Lê Dương Minh Thiên - 22521386

Trường Đại học Công Nghệ Thông Tin
Thành phố Hồ Chí Minh, Việt Nam.

Tóm tắt nội dung

Đối với các nền tảng thương mại điện tử như Shopee, việc đảm bảo sự hài lòng của khách hàng thông qua chất lượng sản phẩm và dịch vụ là rất quan trọng. Shopee cam kết mang đến trải nghiệm tốt nhất cho người dùng, từ khâu tìm kiếm sản phẩm cho đến giao hàng, bao gồm cả đóng gói và chất lượng sản phẩm.Đánh giá và nhận xét của khách hàng là thông tin quý giá giúp các đối tác và người bán hiểu rõ nhu cầu của người tiêu dùng. Những phản hồi này không chỉ phản ánh chất lượng sản phẩm mà còn bao gồm các yếu tố như dịch vụ giao hàng, đóng gói, và phương thức thanh toán. Để xử lý và phân tích các đánh giá này theo thời gian thực, dự án sử dụng Pyspark và Kafka, giúp phân tích sắc thái các bình luận sản phẩm và cải thiện chất lượng trải nghiệm người dùng.

Keywords: Lời nói độc hại, Phân tích cảm xúc, Dữ liệu lớn, Truyền dữ liệu trực tuyến

1 Giới thiệu

Cùng với sự phát triển mạnh mẽ của công nghệ trong cuộc Cách mạng Công nghiệp lần thứ Tư, sự bùng nổ của các nền tảng thương mại điện tử như Shopee đã thay đổi đáng kể cách thức mua sắm của người tiêu dùng. Trong đó, đánh giá và phản hồi của người dùng đối với các sản phẩm là nguồn thông tin quý giá giúp người bán hiểu rõ nhu cầu và mong muốn của khách hàng. Những đánh giá này không chỉ giúp người bán nắm bắt được chất lượng sản phẩm, dịch vụ giao hàng hay phương thức thanh toán mà còn cung cấp cái nhìn sâu sắc về những cải tiến cần thiết để nâng cao trải nghiệm của khách hàng trong các đơn hàng tiếp theo.

Tuy nhiên, việc xử lý và phân tích những đánh giá này theo thời gian thực lại gặp phải nhiều thách thức, đặc biệt khi dữ liệu đánh giá liên tục thay đổi và có sự đa dạng về ngữ nghĩa. Đồ án của chúng tôi tập trung vào việc phân

tích sắc thái của các bình luận và đánh giá sản phẩm trên Shopee bằng cách sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến. Mục tiêu là phát triển một hệ thống có khả năng phân loại chính xác các phản hồi của người dùng, từ đó giúp người bán nhanh chóng điều chỉnh các chiến lược bán hàng và cải thiện chất lượng dịch vụ.

Bằng cách áp dụng các công cụ như Pyspark và Kafka để xử lý và phân tích dữ liệu theo thời gian thực, hệ thống của chúng tôi sẽ giúp việc đưa ra các quyết định kinh doanh trở nên nhanh chóng và hiệu quả hơn. Những thông tin thu thập được từ các đánh giá sản phẩm không chỉ mang lại lợi ích cho người bán mà còn góp phần xây dựng một không gian mua sắm an toàn và lành mạnh hơn trên Shopee.

2 Các công trình nghiên cứu liên quan

Một trong những nghiên cứu quan trọng trong lĩnh vực phát hiện ngôn từ thù địch và xúc phạm trong tiếng Việt là bài báo "Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data". Bài nghiên cứu này sử dụng kết hợp PhoBERT, một mô hình ngôn ngữ tiếng Việt tiên tiến, với CNN (Convolutional Neural Networks) để phát hiện ngôn từ thù địch và xúc phạm trên các nền tảng mạng xã hội.

PhoBERT, một phiên bản của BERT được huấn luyện riêng cho tiếng Việt, đã được chứng minh là hiệu quả trong việc xử lý các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản. Các tác giả của nghiên cứu này đã sử dụng PhoBERT kết hợp với CNN để phát hiện các bình luận thù địch và xúc phạm trong môi trường trực tuyến. Bằng cách áp dụng phương pháp học sâu, hệ thống có thể học các đặc trưng ngữ nghĩa từ dữ liệu mạng xã hội, giúp phân loại chính xác các bình luận có tính chất tiêu cực.

Ngoài việc sử dụng PhoBERT và CNN, nghiên cứu còn sử dụng dữ liệu streaming từ các nền tảng mạng xã hội, điều này giúp hệ thống có thể phân tích dữ liệu theo thời gian thực, nhận diện kịp thời các bình luận có nội dung thù địch. Điều này đặc biệt quan trọng khi xử lý khối lượng lớn dữ liệu và các bình luận xuất hiện liên tục trên các nền tảng như Facebook, Twitter, hay các mạng xã hội khác.

Kết quả từ nghiên cứu này cung cấp một phương pháp hiệu quả cho việc phát hiện ngôn từ thù địch trong tiếng Việt, đồng thời mở ra hướng đi mới trong việc áp dụng các mô hình học sâu và dữ liệu streaming để giải quyết vấn đề này. Nghiên cứu đã đóng góp quan trọng vào việc xây dựng các hệ thống tự động phát hiện ngôn từ thù địch trong các ứng dụng thực tế, đồng thời cung cấp một công cụ hữu ích cho các nền tảng mạng xã hội trong việc duy trì môi trường trực tuyến an toàn và lành mạnh.

Một nghiên cứu quan trọng trong việc phát hiện ngôn từ thù địch là bài báo "Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data". Bài báo này kết hợp mô hình PhoBERT và CNN để phát hiện các bình luận thù địch và xúc phạm trên các nền tảng mạng xã hội tiếng Việt. Các tác giả sử dụng dữ liệu streaming để phân tích và phân loại

các bình luận theo thời gian thực, giúp cải thiện chất lượng trải nghiệm người dùng trên các nền tảng trực tuyến.

Bài nghiên cứu "A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts"cũng đóng góp bộ dữ liệu quy mô lớn cho việc phát hiện ngôn từ thù địch trong tiếng Việt. Bộ dữ liệu này được thu thập từ các nền tảng mạng xã hội như Facebook và Zalo, giúp phát triển các mô hình học máy nhằm nhận diện ngôn từ thù địch trong các bình luận mạng xã hội. Bộ dữ liệu này đã thúc đẩy việc nghiên cứu và xây dựng các hệ thống tự động phát hiện ngôn từ thù địch trong môi trường trực tuyến.

3 Bộ dữ liệu

3.1 Giới thiệu bài toán

Trong môi trường thương mại điện tử như Shopee, các đánh giá của người tiêu dùng rất quan trọng để cải thiện chất lượng sản phẩm và dịch vụ. Tuy nhiên, một số bình luận có thể chứa ngôn từ thù địch hoặc xúc phạm, ảnh hưởng đến trải nghiệm người dùng. Bài toán trong đồ án này là phát triển hệ thống tự động phân loại các bình luận thù địch và xúc phạm trên Shopee bằng các kỹ thuật xử lý ngôn ngữ tự nhiên. Hệ thống sẽ sử dụng Pyspark và Kafka để xử lý dữ liệu theo thời gian thực và giúp người bán nhanh chóng điều chỉnh chiến lược để cải thiện dịch vụ.

Input: Dữ liệu đầu vào là các bình luận sản phẩm từ Shopee, được thu thập qua API Shopee và truyền qua Kafka topic. Dữ liệu này bao gồm văn bản các bình luận và các thông tin liên quan như ID sản phẩm, ID người dùng.

Output: Kết quả đầu ra là nhãn phân loại của các bình luận (thù địch, xúc phạm, hoặc bình thường), được gửi qua Kafka topic cho các bước xử lý tiếp theo.

3.2 Tổng quan

Bộ dữ liệu mà tôi sử dụng được lấy từ nghiên cứu Vietnamese Hate Speech Detection (Lưu et al., 2021) [1]. Tập dữ liệu ViHSD bao gồm 33.400 bình luận thu thập từ các mạng xã hội, được chia thành ba phần: train, dev và test. Mỗi bình luận trong bộ dữ liệu này sẽ được phân loại vào một trong ba nhóm: CLEAN (bình thường), OFFENSIVE (xúc phạm) hoặc HATE (thù địch). Tuy nhiên, có sự không cân đối rõ rệt về số lượng bình luận được gán nhãn CLEAN so với các bình luận thuộc nhãn OFFENSIVE và HATE.

Tôi nhận thấy rằng cả hai nhóm bình luận OFFENSIVE và HATE đều chứa đựng những nội dung độc hại, gây ảnh hưởng tiêu cực đến người dùng. Do đó, chúng tôi quyết định hợp nhất hai nhãn OFFENSIVE và HATE thành một nhãn duy nhất là HARMFUL (độc hại). Bảng 1 cung cấp cái nhìn tổng quan về bộ dữ liệu này.

3.3 Tiền xử lí dữ liệu

Các kỹ thuật tiền xử lý dữ liệu luôn đóng vai trò quan trọng trong các nhiệm vụ phân loại dữ liệu từ các mạng xã hội Việt Nam nói chung và trong nhiệm vụ phát hiện ngôn từ thù địch nói riêng. Các bình luận bằng tiếng Việt trên mạng xã hội thường chứa các ký tự và từ ngữ liên quan đến các sắc thái cảm xúc được thể hiện theo nhiều cách khác nhau, khiến việc nhận diện, phân biệt và trích xuất thông tin trở nên khó khăn. Để đạt hiệu quả tốt nhất, tôi chia làm 2 giai đoạn:

Giai đoạn 1: Trong giai đoạn này, tôi sẽ loại bỏ các biểu tượng emoji, đường link, URL, ký tự đặc biệt, hashtag, từ lặp lại và khoảng trắng dư thừa. Bên cạnh đó, tôi sẽ chuẩn hóa các từ viết tắt trong tiếng Việt. Ví dụ: "hp"sẽ được chuyển thành "hạnh phúc", "kakak"sẽ thành "kaka"(tiếng cười), và "đc"sẽ là "được". Việc chuẩn hóa này giúp cải thiện tính chính xác khi phân tích các bình luận.

Giai đoạn 2: rong giai đoạn tiếp theo, tôi sẽ sử dụng thư viện Pyvi để phân tách các bình luận thành các từ hoặc cụm từ có ý nghĩa. Sau đó, các stopwords, những từ không mang nhiều thông tin hoặc ý nghĩa trong câu, sẽ được loại bỏ. Để thực hiện bước này, chúng tôi sử dụng bộ từ điển stopword tiếng Việt [2] để lọc các stopwords khỏi các bình luận.

4 Phương pháp

4.1 Mô hình phân loại

- **Logistic Regression**: Là một phương pháp thống kê được sử dụng để dự đoán mối quan hệ giữa một biến phụ thuộc nhị phân (có hai giá trị, chẳng hạn như 0 và 1) và một hoặc nhiều biến độc lập (có thể là rời rạc hoặc liên tục).
- **Decision Tree**: Là một mô hình học máy dùng để ra quyết định thông qua việc phân chia dữ liệu dựa trên các thuộc tính. Nó được biểu diễn dưới dạng một cấu trúc cây, với nút gốc đại diện cho dữ liệu ban đầu, các nút nội thể hiện các thuộc tính, và các nút lá cho kết quả cuối cùng. Quy trình phân tách dựa trên các điều kiện cho phép cây quyết định phân loại dữ liệu hoặc dự đoán giá trị.

4.2 Công nghệ sử dụng

- **Apache Spark:** Xử lý dữ liệu lớn và huấn luyện mô hình máy học.
- **Apache Kafka:** Quản lý luồng dữ liệu thời gian thực.

4.3 Quy trình xây dựng hệ thống

4.3.1 Kiến trúc hệ thống

4.3.2 Triển khai hệ thống

- *Huấn luyện mô hình dựa trên bộ dữ liệu có sẵn*: Quá trình huấn luyện mô hình Decision Tree bao gồm các bước chính như sau:

  - **Chuẩn bị dữ liệu**: Tách từ các bình luận bằng thư viện pyvi và chuyển dữ liệu thành dạng DataFrame với schema xác định. Tạo bộ dữ liệu huấn luyện và kiểm tra từ các câu và nhãn.
  - **Tiền xử lý dữ liệu**: Sử dụng UDF để tách từ và tạo n-grams từ các từ đã tách. Tính toán các đặc trưng TF-IDF cho các n-grams.
  - **Xây dựng mô hình Decision Tree**: Khởi tạo mô hình Decision Tree với các tham số như độ sâu cây tối đa (maxDepth) và số lượng tối thiểu các mẫu trong mỗi nút (minInstancesPerNode).
  - **Tối ưu hóa tham số**: Sử dụng CrossValidator để tối ưu hóa các tham số mô hình với phương pháp đánh giá độ chính xác (accuracy).
  - **Huấn luyện mô hình**: Tạo pipeline gồm các bước tiền xử lý và huấn luyện mô hình trên tập dữ liệu huấn luyện.

- *Thu thập dữ liệu từ Shopee API*: Quá trình thu thập dữ liệu bình luận từ Shopee API bao gồm các bước chính như sau:

  - **Khởi tạo Kafka Producer**: Cấu hình Kafka Producer để gửi dữ liệu thu thập được vào một topic Kafka.
  - **Lấy dữ liệu từ Shopee API**: Sử dụng API của Shopee để lấy bình luận sản phẩm, bao gồm các thông tin như đánh giá sao, nội dung bình luận, mã đơn hàng và thời gian bình luận. Dữ liệu này được lấy theo từng đợt với giới hạn và offset để thu thập liên tục.
  - **Xử lý và gửi dữ liệu**: Dữ liệu thu thập được sẽ được xử lý, chuyển đổi thời gian từ dạng timestamp sang định dạng thời gian dễ đọc, và gửi vào Kafka để tiếp tục xử lý trong các bước sau.
  - **Lặp lại quá trình**: Mỗi 5 phút, hệ thống sẽ gửi yêu cầu lấy dữ liệu bình luận mới và tiếp tục thu thập dữ liệu cho đến khi có yêu cầu ngừng.

- *Xử lý và phân loại cảm xúc bình luận từ dữ liệu streaming*: Quá trình xử lý và phân loại cảm xúc bình luận từ dữ liệu streaming bao gồm các bước chính như sau:

  - **Khởi tạo SparkSession**: Cấu hình Spark để xử lý dữ liệu streaming từ Kafka với các gói cần thiết, như spark-sql-kafka và kafka-clients.
  - **Đọc dữ liệu từ Kafka**: Dữ liệu được đọc từ Kafka topic rawData, sử dụng schema xác định cấu trúc của dữ liệu, bao gồm các trường như rating, comment, orderID, và thời gian.
  - **Tiền xử lý dữ liệu**: Các bình luận được tiền xử lý (ví dụ chuyển thành chữ thường, loại bỏ stopwords) thông qua một UDF (User Defined Function).
  - **Dự đoán cảm xúc**: Sử dụng hàm phân loại cảm xúc để dự đoán nhãn cảm xúc (positive hoặc negative) cho từng bình luận.
  - **Gửi kết quả vào Kafka**: Kết quả dự đoán được chuyển đổi thành định dạng JSON và gửi vào Kafka topic.