Ứng dụng Spark Streaming và Deep Learning vào phân tích cảm xúc dữ liệu các trang mạng xã hội về vắc xin COVID-19 theo thời gian thực

Trịnh Ngọc Pháp[1,2,*], Trần Nguyễn Anh Khoa[1,2,*], Nguyễn Thị Thanh Kim[1,2,*], Hà Văn Luân[1,2,*], Võ Linh Bảo[1,2,*] và Đỗ Trọng Hợp[1,2,†]

[1]Trường đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh
[2]Đại học Quốc gia Việt Nam, Thành phố Hồ Chí Minh
Email:*{18521227, 18520938, 18520963, 18521062, 18520503}@gm.uit.edu.vn
†{hopdt}@uit.edu.vn

Tóm tắt nội dung Trong nghiên cứu này, chúng tôi xây dựng một hệ thống Spark Streaming nhằm phân loại cảm xúc theo thời gian thực với dữ liệu trên Twitter và Reddit, cụ thể là các tweet và bình luận liên quan đến vắc xin COVID-19, sau đó thống kê và phân tích các kết quả đạt được. Các công việc được thực hiện bao gồm việc tìm ra mô hình tối ưu với hiệu suất tốt nhất cho nhiệm vụ phân tích cảm xúc của cộng đồng mạng về vắc xin COVID-19 và sau đó áp dụng mô hình này để phân tích dữ liệu theo thời gian thực. Nhìn chung, hệ thống này bao gồm hai thành phần chính là thành phần ngoại tuyến và thành phần trực tuyến. Thành phần ngoại tuyến có vai trò phát triển các mô hình phân tích cảm xúc và thành phần trực tuyến là một pipeline dự đoán, thống kê dữ liệu theo thời gian thực được xây dựng trên Spark. Đối với thành phần ngoại tuyến, chúng tôi sử dụng một bộ dữ liệu tự thu thập từ Twitter bao gồm các tweet về phản ứng của người dùng đối với các loại vắc xin COVID-19, sau đó huấn luyện các mô hình trên bộ dữ liệu này theo hai hướng tiếp cận là sử dụng các thư viện Spark MLlib (Machine Learning) và SparkNLP (Deep Learning). Chúng tôi kết hợp phương pháp trích xuất đặc trưng TF-IDF với hai mô hình Machine Learning truyền thống là Naive Bayes và Logistic Regression, so sánh với các thuật toán nhúng từ hiện đại LaBSE, GloVe, DistilBERT và Universal Sentence Encoder kết hợp với mô hình Deep Learning để chọn ra mô hình tốt nhất và áp dụng vào thành phần dự đoán trực tuyến. Kết quả thử nghiệm chỉ ra rằng kiến trúc ClassifierDL (mô hình DNNs) sử dụng phương pháp trích xuất đặc trưng LaBSE đã đạt được hiệu suất tốt nhất và được chúng tôi áp dụng cho thành phần trực tuyến. Đặc biệt, nghiên cứu của chúng tôi là nghiên cứu đầu tiên áp dụng Deep Learning vào đề tài phân tích cảm xúc dữ liệu về vắc xin COVID-19 theo thời gian thực trên công cụ xử lý dữ liệu lớn Apache Spark. Quy trình thu thập dữ liệu trực tuyến được phát triển bằng cách sử dụng Twitter API và Tweepy đối với Twitter, Reddit API và Praw đối với Reddit. Các phân tích trên kết quả phân loại cảm xúc của thành phần trực tuyến cho thấy Pfizer và Moderna là hai loại vắc xin được quan tâm và có tỉ lệ đánh giá tích cực cao hơn các vắc xin còn lại.

Keywords: Phân tích cảm xúc · Xử lí dữ liệu thời gian thực · Twitter · Reddit · vắc xin COVID-19 · Spark Streaming · Machine Learning · Deep Learning.

1 Giới thiệu

COVID-19 là một bệnh hô hấp cấp tính do một loại virus mới SARS-CoV-2 gây ra được phát hiện vào cuối năm 2019 [25]. Vào ngày 11 tháng 3 năm 2020, Tổ chức Y tế Thế giới (WHO) tuyên bố sự bùng phát COVID-19 là một đại dịch. Bắt đầu từ Trung Quốc - quốc gia đông dân nhất thế giới, COVID-19 đã lây lan và giết chết hàng triệu người từ nhiều quốc gia khác nhau, bao gồm Ấn Độ, Ý, Tây Ban Nha, Mỹ, Iran và các quốc gia khác trên toàn thế giới. Trong nỗ lực ngăn chặn và dập tắt đại dịch, hàng loạt loại vắc xin ngăn ngừa COVID-19 đã được các quốc gia nghiên cứu phát triển và xin cấp phép lưu hành khẩn cấp. Trong khi một số người lạc quan và tin tưởng vào các loại vắc xin, một số khác lại lo sợ về tính hiệu quả cũng như mức độ an toàn của các loại vắc xin này. Twitter và Reddit là hai trong số những trang mạng xã hội được sử dụng rộng rãi, qua đó mọi người có thể bày tỏ suy nghĩ, đánh giá của mình hay thảo luận về một vấn đề nào đó một cách đa dạng với những trạng thái hoặc bình luận ở dạng văn bản. Phân tích cảm xúc (sentiment analysis) là công nghệ được sử dụng để đo lường xúc cảm trong thông điệp truyền tải dựa vào những đặc điểm được lập trình sẵn dựa trên thang điểm mặc định trong hệ thống, có sự tác động của ngữ cảnh, không gian, thời gian. Phân tích cảm xúc là một trong những lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên [15,28].

Nhiều người đã sử dụng các mạng xã hội như Twitter và Reddit để bày tỏ ý kiến và thái độ của họ đối với COVID-19 cũng như các loại vắc xin đang được lưu hành. Do đó, tầm quan trọng của các dòng đăng tải trên mạng xã hội đã tăng lên hơn bao giờ hết. Trong hoàn cảnh đó, một số nhà nghiên cứu đã áp dụng phân tích cảm xúc để nghiên cứu ý kiến của mọi người về COVID-19. Ví dụ, Samuel và cộng sự (2020) [22] cũng như Ali và cộng sự (2020) [19] đã sử dụng các mô hình Machine Learning để phân loại các tweet về coronavirus. Mặt khác, quy mô dữ liệu trong các mạng xã hội chẳng hạn như Twitter đang tăng lên theo cấp số nhân [13] nên phân tích cảm xúc theo thời gian thực được coi là một trong những lĩnh vực nghiên cứu khó nhất, đòi hỏi các công cụ mạnh mẽ về phân tích dữ liệu lớn như Apache Spark [24]. Các nghiên cứu gần đây đã sử dụng Machine Learning với công nghệ dữ liệu lớn để thực nghiệm. Ví dụ, Rath và cộng sự (2017) [17] đã phát triển một hệ thống phân tích cảm xúc theo thời gian thực dựa trên quảng cáo theo đối tượng cụ thể. Das và cộng sự (2018) [4] cũng đã phát triển một hệ thống phân tích cảm xúc về giá cổ phiếu từ dữ liệu Twitter theo thời gian thực. Các nghiên cứu trước đây đã áp dụng các mô hình Machine Learning để nghiên cứu, phân tích thái độ và ý kiến về coronavirus bằng cách sử dụng dữ liệu thu thập được từ Twitter. Tuy nhiên, chưa có nghiên cứu nào sử dụng mô hình Deep Learning để áp dụng vào đề tài phân tích cảm xúc dữ liệu theo thời gian thực, điều này thúc đẩy chúng tôi giới thiệu một hệ thống mới với việc kết hợp mô hình Deep Learning và công cụ xử lý dữ liệu lớn Spark để phân tích cảm xúc theo thời gian thực dữ liệu Twitter, Reddit về vắc xin COVID-19. Kết quả dự đoán có thể sẽ có ích lợi mang tính tham khảo đối với cộng đồng cũng như các tổ chức chăm sóc sức khỏe, ngành y tế và các tổ chức xã hội.

Các đóng góp của bài báo có thể được tóm tắt như sau:

- Áp dụng phương pháp trích xuất đặc trưng TF-IDF, mô hình nhúng từ LaBSE, GloVe, DistilBERT và Universal Sentence Encoder.
- So sánh các mô hình Machine Learning và Deep Learning, tìm ra mô hình tối ưu để áp dụng dự đoán cảm xúc về vắc xin COVID-19 theo thời gian thực.
- Áp dụng Deep Learning vào hệ thống để dự đoán cảm xúc các tweet, bình luận Reddit về các loại vắc xin COVID-19 theo thời gian thực trên công cụ xử lý dữ liệu lớn Apache Spark.

Bài toán của chúng tôi được mô tả như sau:

- **Đầu vào**: Một câu tweet/bình luận Reddit tiếng Anh liên quan đến các loại vắc xin (vừa được đăng tải).
- **Đầu ra**: Nhãn cảm xúc của tweet/bình luận đó (tích cực hoặc tiêu cực hoặc trung tính).

Phần còn lại của bài viết này được tổ chức như sau. Các công trình liên quan được trình bày trong Phần 2. Hệ thống dự đoán cảm xúc theo thời gian thực cùng các phương pháp tiếp cận sẽ được giới thiệu trong Phần 3. Quá trình thực nghiệm và kết quả sẽ được đánh giá, phân tích trong Phần 4. Cuối cùng, kết luận và hướng phát triển sẽ được trình bày trong Phần 5.

2 Công trình liên quan

2.1 Các công trình phân tích cảm xúc

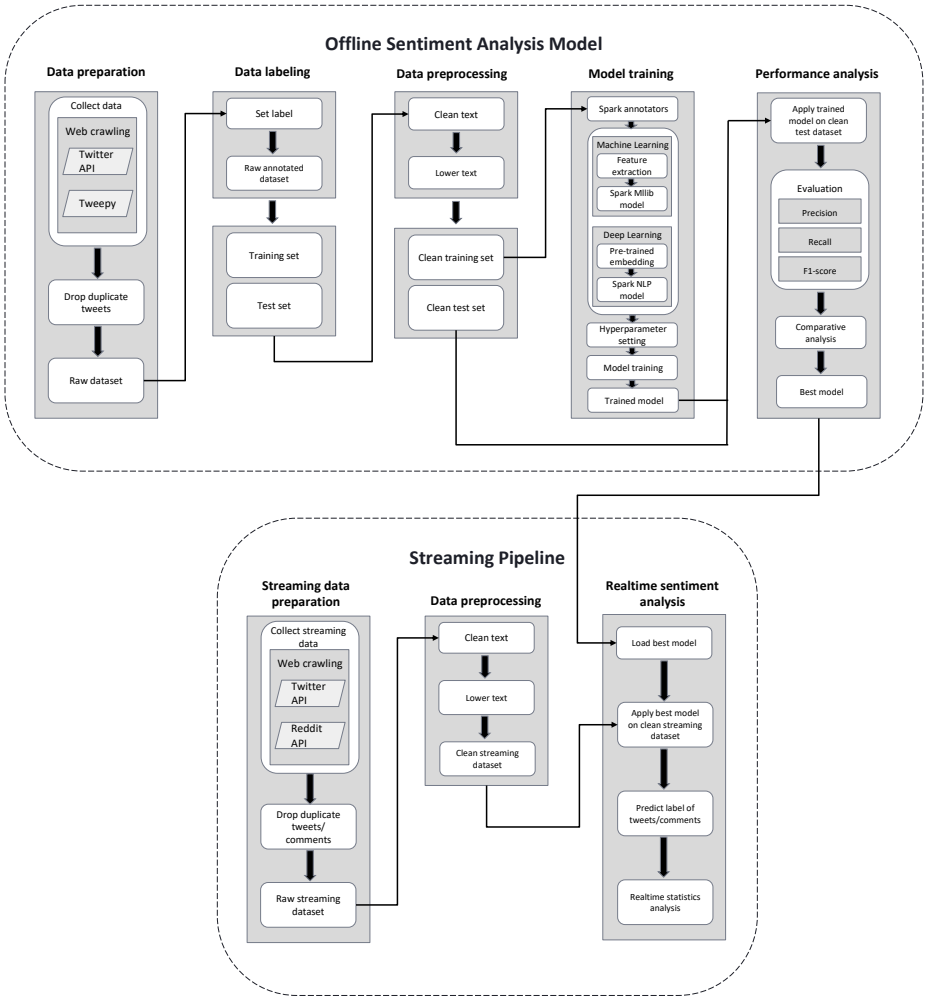
Goel và cộng sự (2016) [8] sử dụng SentiWordNet cùng với Naive Bayes để cải thiện độ chính xác của việc phân loại các tweet. Rajput và cộng sự (2020) [20] đã sử dụng kỹ thuật tần suất từ và phân tích cảm xúc các tweet về sự bùng phát COVID-19. Các tác giả đã sử dụng unigram, bigram và trigram để mô tả tỷ lệ của một từ, hai từ và ba từ. Về phương pháp phân tích cảm xúc, họ sử dụng TextBlob - một thư viện trong Python để phân loại các tweet thành tích cực (positive), tiêu cực (negative) và trung tính (neutral). Bhat và cộng sự (2020) [2] đã thu thập dữ liệu Twitter bằng cách lọc ra các tweet bao gồm hai hashtag #COVID-19 và #Coronavirus, sau đó áp dụng phân tích cảm xúc cho dữ liệu này để phân loại chúng thành tích cực, tiêu cực và trung tính. Ngoài ra, Dubey (2020) [5] đã thu thập các tweet về COVID-19 từ 11/03/2020 đến 31/03/2020, tác giả đã phân loại cảm xúc của mọi người về COVID-19 từ các quốc gia khác nhau trên thế giới bằng NRC Lexicon thành tám hình thái cảm xúc. Manguri và cộng sự (2020) [14] đã thu thập dữ liệu tweet về coronavirus trong 7 ngày từ 09/04/2020 đến 15/04/2020 và sử dụng TextBlob để phân loại các tweet thành tích cực, tiêu cực và trung tính, kết quả cho thấy cảm xúc trung tính chiếm tỷ lệ cao nhất. Gần đây, Na và cộng sự (2021) [16] đã thu thập các bài đăng trên tweet của cư dân Vương quốc Anh và Hoa Kỳ từ Twitter API trong thời gian xảy ra đại dịch và thiết kế các thí nghiệm để giải đáp các vấn đề liên quan đến việc tiêm chủng vắc xin.

2.2 Các hệ thống dự đoán theo thời gian thực

Một số nhà nghiên cứu đã áp dụng kỹ thuật Machine Learning và Big Data trên dữ liệu tweet để thực hiện các thí nghiệm theo thời gian thực. Ví dụ, Elzayady và cộng sự (2018) [6] thực hiện phân tích cảm xúc sử dụng Apache Spark và đề xuất một số mô hình Machine Learning và tiền xử lý để đạt được kết quả cao hơn. Ahmed và cộng sự (2020) [1] đã giới thiệu một hệ thống thời gian thực để dự đoán bệnh tim từ các dòng tweet trực tuyến dựa trên Apache Spark và Apache Kafka. Nhóm tác giả đã áp dụng các thuật toán Machine Learning thông thường bao gồm Decision Tree, SVM, Random Forest và Linear Regression, kết hợp kỹ thuật cross-validation cùng grid search để tối ưu hóa các mô hình và đạt được hiệu suất tốt nhất. Kết quả cho thấy mô hình Random Forest đạt hiệu suất tốt nhất và được sử dụng để dự đoán tình trạng bệnh tim từ các tweet của bệnh nhân trong thời gian thực. Zaki và cộng sự (2020) [27] cũng đã phát triển một framework để thu thập, xử lý, dự đoán và trực quan hóa dữ liệu Twitter. Các tác giả cũng tạo ra một mô hình phân tích thời gian thực để dự đoán tâm lý của người Iraq từ các tweet theo thời gian thực dựa trên Apache Spark. Ngoài ra, Kilinc (2019) [10] đã giới thiệu một hệ thống thời gian thực để phát hiện các tài khoản Twitter giả mạo dựa trên Apache Spark, tác giả đã sử dụng Spark MLlib để phát triển các mô hình Machine Learning khác nhau, bao gồm Decision Tree, SVM và Naive Bayes. Kabir và cộng sự (2020) [9] phát triển một ứng dụng trực tiếp để quan sát các tweet được đăng tải từ quốc gia Hoa Kỳ trong đại dịch COVID-19, các tác giả đã tạo ra nhiều phân tích dữ liệu khác nhau trong một khoảng thời gian để nghiên cứu những thay đổi về chủ đề, tính chủ quan và cảm xúc của con người. Ngoài ra, các tác giả cũng chia sẻ bộ dữ liệu Twitter CoronaVis (các tweet ở tại quốc gia Hoa Kỳ).

3 Hệ thống phân tích cảm xúc theo thời gian thực

Hệ thống phân tích cảm xúc theo thời gian thực được đề xuất bao gồm hai thành phần chính: một mô hình phân tích cảm xúc được huấn luyện trên dữ liệu có nhãn (ngoại tuyến) và một hệ thống phân tích cảm xúc theo thời gian thực (trực tuyến) như thể hiện trong Hình 1.



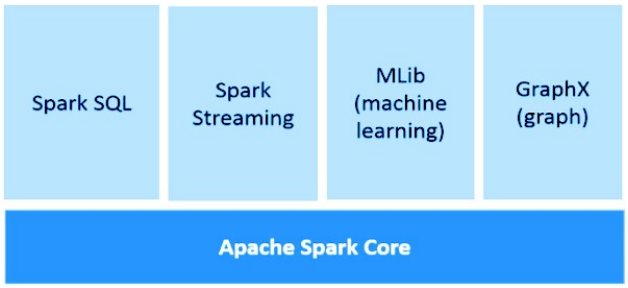
Hình 1: Kiến trúc của hệ thống phân tích cảm xúc theo thời gian thực.

3.1 Apache Spark

Vì các hệ thống xử lý dữ liệu truyền thống gặp các vấn đề về khả năng mở rộng dữ liệu và không xử lý được dữ liệu trực tuyến với khối lượng lớn, do đó hệ thống xử lý dữ liệu lớn ra đời để giải

quyết vấn đề này. Apache Spark [1] là một framework mã nguồn mở tính toán cụm, được phát triển sơ khởi vào năm 2009 bởi AMPLab. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay. Tốc độ xử lý nhanh của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM. Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming). Apache Spark bao gồm Spark Core và Bộ thư viện. Spark Core thực thi và quản lý công việc bằng cách cung cấp trải nghiệm liền mạch cho người dùng. Người dùng phải gửi công việc tới Spark Core và Spark Core đảm nhiệm việc xử lý, thực thi và trả lời lại cho người dùng qua API Spark Core bằng các ngôn ngữ lập trình khác nhau như Scala, Python, Java và R. Bộ thư viện Spark hỗ trợ các công cụ cấp cao bao gồm Spark SQL cho SQL và xử lý dữ liệu có cấu trúc, MLlib dành cho các mô hình Machine Learning, GraphX để xử lý đồ thị và Structured Streaming để tính toán gia tăng và xử lý luồng.

Trong đồ án này, chúng tôi sử dụng Spark SQL và Structured Streaming để xử lý dữ liệu Twitter theo thời gian thực, MLlib để xây dựng các mô hình Machine Learning trong Spark. Chúng tôi sử dụng Pyspark [2] là một giao diện cho Apache Spark trong ngôn ngữ Python để triển khai hệ thống phân tích cảm xúc theo thời gian thực.



Hình 2: Thành phần của Apache Spark.

3.2 Thành phần ngoại tuyến

Thành phần ngoại tuyến đã được phát triển để huấn luyện và kiểm tra các mô hình nhằm tìm ra mô hình tối ưu để áp dụng trong pipeline phân tích cảm xúc trực tuyến. Các mô hình học máy đã được sử dụng là Naive Bayes [26] và Logistic Regression [11] trong Spark MLlib kết hợp với TF-IDF [21]. Ngoài ra, chúng tôi cũng sử dụng một mô hình Deep Learning là ClassifierDL (DNNs) kết hợp với các mô hình nhúng từ như GloVe [18], DistilBERT [23], LaBSE [7] (Language-agnostic BERT Sentence Embedding) và USE [3] (Universal Sentence Encoder) do SparkNLP [12] cung cấp. Các mô hình này được huấn luyện và đánh giá bằng cách sử dụng bộ dữ liệu tweet đã được thu thập về các loại vắc xin COVID-19. Các phần sau chúng tôi sẽ trình bày về bộ dữ liệu huấn luyện, các bước tiền xử lý dữ liệu, các mô hình đề xuất và cách đánh giá các mô hình.

3.2.1 Bộ dữ liệu huấn luyện Bộ dữ liệu được thu thập từ Twitter bao gồm các tweet về vắc xin COVID-19. Trích chọn câu với các hashtag là tên của các loại vắc xin COVID-19 đang phổ

[1] Apache Spark: https://spark.apache.org/
[2] Pyspark: https://spark.apache.org/docs/latest/api/python/

biến hiện nay (Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/Astra-Zeneca, Covaxin, Sputnik V). Chúng tôi sử dụng thư viện Vader để gán nhãn, sau đó kiểm tra và hiệu chỉnh thủ công để việc gán nhãn đạt độ chính xác cao. Kết quả thu được 10,000 dòng dữ liệu (tỉ lệ neg-neu-pos: 0.17-0.48-0.35). Chúng tôi phân chia dữ liệu thành hai tập: tập huấn luyện và tập kiểm thử theo tỉ lệ 9-1 với tỉ lệ nhãn đồng đều ở mỗi tập.

3.2.2 Tiền xử lý dữ liệu Việc tiền xử lý dữ liệu là rất quan trọng trong bất kỳ hệ thống phân tích dữ liệu trên mạng xã hội nào vì nó ảnh hưởng trực tiếp đến độ phức tạp của dữ liệu. Mặc dù mạng xã hội được coi là mỏ vàng dữ liệu, đây lại là nguồn dữ liệu có dạng văn bản phức tạp nhất vì bao gồm nhiều liên kết, hashtag, ký hiệu đặc biệt, biểu tượng cảm xúc và nhiều loại khác. Do đó, dữ liệu Twitter, Reddit thu thập được đã được tiền xử lý bằng các bước sau:

Loại bỏ nhiễu: Trong giai đoạn này, việc loại bỏ nhiễu được thực hiện theo các bước:

- Xoá các ký tự @User (đề cập người dùng) trong đoạn text.
- Xóa ký tự # khỏi các Hashtag trong đoạn text.
- Xóa URL, Email. Trong bước này, chúng tôi đã loại bỏ các liên kết được đính kèm trong đoạn text.
- Loại bỏ ký hiệu RT (re-tweet) ở đầu đoạn text.
- Loại bỏ các ký hiệu đặc biệt, các số và loại bỏ các khoảng trắng thừa trong đoạn text.
- Chuyển biểu tượng cảm xúc (emoji) thành kí tự.

Đưa về chữ viết thường: Chuyển các chữ trong đoạn text bao gồm chữ in hoa và chữ viết thường về thành chữ viết thường. Ví dụ: "We love COVAXIN" được chuyển thành "we love covaxin".

Đưa về từ gốc: Bước này chuyển các từ về dạng ban đầu của chúng. Ví dụ: các từ “loved”, “loving” sẽ được rút gọn thành “love”.

Tách từ: Chúng tôi tách từ trong câu theo khoảng trắng.

3.2.3 Mô hình Machine Learning Đầu tiên, chúng tôi sử dụng hai mô hình Machine Learning trong Spark MLlib là Naive Bayes và Logistic Regression kết hợp với TF-IDF. Logistic Regression và Naive Bayes Classification đều là thuật toán phân loại được dùng để gán các đối tượng cho một tập hợp giá trị rời rạc. Trong đó, Logistic Regression dùng hàm sigmoid để đưa ra đánh giá theo xác suất, Naive Bayes Classification dự đoán dựa trên tính toán xác suất áp dụng định lý Bayes. TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp nổi tiếng được sử dụng để đánh giá mức độ quan trọng của một từ trong tài liệu được sử dụng để truy xuất thông tin và xử lý ngôn ngữ tự nhiên, trong đó thành phần TF thể hiện tần suất xuất hiện của một từ trong một văn bản, thành phần IDF thể hiện nghịch đảo tần suất của văn bản, dùng để đánh giá mức độ quan trọng của một từ trong văn bản. Mục tiêu của TF-IDF là tính toán tần suất từ trong văn bản trong một kho tài liệu khổng lồ.

3.2.4 Mô hình Deep Learning Mục tiêu chính của chúng tôi là áp dụng mô hình Deep Learning trên công cụ xử lý dữ liệu lớn Apache Spark, điều mà các công trình liên quan trước đây chưa thực hiện vì Spark chưa hỗ trợ thư viện cho các mô hình Deep Learning. Do đó, chúng tôi sử dụng SparkNLP - một thư viện mã nguồn mở cho xử lý ngôn ngữ tự nhiên trên Python, Java và Scala, được xây dựng dựa trên Apache Spark và thư viện Spark ML (MLlib), hỗ trợ phát triển mô hình Deep Learning trực tiếp trên Spark. Cụ thể, chúng tôi sử dụng mô hình ClassifierDL, một kiến trúc dùng cho bài toán phân loại văn bản nhiều lớp. ClassifierDL sử dụng mô hình học sâu (Deep Neural Network - DNNs) được xây dựng bên trong TensorFlow và hỗ trợ lên đến 100 lớp kết hợp với các mô hình nhúng từ được cung cấp bởi SparkNLP như:

- LaBSE (Language-agnostic BERT Sentence Embedding) hỗ trợ đến 109 ngôn ngữ, LaBSE mã hóa văn bản thành các vector nhiều chiều. Mô hình này được đào tạo và tối ưu hóa để tạo ra các biểu diễn tương tự của 1 câu bất kỳ.
- USE (Universal Sentence Encoder) là mô hình được đào tạo và tối ưu hóa cho văn bản dài hơn một từ, chẳng hạn như câu, cụm từ hoặc đoạn văn ngắn. Mô hình được đào tạo dựa trên nhiều nguồn dữ liệu và nhiều nhiệm vụ khác nhau với mục đích đáp ứng một cách linh hoạt nhiều loại nhiệm vụ hiểu ngôn ngữ tự nhiên.
- DistilBERT được phát hành bởi huggingface.co là phiên bản chắt lọc của BERT. DistilBERT tận dụng sự chắt lọc kiến thức trong giai đoạn trước khi đào tạo và cho thấy rằng có thể giảm kích thước của một mô hình BERT xuống 40% trong khi vẫn giữ được 97% khả năng hiểu ngôn ngữ và nhanh hơn 60%.
- GloVe (Global Vector) là một mô hình biểu diễn từ phân tán, bằng cách ánh xạ các từ vào một không gian ngữ nghĩa mà khoảng cách giữa các từ có liên quan đến sự tương đồng về ngữ nghĩa.

3.3 Thành phần trực tuyến

Thành phần trực tuyến là một pipeline dự đoán cảm xúc các tweet, bình luận Reddit về vắc xin COVID-19 trong thời gian thực. Hai công việc chính trong phần này là thu thập dữ liệu trực tuyến trên Twitter, Reddit và phân tích cảm xúc dữ liệu này theo thời gian thực.

3.3.1 Thu thập dữ liệu trực tuyến từ Twitter và Reddit Trong bước này, chúng tôi đã sử dụng Twitter API và Reddit API để thu thập các tweet, bình luận bao gồm các hashtag về các loại vắc xin nổi tiếng nhất hiện nay (Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/Astra-Zeneca, Covaxin, Sputnik V). Twitter API và Reddit API được sử dụng để truy xuất dữ liệu về vắc xin COVID-19 được tạo ra trong thời gian thực. Để kết nối với API và truy xuất dữ liệu, chúng tôi đã sử dụng hai thư viện trong Python là Tweepy để thu thập dữ liệu Twitter và Praw để thu thập dữ liệu Reddit. Chúng tôi đã sử dụng kết nối HTTP liên tục và ủy quyền người dùng được hỗ trợ bởi giao thức OAuth.

3.3.2 Phân tích cảm xúc theo thời gian thực Spark Streaming xử lý các dữ liệu được thu thập theo thời gian thực, chuyển đổi chúng thành các vector để làm đầu vào cho mô hình tốt nhất đã được huấn luyện ở phần ngoại tuyến, sau đó mô hình này sẽ dự đoán cảm xúc của người dùng đăng tải các tweet hay bình luận tương ứng. Về cơ bản, các bước xử lý và dự đoán theo thời gian thực được giải thích như sau:

- Đối với bước xử lý, Spark Streaming truy xuất các tweet, bình luận Reddit và thực hiện các bước tiền xử lý cần thiết. Sau đó, trích xuất đặc trưng của các dữ liệu để gửi chúng theo cấu trúc vector đến mô hình đã được huấn luyện.
- Đối với bước dự đoán, Spark sử dụng mô hình dự đoán tốt nhất thu được trong giai đoạn ngoại tuyến để phân loại sắc thái của mỗi văn bản về vắc xin COVID-19 thành ba nhãn: tích cực, tiêu cực, trung tính và xuất ra kết quả dự đoán cũng như các phân tích thống kê theo thời gian thực.

4 Kết quả thực nghiệm và đánh giá

4.1 Thành phần ngoại tuyến

4.1.1 Độ đo đánh giá Chúng tôi đã sử dụng các độ đo Precision, Recall và F1-score để đánh giá hiệu suất phân loại của các mô hình, trong đó TP là true positive, TN là true negative, FP là

false positive và FN là false negative được đưa ra trong các biểu thức dưới đây:

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

$$F1-score = \frac{2*precision*recall}{precision+recall}$$

4.1.2 Thông số mô hình Bằng cách tiếp cận Spark MLlib, chúng tôi triển khai các mô hình Machine Learning bao gồm Naive Bayes với tham số smoothing = 111 và Logistic Regression với các tham số maxIter = 10, regParam = 0.01. Bên cạnh đó, thư viện SparkNLP được sử dụng để phát triển các mô hình Deep Learning là ClassifierDL kết hợp với các mô hình nhúng từ LaBSE, GloVe, DistilBERT và Universal Sentence Encoder. Các thông số của từng mô hình Deep Learning được liệt kê ở Bảng 1 bên dưới. Các thử nghiệm đã được thực hiện bằng cách sử dụng thư viện Pyspark 3.1.2 và SparkNLP 3.1.3 với Python 3.7 trên Google Colab.

Bảng 1: Chi tiết thông số các mô hình Deep Learning.

Model	Batch Size	Embed Size	Num Epochs	Learning Rate
USE + ClassifierDL	64	512	10	5e-03
DistilBERT + ClassifierDL	64	768	10	5e-03
GloVe + ClassifierDL	32	300	10	3e-03
LaBSE + ClassifierDL	64	768	1	5e-03

4.1.3 Kết quả phân loại của các mô hình Bảng 2 thể hiện kết quả phân loại của các mô hình ngoại tuyến, trong đó LaBSE kết hợp với ClassifierDL cho kết quả tốt nhất với 0.78 F1-score, mô hình Naive Bayes kết hợp với TF-IDF cho hiệu suất kém nhất với 0.48 F1-score. Có thể thấy sự chênh lệch rõ rệt về hiệu suất phân loại của mô hình Machine Learning và Deep Learning, mô hình tốt nhất theo kiến trúc Deep Learning cao hơn mô hình tốt nhất của Machine Learning 7% ở F1-score. Kết quả này tương ứng với các nghiên cứu trong những năm gần đây đã chỉ ra các mô hình Deep Learning và đặc biệt là các mô hình sử dụng kiến trúc BERT thường cho kết quả tốt hơn so với các mô hình Machine Learning trong hầu hết nhiệm vụ NLP.

Bảng 2: Kết quả huấn luyện các mô hình.

Model	Precision	Recall	F1-score
TF-IDF + Naive Bayes	0.78	0.50	0.48
TF-IDF + Logistic Regression	0.73	0.70	0.71
USE + ClassifierDL	0.73	0.72	0.72
DistilBERT + ClassifierDL	0.75	0.71	0.73
GloVe + ClassifierDL	0.76	0.73	0.74
LaBSE + ClassifierDL	0.79	0.77	0.78

Bảng 3 mô tả kết quả chi tiết của mô hình ngoại tuyến tốt nhất. Từ kết quả chi tiết có thể thấy hiệu suất của mô hình ổn định trên cả ba nhãn, trong đó kết quả cao nhất trên nhãn Neutral (trung tính) với 0.79 F1-score, kế tiếp là Negative (tiêu cực) với 0.78 F1-score và cuối cùng là Positive (tích cực) với 0.76 F1-score. Có sự chênh lệch đáng kể về phân bố giữa các nhãn trong tập dữ liệu nhưng hiệu suất của mô hình giữa các nhãn chênh lệch không nhiều.

Bảng 3: Chi tiết kết quả huấn luyện mô hình LaBSE + ClassifierDL.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.82	0.75	0.78	170
Neutral	0.77	0.81	0.79	480
Positive	0.77	0.75	0.76	350
Accuracy			0.78	1000
Macro avg	0.79	0.77	0.78	1000
Weighted avg	0.78	0.78	0.78	1000

4.2 Thành phần trực tuyến

4.2.1 Tốc độ thu thập dữ liệu trực tuyến Trong quá trình phân tích dữ liệu theo thời gian thực, tốc độ thu thập dữ liệu là một vấn đề được quan tâm hàng đầu. Do đó, chúng tôi tiến hành thực nghiệm đo thời gian mà hệ thống thu thập được các dữ liệu trên hai trang mạng xã hội Twitter và Reddit. Cụ thể, thời gian khi thu thập 100 câu tweet và bình luận Reddit ở cùng thời điểm bắt đầu quá trình thu thập là 4 phút 56 giây đối với dữ liệu Twitter và 33 phút 22 giây đối với dữ liệu Reddit. Tốc độ này phụ thuộc vào số lượng người dùng đăng tải các câu trạng thái, bình luận tại thời điểm thu thập. Tuy nhiên, có thể thấy tốc độ thu thập dữ liệu Twitter nhanh hơn nhiều lần so với Reddit do cộng đồng sử dụng đông hơn cũng như người dùng thường đề cập tới vấn đề vắc xin COVID-19 trên Twitter nhiều hơn.

4.2.2 Độ trễ của mô hình phân tích cảm xúc theo thời gian thực Bên cạnh tốc độ thu thập dữ liệu, độ trễ cũng là một vấn đề rất được quan tâm ở các hệ thống xử lý dữ liệu theo thời gian thực. Ở đồ án này, chúng tôi tiến hành kiểm tra độ trễ của hệ thống phân tích cảm xúc theo thời gian thực mà chúng tôi đề xuất bằng cách tính thời gian từ khi các câu tweet, bình luận được người dùng đăng tải đến khi mô hình dự đoán nhãn cảm xúc cho các câu đó. Cụ thể, khi thu thập các tweet, bình luận Reddit, ngoài thuộc tính "text" chứa đoạn văn bản được người dùng đăng tải hay "user" chứa thông tin tên của người dùng, chúng tôi còn trích xuất thuộc tính "created_at" để lưu lại thời gian các câu này được đăng tải lên mạng xã hội. Khi đưa các câu tweet, bình luận theo thời gian thực này qua hệ thống dự đoán cảm xúc, bên cạnh việc xuất ra nhãn cảm xúc của các câu này, chúng tôi thêm vào dataframe kết quả một thuộc tính "predicted_at" ghi lại thời gian mô hình dự đoán xong nhãn các câu tương ứng. Cả hai thuộc tính "created_at" và "predicted_at" đều được chuyển thành định dạng thời gian Unix [3]. Qua đó, chúng tôi đưa ra một độ đo Latency để đánh giá độ trễ của mô hình phân tích theo thời gian thực được tính theo giây với công thức:

$$latency = t_{predicted\_at} - t_{created\_at}$$

[3] Unix timestamp: https://www.unixtimestamp.com/

Kết quả Latency trung bình của hệ thống khi dự đoán khoảng 100 câu tweet, bình luận theo thời gian thực được chúng tôi thực nghiệm là: Latency = 53.49 giây/câu đối với dữ liệu Twitter và Latency = 76.14 giây/câu đối với dữ liệu Reddit. Kết quả này chưa thực sự nhanh do điều kiện cấu hình hệ thống, bộ nhớ, dung lượng máy tính, tuy nhiên kết quả này đủ đáp ứng được nhu cầu phân tích cảm xúc theo thời gian thực ở mức độ cơ bản.

4.2.3 Kết quả dự đoán cảm xúc trực tuyến Chúng tôi thực hiện một số thống kê trên dữ liệu streaming từ Twitter và Reddit để đánh giá hệ thống. Bảng 4 thể hiện tỉ lệ nhãn cảm xúc mà hệ thống dự đoán trên khoảng 800 điểm dữ liệu, có thể thấy được số lượng các nhãn có cảm xúc trung tính chiếm tỉ lệ lớn nhất như được mong đợi.

Bảng 4: Thống kê số lượng các nhãn trên mỗi nguồn thu thập.

	Neutral	**Positive**	**Negative**	**Total**
Twitter	471	247	91	809
Reddit	570	182	46	798

Bảng 5 mô tả số lượng tweet và số lượng bình luận Reddit cùng với tỉ lệ các nhãn trên mỗi loại vắc xin. Trong số bảy loại vắc xin được sử dụng để thu thập dữ liệu, số lượng tweet và bình luận liên quan đến vắc xin Pfizer và Moderna chiếm nhiều nhất ở cả hai trang mạng xã hội Twitter và Reddit, điều này cho thấy hai loại vắc xin đến từ Hoa Kỳ đang được người dùng quan tâm nhiều nhất trên các mạng xã hội này, đặc biệt là Pfizer. Ngược lại, Sinovac và Sinopharm là hai loại vắc xin ít được nhắc tới nhất. Các số liệu cũng chỉ ra tỉ lệ cảm xúc tích cực ở các vắc xin Pfizer và Moderna cao hơn hầu hết các loại vắc xin còn lại trên cả hai nguồn, điều đó cũng minh chứng cho việc hai loại vắc xin của Hoa Kỳ này rất tốt, được đông đảo các quốc gia trên thế giới công nhận ở thời điểm hiện tại.

Bảng 5: Thống kê số lượng tweet và bình luận Reddit và tỉ lệ các nhãn trên mỗi vắc xin.

Vaccine	**Twitter**				**Reddit**			
	Neutral	Positive	Negative	Total	Neutral	Positive	Negative	Total
Pfizer	0.5649	0.3181	0.1168	462	0.7305	0.2079	0.0614	553
Moderna	0.5909	0.3465	0.0625	176	0.6854	0.3004	0.0140	213
Covaxin	0.9130	0.0724	0.0144	69	0.8194	0.0972	0.0833	72
Sputnik V	0.2656	0.7031	0.0312	64	0.7727	0.1818	0.0454	44
AstraZeneca	0.8113	0.1320	0.0566	53	0.8205	0.1282	0.0512	39
Sinovac	0.8421	0.1578	-	19	0.9090	0.0909	-	11
Sinopharm	0.8333	0.1666	-	12	0.9	0.1	-	10

5 Kết luận và hướng phát triển

Đồ án này đã trình bày một hệ thống dự đoán cảm xúc theo thời gian thực trên dữ liệu trực tuyến của Twitter và Reddit về vắc xin COVID-19. Hệ thống đề xuất được phát triển bằng cách sử dụng

Twitter API, Reddit API, Pyspark, Spark MLlib và SparkNLP. Hệ thống này bao gồm hai thành phần là một mô hình phân tích cảm xúc ngoại tuyến và một pipeline dự đoán trực tuyến. Ở thành phần ngoại tuyến, chúng tôi huấn luyện, đánh giá mô hình trên bộ dữ liệu bao gồm các tweet liên quan đến vắc xin COVID-19, sau đó tìm ra mô hình tốt nhất, mô hình này sẽ được sử dụng để dự đoán cảm xúc cho các tweet, bình luận được thu thập theo thời gian thực. Chúng tôi đã cài đặt thành công sáu mô hình, trong đó mô hình Machine Learning là Naive Bayes và Logistic Regression lần lượt kết hợp với phương pháp nhúng từ TF-IDF, mô hình Deep Learning là ClassiferDL lần lượt kết hợp với bộ nhúng từ GloVe, DistilBERT, LaBSE và USE. Kết quả thực nghiệm đã chứng minh rằng mô hình Deep Learning với kiến trúc mô hình hiện đại, phức tạp hơn đã cho kết quả tốt hơn, cụ thể là mô hình ClassiferDL kết hợp với LaBSE đã đạt được hiệu suất tốt nhất so với các mô hình còn lại với 0.78 F1-score. Thành phần còn lại là một pipeline dự đoán trực tuyến được sử dụng để dự đoán cảm xúc các tweet, bình luận về chủ đề vắc xin trong thời gian thực. Cụ thể, pipeline này sử dụng Twitter API và Reddit API để thu thập các tweet, bình luận về vắc xin COVID-19 theo thời gian thực, sau đó đưa dữ liệu này vào Spark để xử lý và áp dụng mô hình tốt nhất đã được huấn luyện trước là LaBSE kết hợp ClassifierDL để dự đoán phân loại cảm xúc các tweet, bình luận này. Qua đó, chúng tôi đã xây dựng thành công một hệ thống xử lý dữ liệu thời gian thực và áp dụng mô hình Deep Learning trên công cụ xử lý dữ liệu lớn Apache Spark, điều mà các công trình nghiên cứu trước đây chưa thực hiện.

Trong tương lai, chúng tôi mong muốn cải thiện các mô hình để đạt hiệu suất phân loại tốt hơn. Các phương pháp đề ra có thể là: nghiên cứu cách tiền xử lý dữ liệu tốt hơn, tinh chỉnh thông số mô hình, thử nghiệm các bộ nhúng từ khác, áp dụng kiến trúc mô hình Deep Learning khác. Bên cạnh đó, chúng tôi cũng lên kế hoạch mở rộng phạm vi nghiên cứu, thực hiện công việc này trên ngôn ngữ tiếng Việt với nhiều thách thức, hứa hẹn và mở rộng nguồn dữ liệu trực tuyến vì ngoài Twitter và Reddit, còn có nhiều mạng xã hội hoặc nguồn dữ liệu trực tuyến cũng đang rất phổ biến. Ngoài ra, trong quá trình huấn luyện mô hình ngoại tuyến và thử nghiệm hệ thống trực tuyến, chúng tôi cũng gặp các vấn đề về bộ nhớ, dung lượng và cấu hình hệ thống. Do đó chúng tôi cũng hy vọng có thể giải quyết các vấn đề này để tăng hiệu suất cũng như tốc độ của hệ thống mà chúng tôi đã xây dựng.

Tài liệu

1. Ahmed, H., Younis, E.M., Hendawi, A., Ali, A.A.: Heart disease identification from patients' social posts, machine learning solution on spark. Future Generation Computer Systems **111**, 714–722 (2020)
2. Bhat, M., Qadri, M., Noor-ul Asrar Beg, M.K., Ahanger, N., Agarwal, B.: Sentiment analysis of social media response on the covid19 outbreak. Brain, Behavior, and Immunity **87**, 136 (2020)
3. Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R.: Universal sentence encoder (2018)
4. Das, S., Behera, R.K., Rath, S.K., et al.: Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia computer science **132**, 956–964 (2018)
5. Dubey, A.D.: Twitter sentiment analysis during covid-19 outbreak. Available at SSRN 3572023 (2020)
6. Elzayady, H., Badran, K.M.S., Salama, G.: Sentiment analysis on twitter data using apache spark framework. 2018 13th International Conference on Computer Engineering and Systems (ICCES) pp. 171–176 (2018)
7. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding (2020)

8. Goel, A., Gautam, J., Kumar, S.: Real time sentiment analysis of tweets using naive bayes. 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) pp. 257–261 (2016)
9. Kabir, M.Y., Madria, S.: Coronavis: A real-time covid-19 tweets data analyzer and data repository (2020)
10. Kılınç, D.: A spark-based big data analysis framework for real-time sentiment prediction on streaming data. Software: Practice and Experience **49**(9), 1352–1364 (2019)
11. Kleinbaum, D.G., Klein, M.: Logistic regression: a self-learning text. Springer (2010)
12. Kocaman, V., Talby, D.: Spark nlp: Natural language understanding at scale (2021)
13. Koval, M.J., Lawton, W.W., Tyler, J.G., Winters, S.L.: Data stream protocol for multimedia data streaming data processing system (Aug 16 1994), uS Patent 5,339,413
14. Manguri, K.H., Ramadhan, R.N., Amin, P.R.M.: Twitter sentiment analysis on worldwide covid-19 outbreaks. Kurdistan Journal of Applied Research pp. 54–65 (2020)
15. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal **5**(4), 1093–1113 (2014). https://doi.org/https://doi.org/10.1016/j.asej.2014.04.011, https://www.sciencedirect.com/science/article/pii/S2090447914000550
16. Na, T., Cheng, W., Li, D., Lu, W., Li, H.: Insight from nlp analysis: Covid-19 vaccines sentiments on social media (2021)
17. Nair, L.R., Shetty, S.D., Shetty, S.D.: Streaming big data analysis for real-time sentiment based targeted advertising. International Journal of Electrical and Computer Engineering **7**(1), 402 (2017)
18. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
19. Ra, M., Ab, B., Kc, S.: Covid-19 outbreak: Tweet based analysis and visualization towards the influence of coronavirus in the world (2020)
20. Rajput, N.K., Grover, B.A., Rathi, V.K.: Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv preprint arXiv:2004.03925 (2020)
21. Robertson, S.: Understanding inverse document frequency: On theoretical arguments for idf. Journal of Documentation - J DOC **60**, 503–520 (10 2004). https://doi.org/10.1108/00220410410560582
22. Samuel, J., Ali, G.G.M.N., Rahman, M.M., Esawi, E., Samuel, Y.: Covid-19 public sentiment insights and machine learning for tweets classification. Information **11**(6) (2020). https://doi.org/10.3390/info11060314, https://www.mdpi.com/2078-2489/11/6/314
23. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
24. Spark, A.: Apache spark. Retrieved January **17**, 2018 (2018)
25. Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., Zhang, S., Tsamlag, L., Shang, M., Huang, J., Wang, Y., Xu, G., Shen, T., Zhang, X., Cai, Y.: Phase-adjusted estimation of the number of coronavirus disease 2019 cases in wuhan, china. Cell discovery **6**(1), 10–10 (2020)
26. Watson, T.J.: An empirical study of the naive bayes classifier (2001)
27. Zaki, N.D., Hashim, N.Y., Mohialden, Y.M., Mohammed, M.A., Sutikno, T., Ali, A.H.: A real-time big data sentiment analysis for iraqi tweets using spark streaming. Bulletin of Electrical Engineering and Informatics **9**(4), 1411–1419 (2020)
28. Zhang, L., Liu, B.: Sentiment Analysis and Opinion Mining, pp. 1152–1161. Springer US, Boston, MA (2017). https://doi.org/10.1007/978-1-4899-7687-1$_9$07, https://doi.org/10.1007/978-1-4899-7687-1_907