ký tự thay vì từ với việc sử dụng Ngrams. FastText công bố vào tháng 8 năm 2016 và cung cấp mã nguồn mở cho cộng đồng sử dụng với mô hình huấn luyện trước cho 294 ngôn ngữ.

Module 3 - phân lớp: Từ dữ liệu đã được huấn luyện, tiến hành phân lớp dữ liệu đầu vào và cho ra kết quả về nhu cầu.

3. Kết quả và bàn luận

Sau khi thu thập dữ liệu trên mạng xã hội thông qua fanpage của VNPT, chúng tôi tiến hành xử lý và thu thập được hơn 5.000 câu bình luận về các nhu cầu của người dùng, tập này được dùng cho mô hình, giải thuật phục vụ cho nghiên cứu.

Chúng tôi dùng tập huấn luyện gán nhãn dữ liệu cho 5040 câu, như mô tả ở Bảng 1. Về thử nghiệm, chúng tôi chọn 20% câu ngẫu nhiên từ dữ liệu trên (1008 câu) để kiểm thử cho các thuật toán đánh giá với mong muốn mang lại độ tin cậy cho mô hình đã đề xuất.

Sau khi chúng tôi tiến hành thực nghiệm trên mô hình đề xuất, kết quả đánh giá và thử nghiệm của mô hình được mô tả ở Bảng 2 và minh họa ở Hình 4.

Kết quả thực nghiệm các giải thuật trên cùng một tập dữ liệu cho thấy:

- Giải thuật NaiveBayes có thời gian huấn luyện nhanh nhất, đúng với ưu điểm của giải thuật này với thời gian 0,07s. Nhưng kết quả về độ chính xác của giải thuật này không cao, chỉ đạt mức 83,4% với độ lệnh chuẩn khá lớn 16,6%.
- Giải thuật SVM cho thời gian huấn luyện chậm hơn giải thuật NaiveBayes (0,25s), tuy nhiên giải thuật này lại đạt được độ chính xác cao hơn, lên đến 85,4% với độ lệch chuẩn 14,6%.
- Giải thuật FastText cho thời gian huấn luyện chậm nhất với 0,95s, bởi vì FastText biểu thị mỗi từ dưới dạng n-gam ký tự, điều này giúp nắm bắt ý nghĩa các từ ngắn hơn. Nhưng đây lại là

giải thuật cho độ chính xác cao nhất lên đến 90,1% với độ lệch chuẩn thấp nhất 9,9%. Trong phạm vi nghiên cứu này, chúng tôi đề xuất FastText là giải pháp được lựa chọn áp dụng để cài đặt cho mô hình đề xuất và giải quyết bài toán phân loại nhu cầu được đề ra.

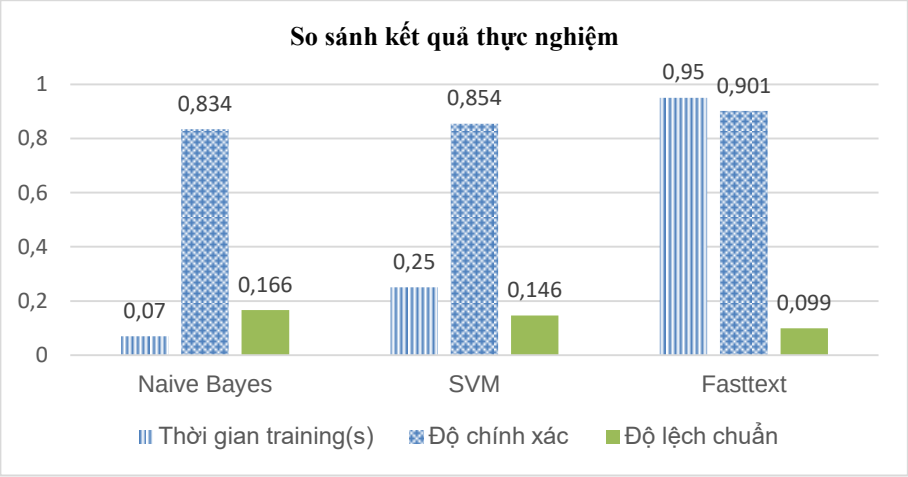
Bảng 1. *Thống kê dữ liệu thu thập trên các nhãn*

STT	Nhãn dữ liệu	Tổng số dữ liệu
1	MYTV	460
2	MOBILE	2861
3	INTERNET	1719

Bảng 2. *Kết quả đánh giá thực nghiệm*

Giải thuật	Huấn luyện	Kiểm thử	Kết quả		
			Thời gian training(s)	Độ chính xác	Độ lệch chuẩn
Naive Bayes	5040	1008	0,07	83,4%	16,6%
SVM	5040	1008	0,25	85,4%	14,6%
Fasttext	5040	1008	0,95	90,1%	9,9%

Với kết quả nghiên cứu này, chúng tôi đánh giá những ưu điểm và hạn chế sau. Mô hình xây dựng đơn giản, thân thiện, rõ ràng và không phức tạp cho người phát triển hệ thống. Thông qua giải pháp đề xuất, chúng tôi đã xác định và đánh giá được khá chính xác nhu cầu về Intermet, Mobile, MyTV. Nghiên cứu này đã tạo được một dataset về VNPT, có thể làm nền tảng cho sử dụng và mở rộng sau này. Việc áp dụng giải pháp FastText để phân loại nhu cầu về lĩnh vực VNPT bằng tiếng Việt là nghiên cứu mới chưa được thực hiện và mang lại kết quả tốt. Việc đánh giá và so sánh các giải pháp được thực hiện có hệ thống.



Hình 4. *Minh họa kết quả thực nghiệm*

Tuy nhiên, việc phải xây dựng dữ liệu huấn luyện mất khá nhiều thời gian. Giải pháp đề xuất mới chỉ dừng lại xử lý dữ liệu trên những câu đơn giản như có từ khóa hoặc nội dung liên quan. Mẫu dữ liệu vẫn còn hạn chế và có thể ảnh hưởng đến độ chính xác của thực nghiệm.

4. Kết luận

Nghiên cứu này đã thu thập và tạo ra được Dataset chứa hơn 5.000 câu phản hồi về các dịch vụ trong lĩnh vực viễn thông bằng tiếng Việt. Dataset này được sử dụng trong quá trình nghiên cứu và có thể được sử dụng cho những ai cần nghiên cứu về sau. Dữ liệu này cũng có thể được mở rộng và bổ sung thêm vào từ điển các tri thức thu thập được từ các bài viết, bài báo trên mạng xã hội cho quá trình nghiên cứu tiếp theo. Nghiên cứu này đã đề xuất mô hình được sử dụng dựa

trên tiếp cận FastText, đồng thời đánh giá, so sánh với các thuật toán Naïve Bayes và SVM trong các pha của mô hình đề xuất. Kết quả đánh giá nhu cầu của người dùng đã được đánh giá qua dữ liệu thực nghiệm thu thập thực tế của ngành VNPT. Kết quả đánh giá thu được có độ chính xác lên đến 90,1% là trong khoảng chấp nhận được với môi trường thử nghiệm. Điều này chứng tỏ giải pháp đề xuất có khả năng ứng dụng cao và áp dụng vào việc xác định nhu cầu người dùng thực tế tại VNPT nhằm nâng cao hiệu quả sản xuất và kinh doanh.

Tuy nhiên, giải pháp đề xuất chủ yếu tập trung vào xác định các nhu cầu của khách hàng về các dịch vụ viễn thông trên dữ liệu chứa một số loại câu đơn giản được huấn luyện, trùng lặp về từ khóa. Trong tương lai, chúng tôi sẽ mở rộng thêm các loại câu phức tạp hơn trong tiếng Việt, các câu kèm biểu tượng và mở rộng phân tích cảm xúc khách hàng về các dịch vụ chúng tôi cung cấp, từ đó có thể làm căn cứ để phục vụ công tác hậu kiểm, chăm sóc sau bán hàng và điều chỉnh các chiến lược phát triển của ngành. Xây dựng hệ thống hỏi đáp và hỗ trợ các hệ thống HelpDesk cũng là một định hướng nghiên cứu trong thời gian tới.

Lời cám ơn

Nghiên cứu này được tài trợ bởi Quỹ Phát triển Khoa học và Công nghệ Quốc gia (NAFOSTED) trong đề tài mã số 503.01-2021.14.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] A. S. Al-Malaise and F. Saleem, "Impact of Artificial Intelligence on Social Media Networks," *Journal of Electrical Systems*, vol. 20, no. 9, pp. 2112-2118, 2024.

[2] C. Gerling and S. Lessmann, "Leveraging AI and NLP for Bank Marketing: A Systematic Review and Gap Analysis," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2411.14463. [Accessed January 11, 2024].

[3] Blog.fastText, "FastText," 2016. [Online]. Available: https://fasttext.cc/docs/en/english-vectors.html, [Accessed January 11, 2024].

[4] L. P. Hoang, H. T. Le, H. V. Tran, T. C. Phan, D. M. Van, P. A. Le, D. T. Nguyen, and C. Pong-inwong, "Does Evaluating Peer Assessment Accuracy and Taking It into Account in Calculating Assessor's Final Score Enhance Online Peer Assessment Quality?" *Education and Information Technologies*, vol. 27, pp. 4007-4035, 2022, doi: 10.1007/s10639-021-10763-1.

[5] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Journal Expert Systems with Applications: An International Journal archive*, vol. 36, no. 7, pp. 10760- 10773, 2009, doi: 10.1016/j.eswa.2009.02.063.

[6] P. D. Turney, "Thumbs up or thumbs down, semantic orientation applied to unsupervised classification of reviews," *Proc. of the 40th ACL, Philadelphia*, July 2002, pp. 417-424.

[7] B. Pang and L. Lee, and S. Vaithyanathan, "Thumbs up Sentiment classification using machine learning techniques," *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002, pp. 79-86, doi: 10.3115/1118693.1118704.

[8] K. Dave, S. Lawrence, and D. Pennock, "Mining the peanut gallery Opinion extraction and semantic classification of product reviews," *WWW '03: Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519-528, doi: 10.1145/775152.775226.

[9] M. Taboada, C. Anthony, and K. Voll, "Methods for Creating semantic orientation Dictionaries," in *Proceedings of 5th international conference on language resources and evaluation (LREC'06)*, May, 2006, Genoa, Italy, pp. 427-432.

[10] P. Beineke, T. Hastie, and S. Vaithyanathan, "The sentimental factor: Improving review classification via human-provided information," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, June, 2004, Barcelona, Spain, pp. 263-270.

[11] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees," in T. B. Ho, D. Cheung, H. Liu, (eds) *Advances in Knowledge Discovery and Data Mining*. PAKDD 2005, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol. 3518, pp. 301-311, 2005, doi: 10.1007/11430919_37.

[12] D. Dinh, K. Hoang, and V. T. Nguyen, "Vietnamese Word Segmentation," in *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan 2001, pp. 1-8.
[13] C. T. Nguyen and X. H. Phan, "JVnTextPro: A Java-based Vietnamese Text Processing Tool," 2010. [Online]. Available: http://jvntextpro.sourceforge.net/. [Accessed October 5, 2024].
[14] H. P. Le and T. M. H. Nguyen, "vnTokenizer," 2020. [Online]. Available: https://github.com/vuthaihoc/vntokenizer4.1. [Accessed October 5, 2024].
[15] V. Anh, "Underthesea," 2017. [Online]. Available: https://github.com/undertheseanlp/underthesea. [Accessed October 5, 2024].
[16] Medium, "FastText," 2018. [Online]. Available: https://medium.com/@mariamestre/fasttext-stepping-through-the-code-259996d6ebc4. [Accessed October 5, 2024].