NGHIÊN CỨU MÔ HÌNH PHÁT HIỆN CHỦ ĐỀ BÌNH LUẬN VÀ PHÂN TÍCH CẢM XÚC NGƯỜI DÙNG ĐIỆN THOẠI TRÊN DỮ LIỆU LỚN

Lê Ngọc Lợi[*], Lê Thị Thanh Ngân[*], Nguyễn Thái Anh[*], Phan Hồ Viết Trường[*]

TÓM TẮT

Phân tích cảm xúc người dùng là bài toán không còn quá xa lạ trong lĩnh vực xử lý ngôn ngữ tự nhiên. Đầu vào là tập hợp các văn bản, và phân loại cảm xúc sẽ là kết quả của văn bản đầu vào. Bài toán này được ứng dụng trong rất nhiều lĩnh vực khác nhau như giáo dục, thương mại điện tử, hay nhà hàng khách sạn. Nhưng hiện có ít nghiên cứu về phân tích cảm xúc theo chủ đề mà người dùng bình luận trên tập dữ liệu lớn. Nghiên cứu này sử dụng mô hình phân lớp logistic regression để dự báo cảm xúc của người dùng với tập dữ liệu lên đến gần 8 triệu dòng bình luận đánh giá bằng tiếng Anh của trang bán hàng trên website bán hàng Amazon.com kết hợp với mô hình phát hiện chủ đề Latent Dirichlet Allocation, là một mô hình thống kê nhằm đưa ra các chủ đề ẩn trong tập dữ liệu văn bản.

Từ khóa: *Latent Dirichlet Allocation (LDA), semantic analysis, phân tích chủ đề, dữ liệu lớn*

1. Giới thiệu

Xử lý ngôn ngữ tự nhiên có hai nhóm: hiểu ngôn ngữ và tạo ra ngôn ngữ của con người (Su và cộng sự, 2020). Tuy nhiên, nhận dạng giọng nói, trả lời các câu hỏi như chat bot hay là dịch ngôn ngữ bằng google dịch đều sử dụng NLP. Hai lĩnh vực quan trọng của quá trình là phân tích cảm xúc con người và nhận biết cảm xúc con người (Fang & Zhan, 2015; Nandwani & Verma, 2021). Phát hiện cảm xúc con người là bài toán xác định những cảm xúc của con người như buồn, vui, lo lắng, tức giận, chán nản,… (Dzedzickis và cộng sự, 2020). Ngược lại, phân tích những cảm xúc con người là bài toán đánh giá dữ liệu nằm trong nhóm giá trị định nghĩa trước. Trong giáo dục, ban giám hiệu nhà trường có thể biết được cảm nhận của học sinh hay sinh viên khi tham gia lớp học đó và đánh giá phương pháp dạy đúng đắn của các thầy giáo, cô giáo. Thông qua việc đánh giá cảm xúc, nhà trường có thể biết học sinh, sinh viên có hứng thú với một môn học nào đó hay không (Barron-Estrada, 2019). Trong kinh doanh, các nhà cung cấp sử dụng mạng xã hội để quảng bá các sản phẩm của họ và thu hút phản hồi thông qua nhận xét của khách hàng. Từ nhận xét của khách hàng, doanh nghiệp nắm bắt được mức độ hài lòng của khách hàng, từ đó tăng doanh số bán hàng (Al Ajrawi và cộng sự, 2021).

[*] Khoa Công nghệ Thông tin, Trường Kỹ Thuật và Công Nghệ Văn Lang, Đại học Văn Lang, TPHCM

Tuy vậy, các công trình về phân tích cảm xúc của người dùng sản phẩm theo chủ đề bình luận là rất ít. Nhu cầu thu thập và phân tích dữ liệu để rút ra những kiến thức từ nguồn bình luận sản phẩm rất cần thiết để từ đó cải tiến chất lượng sản phẩm. Với việc mua sắm trực tuyến ngày càng tăng, các website bán hàng cho phép doanh nghiệp thu thập những bình luận về sản phẩm, dịch vụ nhằm đưa ra những định hướng tốt hơn. Những năm gần đây, nhiều doanh nghiệp đã cung cấp công cụ để phân tích phản hồi của khách hàng trên các website từ đó đưa ra thang điểm và mức độ hài lòng của người dùng về sản phẩm, dịch vụ qua những câu văn và bình luận (Naseem và cộng sự, 2021). Tuy nhiên, thước đo này chưa đưa ra được vấn đề chi tiết của sản phẩm để doanh nghiệp kịp thời tìm những giải pháp, ra quyết định, hướng đi tốt hơn.

Trong công trình này, chúng tôi sẽ sử dụng mô hình logistic regression phân tích cảm xúc của người dùng với tập dữ liệu lên đến gần 8 triệu dòng bình luận đánh giá bằng tiếng Anh của trang bán hàng trên website bán hàng Amazon.com kết hợp mô hình Latent Dirichlet Allocation (LDA), là kiểu mô hình thống kê và phát hiện các chủ đề tiềm ẩn ở trong tập dữ liệu. Qua đó có thể nhận biết những vấn đề mà khách hàng quan tâm giúp nhà quản trị đưa ra những khuyến nghị phù hợp để phát triển sản phẩm tốt hơn.

2. Tổng quan nghiên cứu

2.1. Tổng quan về các nghiên cứu trong và ngoài nước

Phân tích cảm xúc của người dùng là một dạng xử lý ngôn ngữ tự nhiên thông qua việc sử dụng những giải thuật trong học máy để xây dựng những ứng dụng giúp cho nhà kinh doanh hiểu được mối quan tâm của khách hàng về sản phẩm đang dùng dễ dàng hơn.

Tác giả Wankhade và cộng sự (2022) đã có một bài báo về một cuộc khảo sát về các phương pháp phân tích tình cảm, ứng dụng và thách thức. Tuy nhiên còn nhiều thách thức trong quy trình đánh giá và phân tích cảm xúc Những thách thức này tạo ra trở ngại để giải thích chính xác tình cảm và xác định cực tình cảm thích hợp. Phân tích cảm xúc xác định và trích xuất thông tin chủ quan từ văn bản bằng cách ứng dụng xử lý ngôn ngữ tự nhiên và khai thác văn bản. Bài báo này thảo luận tổng quan đầy đủ về phương pháp hoàn thành nhiệm vụ này cũng như các ứng dụng của phân tích tình cảm và sau đó sẽ đưa ra những đánh giá, so sánh và đối chiếu các phương pháp tiếp cận được sử dụng để có được sự hiểu biết toàn diện về ưu điểm và nhược điểm của chúng. Cuối cùng, những thách thức của phân tích tình cảm được xem xét để xác định các hướng đi trong tương lai.

Trong bài báo về phân tích tình cảm dựa trên dữ liệu Twitter, Ficamos và Yan (2016) đề xuất phương pháp phân tích tình cảm dựa vào các chủ đề được trích xuất từ tập văn bản và ước tính tình cảm theo chủ đề với thuật toán học máy. Kết quả đạt được giúp các thuật toán giảm bớt độ độ phức tạp và các bước tiền xử lý. Tuy nhiên công trình cần nghiên cứu thêm về trích xuất chủ đề và phương pháp xử lý trung lập.

Bài báo của Bahrawi (2019) đề xuất phương pháp phân tích tình cảm với nguồn dữ liệu từ Twitter bằng mô hình Random Forest. Nghiên cứu có độ chính xác 75%.

Bài báo phân tích chủ đề và tình cảm dựa trên các bài đánh giá về Omni-Channel của Kim và Yoo (2021) phân tích tần số được thực hiện và phân tích LDA (Latent Dirichlet Allocation) ứng dụng vào để phân tích dữ liệu lớn về phản hồi phản ứng của người đánh giá. Phân tích đặc điểm của từng chủ đề về cảm xúc tích cực hoặc tiêu cực. Tuy nhiên phân tích chủ đề với LDA là phương pháp cơ học trích xuất thông qua xác xuất của từ trong một câu nên có giới hạn trong toàn bộ nội dung.

Bài báo về phân tích cảm xúc cho khóa học từ xa của Osmanoğlu và cộng sự (2020) dùng phương pháp Likert và máy học để phân tích. Với tỷ lệ chính xác 0,775 thông qua thuật toán hồi quy logistic đã đạt kết quả đúng. Tỷ lệ thành công trong nghiên cứu dao đọng từ 42% đến 85%, ưu điểm lớn nhật của phân tích tình cảm là những đánh giá không tuân theo quy tắc ngữ pháp.

Nghiên cứu của Farkhod và cộng sự (2021) sử dụng LDA để khai thác các chủ đề ẩn và áp dụng mô hình TDS để xác định cảm xúc của từng chủ đề. Phương pháp TDS cho thấy chỉ đạt được kết quả chính xác về hai khả năng phản hồi là tích cực và tiêu cực.

Nghiên cứu của Al Ajrawi và cộng sự (2021) đã dung mô hình đa ngôn ngữ để tìm những cảm xúc tích cực trong các bình luận trên mạng xã hội. Mô hình cho kết quả tốt hơn so với kỹ thuật dùng không gian vectơ.

Tại Việt Nam cũng có nhiều nghiên cứu về phân tích cảm xúc như các công trình sau:

Bài báo về phân tích ý kiến của Trần Thị Ngọc Thảo và cộng sự (2014) đã sử dụng kỹ thuật supported vector machine kết hợp với thư viện WEKA để xây dựng một ứng dụng phân tích ý kiến của các bình luận bằng tiếng Anh về mỹ phẩm dành cho phụ nữ. Nghiên cứu này chưa thực hiện trên các dạng câu khác nhau như câu phủ định, câu so sánh và câu điều kiện.

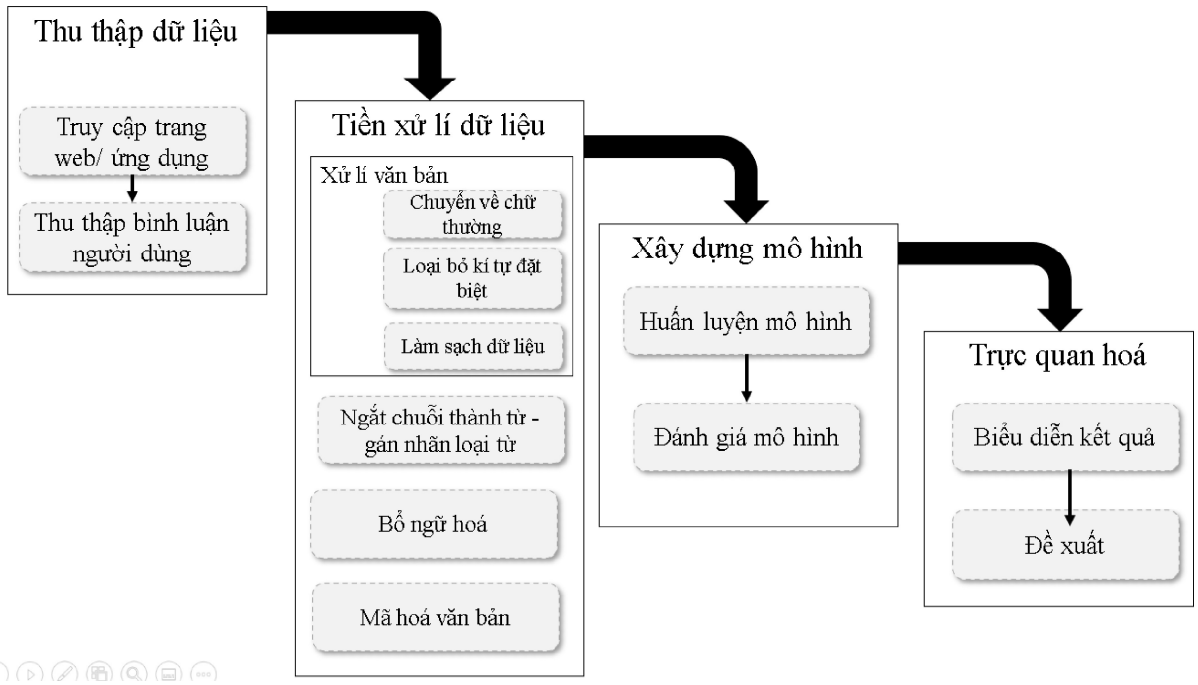
Bài báo về phân tích ý kiến của Thái Kim Phụng và cộng sự (2020) thử nghiệm nhiều mô hình học máy với tập dữ liệu huấn luyện để dự đoán cảm xúc cho toàn bộ tập dữ liệu có kết quả chính xác nhất. Công trình này cần mở rộng để thu thập đánh giá trên nhiều sản phẩm, bổ sung thêm nhiều cấp độ của thang phân tích cảm tính.

Báo cáo hội thảo khoa học của Nguyễn Thành Thủy và Trần Thị Châu Giang (2019) ứng dụng học máy trong phân tích ý kiến khách hàng dựa trên văn bản tiếng Việt trong dịch vụ khách sạn. Trong đó, giải pháp tập trung cải tiến: tiền xử lý, chuẩn hóa, gán lại nhãn cho dữ liệu huấn luyện bằng kỹ thuật Error Analysis; Tăng cường dữ liệu huấn luyện bằng từ điển cảm xúc; Sử dụng 5-FoldCV, Confusion Matrix để kiểm soát overfitting và underfitting và kiểm thử mô hình; Tunning Hyper Parameter để tối ưu hóa tham số mô hình; Ensemble Methods kết hợp sub-models để đưa ra mô hình học máy có sự kết hợp

cao về hiệu suất. Đối với model này, tác giả nhận thấy khá hiệu quả với độ chính xác đạt đến 96,03 %.

A. Giải quyết bài toán nghiên cứu

Trong bài viết này, chúng tôi quyết định sử dụng hồi quy Logistic và phân lớp Naive Bayes để tiến hành đào tạo phân tích cảm xúc của người dùng sản phẩm điện thoại. Ngoài việc giảm thiểu số lượng tính toán cho các vấn đề với phạm vi lên đến 8 triệu dòng dữ liệu, cũng như để xác định các chủ đề xuất hiện trong nhận xét của người dùng, mô hình được chúng tôi sử dụng là phân tích Latent Dirichlet Allocation (LDA) và tính toán phân tán Apache Spark.



Hình 1. Mô hình tổng quan phân tích cảm xúc người dùng

B. Mô hình LDA

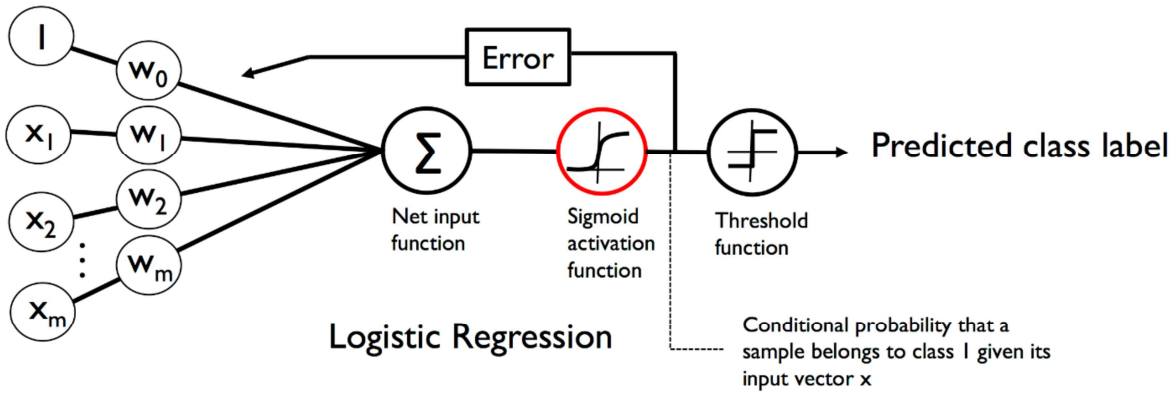
Mô hình LDA là một trong những mô hình chủ đề ẩn để trích xuất các chủ đề từ một tập ngữ liệu nhất định (Liu và cộng sự, 2016). Mỗi tài liệu được tạo thành từ nhiều từ khác nhau và mỗi chủ đề cũng có nhiều từ khác nhau thuộc về nó.

C. Thuật toán Hồi quy logistic

Là một giải thuật máy học có giám sát được ứng dụng để phân loại dữ liệu được trình bày trong hình 2 (Liu, 2018; Dreiseitl & Ohno-Machado, 2002). Để phân loại, một mô hình hoc máy thường bao gồm các thành phần sau (Jurafsky & Martin, 2008):

1. Biểu diễn đặc điểm của đầu vào: đối với mỗi quan sát đầu vào ($x^{(i)}$), điều này sẽ được đại diện bởi các vectơ đặc trưng [x_1, x_2,..., x_n]

2. Hàm phân loại: tính toán lớp ước tính. Hàm sigmoid được sử dụng để phân loại.
3. Một hàm mục tiêu: công việc của hàm mục tiêu là giảm thiểu sai số của các ví dụ huấn luyện. Hàm Entropy (cross-entropy) chéo thường được sử dụng cho mục đích này.
4. Hàm tối ưu hóa: sử dụng để tối ưu hóa hàm mục tiêu và giảm tốc độ dốc ngẫu nhiên.



Hình 2. Mô hình hồi quy logistic

Biểu diễn mô hình hồi quy với:

- $x = [1, x_1, x_2,..., x_n]$ là các biến độc lập biểu thị cho các đặc trưng trong data set.
- $w = [w_0, w_1, w_2,..., w_n]$ là các tham số cần học, với là hệ số bias b.
- Mạng: input $\sum = w^T x$

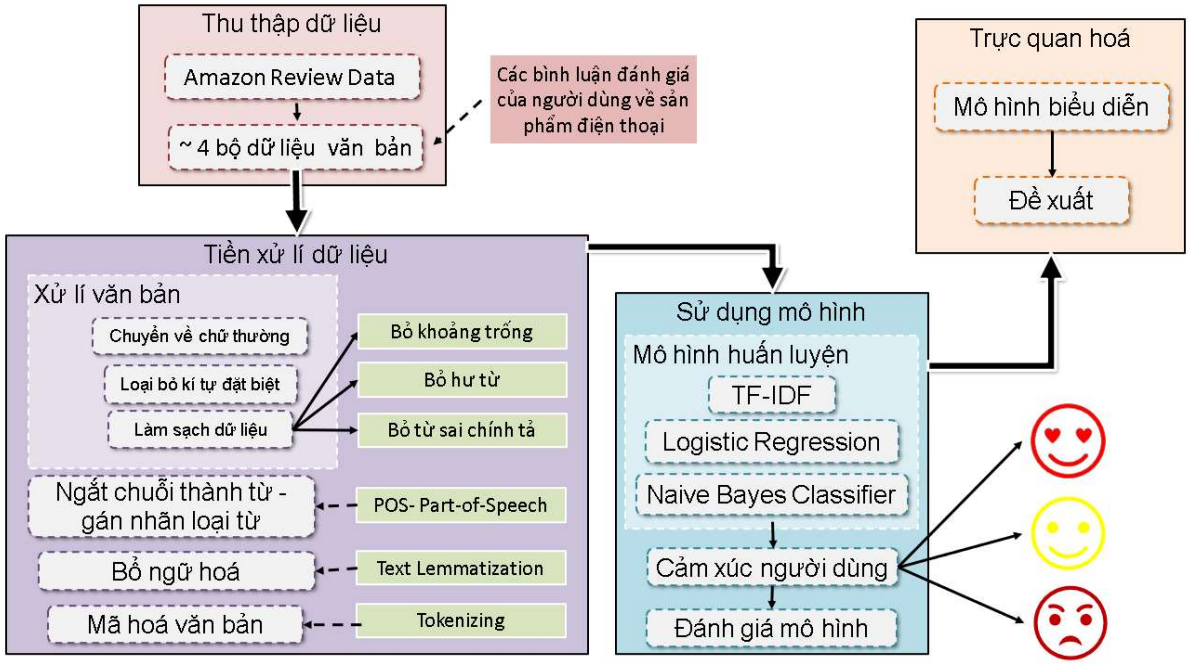
D. Thuật toán naïve bayes

Naive Bayes là một trong những thuật toán phân loại Bayes cổ điển, có cấu trúc thuật toán đơn giản và hiệu quả tính toán cao (Chen và cộng sự, 2021). Thuật toán sẽ tính xác suất của và xuất kết quả với xác suất cao nhất theo công thức 1. Nó được ứng dung trong phân tích dữ liệu và các lĩnh vực khác vì cấu trúc thuật toán nhanh và đơn giản.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

3. Phương pháp nghiên cứu

Sơ đồ tổng quát mô hình đề xuất của chúng tôi được trình bày trong hình 3. Hệ thống gồm 4 thành phần là (1) thu thập dữ liệu; (2) tiền xử lý dữ liệu; (3) sử dụng mô hình; (4) trực quan hóa.



Hình 3. Tổng quan mô hình phân tích cảm xúc người dùng

3.1. Thu thập dữ liệu

Dữ liệu lấy từ trang Amazon Review Data (2018). Chúng tôi chia thành 4 tập dữ liệu với kích thước khác nhau:

- Bộ 1 với hơn 100.000 dòng dữ liệu dùng để xác định chủ đề và phân tích tình cảm cũng như kiểm tra tốc độ thời gian để so sánh Pandas và Pyspark.
- Bộ 2 với hơn 250.000 dòng dữ liệu dùng để xác định chủ đề và phân tích cảm xúc đồng thời kiểm tra tốc độ thời gian để so sánh giữa Pandas so với Pyspark.
- Bộ 3 với khoảng 3.500.000 dòng dữ liệu dùng để xác định chủ đề và phân tích cảm xúc với Pyspark.
- Bộ 4 với khoảng 8.000.000 dòng dữ liệu để xử lý dữ liệu khi thực hiện xác định chủ đề và phân tích cảm xúc với Pyspark.

Biểu diễn dữ liệu một bài đánh giá, bình luận trên 1 dòng dữ liệu trong JSON, sau đó chúng tôi nhận thấy trong tập dữ liệu có những trường không dùng đến trong mô hình phân tích cảm xúc, cũng như không đưa ra dự đoán đánh giá tích cực, còn gây tốn bộ nhớ lưu trữ, nếu xử lí ở dữ liệu càng lớn sẽ rất tốn thời gian nên tiến hành xoá các trường dữ liệu không cần thiết để giảm tải việc tính toán của máy. Tiếp theo, chúng tôi gắn nhãn dữ liệu dựa trên xếp hạng dẫn đến điểm xếp hạng nhỏ hơn 3 có ý nghĩa tiêu cực (Negative) và ngược lại xếp hạng lớn hơn 3 có ý nghĩa tích cực (Positive) ở mức bằng 3 có nghĩa là trung tính (Neutral). Sau khi gán nhãn dựa điểm đánh giá ở các bình luận, chúng tôi chỉ định và xếp hạng nhị phân cho bộ dữ liệu, với 1: Positive và 0: Negative.

3.2. Tiền xử lý dữ liệu

– Xử lý văn bản: Đầu tiên, chúng tôi sẽ đổi tất cả các chữ hoa sang chữ thường vì chữ hoa không thể nào phân biệt được dữ liệu vào khi chữ hoa khác chữ thường về ngữ nghĩa và từ đó kết quả bị ảnh hưởng khi dự đoán. Việc xóa các kí tự đặc biệt, các biểu tượng sẽ giúp cho quá trình phân tính không bị nhiễu. Cuối cùng để cho dữ liệu trở nên sạch hơn, giúp cho hệ thống máy tính dễ dàng phân tích, việc cuối cùng chúng ta cần làm làm loại bỏ khoảng trống, loại bỏ các hư từ, và bỏ các từ sai chính tả.

– Tính TF, IDF: Nhằm mục đích giảm các từ có mức độ quan trọng cho việc thực hiện training dữ liệu với mô hình LDA ta cần tính TF và IDF.

– Ngắt chuỗi thành từ và gán nhãn kiểu từ nhằm định vị các từ trong câu. Đây là bước phân tách các câu, các đoạn văn hoặc những văn bản thành các đơn vị nhỏ hơn. Chúng tôi chủ yếu gán nhãn các loại từ như: danh từ (Noun), tính từ (adjectives), động từ (Verb). Sau khi gắn thẻ văn bản chúng tôi chỉ giữ lại những văn bản có thẻ phù hợp.

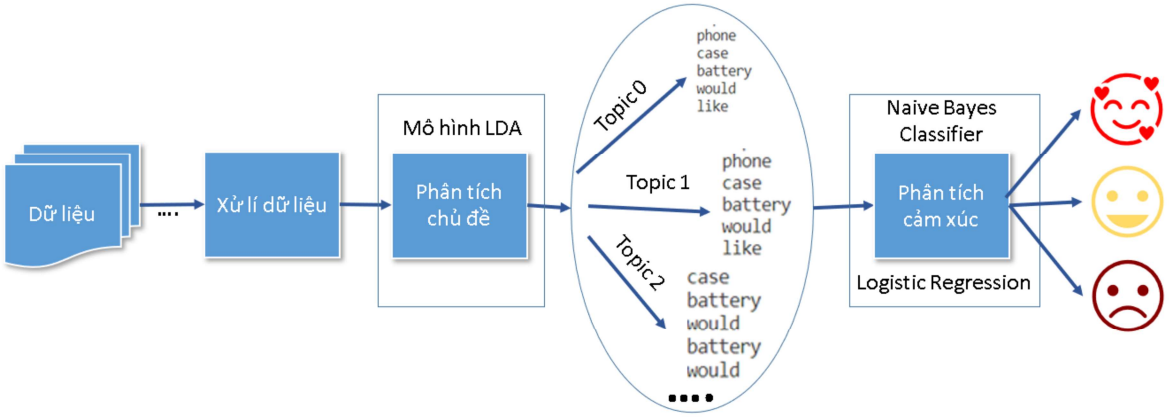
– Bổ ngữ hóa: tổng hợp lại các dạng từ với nhau, nó liên kết các từ có nghĩa tương tự với một từ và ánh xạ các từ khác nhau vào một gốc. Ví dụ như: “walk”, “walked” và “walking”.

– Mã hoá văn bản: chúng tôi tiến hành mã hóa văn bản sau khi bổ ngữ. Những từ này sẽ được chuyển thành ID và giúp cho việc mô hình hóa theo các quy tắc khác nhau cho trước.

Để quá trình huấn luyện diễn ra dễ dàng và hiểu quả thì việc chuẩn bị dữ liệu phù hợp là một trong các yếu tố cần thiết. Vì ở nhiều mô hình huấn luyện với dữ liệu lớn thì cần rất nhiều thời gian nên nếu dữ liệu không thích hợp thì sẽ gây ra vô cùng tốn thời gian cho quá trình huấn luyện.

3.3. Huấn luyện dữ liệu

Với mô hình LDA, mô hình sẽ tạo ra những chủ đề ẩn ứng với các thành phần trong sản phẩm mà khách hàng đang quan tâm. Sau đó, chúng tôi sử dụng mô hình Logistic Regression để phân loại cảm xúc người dùng được trình bày trong hình 4.



Hình 4. Tổng quan các mô hình phát hiện chủ đề và phân tích cảm xúc người dùng

4. Thực nghiệm

Mô hình của chúng tôi được cài đặt trên máy tính có cấu hình như CPU 11th Gen Intel® Core™ i7-11800h @ 2.30GHz (16 CPUs), ~2.3GHz, RAM: 16 GB. Máy tính cài phần mềm Sublime Text, python 3.9.0

Tổ chức thư mục

- Thư mục mã nguồn được tổ chức bao gồm 2 thư mục là data và code.
- Trong đó các thư mục có ý nghĩa như sau:
- Thư mục data chứa các tập dữ liệu dùng để thực hiện đề tài
- Thư mục code bao gồm 3 file *.py bao gồm 1 file xử lý phân tích topic, 1 file xử lý và cho ra kết quả đánh giá cảm xúc thông qua bình luận và 1 file thể hiện sự so sánh giữa 2 mô hình là LRM và NBM

A. Kết quả tiền xử lý dữ liệu

Chúng tôi thu thập 10 triệu dòng bình luận đánh giá bằng tiếng Anh của trang bán hàng trên website bán hàng Amazon.com dữ liệu được biểu diễn trong file.cvs như sau:

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
1	A1YX2RBMS1L9I	1E+08	Andrea Busch	[0, 0]	Saw this same	5	Great product	1353542400	11 22, 2012
2	A180NNPPKWC0	1E+08	Aniya pennington	[3, 3]	case fits perfec	5	Perfect	1374105600	07 18, 2013
3	A3HVRXV0LVJN7	1E+08	BiancaNicole	[4, 4]	Best phone ca:	5	A++++	1358035200	01 13, 2013
4	A292527VPX98P	1E+08	Cebell	[0, 1]	It may look cut	1	Do NOT GET IT!	1353888000	11 26, 2012
5	A1BJGDS0L1IO6I	1E+08	cf "t"	[0, 3]	ITEM NOT SEN	1	ITEM NOT SENT	1359504000	01 30, 2013
6	ANG01NK4RXCIS	1E+08	Charles Kodi Bon	[1, 1]	this is a cute ca	2	meh	1389139200	01 8, 2014

Hình 5. Biểu diễn dữ liệu thu thập được trong file.csv

B. Kết quả đánh giá mô hình phân tính chủ đề LDA

Sử dụng thuật toán Latent Dirichlet Allocation trong nghiên cứu này cho ta thấy kết quả huấn luyện được thể hiện như sau:

Với mô hình LDA chúng tôi nhận được kết quả sau:

Topic 0:	Topic 1:	Topic 2:	Topic 3:	Topic 4:
phone	phone	phone	phone	phone
case	case	case	case	case
battery	battery	battery	battery	battery
would	would	would	would	would
like	like	like	like	like

Hình 6. Kết quả khi huấn luyện qua mô hình LDA

Bảng 1. Kết quả thời gian huấn luyện mô hình LDA qua 3 bộ dữ liệu (đơn vị: phút)

Bộ dữ liệu	Latent Dirichlet Allocation
review_phone_250k	7 phút
review_phone_3tr	45 phút
review_phone_8tr	125 phút

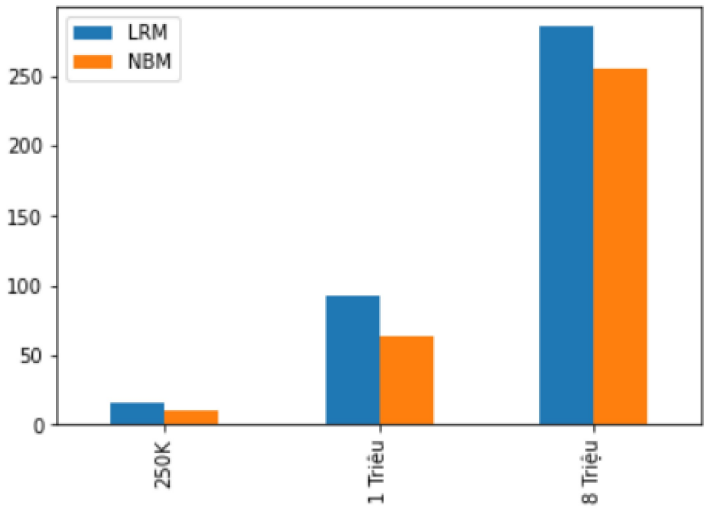
Kết quả ứng dụng kỹ thuật LDA đã cho ra kết quả cho tập dữ liệu với khoảng hơn 8 triệu dòng bình luận với thời gian là 125 phút.

C. Kết quả đánh giá mô hình phân tính cảm xúc

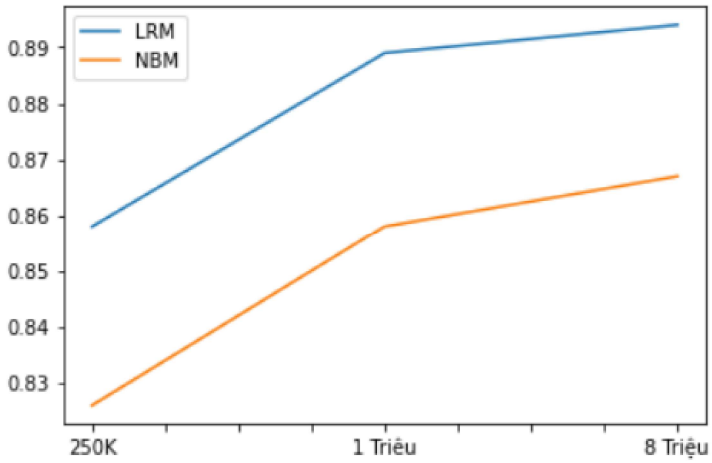
Hai mô hình được dùng để phân tích cảm xúc gồm: Logistic Regression, Naive Bayes Classifier cho ra kết quả như sau:

Bảng 2. Kết quả huấn luyện Logistic Regression và Naive Bayes Classifier qua ba bộ dữ liệu

Bộ dữ liệu	review_phone_250k		review_phone_3tr		review_phone_8tr	
Mô hình	Thời gian	Độ chính xác	Thời gian	Độ chính xác	Thời gian	Độ chính xác
Logistic Regression	15 phút	0.858	92 phút	0.889	285 phút	0.894
Naive Bayes Classifier	10 phút	0.826	64 phút	0.858	255 phút	0.867



Hình 7. Biểu đồ thời gian của Logistic Regression và Naive Bayes Classifier



Hình 8. Biểu đồ độ chính xác của Logistic Regression và Naive Bayes Classifier

Kết quả ứng dụng kỹ thuật LDA với phương pháp Logistic Regression, với độ chính xác lên đến **89,4%** cho tập dữ liệu với khoảng hơn 8 triệu dòng bình luận. Các chỉ số về độ chính trên từng tập review_phone_250k, review_phone_3tr là khá cao và tương quan, cho thấy mô hình máy học được xây dựng là đáng tin cậy.

5. Kết luận

Trong nghiên cứu này, chúng tôi đã xây dựng giải pháp phân tích cám xúc người dùng di động dựa trên dữ liệu từ 5 đến 10 triệu dòng bình luận được thu thập từ website Amazon. Mô hình LDA dùng để phát hiện các chủ đề mà người dùng đang bình luận và trực quan hóa dữ liệu nhằm cho các tổ chức có thể hiểu được tâm lý người dùng. Tiếp theo, chúng tôi dùng mô hình logistic regression để phân loại cảm xúc người dùng thành ba lớp

là positive, negative và neutral dựa trên các bình luận về sản phẩm để từ đó các công ty có thể ra quyết định những cải tiến phù hợp nhằm làm cho sản phẩm, dịch vụ ngày tốt hơn. Hướng phát triển tương lai, chúng tôi sẽ cải tiến mô hình phân tích kèm với điểm số nhận xét, và phân tích ngữ nghĩa câu bình luận nhằm tăng độ chính xác của mô hình đề xuất.

Tài liệu tham khảo

Su, S. Y., Chuang, Y. S., & Chen, Y. N. (2020). Dual Inference for Improving Language Understanding and Generation. *Findings of the Association for Computational Linguistics: EMNLP 2020,* 4930-4936.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1), 1-14.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1), 81.

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, *20*(3), 592.

Barron-Estrada, M. L., Zatarain-Cabada, R., & Bustillos, R. O. (2019). Emotion Recognition for Education using Sentiment Analysis. *Research in Computing Science*, *148*(5), 71-80.

Al Ajrawi, S., Agrawal, A., Mangal, H., Putluri, K., Reid, B., Hanna, G., & Sarkar, M. (2021). WITHDRAWN: Evaluating business Yelp's star ratings using sentiment analysis.

Naseem, S., Mahmood, T., Asif, M., Rashid, J., Umair, M., & Shah, M. (2021). Survey on sentiment analysis of user reviews. In *2021 International Conference on Innovative Computing (ICIC)* (pp. 1-6). IEEE.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731-5780.

Ficamos, P., & Yan, L. I. U. (2016). A topic based approach for sentiment analysis on Twitter data. *International Journal of Advanced Computer Science and Applications*, *7*(12).

Bahrawi, B. (2019). Sentiment analysis using random forest algorithm-online social media based. *Journal of Information Technology and Its Utilization*, *2*(2), 29-33.

Kim, S. H., & Yoo, B. K. (2021). Topics and sentiment analysis based on reviews of omni-channel retailing. *Journal of Distribution Science*, *19*(4), 25-35.

Osmanoğlu, U. Ö., Atak, O. N., Çağlar, K., Kayhan, H. & Can, T. (2020). Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach. Journal of Educational Technology and Online Learning, *3*(1), 31-48. DOI: 10.31681/jetol.663733

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. *Applied Sciences*, *11*(23), 11091.

Trần Thị Ngọc Thảo, Nguyễn Ngọc Kim Liên, & Ngô Minh Vương (2014). Phân tích ý kiến của nhận xét tiếng anh dựa trên phương pháp học máy. *Tạp chí Khoa học và Công nghệ, 52(4D)*,142-155

Thái Kim Phụng, Nguyễn An Tế, & Trần Thị Thu Hà (2020). Tiếp cận phương pháp máy học trong khai thác ý kiến khách hàng trực tuyến. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, *30*(10), 27-41.

Nguyễn Thành Thủy, Trần Thị Châu Giang (2019). Một mô hình học máy trong phân tích ý kiến khách hàng dựa trên văn bản tiếng Việt: Bài toán dịch vụ Khách sạn. *Kỷ yếu Hội thảo quốc gia 2019 "CNTT và ứng dụng trong các lĩnh vực"*.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1-22.

Liu, L. (2018). Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)* (pp. 157-160). IEEE.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352-359.

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing*. Prentice Hall, Upper Saddle River, NJ.

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 1-12.