COMPARISON OF MACHINE LEARNING ALGORITHMS FOR SENTIMENT ANALYSIS OF VIETNAMESE YOUTUBE SUBTITLES

Nguyen Trong Tu*, Nguyen Trung Tin
Le Quy Don Technical University

ARTICLE INFO

Received: 21/2/2024
Revised: 23/5/2024
Published: 24/5/2024

KEYWORDS

Machine learning
YouTube subtitles
Sentiment analysis
Subtitle classification
Algorithm comparison

ABSTRACT

Currently, YouTube has become one of the most significant online platforms, with billions of hours of video uploaded every day, attracting a vast user base. Recently, foreign reactionary forces and extremist organizations have exploited YouTube to disseminate videos undermining the Party, the State, and the Vietnamese military. This study focuses on analyzing Vietnamese subtitles collected from YouTube. By using machine learning algorithms, it conducts sentiment analysis and categorizes the subtitles of videos. This research provides a profound insight into the emotions and perspectives of the online community regarding content on YouTube, particularly those related to politics and society. The results of the study among four machine learning algorithms include Naive Bayes, Random Forest, Support Vector Machine, and Logistic Regression. Among them, the Random Forest algorithm has achieved the highest accuracy rate of 81%, surpassing the other three algorithms in analyzing the sentiments of subtitles from YouTube videos with negative content.

SO SÁNH CÁC THUẬT TOÁN HỌC MÁY CHO PHÂN TÍCH TÌNH CẢM PHỤ ĐỀ YOUTUBE TIẾNG VIỆT

Nguyễn Trọng Tú*, Nguyễn Trung Tín
Trường Đại học Kỹ thuật Lê Quý Đôn

THÔNG TIN BÀI BÁO

Ngày nhận bài: 21/2/2024
Ngày hoàn thiện: 23/5/2024
Ngày đăng: 24/5/2024

TỪ KHÓA

Học máy
Phụ đề YouTube
Phân tích cảm xúc
Phân loại phụ đề
So sánh thuật toán

TÓM TẮT

Hiện nay, YouTube đã trở thành một trong những nền tảng trực tuyến quan trọng, với hàng tỷ giờ video được tải lên mỗi ngày, thu hút đông đảo người dùng. Gần đây, các lực lượng phản động và các tổ chức cực đoan từ nước ngoài đã tận dụng YouTube để lan truyền video chống phá Đảng, Nhà nước và Quân đội Việt Nam. Nghiên cứu này tập trung vào phân tích các phụ đề Tiếng Việt được thu thập từ YouTube. Bằng cách sử dụng các thuật toán học máy, thực hiện phân tích cảm xúc và phân loại phụ đề của các video. Nghiên cứu này mang lại cái nhìn sâu sắc về tâm trạng và quan điểm của cộng đồng mạng đối với nội dung trên YouTube, đặc biệt là những nội dung liên quan đến chính trị và xã hội. Kết quả của nghiên cứu giữa bốn thuật toán học máy, thuật toán Random Forest đã đạt tỷ lệ chính xác cao nhất là 81%, vượt trội so với ba thuật toán khác trong phân tích cảm xúc của các phụ đề từ video YouTube có nội dung tiêu cực.

DOI: https://doi.org/10.34238/tnu-jst.9741

* Corresponding author. *Email: trongtu189@gmail.com*

1. Giới thiệu

Sự phát triển nhanh chóng của Internet và mạng xã hội, như Facebook, Twitter và YouTube, đã mang lại những lợi ích to lớn, xác nhận vai trò quan trọng trong đời sống xã hội, tạo ra một môi trường phong phú để cung cấp, chia sẻ, trao đổi và khai thác thông tin cho cộng đồng. Sự gia tăng về cấp độ, mật độ, tần suất và lưu lượng đăng video trên YouTube của những thế lực này đang diễn ra một cách đáng kể. Họ sử dụng những chiêu thức và thủ đoạn tinh vi, thâm hiểm để thu hút sự quan tâm và theo dõi từ cộng đồng mạng. Qua đó, họ thực hiện các biện pháp tuyên truyền, kích động, xuyên tạc thông tin, và chống phá một cách quyết liệt. Các thủ đoạn mới xuất hiện, như việc thực hiện Live stream trực tiếp để kêu gọi cộng đồng mạng can thiệp vào nội bộ hoặc tham gia bình luận trái chiều trên mạng xã hội. Họ cũng thường xuyên làm mới thông tin cũ, bịa đặt thông tin mới nhằm chống phá Quân đội và gây nhiễu loạn trong cộng đồng mạng. Những hành động này khiến một phần cư dân mạng mất phương hướng, lầm tưởng rằng đó là sự thật, dẫn đến hoài nghi và thiếu niềm tin vào Đảng và chế độ.

Phát hiện cảm xúc là một phương pháp nhằm xác định và phân loại các loại cảm xúc riêng biệt của con người, như sự tức giận, vui mừng hoặc chán nản. Cụm từ "phát hiện cảm xúc", "điện toán cảm xúc", "phân tích cảm xúc" và "nhận dạng cảm xúc" đôi khi được sử dụng thay thế cho nhau, như đã được mô tả trong nghiên cứu của Munezero và đồng nghiệp [1]. Phân tích tình cảm là quá trình đánh giá và tách rời thông tin về ý kiến, cảm xúc và tâm trạng liên quan đến một đối tượng, thường được diễn đạt dưới dạng văn bản. Phương pháp này nhằm trích xuất các đặc tính và thành phần quan trọng từ văn bản, từ đó xác định xem phụ đề đó được phân loại là tích cực hay tiêu cực. Phân tích tình cảm thường sử dụng một loạt các thuật toán học máy như Naive Bayes [2], Random Forest [3], Support Vector Machine (SVM), Logistic Regression [4], và nhiều thuật toán khác.

Trong thời đại số hóa ngày nay, nghiên cứu về phân tích tình cảm, đặc biệt là từ dữ liệu văn bản trên các nền tảng trực tuyến như YouTube, đã thu hút sự quan tâm lớn từ cộng đồng nghiên cứu. Cùng với sự phát triển của học máy và xử lý ngôn ngữ tự nhiên, nhiều phương pháp và ứng dụng đã được đề xuất để hiểu và phân loại tình cảm từ ý kiến của người dùng.

Trong loạt tài liệu nghiên cứu hiện đại, Cha et al. [5] đã thực hiện một nghiên cứu so sánh và kết hợp các phương pháp phân tích tình cảm, đánh giá sự hiệu quả của chúng. Medhat et al. [6] thực hiện một khảo sát toàn diện về các thuật toán và ứng dụng trong lĩnh vực phân tích tình cảm, đặt ra những xu hướng và thách thức hiện nay. Chong et al. [7] tập trung vào xử lý ngôn ngữ tự nhiên để phát triển phương pháp phân tích tình cảm tiên tiến.

Các nghiên cứu về ứng dụng của phân tích tình cảm trên YouTube cũng đã thu hút sự chú ý. Bhuiyan et al. [8] và Novendri et al. [9] đều thảo luận về việc áp dụng phân tích tình cảm cho các ý kiến người dùng trên YouTube, trong khi Tafesse [10] nghiên cứu về cách tối ưu hóa video trên YouTube ảnh hưởng đến lượt xem và hiệu suất tiếp thị. Ngoài ra, có những nghiên cứu như Das et al. [11] và Bakshi et al. [12] đã đưa ra cái nhìn sâu sắc về lĩnh vực tích hợp của tính toán cảm xúc và phân tích tình cảm.

Nghiên cứu của M. Cliche [13] có tựa đề "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", đã đạt được kết quả ấn tượng trong cuộc thi SemEval-2017 Task 4 về phân tích cảm xúc trên Twitter. Mô hình sử dụng kỹ thuật ensemble của LSTMs và CNNs với nhiều phép toán tích chập đã đứng đầu bảng xếp hạng với F1-score đạt 0,685. Nghiên cứu này tập trung vào việc kết hợp hai mô hình học sâu này để nâng cao hiệu suất phân loại cảm xúc trên Twitter.

Một nghiên cứu khác cũng đạt được kết quả tốt đó là "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis" [14]. Mô hình này sử dụng LSTM kép (Bi-LSTM) kết hợp với cơ chế attention để phân tích cảm xúc trên các mức độ khác nhau của thông điệp và dựa trên chủ đề. Với F1-score đạt 0,677, nghiên cứu này đã đóng góp vào việc nâng cao hiểu biết và kỹ năng trong lĩnh vực phân tích cảm xúc trên mạng xã hội như Twitter.

Dựa trên các nghiên cứu liên quan, đã xác định sáu phương pháp hiệu quả nhất trong việc phân tích cảm tính từ các phụ đề tiêu cực, bao gồm: bốn thuật toán học máy Naive Bayes, SVM, Logistic Regression, Random Forest và hai kiến trúc mạng học sâu CNN, LSTM. Trong nghiên cứu này, tất cả sáu phương pháp này đã được áp dụng để đánh giá hiệu suất của thuật toán tốt nhất trong việc phân loại tình cảm trong nội dung tiêu cực của các phụ đề. Hy vọng rằng thông qua việc thử nghiệm các phương pháp và thuật toán này, chúng ta sẽ có cái nhìn sâu sắc hơn về cách mà công chúng phản ứng và tương tác với thông điệp trên các nền tảng truyền thông xã hội, dựa trên trí tuệ nhân tạo.

2. Phương pháp nghiên cứu

2.1. Đối tượng nghiên cứu

Nghiên cứu này tập trung chủ yếu vào việc phân tích cảm xúc được thể hiện trong các phụ đề trên mạng xã hội YouTube, đặc biệt là trong việc phân loại và nghiên cứu các phụ đề chứa nội dung tiêu cực. Mục tiêu chính của nghiên cứu là hiểu rõ hơn về sự phản ứng của công chúng đối với các sự kiện cụ thể. Dữ liệu nghiên cứu sẽ tập trung vào các từ khóa và vấn đề nổi bật hiện nay như "Chính trị, Xã hội, Thời sự, Oan sai, Đảng và Nhà nước,..." và được lưu trữ trong định dạng CSV. Nghiên cứu sẽ chi tiết hóa phân tích các phản ứng và ý kiến tiêu cực xuất hiện trong phụ đề của các video trên kênh này, giúp đưa ra cái nhìn sâu sắc về cách cộng đồng mạng tương tác với các sự kiện và chủ đề nhất định.

2.2. Mô hình cơ bản

2.2.1. Mô hình học máy

Có nhiều phương pháp để giải quyết bài toán nhận diện phát ngôn thù hận, với các mô hình học máy là phương pháp cơ bản nhất. Dưới đây là một số thuật toán áp dụng cho bài toán trên.

Multinomial Naive Bayes: Thuật toán này dự đoán và phân loại dữ liệu dựa trên dữ liệu và số liệu thống kê có thể quan sát được, sử dụng định lý Bayes của lý thuyết xác suất. Multinomial Naive Bayes là một thuật toán học có giám sát được sử dụng phổ biến trong học máy vì nó tương đối dễ huấn luyện và đạt hiệu suất cao.

Hồi quy logistic: Đây là một thuật toán phân loại nhị phân, nó là một phương pháp đơn giản, nổi tiếng và quan trọng trong lĩnh vực học máy. Ngoài ra, thuật toán này còn được sử dụng trong ứng dụng học máy để phân loại dữ liệu sau dựa trên dữ liệu trước đó. Bằng cách phân tích mối quan hệ giữa tất cả các biến độc lập hiện có, mô hình hồi quy logistic dự đoán một biến dữ liệu phụ thuộc. Trong xử lý ngôn ngữ tự nhiên, phương pháp này yêu cầu trích xuất các đặc trưng thủ công từ dữ liệu để phân loại văn bản.

Decision Tree: Đây là một thuật toán học có giám sát, nó là phương pháp phân loại mạnh mẽ và phổ biến nhất. Thuật toán cây quyết định còn được gọi là cây cấu trúc, trong đó mỗi nút đại diện cho một phép thử trên một thuộc tính, mỗi nhánh là kết quả của phép thử và mỗi nút lá là một nhãn lớp. Cách tiếp cận này sử dụng các quy tắc cơ bản từ dữ liệu huấn luyện để dự đoán lớp hoặc giá trị của biến mục tiêu. Cụ thể, bản ghi bắt đầu từ gốc của cây và so sánh thuộc tính với thuộc tính nút tại mỗi nhánh trong cây quyết định trước khi dự đoán nhãn lớp cuối cùng trong nút lá.

Random forest: Đây là một phương pháp học có giám sát được sử dụng để giải quyết các nhiệm vụ phân loại và hồi quy. Nó được xây dựng trên nhiều bộ cây quyết định và đầu ra của thuật toán này dựa trên quyết định tổng hợp trên các cây quyết định mà nó tạo ra bằng phương thức biểu quyết.

Support Vector Machine (SVM): SVM là một phương pháp học máy được sử dụng chủ yếu trong các bài toán phân loại và hồi quy. Đây là một thuật toán học có giám sát, có khả năng tìm ra ranh giới quyết định tối ưu giữa các lớp hoặc dự đoán một giá trị liên tục. SVM thường được sử dụng cho các bài toán phân loại tuyến tính, nơi mục tiêu là tìm ra một siêu phẳng tốt nhất để phân tách giữa các điểm dữ liệu thuộc các lớp khác nhau.

2.2.2. Mô hình học sâu

CNN (Convolutional Neural Network) là một phương pháp học sâu, kiến trúc của CNN có thể áp dụng trong nhiều lĩnh vực từ nhận dạng hình ảnh đến xử lý ngôn ngữ tự nhiên. Mục tiêu của CNN là giảm kích thước của dữ liệu mà vẫn giữ được các đặc trưng quan trọng trong quá trình xử lý, đảm bảo độ chính xác cao hơn cho các dự đoán.

LSTM (Long Short-Term Memory) là một phương pháp sử dụng mạng neural để học và dự đoán các mẫu trong dữ liệu chuỗi. Trong mạng neural, việc cập nhật các trọng số thông qua thuật toán backpropagation có thể gặp vấn đề như đạo hàm tiêu biến hoặc phát triển vượt quá mức. Kiến trúc bộ nhớ dài hạn (LSTM) là một phiên bản cải tiến của Mạng Nơ-ron Tái phát (RNN), giúp khắc phục vấn đề đạo hàm tiêu biến bằng cách sử dụng thêm một trạng thái tái phát gọi là ô nhớ. Mô hình LSTM cung cấp khả năng học chuỗi dữ liệu trải dài trong thời gian dài, từ đó làm cho nó trở thành một kỹ thuật phù hợp cho nhiệm vụ phân tích cảm xúc. Bằng cách kết hợp các RNN tiến và lùi lại với nhau, ta tạo thành một tensor duy nhất để tăng hiệu suất của mô hình dựa trên LSTM. Ngoài khả năng hai chiều, nhiều lớp LSTM có thể được xếp chồng lên nhau để tăng hiệu suất hơn nữa.

2.3. Tổng quan về bộ dữ liệu

Trong nghiên cứu này, chúng tôi sử dụng ngôn ngữ lập trình Python kết hợp với API YouTube của Google để tự động tải dữ liệu, thay vì phải thủ công truy cập và tải phụ đề từng video. Việc tích hợp API YouTube không chỉ giúp tiết kiệm thời gian mà còn mang lại sự thuận tiện trong quá trình thu thập dữ liệu. Tuy nhiên, hạn chế của tài khoản miễn phí là chỉ cho phép tải xuống tối đa 1000 video trong vòng 24 giờ. Để vượt qua hạn chế này, chúng tôi triển khai 10 tài khoản Gmail khác nhau để tăng tốc quá trình thu thập dữ liệu phụ đề.

Sau quá trình thu thập dữ liệu phụ đề, chúng tôi đã loại bỏ những video không có phụ đề và giữ lại một tập hợp gồm 7180 video đã được phụ đề. Mục tiêu của nhóm nghiên cứu là xây dựng một bộ dữ liệu phụ đề YouTube có chất lượng để phục vụ cho các nghiệp vụ cụ thể của cơ quan đơn vị. Danh sách các video đã được phân chia đều cho mỗi thành viên trong nhóm để thực hiện công việc đánh nhãn, và quá trình này đã kéo dài trong khoảng 3 tuần với sự tham gia của 21 thành viên.

Trong quá trình đánh nhãn dữ liệu phụ đề trên YouTube, chúng tôi đã sử dụng một số tiêu chí nhất định để gán nhãn tích cực, tiêu cực và trung tính. Đầu tiên, chúng tôi xem xét nội dung của video để xác định tính chất của nó. Video mang tính hướng dẫn, giáo dục hoặc mang lại giá trị cho người xem thường được gán nhãn tích cực, trong khi video chứa nội dung không phù hợp, gây căng thẳng hoặc phản cảm thường được gán nhãn tiêu cực. Các video không rõ ràng hoặc không đủ cơ sở để đánh giá được gán nhãn trung tính.

Tiếp theo, chúng tôi xem xét ngôn ngữ và ngữ cảnh trong phụ đề để hiểu ý nghĩa và tư duy của nội dung. Từ ngữ và biểu cảm tích cực như "tuyệt vời", "hấp dẫn", "tôn trọng" thường được liên kết với video tích cực, trong khi từ ngữ tiêu cực như "khó chịu", "phản bội", "thất vọng" thường được liên kết với video tiêu cực.

Cuối cùng, chúng tôi đánh giá phản ứng của người xem đối với video, sử dụng các phản hồi tích cực như số lượt xem, lượt thích và bình luận tích cực để xác định video tích cực. Đối với các trường hợp mâu thuẫn hoặc không rõ ràng, chúng tôi đã thực hiện sự phân tích cẩn thận và thảo luận giữa các thành viên trong nhóm để ra quyết định cuối cùng. Mỗi thành viên trong nhóm được giao khoảng 50 video mỗi ngày, đảm bảo sự phân công công việc một cách hợp lý và hiệu quả.

2.4. Quy trình tiền xử lý dữ liệu

Các kỹ thuật tiền xử lý dữ liệu luôn đóng một vai trò quan trọng trong các nhiệm vụ phân loại dữ liệu phụ đề Tiếng Việt trên các video Youtube. Việc tiền xử lý có tác động đáng kể đối với việc trích xuất thông tin từ dữ liệu. Vì vậy, tiến hành xây dựng quy trình tiền xử lý dữ liệu để cải thiện chất lượng của bộ dữ liệu, nhằm trích xuất các đặc trưng có giá trị trước khi sử dụng chúng để huấn luyện các mô hình phân loại.

Chuyển đổi thành chữ thường: Tất cả các ký tự của tất cả các bình luận trong bộ dữ liệu đều được chuyển đổi thành chữ thường. Việc thực hiện điều này để tránh Python nhận biết hai từ giống hệt nhau nhưng khác nhau về chữ hoa.

Xóa khoảng trắng dư thừa: Loại bỏ các khoảng trắng không cần thiết trong phụ đề, phục vụ mục đích làm cho dữ liệu gọn gàng hơn.

Xóa liên kết: Loại bỏ các liên kết đến trang web trong bình luận, vì chúng không đóng góp vào ý nghĩa của nội dung.

Chuẩn hóa Unicode: Thực tế cho thấy nhiều từ tiếng Việt trong bộ dữ liệu giống nhau nhưng Python nhận biết chúng là khác nhau do sự khác biệt về Unicode. Lý do là có nhiều định dạng biến đổi Unicode (UTF) như UTF-8, UTF-16, UTF-32 được sử dụng rộng rãi, do đó chúng ta nên chuẩn hóa thành định dạng chung, như UTF-8.

Xóa ký tự dư thừa: Loại bỏ các ký tự dư thừa mà người dùng cố ý tạo ra.

Chuẩn hóa ký tự có dấu: Do sự không đồng nhất trong cách đặt dấu trong Tiếng Việt, nên ta sẽ chuẩn hóa chúng trong các bình luận theo các quy tắc sau:

- Nếu chỉ có một nguyên âm, dấu thanh sẽ nằm trên nguyên âm đó. Ví dụ: má, lá, mê.
- Nếu có hai nguyên âm, dấu thanh sẽ nằm trên nguyên âm đầu tiên. Ví dụ: lóa, qùa.
- Nếu có ba nguyên âm hoặc hai nguyên âm kèm theo một phụ âm, dấu thanh sẽ nằm trên nguyên âm thứ hai. Ví dụ: khuỷu, quán.
- "ê" và "ơ" là đặc biệt vì dấu phụ luôn ở trên họ, ví dụ: khuyển, quở.

Tách từ: Quá trình tách từ trong nghiên cứu được thực hiện bằng cách sử dụng bộ tách từ của underthesea, một thư viện NLP mạnh mẽ cho ngôn ngữ Tiếng Việt. Bộ tách từ này giúp chia câu nhập vào thành các đơn vị từ hoặc cụm từ có ý nghĩa, tạo nền tảng cho các pha tiếp theo của quy trình xử lý ngôn ngữ tự nhiên.

Xóa từ dừng (stopwords): Các từ dừng thường là những từ phổ biến như "là", "và", "một",... không mang lại nhiều ý nghĩa khi phân tích cảm xúc. Việc loại bỏ chúng giúp tập trung vào các từ quan trọng hơn trong quá trình phân tích, làm cho kết quả trở nên chính xác hơn và dễ hiểu hơn.

2.5. Trích xuất đặc trưng

Trong quá trình trích xuất đặc trưng văn bản, một trong những phương pháp cơ bản nhất là Term Frequency (TF). Theo phương pháp này, mỗi từ trong văn bản được ánh xạ tới một số biểu thị số lần xuất hiện của từ đó trong toàn bộ kho ngữ liệu. Các phương pháp mở rộng thường sử dụng tần số từ dưới dạng trọng số theo tỷ lệ boolean hoặc logarit. Kết quả của quá trình này là mỗi tài liệu được biểu diễn bằng một vectơ chứa tần suất xuất hiện của các từ trong tài liệu đó.

TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp đánh giá tầm quan trọng của từng từ trong một tài liệu và so sánh với bộ sưu tập tài liệu. Phương pháp này đo lường tần suất xuất hiện của một từ trong một tài liệu và so sánh nó với số lượng tài liệu mà từ đó xuất hiện. TF-IDF là một công cụ mạnh mẽ trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo, giúp hiểu và đánh giá nội dung của văn bản một cách chính xác và hiệu quả. Cách tính cho TF-IDF dựa trên công thức (1):

$$W\ (d,t) = TF\ (d,t) * \log(\frac{N}{df(t)}) \quad (1)$$

Trong đó, W(d, t) = Số lần thuật ngữ "t" xuất hiện trong tài liệu "d". df(t) là Tần suất tài liệu nghịch đảo của thuật ngữ t. TfidfVectorizer chuyển đổi một tập hợp các tài liệu thô thành một ma trận các tính năng TF-IDF.

Bag-of-Words (BoW)

Ngoài ra, chúng ta cũng sử dụng phương pháp Bag-of-Words (BoW) để biểu diễn văn bản. BoW là một phương pháp đơn giản nhưng hiệu quả, trong đó mỗi tài liệu được biểu diễn bằng một vectơ chứa tần suất xuất hiện của các từ trong tài liệu đó. BoW không quan tâm đến thứ tự

của từ, chỉ quan trọng về việc từ đó có xuất hiện trong tài liệu hay không. Từ đó, chúng ta có thể sử dụng CountVectorizer để chuyển đổi dữ liệu văn bản thành ma trận các đặc trưng BoW.

Cả ba phương pháp trích xuất đặc trưng này đều quan trọng trong việc hiểu và biểu diễn nội dung của văn bản trong nghiên cứu của chúng ta.

2.6. Thuật toán nhúng

Nhúng từ là một kỹ thuật học tính năng trong đó mỗi từ hoặc cụm từ vựng được ánh xạ tới một vectơ N chiều của các số thực.

GloVe, hoặc Global Vectors, là một kỹ thuật nhúng từ mạnh mẽ khác. Nó đào tạo vectơ chiều cao cho mỗi từ dựa trên từ xung quanh trong một kho văn bản lớn. GloVe sử dụng các véc-tơ từ được đào tạo trước từ ngữ liệu lớn và cung cấp các vectơ hóa từ với kích thước khác nhau (100, 200, 300).

2.7. Sử dụng các công cụ hỗ trợ

Nghiên cứu này sử dụng một loạt các công cụ để thực hiện phân tích tình cảm, bao gồm cả phần cứng và phần mềm:

Máy Tính: Sử dụng để chạy mã nguồn cho quá trình huấn luyện mô hình và thực hiện phân tích tình cảm trên dữ liệu thu thập từ phụ đề trên YouTube.

PyCharm Community: Môi trường phát triển tích hợp (IDE) được sử dụng để viết, thực thi và quản lý mã nguồn Python cho việc phân tích dữ liệu.

Dữ liệu thu thập từ phụ đề trên Youtube: Dữ liệu phụ đề từ các video trên YouTube được thu thập bằng cách sử dụng API YouTube và các công cụ lập trình Python. Dữ liệu này sẽ được sử dụng để huấn luyện mô hình và thực hiện phân tích tình cảm.

Scikit-learn: Thư viện máy học và phân tích dữ liệu Python, Scikit-learn cung cấp các công cụ và thuật toán tiêu biểu để triển khai mô hình phân loại và đánh giá hiệu suất của chúng.

Các mô hình máy học thực nghiệm trong nghiên cứu được triển khai bằng cách sử dụng thư viện máy học Scikit-learn do Python cung cấp. Scikit-learn là một thư viện đầy đủ tính năng và dễ sử dụng để triển khai các mô hình học máy trong phân loại văn bản. Scikit-learn cung cấp nhiều tính năng và công cụ cho phân loại văn bản, bao gồm: "Mô hình phân loại, Tối ưu hóa siêu tham số, Đánh giá mô hình".

Tối ưu hóa siêu tham số

Để đạt hiệu suất tối ưu, Scikit-learn cung cấp công cụ GridSearchCV. Công cụ này cho phép tìm kiếm siêu tham số tối ưu cho các mô hình học máy. GridSearchCV duyệt qua tất cả các giá trị trong các tập hợp siêu tham số đã được định nghĩa, huấn luyện mô hình với mỗi bộ siêu tham số để tìm ra bộ siêu tham số tối ưu cho mô hình phân loại.

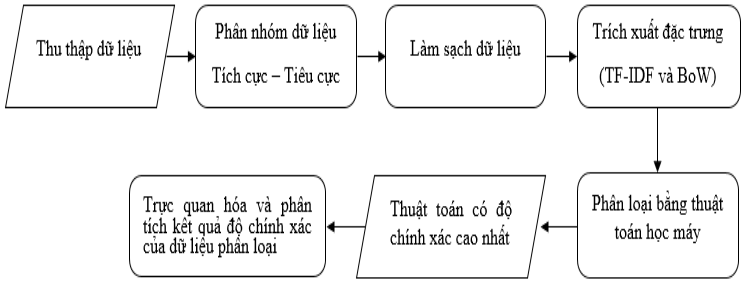
Đánh giá hiệu suất

Đánh giá hiệu suất của mô hình là bước quan trọng. Scikit-learn cung cấp các công cụ đánh giá như cross-validation và các độ đo như accuracy, precision, recall, và F1-score. Những chỉ số này giúp đánh giá độ chính xác và hiệu suất của mô hình phân loại trong việc xác định cảm xúc của các phụ đề YouTube.

2.8. Luồng phương pháp

Quy trình được thực hiện trong nghiên cứu này bao gồm việc Sử dụng API YouTube và các công cụ lập trình Python để thu thập dữ liệu từ phụ đề trên YouTube. Sau đó được phân nhóm, đánh nhãn dữ liệu tích cực và tiêu cực. Bước vô cùng quan trọng đó là dữ liệu thu thập được sẽ được xử lý và làm sạch để loại bỏ các nhiễu và thông tin không mong muốn (như tiền xử lý văn bản Tiếng Việt, loại bỏ từ dừng). Tiếp theo chia dữ liệu thành tập huấn luyện và tập kiểm thử để đảm bảo tính đa dạng và khả năng tổng quát của mô hình. Sử dụng bốn thuật toán phân loại, bao

gồm SVM, Random Forest, Logistic Regression và Naive Bayes để đánh giá và phân loại dữ liệu. Và cuối cùng đánh giá kết quả của mỗi thuật toán để xác định thuật toán nào đạt hiệu suất tốt nhất.



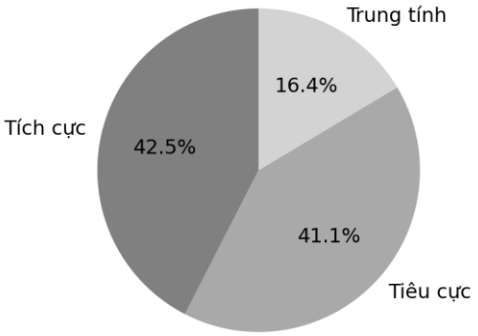
Hình 1. *Mô hình triển khai phân loại cảm xúc phụ đề YouTube*

Đối với bốn thuật toán này, kết quả cuối cùng sẽ được phân tích để xác định thuật toán nào là tốt nhất như trong Hình 1.

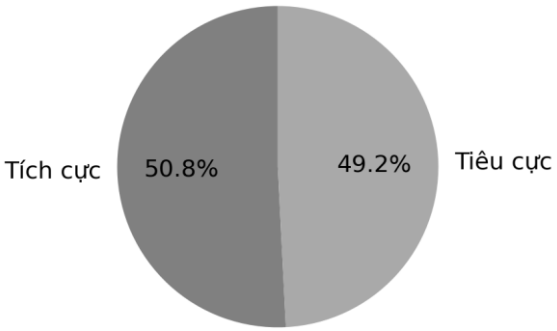
3. Kết quả và thảo luận

Trong nghiên cứu này, chúng tôi sử dụng dữ liệu từ các phụ đề Tiếng Việt trên YouTube. Phụ đề (Subtitles) là văn bản xuất hiện trên màn hình trong quá trình phát video, giúp người xem hiểu được nội dung của video mà không cần phải nghe âm thanh. Các phụ đề thường bao gồm các dòng văn bản chứa lời thoại, thông tin cần thiết hoặc mô tả về âm thanh và hành động trong video. Điều này giúp người xem có thể theo dõi và hiểu rõ hơn về nội dung của video một cách thuận tiện.

Bộ dữ liệu bao gồm tổng cộng 7.180 phụ đề Tiếng Việt sau quá trình lọc bỏ những video không có phụ đề. Sau đó dữ liệu được nhóm lại và chia thành ba lớp: 3048 phụ đề được đánh nhãn tích cực, 2952 phụ đề được đánh nhãn tiêu cực và 1180 phụ đề được đánh nhãn trung tính. Hình 2 thể hiện kết quả của dữ liệu đã được nhóm thành ba lớp. Sau khi việc phân loại các lớp hoàn tất, lớp trung tính đã được loại bỏ hoặc xóa bỏ.



Hình 2. *Dữ liệu trong ba lớp (tích cực, tiêu cực và trung tính)*



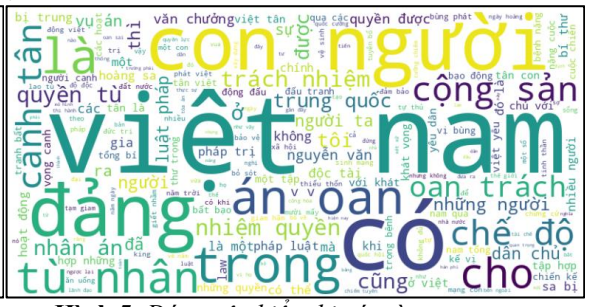
Hình 3. *Biểu đồ phân phối phụ đề tích cực và tiêu cực*

Hình 3 thể hiện kết quả sau khi loại bỏ lớp trung tính khỏi dữ liệu đã được phân cụm. Cụ thể, hình 3 cho thấy rằng tỷ lệ dữ liệu thuộc lớp tiêu cực là 50,8% và lớp tích cực là 49,2%. Sau khi phân nhóm dữ liệu xong ta sẽ làm các tiền xử lý văn bản Tiếng Việt và loại bỏ từ dừng, những từ không có giá trị. Số lượng token trung bình của mỗi phụ đề là khoảng 1145, tổng số từ vựng trong tập dữ liệu sau khi tiền xử lý là 304.524 từ.

Trong nghiên cứu này tôi so sánh hai phương pháp trích xuất đặc trưng TF-IDF và BoW cho các thuật toán học máy, sử dụng phương pháp GloVe cho các mô hình học sâu và so sánh thời gian huấn luyện và các độ đo như accuracy, precision, recall, và F1-score giúp đánh giá độ chính xác và hiệu suất của mô hình phân loại trong việc xác định cảm xúc của các phụ đề YouTube.

Dữ liệu này sẽ được hiển thị dưới dạng đám mây từ (wordcloud), một kỹ thuật trực quan hóa dữ liệu thường được sử dụng để biểu thị tần suất xuất hiện của các từ trong một tập hợp văn bản. Các từ được hiển thị với kích thước tương ứng với tần suất của chúng: từ nào xuất hiện nhiều lần sẽ có kích thước lớn hơn. Hình 4 và Hình 5 minh họa việc sử dụng đám mây từ để phân tích các phụ đề YouTube, điều này giúp người xem nhanh chóng nắm bắt được những từ quan trọng và phổ biến nhất trong tập dữ liệu tích cực và tiêu cực.

Hình 4. *Đám mây hiển thị các từ quan trọng được sử dụng trong tập dữ liệu tích cực*



Hình 5. *Đám mây hiển thị các từ quan trọng được sử dụng trong tập dữ liệu tiêu cực*

Sau khi dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỉ lệ 80-20, mỗi tập dữ liệu sẽ được sử dụng để huấn luyện và kiểm tra hiệu suất của các thuật toán phân loại khác nhau, bao gồm Naïve Bayes [2], Random Forest [3], SVM và Logistic Regression [4]. Trong quá trình đầu tiên, mô hình Logistic Regression đã đạt được kết quả độ chính xác là 79,16% trên tập kiểm tra. Tiếp theo, tập dữ liệu đã được phân chia sẽ được sử dụng để huấn luyện và kiểm tra lại, kết quả độ chính xác của mô hình Logistic Regression là 79%. Kết quả của phân loại sử dụng Logistic Regression được thể hiện trong Hình 6.

	precision	recall	f1-score	support
0	0.79	0.82	0.81	127
1	0.79	0.76	0.77	113
accuracy			0.79	240
macro avg	0.79	0.79	0.79	240
weighted avg	0.79	0.79	0.79	240

Hình 6. *Kết quả chính xác với Logistic Regression TF-IDF trên tập dữ liệu kiểm thử*

Tiếp theo, trong thuật toán Random Forest, bước đầu tiên sau khi dữ liệu được chia là xác định max_depth để biết đến đâu là độ sâu tối đa của cây quyết định, kết quả là có giá trị là 5. Sau đó, dữ liệu được huấn luyện và kiểm thử bằng cách sử dụng max_depth đã xác định, với kết quả cuối cùng là độ chính xác của thuật toán Random Forest là 81%. Kết quả của quá trình phân loại bằng thuật toán Random Forest được xem trong Hình 7.

	precision	recall	f1-score	support
0	0.82	0.83	0.82	127
1	0.81	0.79	0.80	113
accuracy			0.81	240
macro avg	0.81	0.81	0.81	240
weighted avg	0.81	0.81	0.81	240

Hình 7. *Kết quả chính xác với Random Forest TF-IDF trên tập dữ liệu kiểm thử*

Trong thuật toán Naive Bayes, dữ liệu sau khi được huấn luyện và kiểm thử có độ chính xác là 80%. Kết quả của quá trình phân loại bằng thuật toán Naive Bayes được xem trong Hình 8.

Trong thuật toán SVM, dữ liệu sau khi được huấn luyện và kiểm thử có độ chính xác là 77%. Kết quả của quá trình phân loại bằng thuật toán Random Forest được xem trong Hình 9.

	precision	recall	f1-score	support
0	0.77	0.89	0.83	127
1	0.85	0.71	0.77	113
accuracy			0.80	240
macro avg	0.81	0.80	0.80	240
weighted avg	0.81	0.80	0.80	240

Hình 8. *Kết quả chính xác với Naive Bayes TF-IDF trên tập dữ liệu kiểm thử*

	precision	recall	f1-score	support
0	0.80	0.74	0.77	127
1	0.73	0.80	0.76	113
accuracy			0.77	240
macro avg	0.77	0.77	0.77	240
weighted avg	0.77	0.77	0.77	240

Hình 9. *Kết quả chính xác với SVM TF-IDF trên tập dữ liệu kiểm thử*

Dựa trên nghiên cứu của Vu et al. [15] đã được trình bày về việc cải thiện phân tích phụ thuộc tiếng Việt bằng cách sử dụng các biểu diễn từ phân tán. Hệ thống phân tích của họ đạt được độ chính xác là 76,29% điểm gắn kết không được gán nhãn hoặc 69,25% điểm gắn kết được gán nhãn. Đây được coi là trình phân tích phụ thuộc chính xác nhất cho ngôn ngữ tiếng Việt so với các hệ thống khác được huấn luyện và kiểm tra trên cùng một tập dữ liệu treebank [16], [17], [18]. Các biểu diễn từ phân tán được tạo ra bởi hai mô hình học không giám sát gần đây đó là mô hình Skip-gram và mô hình GloVe. Các tác giả cũng đã chỉ ra rằng các biểu diễn phân tán được tạo ra bởi mô hình GloVe có hiệu suất tốt hơn so với các biểu diễn được tạo ra bởi mô hình Skip-gram khi được sử dụng trong phân tích phụ thuộc. Đối với mô hình học sâu, chúng tôi sử dụng Glove đã được huấn luyện trên bộ dữ liệu Treebank gồm 2.700 câu chứa 34.884 token từ, để thực hiện vector hoá từ, làm tham số đầu vào cho mô hình CNN, LSTM [13], [14] như bảng 1.

Bảng 1. *Thống kê một số giá trị cho các mô hình học sâu*

STT	Mô hình	Giá trị các tham số
1	CNN	Filters: 64; Kernel_size: (3, 3); Strides: (1, 1); Padding: 'same'; Activation: 'relu'; Pool_size: (2, 2); Dropout: 0.25; Loss: 'categorical_crossentropy'; Optimizer: Adam optimizer; Epochs: 10; Batch_size: 32
2	LSTM	Units: 128; Dropout: 0.2; Recurrent_dropout: 0.2; Activation: 'relu'; Loss: 'categorical_crossentropy'; Optimizer: Adam optimizer; Epochs: 10; Batch_size: 32

Mô hình CNN trong nghiên cứu được xây dựng với các lớp Embedding, Conv1D, GlobalMaxPooling1D, Dense và Dropout. Lớp Embedding biểu diễn từ vựng thành các vectơ có số chiều là 100. Một lớp Conv1D với 128 bộ lọc, kích thước 5 và hàm kích hoạt ReLU được thêm vào mô hình để trích xuất đặc trưng từ dữ liệu văn bản. Lớp GlobalMaxPooling1D được sử dụng để trích xuất đặc trưng quan trọng nhất từ ma trận đặc trưng đầu ra của lớp Conv1D. Một lớp Dense với 128 đơn vị và hàm kích hoạt ReLU được thêm vào để học các mối quan hệ phi tuyến tính giữa các đặc trưng. Một lớp Dropout với tỷ lệ dropout là 0,5 được sử dụng để tránh hiện tượng quá khớp. Một lớp Dense với một đơn vị và hàm kích hoạt sigmoid được thêm vào để đưa ra dự đoán nhị phân về lớp đích. Mô hình được biên dịch bằng thuật toán tối ưu hóa Adam và hàm mất mát là entropy chéo nhị phân, cùng với độ chính xác được sử dụng làm độ đo đánh giá hiệu suất. Quá trình huấn luyện của mô hình được giám sát bằng một sự kiện EarlyStopping, giúp dừng quá trình huấn luyện nếu không có sự cải thiện đáng kể trong việc đánh giá trên tập validation sau một số lượng epochs được chỉ định (trong trường hợp này là 10 epochs).

Mô hình LSTM trong nghiên cứu được xây dựng với các lớp Embedding, LSTM và Dense. Lớp Embedding biểu diễn từ vựng thành các vectơ có số chiều là 100. Lớp LSTM với 128 đơn vị được sử dụng để học các mẫu dữ liệu dựa trên thông tin chuỗi. Cuối cùng, một lớp Dense với một đơn vị và hàm kích hoạt sigmoid được thêm vào để đưa ra dự đoán nhị phân về lớp đích. Mô hình được biên dịch bằng thuật toán tối ưu hóa Adam và hàm mất mát là entropy chéo nhị phân, cùng với độ chính xác được sử dụng làm độ đo đánh giá hiệu suất. Quá trình huấn luyện của mô hình được giám sát bằng một sự kiện EarlyStopping, giúp dừng quá trình huấn luyện nếu không có sự cải thiện đáng kể trong việc đánh giá trên tập validation sau một số lượng 10 epochs.

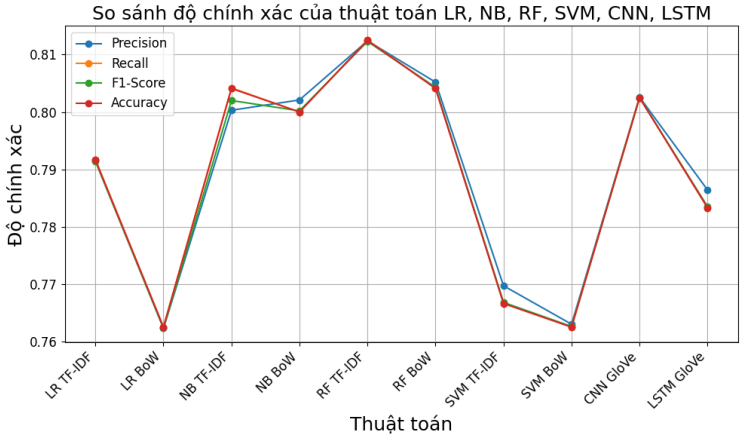
Bảng 2. *Kết quả so sánh các thuật toán học máy, học sâu*

TT	Thuật toán	Phương pháp	Thời gian (s)	Precision	Recall	F1-Score	Accuracy
1	LR	TF-IDF	0,0509	0,7916	0,7917	0,7914	0,7916
		BoW	0,5911	0,7623	0,7625	0,7623	0,7625
2	NB	TF-IDF	0,1861	0,8003	0,8042	0,8020	0,8041
		BoW	0,3805	0,8021	0,8000	0,8002	0,8000
3	RF	TF-IDF	**1,390**	**0,8124**	**0,8125**	**0,8123**	**0,8125**
		BoW	1,080	0,8052	0,8042	0,8043	0,8041
4	SVM	TF-IDF	1,6201	0,7697	0,7667	0,7668	0,7666
		BoW	0,7325	0,7630	0,7625	0,7626	0,7625
5	CNN	GloVe	5,2073	0,8026	0,8025	0,8025	0,8025
6	LSTM	GloVe	5,888	0,7864	0,7833	0,7835	0,7833

Có thể thấy trong Bảng 2, thuật toán tốt nhất là Random Forest khi sử dụng TF-IDF đạt độ chính xác cao nhất 81,25%, vượt trội so với các thuật toán khác. Tiếp theo là Naïve Bayes với 80%, Logistic Regression và Support Vector Machine thấp hơn, độ chính xác lần lượt là 79% và 77%.

Khi sử dụng BoW, Random Forest đạt độ chính xác cao nhất đạt 80,41%, vượt trội so với các thuật toán khác. Tiếp đó là Naïve Bayes đạt độ chính xác 80% và Logistic Regression và Support Vector Machine có độ chính xác bằng nhau 76%. Tuy nhiên, khi sử dụng phương pháp BoW, độ chính xác của cả ba thuật toán Logistic Regression, Naïve Bayes và Logistic Regression đều giảm xuống. Trong khi đó, SVM hiển thị sự ổn định với độ chính xác khá cao khi sử dụng cả hai phương pháp trích xuất đặc trưng.

Về tổng thể, Random Forest kết hợp với TF-IDF là sự kết hợp tốt nhất tại thời điểm nghiên cứu, mang lại hiệu suất cao cả về độ chính xác và thời gian thực hiện. Đối với hai mô hình học sâu CNN và LSTM sử dụng phương pháp GloVe, chúng cho thấy độ chính xác tương đối cao, tuy nhiên thời gian thực hiện của chúng lớn hơn đáng kể so với các thuật toán học máy truyền thống.



Hình 10. *Biểu đồ so sánh độ chính xác*

Biểu đồ Hình 10 trên đây so sánh độ chính xác của bốn thuật toán học máy Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), SVM khi được áp dụng trên hai phương pháp trích xuất đặc trưng là TF-IDF, Bag-of-Words (BoW) và hai mô hình học sâu CNN, LSTM sử dụng phương pháp GloVe.

4. Kết luận

Trong cả hai phương pháp trích xuất đặc trưng TF-IDF và BoW, Random Forest đã thể hiện sự vượt trội với độ chính xác cao nhất đạt được là 81,25% và 80,41%. Điều này cho thấy Random Forest có khả năng phân loại tốt nhất trong việc nhận diện cảm xúc từ các phụ đề tiêu cực khi sử dụng phương pháp này. Naïve Bayes đều đạt độ chính xác 80% với cả hai phương pháp, Random Forest và Support Vector Machine thấp nhất.

Trong hai mô hình học sâu, CNN và LSTM sử dụng phương pháp GloVe, CNN có độ chính xác cao nhất lần lượt là 80,25% và 78,33%. Điều này cho thấy mô hình CNN hiệu quả hơn trong việc phân loại tình cảm từ các phụ đề tiêu cực so với mô hình LSTM.

Random Forest kết hợp với TF-IDF không chỉ có độ chính xác cao mà còn có thời gian thực hiện nhanh, chỉ khoảng 1,3 giây. Trong khi đó, SVM khi sử dụng TF-IDF mất nhiều thời gian nhất gần 1,6 giây. Logistic Regression, mặc dù có độ chính xác thấp hơn, nhưng thời gian thực hiện nhanh hơn so với các thuật toán khác. Trong hai mô hình học sâu sử dụng phương pháp GloVe, CNN có thời gian thực hiện nhanh hơn so với LSTM, với CNN mất khoảng 5,2 giây trong khi LSTM mất khoảng 5,8 giây.

Dựa trên các kết quả này, Random Forest là lựa chọn tốt nhất cho việc phân loại tình cảm trong nội dung tiêu cực của các phụ đề tại thời điểm nghiên cứu. Random Forest không chỉ đạt được độ chính xác cao mà còn có thời gian thực hiện nhanh so với các thuật toán khác.

Đề xuất cho nghiên cứu tiếp theo:

Tối ưu hóa hiệu suất của các thuật toán bằng cách thêm các tham số tinh chỉnh để cải thiện độ chính xác.

Mở rộng quy mô của dữ liệu nghiên cứu để đảm bảo tính độ chính xác cao hơn trong kết quả đạt được.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, 2014.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Opinion Mining and Sentiment Analysis" *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 866–879, June 2008, doi: 10.1109/TKDE.2008.90.

[3] X. Song, X. Liang, and Y. Ma, "A Sentiment Analysis Approach to Predict Stock Market Trends," *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018, pp. 3596–3602, doi: 10.1109/BigData.2018.8621985.

[4] B. Pang and L. Lee, "Sentiment Analysis and Opinion Mining: A Survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 1299–1323, Sept. 2016, doi: 10.1109/TKDE.2015.2476522

[5] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods" in *Proc. first ACM conference on Online social networks*, New York, USA, 2013, pp. 27–38, doi: 10.1145/2512938.2512951.

[6] W. Medhat, *et al., Sentiment analysis algorithms and applications: A survey*, Elsevier, 2020.

[7] W. Y. Chong, *et al., Natural Language Processing for Sentiment Analysis*, IEEE, 2019.

[8] H. Bhuiyan, K. J. Oh, M. K. Hong, and G. S. Jo, "An unsupervised approach for identifying the Infobox template of wikipedia article," in 18th *International Conference on Computational Science and Engineering (CSE)*, 2015, IEEE, pp. 334-338.

[9] R. Novendri *et al.*, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *Bulletin of Computer Science and Electrical Engineering*, 2020.

[10] W. Tafesse, *YouTube marketing: how marketers' video optimization practices influence video views*, Internet Research, Emerald Publishing Limited, 2020.

[11] D. Das, *et al., Affective Computing and Sentiment Analysis*, Springer, 2018.

[12] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in *Proc. 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, IEEE, pp. 452–455.

[13] M. Cliche, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs," *arXiv:1704.06125* [cs.CL], Apr. 2017, doi: 10.48550/arXiv.1704.06125.

[14]C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, Aug. 2017, pp. 747–754.

[15]C. M. Vu, T. A. Luong, and L. H. Phuong, "Improving Vietnamese Dependency Parsing Using Distributed Word Representations," in *Proceedings of the 6th International Symposium on Information and Communication Technology (SoICT)*, Hue, Vietnam, 2015, doi: 10.1145/2833258.2833296.

[16]P. T. Nguyen, L. V. Xuan, T. M. H. Nguyen, V. H. Nguyen, and P. Le-Hong, "Building a large syntactically-annotated corpus of Vietnamese," in *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP*, Suntec City, Singapore, 2009, pp. 182–185.

[17]T.-L. Nguyen, V.-H. Nguyen, T.-M.-H. Nguyen, and P. Le-Hong, "Building a treebank for Vietnamese dependency parsing," in *Proceedings of RIVF*, IEEE, 2013, pp. 147–151.

[18] VLSP Project, "Resources for Vietnamese," 2024. [Online]. Available: https://vlsp.hpda.vn/demo/. [Accessed Mar. 6, 2024].