

Phân Tích Cảm Xúc Về COVID-19 Dựa Trên Dữ Liệu Twitter Theo Thời Gian Thực

Giảng viên hướng dẫn: TS. Đỗ Trọng Hợp

Thực hiện: Nhóm 10 - Smiley Team

Lê Anh Tuấn¹, Nguyễn Công Danh¹, Võ Hoàng Vũ¹

Đại Học Công Nghệ Thông Tin
Đại Học Quốc Gia Thành Phố Hồ Chí Minh
Thành Phố Hồ Chí Minh, Việt Nam

¹{tuanla.14, danhnc.14, vuvh.14}@grad.uit.edu.vn
CH1902037, CH1902029, CH1902039

Tóm tắt. Trong thời đại hiện nay, sự phát triển của các kỹ thuật và công nghệ xử lý dữ liệu lớn, học máy và xử lý ngôn ngữ tự nhiên thúc đẩy sự phát triển của các ứng dụng thông minh, hỗ trợ phân tích cảm xúc, hoặc khai thác ý kiến của con người về một số vấn đề nhất định. Điều này đã trở thành một lĩnh vực thú vị cho việc nghiên cứu, kinh doanh, hoặc phục vụ xã hội. Trong đó, nguồn dữ liệu từ các mạng xã hội là một nguồn dữ liệu phong phú đang được khai thác. Trong phạm vi đồ án môn học, nhóm tìm hiểu và xây dựng ứng dụng phân tích cảm xúc của cộng đồng về COVID-19 dựa trên dữ liệu Twitter theo thời gian thực, áp dụng Apache Spark và Kafka. Hệ thống đề xuất được xây dựng phục vụ cho mục đích học tập, nghiên cứu. Bên cạnh đó, hệ thống có thể áp dụng vào thực tế, giúp chúng ta nắm được cảm xúc của cộng đồng, tín hiệu tích cực hoặc tiêu cực tăng hay giảm thông qua các số liệu thống kê, trong thời gian đại dịch COVID-19 đang diễn ra trên khắp toàn cầu. Kiến trúc hệ thống đề xuất gồm có: Tweepy truy cập Twitter streaming API để thu dữ liệu liên quan đến COVID-19 và lưu vào Apache Kafka; Spark Streaming đọc dữ liệu từ Kafka, đưa vào Spark MLlib và Spark DataFrame xử lý và phân tích; Kết quả phân tích, dùng các mô hình học máy bao gồm: Decision Tree, Random Forest, Logistic Regression và Naive Bayes, được lưu vào Kafka; Dữ liệu được phân tích sẽ qua Spark Streaming và Spark SQL xử lý, truy vấn và trích xuất dữ liệu báo cáo và thống kê, sau đó ghi dữ liệu vào hệ thống tập tin với định dạng CSV; Streamlit đọc dữ liệu từ hệ thống tập tin, dùng thư viện Plotly xử lý và hiển thị thông tin gồm các bảng, biểu đồ trực quan lên giao diện web. Bên cạnh đó, các chức năng được viết theo dạng mô-đun, tham số cấu hình được lưu tập trung, để dành cho việc tích hợp thêm chức năng, triển khai và mở rộng hệ thống cho các dịch vụ khác. Ngoài ra, nhà quản trị có thể thực hiện từ dòng lệnh, thu phát luồng dữ liệu, phân tích và hiển thị kết quả ra bộ nhớ, console, Kafka hoặc tập tin, đồng thời ghi lại nhật ký thu phát, hỗ trợ cho việc chẩn đoán hệ thống nếu gặp sự cố.

Từ khóa: COVID-19, Phân tích cảm xúc, Twitter streaming API, Kafka, Apache Spark, Học máy, Streamlit và Plotly.

1 GIỚI THIỆU

Hiện nay, tình hình dịch bệnh COVID-19 là vấn đề rất được quan tâm, gây ảnh hưởng nghiêm trọng đến đời sống vật chất và tinh thần của người dân trên toàn thế giới. Trong đó, các ý kiến thường được chia sẻ công khai trên mạng xã hội và đây cũng là nguồn được lựa chọn để đánh giá và phân loại cảm xúc, thông qua các bình luận của cộng đồng về đại dịch.

Trong bài viết này, nhóm đề xuất ứng dụng phân tích cảm xúc của cộng đồng về COVID-19 theo thời gian thực dựa trên dữ liệu Twitter, sử dụng nền tảng công nghệ Apache Spark và Kafka. Bài toán phân loại dữ liệu bao gồm các dòng bình luận trên mạng xã hội Twitter, sử dụng các thuật toán học máy được hỗ trợ bởi Apache Spark để phân tích và dự đoán cảm xúc của con người dựa trên các bình luận. Từ đó phân loại được các bình luận và nắm bắt cảm xúc của cộng đồng về đại dịch. Ngoài việc nắm bắt cảm xúc của con người về đại dịch COVID-19, hệ thống giúp chúng ta biết được tín hiệu tích cực khi số lượng bình luận tiêu cực của cộng đồng ngày càng giảm thông qua các biểu đồ thống kê hoặc ngược lại khi tình hình kiểm soát dịch bệnh không khả quan dẫn đến số lượng bình luận tiêu cực ngày càng tăng lên.

Bài viết này được cấu trúc như sau: Phần 1 giới thiệu tổng quan. Phần 2 mô tả các nghiên cứu liên quan đến hệ thống đề xuất. Phần 3 trình bày thông tin chi tiết về thiết kế và triển khai hệ thống được đề xuất. Phần 4 cung cấp các kết quả thực nghiệm, đánh giá độ chính xác của các thuật toán học máy áp dụng vào hệ thống đề xuất. Cuối cùng, kết thúc bài báo này với kết luận và hướng phát triển trong tương lai.

2 CÁC NGHIÊN CỨU LIÊN QUAN

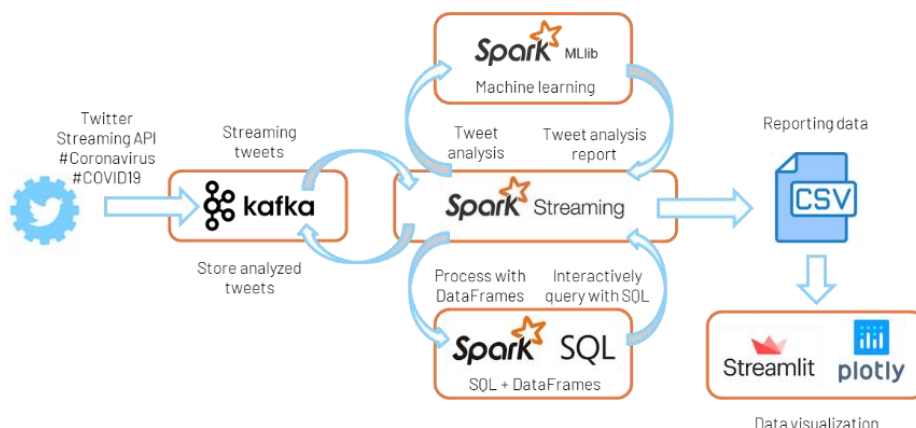
Trong thời đại công nghệ ngày nay, sự phát triển của các kỹ thuật xử lý ngôn ngữ tự nhiên, xử lý dữ liệu lớn và các thuật toán học máy hiện đại, đã tạo ra nhiều ứng dụng thông minh, áp dụng vào nhiều lĩnh vực trong thế giới thực. Đặc biệt, phân tích cảm xúc đã trở thành một chủ đề rất được quan tâm cho lĩnh vực nghiên cứu, kinh doanh và một số lĩnh vực khác. Điều này đề cập đến cảm xúc hoặc suy nghĩ của con người về một số vấn đề nhất định trong đời sống thực tế. Hơn nữa, nó cũng được coi là một ứng dụng trực tiếp để khai thác ý kiến. Một số ứng dụng cụ thể liên quan đến lĩnh vực này gồm có: Phân tích cảm xúc về COVID-19 dùng các thuật toán học máy [1] [2]; Phân tích cảm xúc dùng nền tảng công nghệ Apache Spark [3], nằm trong lĩnh vực mà nhóm đề xuất trong bài viết này.

3 HỆ THỐNG ĐỀ XUẤT

Phần này trình bày thiết kế và triển khai hệ thống được đề xuất, gồm có hai phần chính: Phần 3.1 mô tả thiết kế và kiến trúc hệ thống, và Phần 3.2 trình bày việc hiện thực và triển khai hệ thống được đề xuất.

3.1 Thiết kế và kiến trúc hệ thống

Kiến trúc của hệ thống đề xuất gồm có: Twitter streaming API được truy cập thông qua Tweepy để lấy dữ liệu theo thời gian thực, liên quan đến COVID-19 và lưu vào Kafka; Spark Streaming đọc luồng dữ liệu từ Kafka, đưa vào Spark MLlib và Spark DataFrame để xử lý và phân tích; Dữ liệu sau khi được phân tích, dùng các mô hình học máy, sử dụng các thuật toán bao gồm: Decision Tree, Random Forest, Logistic Regression và Naive Bayes, sau đó được lưu vào Kafka; Dữ liệu này sẽ qua Spark Streaming và Spark SQL để xử lý, truy vấn và trích xuất dữ liệu, và được lưu vào hệ thống tập tin với định dạng CSV; Nền tảng ứng dụng web, Streamlit đọc dữ liệu từ hệ thống tập tin, xử lý và hiển thị lên giao diện web, đồng thời dùng thư viện Plotly để tạo ra các báo cáo trực quan theo dạng bảng, biểu đồ cột để thống kê số liệu theo đơn vị đếm và phần trăm tương ứng với từng phân loại cảm xúc, gồm có: 0 - Negative, 1 - Neutral, và 2 - Positive. Bên cạnh đó, các chức năng được xây dựng có cấu trúc, tham số cấu hình được quản lý riêng biệt, dễ dàng cho việc bảo trì, triển khai và mở rộng cho các chức năng khác. Ngoài ra, nhà quản trị có thể thực hiện thu phát luồng dữ liệu, phân tích và hiển thị kết quả ra bộ nhớ, console, Kafka hoặc tập tin từ dòng lệnh, đồng thời ghi lại nhật ký thu phát, hỗ trợ cho việc chẩn đoán hệ thống.

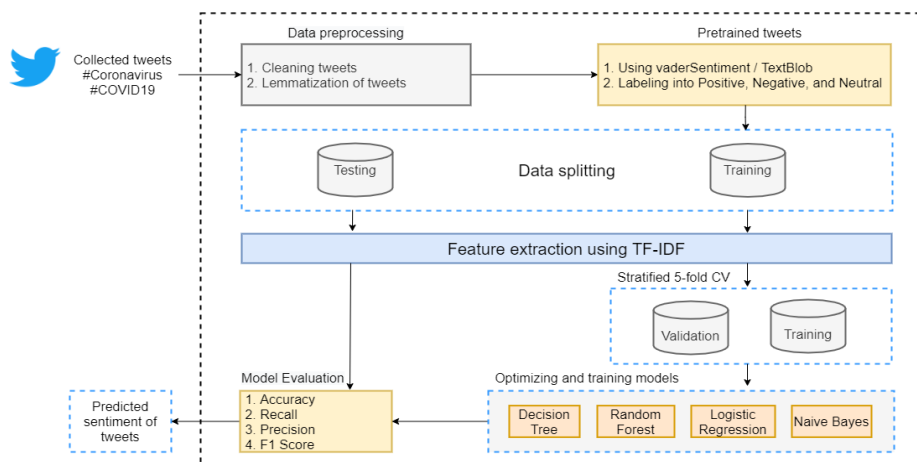


Hình 1. Kiến trúc hệ thống tổng quan.

Hệ thống các thành phần xử lý gồm có ba phần: Hệ thống xử lý offline, online, và ứng dụng web được mô tả như sau:

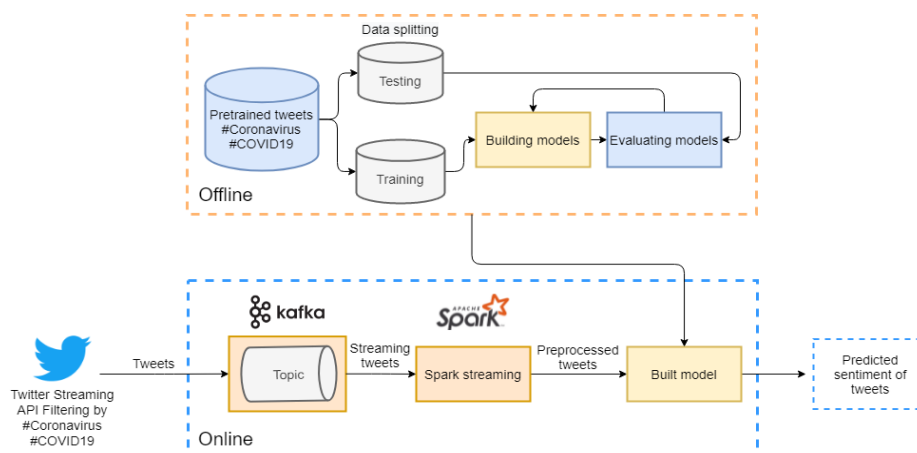
- Hệ thống xử lý offline:** Thành phần xử lý offline dùng để huấn luyện các mô hình học máy, dựa trên dữ liệu liên quan đến COVID-19 được thu thập từ Twitter, sau khi được xử lý, sàng lọc, cân đối qua nhiều công đoạn. Trong đó, các mô hình sử dụng các thuật toán khác nhau bao gồm: Decision Tree, Random Forest, Logistic Regression, và Naive Bayes, nhằm đánh giá và so sánh độ chính xác giữa các mô hình. Từ đó, lựa chọn và trích xuất mô hình phù hợp, có độ chính xác cao nhất, áp dụng vào hệ thống xử lý online. Các bước thực hiện như sau:

- **Bước 1:** Chuẩn bị môi trường thực hiện và dữ liệu thu thập từ mạng xã hội Twitter, với các bình luận liên quan đến COVID-19; Nạp dữ liệu và chuyển đổi dữ liệu, cập nhật lược đồ dữ liệu, đồng thời trích lọc các cột thuộc tính liên quan. Trong đó, bao gồm: cột định danh (id), ngày tháng (date), tên người dùng (user), thông tin vị trí người dùng (location), và lời bình luận (text).
- **Bước 2:** Tiền xử lý dữ liệu, gồm các công đoạn làm sạch dữ liệu, làm gọn từ, chuẩn hóa các từ có các ký tự lặp lại, chuyển đổi ký tự hoa sang ký tự thường, loại bỏ các khoảng trắng dư thừa, các biểu tượng, các số, và các thành phần không cần thiết như e-mail, địa chỉ website, v.v. đồng thời áp dụng kỹ thuật lemmatization, hoặc stemming.
- **Bước 3:** Gán nhãn cho dữ liệu. Trong đó, dùng vaderSentiment và text-blob để xác định các giá trị tương ứng đối với mỗi lời bình luận trong tập dữ liệu đầu vào. Kết quả phân tích nằm trong khoảng từ -1 (tiêu cực nhất) đến 1 (tích cực nhất). Để chuẩn hóa dữ liệu hoặc phân loại các bình luận của tập dữ liệu thuộc vào ba lớp, gồm có: negative (tiêu cực), neutral (trung tính), hay positive (tích cực), cũng như tăng độ chính xác cho việc gán nhãn của tập huấn luyện, chỉ giữ lại các bình luận có giá trị 0 được gán giá trị là neutral, các giá trị gần với -1 là negative (ví dụ: lớn hơn -0.95) và gần với 1 là positive (ví dụ: lớn hơn 0.95). Để thực hiện điều này, cần thu thập một số lượng rất lớn dữ liệu, để có đủ mẫu dữ liệu huấn luyện và kiểm tra sau khi sàng lọc, cân đối số mẫu dữ liệu giữa các lớp phân loại, góp phần tăng độ chính xác. Hơn nữa, việc kiểm tra lại và cập nhật nhãn thủ công để đảm bảo tất cả các bình luận được gán nhãn chính xác là cần thiết. Tuy nhiên, việc này tốn rất nhiều thời gian với tập dữ liệu lớn.
- **Bước 4:** Phân tách dữ liệu đã được gán nhãn thành hai tập dữ liệu. Trong đó, 80% dữ liệu được sử dụng cho tập dữ liệu huấn luyện, và 20% còn lại để kiểm tra mô hình.
- **Bước 5:** Sử dụng kỹ thuật TF-IDF, chuyển đổi, chọn lọc dữ liệu, trích xuất đặc trưng, vectơ hóa trước khi đưa vào huấn luyện mô hình. Ngoài ra, có thể dùng Word2Vec, và một số kỹ thuật khác. Bên cạnh đó, áp dụng cross validation, với numFolds là 5. Trong đó, 4/5 dữ liệu cho tập huấn luyện và 1/5 dữ liệu cho tập kiểm tra. Ngoài ra, sử dụng các tham số khác nhau dùng để tinh chỉnh mô hình, tùy theo mỗi thuật toán hỗ trợ, được sử dụng trong quá trình huấn luyện. Các kết quả huấn luyện mô hình có được, sẽ chọn ra mô hình tốt nhất tương ứng với mỗi thuật toán, được thực hiện với dữ liệu kiểm tra để đánh giá độ chính xác của mô hình. Trong đó, việc đánh giá mô hình dựa vào các tiêu chí gồm có: Accuracy, Precision, Recall và F1 Score.
- **Bước 6:** Trích xuất mô hình, để sử dụng cho hệ thống xử lý online sẽ được trình bày trong phần sau.



Hình 2. Hệ thống xử lý offline.

- **Hệ thống xử lý online:** Thành phần xử lý online, được thực hiện gồm các bước như sau:
 - **Bước 1:** Luồng dữ liệu được thu thập theo thời gian thực, sử dụng Twitter streaming API, được kết nối thông qua Tweepy dùng OAuthHandler, Stream và API, và dữ liệu được lưu vào Kafka.
 - **Bước 2:** Spark Streaming đọc dữ liệu từ Kafka để phân tích và xử lý. Trong đó, các công đoạn chuyển đổi dữ liệu, cập nhật lược đồ dữ liệu, trích lọc các cột thuộc tính liên quan, làm sạch dữ liệu, và tiền xử lý dữ liệu tương tự hệ thống xử lý offline.
 - **Bước 3:** Sử dụng các mô hình học máy được trích xuất để phân tích và dự đoán dữ liệu và lưu trữ kết quả phân tích trở lại Kafka.

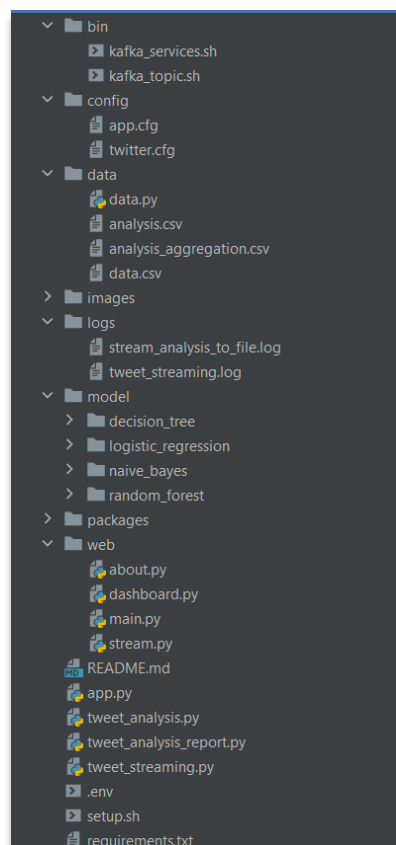


Hình 3. Hệ thống xử lý online.

- **Ứng dụng web:** Các dữ liệu kết quả phân tích từ Kafka, được đưa vào Spark Streaming, sau đó dùng Spark SQL thực hiện các truy vấn và ghi kết quả vào hệ thống tệp, có định dạng CSV. Từ đó, ứng dụng Streamlit đọc dữ liệu từ hệ thống tệp để hiển thị lên giao diện web, và dùng Plotly để tạo các biểu đồ tròn, cột và bảng, trực quan hóa các dữ liệu báo cáo và thống kê. Ngoài ra, từ giao diện web, có thể gọi dịch vụ xử lý để bắt đầu hoặc dùng thu phát phân tích luồng dữ liệu.

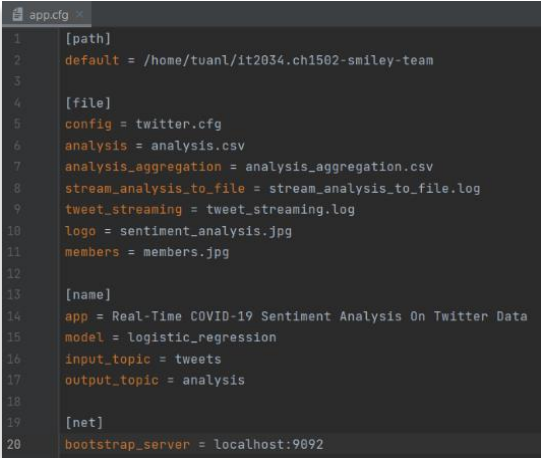
3.2 Triển Khai Hệ thống

Trong phần hiện thực và triển khai, hệ thống xử lý offline được thực hiện trên môi trường Google Colab, sử dụng Apache Spark 3.1.2, Java 11, Python 3.7 và một số thư viện khác. Các bước thực hiện đã được đề cập chi tiết trong phần 3.1. Mặt khác, hệ thống xử lý online và ứng dụng web, và các công cụ quản trị khác, được hiện thực trên máy chủ Ubuntu 18.04 LTS, Java 11, Python 3.8, Spark 3.1.2, Kafka 2.13-2.8.1, Streamlit 0.89.0, Plotly 5.3.1, v.v. và phần cứng CPU core I7, bộ nhớ 16 GB, ổ cứng SSD 512 GB. Hệ thống chương trình có cấu trúc như hình sau:



Hình 4. Cấu trúc chương trình.

- **Công cụ quản trị hệ thống:** Kiểm tra trạng thái, khởi tạo, dừng các dịch vụ Kafka Server và Zookeeper. Tạo, xóa, xem danh sách và nội dung các Kafka topic. Khởi tạo biến môi trường, cài đặt các dịch vụ và các thư viện cần cho hệ thống.
- **Dịch vụ xử lý:** Thực hiện khởi tạo kết nối đến Apache Spark và Kafka; chuyển đổi, trích lọc các thuộc tính liên quan, cập nhật lược đồ dữ liệu; làm sạch dữ liệu và tiền xử lý dữ liệu; phân tích dữ liệu dùng các mô hình học máy đã được trích xuất từ hệ thống xử lý offline, được lưu trong thư mục “model”; và trích xuất kết quả ra console, bộ nhớ, tệp, hoặc Kafka. Dịch vụ streaming để lấy dữ liệu theo thời gian thực từ Twitter và lưu vào Kafka. Các lịch sử luồng dữ liệu được ghi vào tệp và được lưu trong thư mục logs.
- **Dữ liệu:** Dữ liệu báo cáo, thống kê được lưu trong thư mục “data”. Bên cạnh đó, dữ liệu dùng cho việc huấn luyện “data.csv” được thu thập từ Twitter thông qua dịch vụ xử lý thu phát dữ liệu.
- **Cấu hình hệ thống:** Tệp “app.cfg” lưu thông tin các biến hoặc tham số được sử dụng của hệ thống, và “Twitter.cfg” lưu thông tin xác thực để truy cập Twitter. Đề đọc dữ liệu trên các tệp này, sử dụng công cụ hỗ trợ ConfigParser.



```

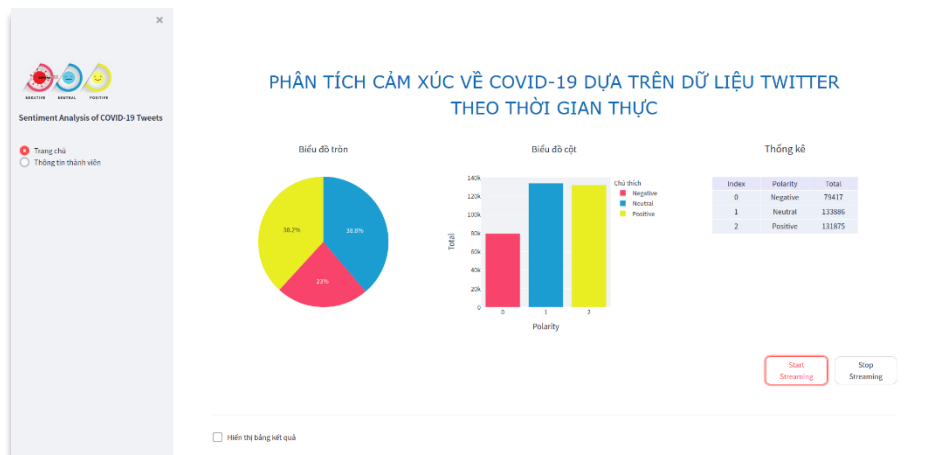
1  [path]
2  default = /home/tuanl/it2034.ch1502-smiley-team
3
4  [file]
5  config = twitter.cfg
6  analysis = analysis.csv
7  analysis_aggregation = analysis_aggregation.csv
8  stream_analysis_to_file = stream_analysis_to_file.log
9  tweet_streaming = tweet_streaming.log
10 logo = sentiment_analysis.jpg
11 members = members.jpg
12
13 [name]
14 app = Real-Time COVID-19 Sentiment Analysis On Twitter Data
15 model = logistic_regression
16 input_topic = tweets
17 output_topic = analysis
18
19 [net]
20 bootstrap_server = localhost:9092

```

Hình 5. Cấu trúc tệp cấu hình.

- **Dịch vụ web:** Nạp dữ liệu từ các tệp “analysis.csv” và “analysis_aggregation.csv”, được lưu trong thư mục “data”, sau đó hệ thống xử lý giao dùng Streamlit và Plotly để hiển thị báo cáo, thống kê lên giao diện web. Bên cạnh đó, các lệnh khởi tạo hoặc dừng xử lý luồng dữ liệu, được kết nối đến các dịch vụ xử lý nếu được gọi thông qua các nút “Start Streaming” hoặc “Stop Streaming” từ giao diện web.

Dưới đây là một số hình ảnh về hệ thống được triển khai. Xem chi tiết demo tại đường dẫn: IT2034.CH1502 - [Google Drive](#).



Hình 6. Giao diện ứng dụng web.

Hiện thị bảng kết quả

Trích lọc cột

	Id	Date	User	Location	Comment	Polarity
6	144459387933737317	2021-10-03 16:23:14	NeverDiever84	<NA>	rt fact according to the gov/health ...	Neutral
7	1444593879336927194	2021-10-03 16:23:13	etbent1	[jNtLd]	rt israel today of citizen are vaccin...	Neutral
8	144459386930839552	2021-10-03 16:23:12	QrenusJem	Southern Hemisphere	rt offering better to take vaccine is ...	Positive
9	1444593869305034625	2021-10-03 16:23:10	Theresa29979716	<NA>	rt israel today of citizen are vaccin...	Neutral
10	144459385842121306	2021-10-03 16:23:09	chabby77	Nottingham, England	rt child are not supposed to die w...	Negative
11	1444593835016553027	2021-10-03 16:23:04	gajjalugirip	HOSPIT	rt how anxious for how had thre...	Neutral
12	1444593826947254013	2021-10-03 16:23:02	ReasonableGuy33	<NA>	rt we tried to warn you with the m...	Negative
13	1444593825087996308	2021-10-03 16:23:01	viahisdevlop	eeeee	rt filtering asymptomatics out for ...	Neutral
14	1444593821904347342	2021-10-03 16:23:01	aToChild	<NA>	rt our call for peter dussak to be h...	Neutral

Chỉ thích

cột	Mô tả
Id	Định danh
Date	Thời gian
User	Tên người dùng
Location	Thông tin vị trí người dùng
Comment	Lời bình luận
Polarity	Kết quả phân tích: 0 - Negative, 1 - Neutral, 2 - Positive

Hình 7. Phần mở rộng.



Hình 8. Thông tin thành viên.

4 THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

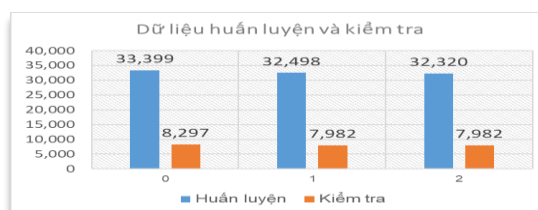
Trong phần này, tiến hành thực nghiệm, đánh giá và so sánh độ chính xác của các mô hình, sử dụng các thuật toán học máy được đề cập trong Phần 3.1, vào hệ thống được đề xuất.

4.1 Chuẩn bị dữ liệu

Dữ liệu thu thập từ Twitter gồm 156,599 mẫu. Dữ liệu này, sau đó được xử lý, sàng lọc, cân đối mẫu dữ liệu giữa các nhãn 0 – Negative, 1 – Neutral, và 2 – Positive nhằm tăng độ chính xác. Dữ liệu sau cùng dùng cho việc huấn luyện và kiểm tra các mô hình, được phân tách thành hai tập dữ liệu. Trong đó, 80% dữ liệu cho huấn luyện và 20% dữ liệu cho việc kiểm tra, như trong bảng sau:

Bảng 1. Dữ liệu huấn luyện và kiểm tra.

Phân loại	Huấn luyện	Kiểm tra
0	33,399	8,297
1	32,498	7,982
2	32,320	7,982



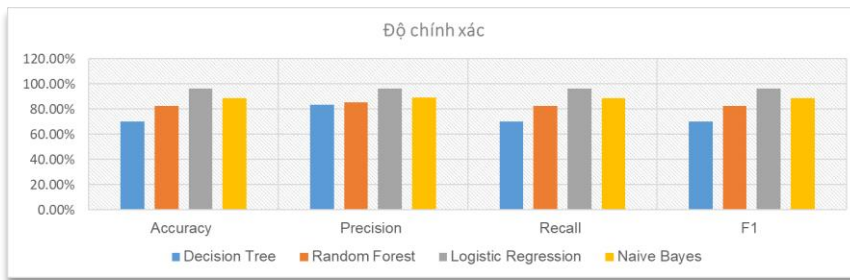
Hình 9. Dữ liệu huấn luyện và kiểm tra.

4.2 Đánh giá kết quả

Phần này trình bày việc đánh giá kết quả và so sánh độ chính xác của các mô hình dùng các thuật toán Decision Tree, Random Forest, Logistic Regression và Naive Bayes. Trong quá trình thực nghiệm, sử dụng kỹ thuật cross validation với numFolds có giá trị là 5, và các tham số tinh chỉnh khác tùy thuộc vào từng thuật toán hỗ trợ. Kết quả thực nghiệm thu được ở bảng sau:

Bảng 2. Độ chính xác của các mô hình học máy.

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree	70.06	83.32	70.06	70.31
Random Forest	82.49	85.37	82.49	82.65
Logistic Regression	96.29	96.34	96.29	96.30
Naive Bayes	88.88	89.27	88.88	88.80



Hình 10. Độ chính xác của các mô hình học máy.

Kết quả ở Bảng 2 cho thấy, mô hình sử dụng thuật toán Logistic Regression cho kết quả tốt nhất, với độ chính xác là 96.29% ứng với numFeatures có giá trị là 262,144. Trong khi đó, do hạn chế về tài nguyên phần cứng, mô hình dùng thuật toán Decision Tree và Random Forest có độ chính xác thấp hơn, được thực hiện với numFeatures có giá trị là 65,536.

5 KẾT LUẬN

Trong bài báo cáo này, nhóm đã tìm hiểu và xây dựng ứng dụng phân tích cảm xúc của cộng đồng về đại dịch COVID-19 dựa trên dữ liệu Twitter theo thời gian thực. Nội dung trình bày gồm có: Giới thiệu các nghiên cứu liên quan; Trình bày thiết kế và kiến trúc hệ thống, triển khai hệ thống; Cuối cùng, đánh giá kết quả thực nghiệm và so sánh độ chính xác giữa các mô hình học máy áp dụng vào hệ thống đề xuất. Hướng phát triển tương lai, tích hợp các nguồn dữ liệu từ các mạng xã hội khác, hỗ trợ xử lý tiếng Việt, tích hợp các kỹ thuật xử lý, nhằm tăng độ chính xác.

TÀI LIỆU THAM KHẢO

- [1] N. Chintalapudi, G. Battineni, and F. Amenta, “Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models,” *Infect. Dis. Rep.*, vol. 13, no. 2, Art. no. 2, Jun. 2021, doi: 10.3390/idr13020032.
- [2] X. Zhang, H. Saleh, E. M. G. Younis, R. Sahal, and A. A. Ali, “Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System,” *Complexity*, vol. 2020, p. e6688912, Dec. 2020, doi: 10.1155/2020/6688912.
- [3] H. Elzayady, K. M. Badran, and G. I. Salama, “Sentiment Analysis on Twitter Data using Apache Spark Framework,” in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, Dec. 2018, pp. 171–176. doi: 10.1109/ICCES.2018.8639195.