Review Sản Phẩm Bằng Bình Luận Người Mua

Võ Gia Bảo[18520502], Nguyễn Đức Hà[18520689], Lê Hoài Nam[18521121]

University of Information Technology – VNUHCM, Vietnam

***Abstract* - Với sự phát triển mạnh mẽ của công nghệ thông tin và Internet, các website Thương mại điện tử ra đời như một phương tiện hữu ích giúp khách hàng thực hiện mua hàng trực tuyến cũng như chia sẻ những trải nghiệm, bình luận và đánh giá sau giao dịch. Vì vậy để có thể thấu hiểu hành vi khách hàng thông qua ý kiến tích cực hay tiêu cực về sản phẩm và dịch vụ được trải nghiệm là một trong những vấn đề quan trọng. Giải pháp cho vấn đề này, nghiên cứu đề xuất phương pháp khai thác ý kiến và phân tích cảm xúc khách hàng thông qua việc thu thập tập dữ liệu là ý kiến bình luận của khách hàng trên shopee và lazada. Sau đó, tiến hành thực nghiệm bằng phương pháp học máy để khai phá ý kiến từ bình luận dạng văn bản của khách hàng và trực quan hóa kết quả. Nhìn vào kết quả biết được sản phẩm đó được đánh giá tốt, xấu như thế nào và từ đó giúp các cửa hàng, nhà quản trị hiểu được các ưu nhược điểm về sản phẩm, dịch vụ để cải thiện chiến lược kinh doanh tốt hơn.**

Keywords*: *NLP, Shopee, Big data, …

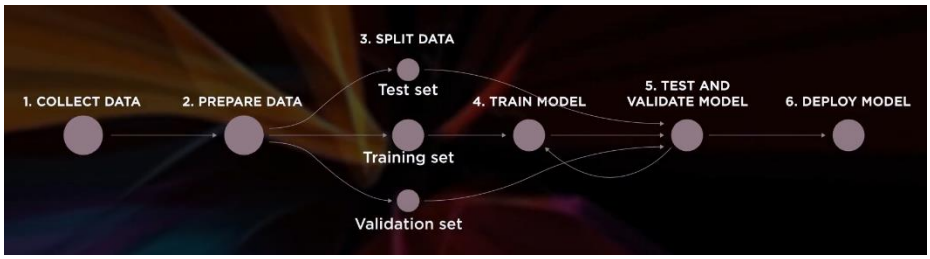
1 Giới thiệu

- Đầu vào: Các bình luận phản hồi của khách hàng về sản phẩm
- Đầu ra: Dự đoán đánh giá sản phẩm

Trong dự án lần này, chúng tôi sử dụng mô hình phân loại cảm xúc với dữ liệu đầu vào là phản hồi của khách hàng về các sản phẩm trên các trang thương mại điện tử, từ đó huấn luyện mô hình dự đoán mức độ tốt xấu của sản phẩm. Chúng tôi tiến hành thử nghiệm trên bộ dữ liệu gồm 30800 phản hồi do nhóm tự thu thập trên các trang thương mại điện tử, cài đặt và đánh giá trên nhiều mô hình học máy và học sâu khác nhau. Với sự hỗ trợ mạnh mẽ của các gói thư viện học máy có sẵn trong Spark và Spark NLP, kết quả cuối cùng đạt được khá khả quan khi accuracy lên tới 0.88.

2 Xây dựng mô hình

Pipeline



Hình 1: Pipeline bài toán

2.1 Collect data

Dữ liệu bình luận đánh giá được chúng tôi thu thập từ trang thương mại điện tử nổi tiếng Shopee Việt Nam. Bộ dữ liệu này rất đa dạng, có khoảng 30800 dòng. Mỗi dòng gồm 2 giá trị là nội dung bình luận tiếng Việt và nhãn của nó (Tốt, Ổn, Tệ).

Ví dụ: Nhận hàng đúng mô tả, tai nghe hay, giọng ấm, nên mua – Tốt

2.2 Prepare data

Dữ liệu sau khi thu thập cần được làm sạch và chuyển đổi về thành một dạng nhất quán nhằm tạo input cho mô hình huấn luyện.

- Loại bỏ các kí tự đặc biệt: '!@#$%^&*()'
- Loại bỏ các chữ số: 0-9
- Loại bỏ các khoảng trắng nhiều hơn 2
- Loại bỏ các comment rỗng
- Chuyển đổi chữ hoa thành chữ thường
- Tách từ trong câu
- Loại bỏ các stopword: chao ôi, trời ơi, bỗng nhiên, cái, thì, …

Ví dụ: Trời ơi, áo dày, vải rất đẹp, lần sau sẽ ủng hộ cho shop → [áo, dày, vải, đẹp, ủng hộ, shop]

2.3 Split data

Sau khi chuẩn bị dữ liệu xong, chúng tôi bắt đầu phân tách bộ dữ liệu 30800 dòng thành hai phần là bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra. Trong đó dữ liệu huấn luyện gồm 30200 dòng và bộ dữ liệu test gồm 600 dòng. Sau phân tách hoàn thành, chúng tôi tiến hành sử dụng bộ dữ liệu huấn luyện để huấn luyện mô hình phân loại cho bài toán của chúng tôi.

3 Cài đặt và đánh giá

3.1 Lựa chọn mô hình và phương thức rút trích đặc trưng.

Model	Features
Logistic Regression	TF-IDF
Random Forest	Count Vetorizer
Decision Tree	DistilBertEmbedding

Bảng 1: Các Phương thức được sử dụng trong bài toán

Chúng tôi thực hiện cài đăt và đánh giá trên các mô hình phân lớp phổ biến trong bài toán phân loại cảm xúc của chúng tôi, tất cả được mô tả trong bảng 2. Về vấn đề rút trích đặc trưng, chúng tôi sử dụng 3 loại chính đó là TF IDF, CountVectorizer và DistilBertEmbedding được pretrained trên tập dữ liệu tiếng Việt.

3.2 Độ đo:

Hiệu suất của hệ thống phân loại cảm xúc sẽ được đánh giá bằng các độ đo phổ biến, đó là độ chính xác (accuracy). Độ chính xác được tính như sau:

$$accuracy = \frac{số\ dự\ đoán\ đúng}{tổng\ số\ dự\ đoán}$$

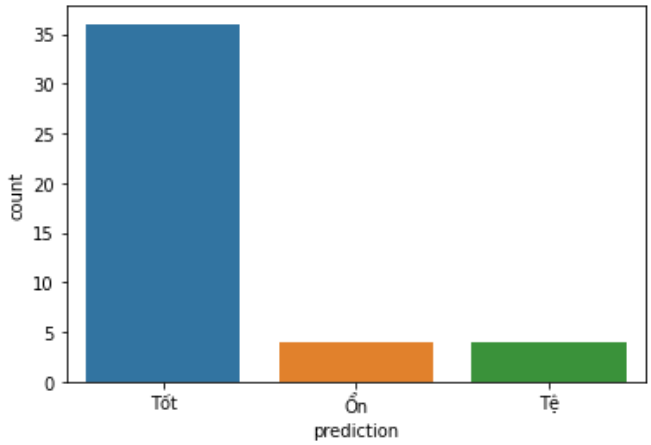
3.3 Kết quả:

Bảng 3 trình bày kết quả sau khi thực nghiệm cài đặt với nhiều mô hình phân lớp và các cách rút trích đặc trưng khác nhau.

	Logistic Regression	Decision Tree	Random Forest	Deep Neural Network
Count Vectorizer	**0.88**	0.7	0.72	-
Count Vectorizer + IDF	0.87	0.68	0.71	-
TF+ IDF	0.80	0.68	0.68	-
Embedding	-	-	-	0.79

Bảng 2: Kết quả thực nghiệm

Nhìn vào bảng trên, nhận thấy rằng với cách rút trích bằng Count Vectorizer thì tất cả các mô hình cho kết quả để vượt trội hơn các trường hợp còn lại. Trong đó, có vẻ như Logistic Regression (LogReg) tỏ ra khá hiệu quả trong bài toán lần này của chúng tôi khi độ chính xác lên tới **0.88**. Ngoài các mô hình máy học phổ biến ra, chúng tôi cũng thực hiện cài đặt một mô hình Deep learning network với đầu vào là vector wordEmbedding được trained trên tập dữ liệu tiếng Việt, không kém kệ so với LogReg khi độ chính xác đạt **0.79**. Hai mô hình máy học phổ biến còn lại như Decision Tree và Random Forest tỏ ra không mấy khả quan, cũng có thể do chúng tôi để tất cả các tham số là mặc định, về vấn đề này chúng tôi sẽ tìm hiểu thêm.



Hình 2: Ví dụ minh họa về đầu ra của hệ thống

3.4 Xây dựng hệ thống review sản phẩm

Sau khi cài đặt và so sánh kết quả giữa các mô hình với nhau, chúng tôi quyết định chọn mô hình LogReg và Count Vectorizer để xây dựng hệ thống review sản phẩm. Sử dụng gói Selenium cho python để thực hiện crawl các phản hồi của khách hàng về sản phẩm. Từ đó, tạo dataframe và đưa vào mô hình LogReg đã được đào tạo từ trước. Cuối cùng là trả về biểu đồ về mức độ tốt xấu của sản phẩm từ các phản hồi đầu vào và đưa ra đề xuất có nên mua sản phẩm đó hay không. Ví dụ minh họa được thể hiện trong hình 2.

4 Thảo luận

Trong báo cáo lần này, chúng tôi đã mô tả hoàn toàn các phương thức chúng tôi đã sử dụng để xây dựng nên một hệ thống dự đoán mức độ tốt của một sản phẩm trên các trang thương mại điện tử. Vì đây là lần đầu tiên chúng tôi nghiên cứu về xử lý ngôn ngữ tự nhiên, nên mọi thứ có vẻ còn chưa tốt cho lắm. Cùng với đó, chúng tôi sẽ tiếp tục phát triển hệ thống này thêm, ngoài việc đề xuất dựa trên phản hồi, chúng tôi sẽ thu thập ý kiến từ phía người dùng, lấy thêm thông tin về các khía cạnh khi chọn và mua một sản phẩm. Ngoài ra, chúng tôi vẫn tiếp tục nghiên cứu và cải thiện độ hiệu quả của mô hình, tìm ra các giải pháp mới hơn, mạnh hơn và tốt hơn.

References

[1] JohnSnowLabs/spark-nlp: State of the Art Natural Language Processing (github.com)

[2] Apache Spark™ - Unified Engine for large-scale data analytics

[3] apache/spark: Apache Spark - A unified analytics engine for large-scale data processing (github.com)

[4] Models - Hugging Face

[5] WebDriver | Selenium

[6] NLP-Vietnamese-progress/sentiment_analysis.md at master · undertheseanlp/NLP-Vietnamese-progress (github.com)