

XÂY DỰNG CHỈ SỐ TỔNG HỢP ĐO LƯỜNG MỨC ĐỘ SỐ HÓA CỦA CÁC NGÂN HÀNG THƯƠNG MẠI VIỆT NAM BẰNG PHƯƠNG PHÁP KHAI THÁC VĂN BẢN

ĐOÀN VĂN THÀNH

Trường Đại học Nguyễn Tất Thành

Nguyen Tat Thanh University

Ngày nhận bài: 5/6/2026

Ngày sửa bài: 22/6/2026

Ngày duyệt đăng bài: 6/7/2026

Tóm tắt:

Chuyển đổi số là xu thế tất yếu và là chiến lược trọng tâm của hệ thống các ngân hàng thương mại (NHTM) Việt Nam. Tuy nhiên, việc thiếu hụt một công cụ đo lường đồng nhất và đa chiều về mức độ số hóa đang là rào cản lớn cho các nghiên cứu thực nghiệm vĩ mô. Bài viết xây dựng Chỉ số số hóa tổng hợp (Digital Banking Index - DBI) cho các NHTM Việt Nam bằng cách ứng dụng phương pháp Khai thác văn bản trên Báo cáo thường niên của 30 NHTM giai đoạn 2015 - 2025, kết hợp kỹ thuật Phân tích thành phần chính (PCA). Kết quả thực nghiệm khẳng định tính đúng đắn, khoa học của phương pháp luận, đồng thời phác họa rõ nét bức tranh chuyển đổi số phân hóa theo 3 giai đoạn tại Việt Nam. Đây là bộ dữ liệu nền tảng, đạt độ tin cậy để ứng dụng vào các mô hình kinh tế lượng đánh giá hiệu quả và rủi ro hệ thống.

Từ khóa: số hóa ngân hàng, Chỉ số số hóa tổng hợp DBI, khai thác văn bản (Text Mining), phân tích thành phần chính (PCA), ngân hàng thương mại Việt Nam.

Developing a composite index to measure the digitalization of Vietnamese commercial banks using text mining

Abstract:

This study develops a composite Digital Banking Index for Vietnamese commercial banks using text mining and principal component analysis. The dataset comprises 330 annual reports from 30 banks during 2015 - 2025. Keyword density was calculated across four dimensions: technological infrastructure, digital channels and customer experience, digital operations, and digital strategy and human resources. Principal component analysis was then applied to generate an objective composite index without subjectively assigned weights. Validation results demonstrate satisfactory reliability and confirm that the index captures the multidimensional nature of banking digitalization. The findings also identify three development stages: basic digitalization during 2015 - 2018, rapid expansion during 2019 - 2022, and deeper but increasingly differentiated digital transformation during 2023 - 2025. The index provides a reliable dataset for subsequent econometric research on banking performance and systemic risk.

Keywords: banking digitalization, Digital Banking Index, text mining, principal component analysis, Vietnamese commercial banks.

1. Đặt vấn đề

Thực hiện "Kế hoạch chuyển đổi số ngành Ngân hàng đến năm 2025, định hướng đến năm 2030" của Ngân hàng Nhà nước, các NHTM Việt Nam đã dịch chuyển mạnh mẽ từ mô hình ngân hàng truyền thống sang mô hình ngân hàng số dựa trên dữ liệu. Việc định lượng chính xác mức độ số hóa của từng định chế tài chính đóng vai trò sống còn đối với ngân hàng lẫn cơ quan quản lý. Mặc dù vậy, tổng quan tại Việt Nam cho thấy, đa số các nghiên cứu thực nghiệm hiện nay đang gặp phải hạn chế lớn về mặt phương pháp luận đo lường. Các tác giả thường sử dụng các biến đại diện đơn lẻ (Proxy variables) mang tính gián tiếp như: Tỷ lệ chi phí công nghệ thông tin trên tổng chi phí, số lượng người dùng ứng dụng di động, hoặc sử dụng biến giả (Dummy variable) bằng 0 hoặc 1 để đánh giá một ngân hàng đã số hóa hay chưa. Cách tiếp cận này chưa thể hiện tính đa chiều của quá trình chuyển đổi số, vốn là sự tổng hợp từ hạ tầng, quy trình, trải nghiệm khách hàng cho đến chiến lược nhân sự.

Để giải quyết khoảng trống phương pháp luận này, đề tài ứng dụng kỹ thuật Khai thác văn bản (Text Mining) - một nhánh tiên tiến của Trí tuệ nhân tạo và Xử lý ngôn ngữ tự nhiên (NLP) - để trực tiếp trích xuất dữ liệu định lượng mang tính khách quan từ văn bản Báo cáo thường niên. Mục tiêu xây dựng và công bố Chỉ số số hóa tổng hợp (DBI) chuẩn mực, khẳng định tính đúng đắn của phương pháp luận trước khi đưa vào các mô hình kiểm định tác động vĩ mô sâu hơn.

2. Cơ sở lý luận và phương pháp nghiên cứu

2.1. Cơ sở lý luận về khung đo lường số hóa ngân hàng

Dựa trên Lý thuyết dựa trên nguồn lực (Resource - Based View - RBV) và các thông lệ quốc tế của Ủy ban Basel, quá trình số hóa ngân hàng không phải là một sự kiện đơn lẻ mà là một năng lực chiến lược cấu trúc từ 4 nền tảng cốt lõi:

- Hạ tầng & Công nghệ nền tảng (Infrastructure): Đo lường mức độ đầu tư và ứng dụng các công nghệ cốt lõi cấu trúc nên hệ thống như: Điện toán đám mây (Cloud); Dữ liệu lớn (Big Data); AI; Blockchain và hệ thống lõi Core Banking thế hệ mới.

- Kinh phân phối & Trải nghiệm khách hàng (Customer Experience): Đo lường sự hiện diện và tương tác của ngân hàng với người dùng trên không gian số, bao gồm các ứng dụng Mobile/Internet Banking, định danh điện tử eKYC, thanh toán qua mã QR và chiến lược hợp kênh Omni-channel.

- Quy trình & Vận hành số (Operations): Đo lường mức độ tự động hóa các hoạt động nội bộ như: ứng dụng robot tự động (RPA); phê duyệt tín dụng trực tuyến; chữ ký số và hợp đồng điện tử nhằm tối ưu hóa chi phí vận hành.

- Chiến lược & Nhân sự số (Strategy & Human Resources): Đo lường tầm nhìn dài hạn của ban lãnh đạo và năng lực thích ứng của đội ngũ nhân sự thông qua văn hóa số, ngân hàng mở (Open Banking) và ngân sách đào tạo kỹ năng công nghệ.

2.2. Phương pháp nghiên cứu và dữ liệu

Thu thập dữ liệu văn bản: Đối tượng nghiên cứu là toàn bộ hệ thống các NHTM cổ phần tại Việt Nam. Mẫu nghiên cứu bao gồm 30 ngân hàng trong giai đoạn 11 năm từ 2015 đến 2025 (tổng cộng 330 quan sát bảng - Firm Year Observations). Nguồn dữ liệu là Báo cáo thường niên chính thức được công bố công khai của các ngân hàng.

Quy trình ứng dụng Text Mining trích xuất từ khóa: Tác giả xây dựng thuật toán Python để xem qua

toàn bộ văn bản của các Báo cáo thường niên. Sử dụng biểu thức chính quy (Regular Expression) để đếm chính xác tần suất xuất hiện của các cụm từ khóa thuộc 4 nền tảng lý thuyết (ví dụ: "trí tuệ nhân tạo", "ngân hàng số", "ekyc", ... nhằm loại bỏ các lỗi đếm trùng ký tự ẩn). Do độ dài của Báo cáo thường niên giữa các ngân hàng và giữa các năm có sự khác biệt rất lớn (dao động từ 80 trang đến hơn 300 trang), việc sử dụng số lượng từ khóa tuyệt đối sẽ gây ra lỗi thiên vị quy mô. Do đó, tác giả tiến hành chuẩn hóa dữ liệu bằng cách tính Mật độ từ khóa (Keyword Density) cho từng nền tảng (j) của báo cáo thường niên ngân hàng (i) trong năm (t):

$$\text{Density}_{\{ijt\}} = \frac{\{\text{từ khóa thuộc trụ cột}\}_{j\}}{\{\text{Tổng số từ trong Báo cáo thường niên}\}_{\{it\}}}$$

Phương pháp Phân tích thành phần chính (PCA) gộp chỉ số: Để tổng hợp 4 chuỗi dữ liệu mật độ từ khóa thành một chỉ số tổng hợp (DBI) duy nhất mà không phụ thuộc vào ý chí chủ quan khi tự gán trọng số, phương pháp Phân tích thành phần chính (PCA) được áp dụng. Công thức toán học của chỉ số DBI có dạng:

$$\text{DBI}_{\{it\}} = w_1 \times Z(\text{Density}_{\{i1t\}}) + w_2 \times Z(\text{Density}_{\{i2t\}}) + w_3 \times Z(\text{Density}_{\{i3t\}}) + w_4 \times Z(\text{Density}_{\{i4t\}})$$

Trong đó: ($Z(\text{Density}_{\{ijt\}})$) là giá trị chuẩn hóa của các biến mật độ; (w_j) là trọng số trích xuất từ vector riêng (Eigenvector) của thành phần chính đầu tiên có giá trị riêng (Eigenvalue) lớn hơn 1; Chỉ số DBI sau đó được chuẩn hóa về thang đo từ 0 đến 1.

3. Kết quả thực nghiệm và thảo luận

3.1. Kiểm định tính đúng đắn của phương pháp luận (PCA Validation)

Kết quả phân tích ma trận tương quan giữa 4 nền tảng mật độ từ khóa đạt hệ số tương quan dương khá cao (từ 0.65 đến 0.82), chứng minh các nền tảng cùng phản ánh chung một xu hướng chuyển đổi số.

Khi chạy mô hình PCA trên phần mềm Stata, thành phần chính đầu tiên (PC1) có Eigenvalue = 2.85 > 1 và có khả năng giải thích đến 71.25% tổng biến thể của toàn bộ dữ liệu gốc. Chỉ số thống kê KMO (Kaiser-Meyer-Olkin) đạt 0.78 > 0.5 và kiểm định bác bỏ Bartlett có ý nghĩa thống kê ở mức 1% ($p < 0.001$). Kết quả định lượng này khẳng định việc tổng hợp dữ liệu bằng phương pháp PCA là hoàn toàn có giá trị, đạt độ tin cậy và cấu trúc chỉ số DBI phản ánh đúng bản chất thực tế.

3.2. Đặc điểm phân hóa của Chỉ số DBI tại Việt Nam

Sau khi tính toán chỉ số DBI từ 0 đến 1 cho 30 ngân hàng giai đoạn 2015 - 2025, kết quả thực nghiệm phác họa rõ nét 3 giai đoạn chuyển đổi số:

- + Giai đoạn 2015 – 2018: Chỉ số DBI trung bình thấp (0.18). Tập trung Internet Banking cơ bản.
- + Giai đoạn 2019 – 2022: Chỉ số DBI bùng nổ (0.55). Động lực từ Covid19, eKYC và quét mã QR.
- + Giai đoạn 2023 – 2025: Chỉ số DBI chiều sâu (0.78). Phân hóa mạnh: Khối Tư nhân đi đầu AI/API.

Sự phân hóa theo cấu trúc sở hữu (Biến điều tiết STATE): Khối NHTM Cổ phần tư nhân quy mô lớn (như Techcombank, VPBank, MB) luôn duy trì điểm số DBI dẫn đầu hệ thống từ năm 2020 trở đi (DBI đạt mức từ 0.82 đến 0.94). Điều này chứng minh cấu trúc quản trị linh hoạt giúp khối tư nhân đón đầu công nghệ nhanh hơn. Trong khi đó, khối NHTM Nhà nước (Big 4) ban đầu có tốc độ tăng DBI chậm hơn ở giai đoạn 2015 - 2019 do các rào cản về quy trình phê duyệt đầu tư vốn công nghệ, nhưng đã có sự bứt phá mạnh mẽ ở giai đoạn 2022 - 2025 nhờ lợi thế sở hữu tệp dữ liệu khách hàng khổng lồ.

4. Kết luận và kiến nghị

4.1. Kết luận

Nghiên cứu đã xây dựng thành công Chỉ số số hóa tổng hợp (Digital Banking Index - DBI) cho hệ thống các ngân hàng thương mại (NHTM) tại Việt Nam giai đoạn 2015 - 2025 dựa trên phương pháp toán học khách quan, giải quyết triệt để các hạn chế mang tính đơn lẻ của các biến đại diện trước đây. Các kết quả kiểm định thống kê khắt khe (bao gồm hệ số KMO = 0.784, kiểm định Bartlett có ý nghĩa ở mức 1% và giá trị Eigenvalue của PC1 đạt 2.85) đều xác nhận tính đúng đắn, khoa học và độ tin cậy của phương pháp luận Khai thác văn bản (Text Mining) kết hợp Phân tích thành phần chính (PCA).

Kết quả thực nghiệm từ bộ dữ liệu không chỉ khẳng định năng lực số hóa của các ngân hàng liên tục tăng trưởng qua 3 giai đoạn mang tính bước ngoặt, mà còn phác họa rõ nét sự phân hóa sâu sắc theo cấu trúc sở hữu. Khối NHTM cổ phần tư nhân quy mô lớn cho thấy sự linh hoạt, chủ động trong việc đón đầu các công nghệ lõi (AI, API), trong khi khối NHTM Nhà nước (Big 4) thể hiện sự bứt phá mạnh mẽ ở giai đoạn sau nhờ lợi thế quy mô dữ liệu. Đóng góp thực tiễn lớn nhất của nghiên cứu là cung cấp một bộ dữ liệu bảng (Panel data) sạch, có độ tin cậy học thuật cao để sẵn sàng ứng dụng vào các mô hình hồi quy vĩ mô tiếp theo, giúp đo lường chính xác tác động của số hóa đến hiệu quả hoạt động và rủi ro hệ thống tài chính.

4.2. Hạn chế của nghiên cứu

Mặc dù đạt được những kết quả khả quan về mặt phương pháp luận, nghiên cứu không tránh khỏi một số hạn chế cốt lõi cần được nhìn nhận một cách khách quan:

- Phương pháp trích xuất văn bản: Việc tính toán mật độ từ khóa từ Báo cáo thường niên mới chỉ phản ánh mức độ cam kết, định hướng chiến lược hoặc hoạt động truyền thông của ban lãnh đạo ngân hàng. Phương pháp này chưa thể đo lường trực tiếp dòng vốn đầu tư thực tế hoặc hiệu quả vận hành thực chất của các công nghệ. Đồng thời, nghiên cứu chưa kiểm soát hoàn toàn hiện tượng "thổi phồng" thông tin công nghệ (Tech-washing) nhằm mục đích làm đẹp báo cáo của các định chế tài chính.

- Tính ngữ cảnh: Kỹ thuật đếm tần suất bằng biểu thức chính quy (Regex) chưa bóc tách được sắc thái ngữ cảnh của câu văn. Do đó, các từ khóa xuất hiện trong câu mang tính phủ định hoặc viết về một dự án công nghệ thử nghiệm thất bại vẫn được ghi nhận điểm số tương đương với các dự án triển khai thành công.

- Phạm vi mẫu và nguồn dữ liệu: Nghiên cứu chỉ giới hạn trong phạm vi 30 NHTM cổ phần. Khối các tổ chức tài chính quan trọng khác như ngân hàng nước ngoài, ngân hàng chính sách hoặc các công ty tài chính công nghệ (Fintech) chưa được đưa vào mô hình. Ngoài ra, việc chỉ phụ thuộc vào một nguồn tài liệu duy nhất là Báo cáo thường niên có thể làm bỏ sót các cập nhật công nghệ liên tục từ truyền thông hoặc đề án nội bộ.

- Tính động của trọng số: Phương pháp PCA áp dụng trong nghiên cứu này tính toán trọng số dựa trên ma trận hiệp phương sai của toàn bộ giai đoạn 11 năm. Việc này vô tình cố định hệ số đóng góp của các nền tảng, trong khi trên thực tế tầm quan trọng của từng nền tảng số hóa có thể thay đổi theo từng thời kỳ chiến lược của ngành ngân hàng.

4.3. Hướng nghiên cứu dự kiến

Để khắc phục các hạn chế nêu trên và nâng cao giá trị khoa học, các nghiên cứu trong tương lai có thể mở rộng theo các hướng sau:

- Ứng dụng các mô hình ngôn ngữ tiên tiến: Thay vì dừng lại ở việc đếm từ khóa, các nghiên cứu tiếp theo nên ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên sâu hơn (phân tích cảm tính - Sentiment Analysis, hoặc các mô hình mã hóa ngữ cảnh dựa trên BERT/GPT) để bóc tách chính xác sắc thái, bối cảnh thực hiện công nghệ của ngân hàng, từ đó kiểm soát hiệu quả hiện tượng "Tech-washing".

- Đa dạng hóa nguồn dữ liệu và đối tượng: Nghiên cứu tương lai có thể tích hợp dữ liệu từ Báo cáo thường niên với các thông cáo báo chí, thông tin từ Ngân hàng Nhà nước và dữ liệu đánh giá trải nghiệm thực tế của người dùng trên các kho ứng dụng (App Store/Google Play). Phạm vi nghiên cứu cũng cần mở rộng sang khối ngân hàng nước ngoài và các công ty Fintech để phác họa bức tranh số hóa toàn diện hơn cho ngành tài chính.

- Áp dụng trọng số động theo thời gian (Dynamic PCA): Chuyển đổi từ mô hình trọng số cố định sang tính toán trọng số động theo từng giai đoạn phát triển cụ thể của thị trường (ví dụ: giai đoạn khởi tạo hạ tầng và giai đoạn bùng nổ ứng dụng công nghệ chiều sâu) để phản ánh chính xác sự dịch chuyển tầm quan trọng của các trụ cột số hóa.

- Mở rộng các kiểm định thực nghiệm sâu hơn: Sử dụng chỉ số DBI đã xây dựng để đưa vào các mô hình kinh tế lượng nâng cao (mô hình GMM hệ thống, hồi quy Ngưỡng hoặc hồi quy bảng động) nhằm phân tích sâu sắc tác động nhân quả, phi tuyến tính của số hóa đến các chỉ số sinh lời (ROA, ROE), năng lực quản trị rủi ro hoặc mức độ lây lan rủi ro hệ thống của các ngân hàng tại Việt Nam ■

Lời cảm ơn:

Để hoàn thành bài nghiên cứu này, tác giả xin gửi lời cảm ơn sâu sắc nhất tới Thầy, Cô khoa Tài chính - Kế toán Trường Đại học Nguyễn Tất Thành đã tận tình hỗ trợ đưa ra những định hướng quý báu và tri ân tới gia đình, đồng nghiệp, bạn bè đã đồng hành, khích lệ hoàn thành nghiên cứu.

TÀI LIỆU THAM KHẢO:

Ngân hàng Nhà nước Việt Nam (2021). Kế hoạch chuyển đổi số ngành Ngân hàng đến năm 2025, định hướng đến năm 2030.

Adrian, T., & Brunnermeier, M. K. (2016). CoVaR. *American Economic Review*, 106(7), 1705-1741.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in profit panel data models. *Journal of Econometrics*, 87(1), 115-143.

Scott, S. V., Van Reenen, J., & Zachariadis, M. (2017). The long-term effect of digital innovation on bank performance: An empirical study of SWIFT adoption in financial services. *Research Policy*, 46(5), 984-1004.