**THIẾT KẾ DATAWAREHOUSE BỘ DỮ LIỆU PHÂN TÍCH CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT**

Võ Văn Nam1\*, Trần Công Hùng2

*1Trường Đại học Công nghệ Thông tin và Truyền thông Việt – Hàn 2NAL Solutions*

*Tác giả liên hệ: Võ Văn Nam\*, namvv.21it@vku.udn.vn*

**TÓM TẮT**

Đề tài này tập trung vào thiết kế kho dữ liệu (Data Warehouse) cho bộ dữ liệu phân tích cảm xúc trong văn bản tiếng Việt, nhằm hỗ trợ các tổ chức trong việc khai thác và hiểu rõ cảm xúc của khách hàng qua các đánh giá. Kho dữ liệu được xây dựng theo mô hình sao (star schema), bao gồm bảng trung tâm chứa thông tin phản hồi và các bảng chiều lưu trữ chi tiết về cảm xúc, người dùng, thời gian, và sản phẩm. Quy trình ETL (Trích xuất, Chuyển đổi, Tải) được áp dụng để xử lý và chuẩn hóa dữ liệu từ các nguồn đánh giá, giúp đảm bảo dữ liệu nhất quán và chính xác. Qua đó, kho dữ liệu cung cấp nền tảng cho các phân tích OLAP, hỗ trợ việc truy vấn và trực quan hóa xu hướng cảm xúc khách hàng theo thời gian và từng danh mục sản phẩm. Kết quả nghiên cứu giúp nâng cao hiệu quả phân tích và ra quyết định trong doanh nghiệp, đồng thời góp phần hoàn thiện ứng dụng kho dữ liệu trong xử lý ngôn ngữ tự nhiên và phân tích cảm xúc.

**TỪ KHOÁ**

*Kho dữ liệu, Phân tích cảm xúc, Xử lý ngôn ngữ tự nhiên, Văn bản tiếng Việt, ETL, OLAP, Mô hình sao, Dữ liệu khách hàng, Học máy, Phân loại cảm xúc.*

**ABSTRACT**

This topic focuses on designing a data warehouse for sentiment analysis datasets in Vietnamese text, to support organizations in exploiting and understanding customer sentiment through reviews. The data warehouse is built according to the star schema model, including a central table containing feedback information and dimension tables storing details about sentiment, users, time, and products. The ETL (Extract, Transform, Load) process is applied to process and standardize data from review sources, helping to ensure data consistency and accuracy. The data warehouse provides a foundation for OLAP analysis, supporting querying and visualization of customer sentiment trends over time and by product category. The research results help improve the efficiency of analysis and decision-making in businesses while contributing to perfecting the application of data warehouses in natural language processing and sentiment analysis.

**KEYWORD**

*Data Warehouse, Sentiment Analysis, Natural Language Processing, Vietnamese Text, ETL, OLAP, Star Model, Customer Data, Machine Learning, and Sentiment Classification.*

1. **GIỚI THIỆU**

Trong bối cảnh cạnh tranh ngày càng cao, việc hiểu rõ cảm xúc và nhu cầu của khách hàng là yếu tố then chốt để các doanh nghiệp nâng cao chất lượng sản phẩm và dịch vụ. Phân tích cảm xúc từ các đánh giá khách hàng trên nền tảng trực tuyến đã trở thành một công cụ hữu ích giúp các doanh nghiệp không chỉ theo dõi phản hồi mà còn dự đoán xu hướng thị hiếu của thị trường. Tuy nhiên, việc lưu trữ và xử lý lượng lớn dữ liệu đánh giá khách hàng là một thách thức lớn, đặc biệt khi dữ liệu phải đáp ứng các yêu cầu về chất lượng, tính nhất quán và sẵn sàng cho các phân tích phức tạp. Kho dữ liệu (Data Warehouse) là một giải pháp lý tưởng cho vấn đề này, cung cấp khả năng lưu trữ và truy xuất dữ liệu một cách có cấu trúc, đáp ứng cho các phân tích chuyên sâu. Đề tài này đề xuất thiết kế một kho dữ liệu cho bộ dữ liệu phân tích cảm xúc trong văn bản tiếng Việt, giúp doanh nghiệp dễ dàng thu thập, quản lý và phân tích các đánh giá khách hàng từ nhiều nguồn. Kho dữ liệu sẽ được thiết kế theo mô hình sao với quy trình ETL (Trích xuất, Chuyển đổi, Tải) giúp đảm bảo dữ liệu luôn sẵn sàng và nhất quán cho việc khai thác. Mục tiêu của nghiên cứu là xây dựng một nền tảng kho dữ liệu mạnh mẽ cho phép các tổ chức dễ dàng phân tích và trực quan hóa dữ liệu cảm xúc, từ đó hỗ trợ ra quyết định chiến lược dựa trên dữ liệu thực tế và phản hồi của khách hàng.

# **PHƯƠNG PHÁP NGHIÊN CỨU**

## 2.1 Xây dựng bộ dữ liệu phân tích cảm xúc

Dữ liệu dùng để huấn luyện mô hình ở đây lấy từ dữ liệu IMDB - gồm 50.000 câu review phim kèm theo cảm xúc (sentiment) tích cực (positive) hoặc tiêu cực (negative). Các câu review này đã được dịch sang tiếng Việt bằng google dịch để phục vụ cho mục đích của mô hình.Bộ dữ liệu được xây dựng thành 1 file VI\_IMDB.csv. Mỗi dòng dữ liệu có 2 thành phần: review - sentiment.

Review: là các phản hồi, đánh giá sản phẩm của khách hàng.

Sentiment: Phân loại cảm xúc của khách hàng được phân loại như sau: ‘Tích cực’ là positive, ‘Tiêu cực’ là negative.

Với cấu trúc như trên, luận văn đã xây dựng được 1 bộ dữ liệu với 50.000 đánh giá. Trong đó có 25000 bình luận tích cực, và 25000 bình luận tiêu cực.

A screenshot of a computer

Description automatically generated

***Hình 1:*** *Mô tả bộ dữ liệu phân tích cảm xúc*

Thông kê dữ liệu được trình bày trong bảng sau:

|  |  |  |
| --- | --- | --- |
| **Bộ dữ liệu** | **Tích cực** | **Tiêu cực** |
| **Huấn luyện** | 20,000 | 20,000 |
| **Đánh giá** | 2,500 | 2,500 |
| **Kiểm tra** | 2,500 | 2,500 |

***Bảng 1:*** *Bộ dữ liệu chưa được xử lý*

Sau khi chúng em thực hiện việc tiền xử lý dữ liệu gồm tách từ và loại bỏ các hư từ cũng như các dấu câu không cần thiết. Chúng em thu được dữ liệu như trong bảng sau:

|  |  |  |
| --- | --- | --- |
| **Bộ dữ liệu** | **Tích cực** | **Tiêu cực** |
| **Huấn luyện** | 20,000 | 20,000 |
| **Đánh giá** | 2,500 | 2,500 |
| **Kiểm tra** | 2,500 | 2,500 |

***Bảng 2:*** *Bộ dữ liệu đã tiền xử lý*

## 2.2 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một trong những bước quan trọng nhất khi giải quyết bất kỳ bài toán nào trong lĩnh vực Học máy. Để mô hình có thể đưa ra kết quả có độ chính xác cao thì bộ dữ liệu luôn cần được xử lý, làm sạch và biến đổi trước khi trở thành dữ liệu huấn luyện cho mô hình học máy.

Đối với luận văn này, dữ liệu input đầu vào là các phản hồi, đánh giá của khách hàng về sản phẩm. Dữ liệu thường không chuẩn, vì thế ta phải tiến hành xử lý dữ liệu:

Loại bỏ các dãy html.

Loại bỏ các dấu ngoặc vuông.

Loại bỏ văn bản nhiễu.

Loại bỏ các ký tự đặc biệt.

Đưa các từ trong văn bản về từ gốc.

Loại bỏ các từ dừng trong tiếng Việt.

Tiếp theo, chúng em tiến hành xử lý bộ dữ liệu. Ở đây chúng em sẽ áp dụng thuật toán Tokenziner đây là một nhánh con trong tập xử lý ngôn ngữ tự nhiên. Tokenziner cho phép ta vectơ hóa một kho ngữ liệu văn bản, bằng cách biến mỗi văn bản thành một chuỗi các số nguyên hoặc thành một vectơ trong đó hệ số cho mỗi mã thông báo có thể là nhị phân, dựa trên số từ, dựa trên tf-idf ...

# **KẾT QUẢ NGHIÊN CỨU**

Biểu đồ và bảng mô tả Dataset: <https://colab.research.google.com/drive/1BjlTHuUTTNOoIikC1CPiiC8OMWuZeBQN#scrollTo=faTxn8uBEJdG>

Trong nghiên cứu này, kho dữ liệu đã được xây dựng và trực quan hóa nhằm cung cấp những hiểu biết sâu sắc về cảm xúc của người dùng trong các đánh giá phim bằng tiếng Việt. Việc xử lý và phân tích dữ liệu đã làm rõ các đặc điểm của từng nhóm cảm xúc (tích cực và tiêu cực) trong bộ dữ liệu, từ đó hỗ trợ các phân tích và ra quyết định dựa trên dữ liệu thực tế. Dưới đây là các kết quả phân tích chính.

**3.1 Phân phối cảm xúc**

Biểu đồ phân phối số lượng đánh giá theo cảm xúc cho thấy rằng bộ dữ liệu được xây dựng với số lượng đánh giá tích cực và tiêu cực cân bằng, với mỗi nhãn có 25,000 đánh giá. Sự cân bằng này rất quan trọng, vì nó giúp mô hình học máy khi huấn luyện không bị lệch về bất kỳ nhãn cảm xúc nào, đồng thời cải thiện độ chính xác khi phân loại cảm xúc của các đánh giá mới.

**3.1 Tích hợp và Tiền xử lý dữ liệu**

Quy trình ETL được thực hiện để đảm bảo dữ liệu đưa vào Data Warehouse là sạch và có chất lượng cao. Cụ thể:

Trích xuất dữ liệu: Dữ liệu từ bộ VI\_IMDB.csv, bao gồm 50.000 đánh giá với các nhãn cảm xúc "Tích cực" và "Tiêu cực", đã được tải vào hệ thống.

Làm sạch dữ liệu: Các yếu tố nhiễu như thẻ HTML, ký tự đặc biệt, và từ dừng được loại bỏ. Điều này giúp giảm thiểu dữ liệu thừa và tăng độ chính xác khi phân tích cảm xúc.

Tách từ và chuẩn hóa từ: Quá trình tokenization được thực hiện để chuẩn hóa văn bản, giúp mô hình hiểu rõ hơn về cấu trúc ngữ pháp và ngữ nghĩa.

Kết quả từ bước tiền xử lý được trình bày trong Bảng 1, cho thấy sự phân phối đều giữa các đánh giá tích cực và tiêu cực sau khi làm sạch.

* 1. **Đặc điểm độ dài đánh giá theo cảm xúc**

Biểu đồ hộp (Box Plot) và Biểu đồ độ dài trung bình của đánh giá theo nhãn cảm xúc cho thấy rõ sự khác biệt về độ dài của các đánh giá tích cực và tiêu cực. Cụ thể:

Các đánh giá tích cực thường có số từ trung bình nhiều hơn so với đánh giá tiêu cực. Điều này có thể được lý giải bởi việc người dùng có xu hướng diễn đạt cảm xúc tích cực một cách chi tiết hơn, trong khi những phản hồi tiêu cực thường ngắn gọn và đi thẳng vào vấn đề.

Ngoài ra, Box Plot cho thấy có một số đánh giá dài bất thường (outliers), đặc biệt là ở các đánh giá tích cực. Đây là các đánh giá đặc biệt mà người dùng có thể đưa ra các chi tiết bổ sung hoặc cảm nhận sâu sắc hơn.

**3.3 Phân phối số lượng từ trong đánh giá**

Biểu đồ phân phối số từ trong các đánh giá thể hiện rằng phần lớn các đánh giá có độ dài từ 10 đến 30 từ, với các đánh giá dài hơn thường nghiêng về phía cảm xúc tích cực. Sự phân bố này cho thấy mối liên hệ giữa độ dài của đánh giá và cảm xúc, khi các đánh giá tích cực thường có số từ trung bình cao hơn.

**3.4 Từ vựng phổ biến trong các đánh giá tích cực và tiêu cực**

Biểu đồ thanh về từ vựng phổ biến và Word Cloud giúp minh họa các từ khóa đặc trưng cho từng nhóm cảm xúc:

Trong các đánh giá tích cực, các từ như *“tuyệt vời”*, *“thích”*, *“hay”* thường xuất hiện với tần suất cao, cho thấy người dùng thường bày tỏ sự hài lòng và ngưỡng mộ đối với sản phẩm hoặc dịch vụ.

Ngược lại, các đánh giá tiêu cực thường sử dụng từ như *“thất vọng”*, *“không thích”*, *“kém”* để bày tỏ sự không hài lòng. Điều này phù hợp với việc người dùng có xu hướng dùng từ ngữ cụ thể và trực tiếp khi phản hồi tiêu cực.

**3.5 Tương quan giữa độ dài đánh giá và cảm xúc**

Biểu đồ tương quan giữa độ dài đánh giá và cảm xúc (Violin Plot) cho thấy độ dài của các đánh giá có liên quan đến nhãn cảm xúc. Các đánh giá tích cực có xu hướng dài hơn và có độ phân tán lớn hơn, cho thấy sự đa dạng trong cách diễn đạt của người dùng khi phản hồi tích cực.

# **KẾT LUẬN**

Trong nghiên cứu này, chúng tôi đã xây dựng thành công một kho dữ liệu (Data Warehouse) dành cho phân tích cảm xúc trong văn bản tiếng Việt, với dữ liệu từ các đánh giá phim IMDB đã được dịch và làm sạch. Việc thiết kế Data Warehouse theo mô hình sao (Star Schema) cùng quy trình ETL chi tiết đã đảm bảo dữ liệu luôn sẵn sàng và nhất quán cho các phân tích phức tạp. Thông qua việc trực quan hóa dữ liệu, chúng tôi đã khám phá sâu hơn về đặc điểm của các đánh giá tích cực và tiêu cực, từ đó rút ra nhiều hiểu biết có giá trị về hành vi và xu hướng cảm xúc của người dùng.

Các biểu đồ và phân tích cho thấy sự cân bằng giữa các đánh giá tích cực và tiêu cực, đặc điểm độ dài của từng loại đánh giá, cũng như các từ vựng phổ biến đặc trưng cho từng nhãn cảm xúc. Những hiểu biết này không chỉ hỗ trợ cho các ứng dụng học máy mà còn đóng vai trò quan trọng trong việc tối ưu hóa dịch vụ và sản phẩm dựa trên phản hồi thực tế từ khách hàng. Qua đó, kho dữ liệu này đã chứng minh được tiềm năng lớn trong việc giúp các tổ chức và doanh nghiệp nắm bắt cảm xúc của người dùng một cách nhanh chóng và chính xác, hỗ trợ ra quyết định chiến lược dựa trên dữ liệu khách quan và thực tế.

# **TÀI LIỆU THAM KHẢO**

[1]. N.V. Xtankêvich, Các loại hình ngôn ngữ, Nhà xuất bản Đại học và trung học chuyên nghiệp Hà Nội.

[2]. Giáo trình “Xử lý ngôn ngữ tự nhiên”, Đinh Điền, NXB Đại học Quốc gia – HCM, năm 2006.

[3]. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[4]. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21

[5]. Binh Thanh Kieu, Son Bao Pham. “Sentiment Analysis for Vietnamese.” Proc. KSE’10, tr. 152-157, 2010.

[6]. Montoyo Andre´s, Martı´nez-Barco Patricio, Balahur Alexandra. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.

[7]. Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Syst Appl 2012;39:10283–91.

[8]. Socher, C. D. Manning, and A. Y. Ng. 2010. Learning continuous phrase representations and syntactic pars- ing with recursive neural networks. In Proceedings of the NIPS-2010 Deep Learning and Unsupervised Fea- ture Learning Workshop.

[9]. Daniel Jurafsky and James H. Martin. “Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.” 1st edition, 2000.

[10]. Binh Thanh Kieu, Son Bao Pham. “Sentiment Analysis for Vietnamese.” Proc. KSE’10, tr. 152-157, 2010.

[11]. Nguyen Thi Duyen, Ngo Xuan Bach, and Tu Minh Phuong, “An Empirical

Study on Sentiment Analysis for Vietnamese.” Proc. ATC’14, tr. 309-314, 2014.

[12]. Bo Pang và Lillian Lee. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” Proc. ACL’04, tr. 271- 278, 2008.

[13]. Le Anh Cuong, Nguyen Thi Minh Huyen, and Nguyen Viet Hung. “Report on Sentiment Analysis Evaluation Campaign: Data and Systems.” Proc. VLSP 2016, 2016.

[14]. Vi Ngo Van, Minh Hoang Van, and Tam Nguyen Thanh. “Sentiment Analysis for Vietnamese using Support Vector Machines with application to Facebook comments.” Proc. VLSP 2016.

[15]. Thien Khai Tran and Tuoi Thi Phan. “Computing Sentiment Scores of Verb Phrases for Vietnamese.” Proc. ROCLING’16, tr. 204-213, 2016.

[16]. Liu, B. (2012). Sentiment analysis and opinion mining. New York, NY: Morgan & Claypool Publishers.

[17]. Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information Fusion, 36(2017), 10-25.

[18]. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1/2), 1-135.

[19]. Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: Review and prospect. International Journal of Computer Applications, 115(9), 31-41.

[20]. Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS Quarterly, 35(3), 553-572.

[21]. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” CoRR 2013