TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN HOÀNG KHÁNH DUY - 52200275
PHAN VĂN QUỐC DINH - 52200209

PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA VĂN HOÁ NƯỚC NGOÀI ĐẾN LỜI BÀI HÁT TIẾNG VIỆT

DỰ ÁN CÔNG NGHỆ THÔNG TIN
KHOA HỌC MÁY TÍNH

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2026

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN HOÀNG KHÁNH DUY-52200275
PHAN VĂN QUỐC DINH-52200209

PHÂN TÍCH SỰ ẢNH HƯỞNG CỦA VĂN HOÁ NƯỚC NGOÀI ĐẾN LỜI BÀI HÁT TIẾNG VIỆT

DỰ ÁN CÔNG NGHỆ THÔNG TIN

KHOA HỌC MÁY TÍNH

Người hướng dẫn
Ths NGUYỄN QUỐC BÌNH

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2026

LỜI CẢM ƠN

Trong quá trình thực hiện dự án Công nghệ thông tin, chúng em đã được Ths. Nguyễn Quốc Bình hỗ trợ về mặt kiến thức và phương pháp nghiên cứu. Thầy đã giúp chúng em hiểu rõ hơn về yêu cầu của Dự án và đưa ra những ý kiến xây dựng giúp chúng em nâng cao chất lượng Dự án của chúng em. Ngoài ra, Thầy Bình còn tạo điều kiện để em có thể thực hiện Dự án này một cách hiệu quả hơn.

Chúng em rất cảm kích và biết ơn thầy trong quá trình thực hiện dự án Công nghệ thông tin. Thầy đã rất tận tình, tâm huyết giúp đỡ và hướng dẫn chúng em trong suốt quá trình làm dự án, từ đó giúp chúng em hiểu rõ hơn về những khái niệm cơ bản và tạo ra được một cách suy nghĩ toán học sâu sắc.

Và đặc biệt hơn là chúng em cũng muốn bày tỏ lòng biết ơn đến Trường Đại học Tôn Đức Thắng đã cung cấp cho chúng em một môi trường học tập và nghiên cứu chuyên nghiệp, đồng thời hỗ trợ chúng em về các tài liệu nghiên cứu để thực hiện Dự án.

TP. Hồ Chí Minh, ngày ... tháng ... năm 20..
Tác giả
(Ký tên và ghi rõ họ tên)

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Ths. Nguyễn Quốc Bình. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

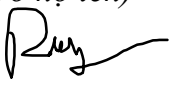
Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày ... tháng ... năm 20..

Tác giả

(Ký tên và ghi rõ họ tên)



Nguyễn Hoàng Khánh Duy

Phan Văn Quốc Dinh

TÓM TẮT

Dự án này với mục đích định lượng và mô tả các biểu hiện ảnh hưởng ngôn ngữ nước ngoài trong lời bài hát tiếng Việt theo thời gian, trên cơ sở tiếp cận khoa học dữ liệu và xử lý ngôn ngữ tự nhiên. Trong bối cảnh lời bài hát được đăng tải phân tán trên nhiều nền tảng, thiếu chuẩn biểu diễn thống nhất và thường không đầy đủ siêu dữ liệu, đề tài đặt trọng tâm vào việc xây dựng một quy trình xử lý dữ liệu có thể tái lập, có khả năng truy vết nguồn và cho phép kiểm chứng các nhận định theo hướng định lượng, thay vì dừng ở nhận xét cảm tính.

Kho dữ liệu lời bài hát tiếng Việt dùng chung của nhóm, được thu thập đa nguồn và hợp nhất theo lược đồ thống nhất. Tập dữ liệu bao gồm cả ca khúc gốc do tác giả Việt Nam sáng tác và ca khúc dịch, chuyển ngữ hoặc thu lại có xuất xứ từ nước ngoài nhưng phổ biến trong không gian âm nhạc Việt Nam. Mỗi bản ghi được tổ chức theo cấu trúc bảng với các trường cốt lõi phục vụ phân tích như tiêu đề, tác giả, lời bài hát, năm, đường dẫn tham chiếu, nhãn nguồn gốc và ghi chú. Nghiên cứu chỉ xử lý phần lời dưới dạng văn bản và không phân tích các yếu tố âm nhạc như âm thanh, hoà âm, giai điệu.

Trên nền dữ liệu đã thống nhất, nhóm triển khai pipeline phân tích theo hướng tách bước và lưu vết trung gian để đảm bảo khả năng tái lập. Ở bước tiền xử lý, lời bài hát được chuẩn hoá theo thư viện teencode, xử lý một số ký tự dễ nhầm và loại nhiễu theo danh sách noise, đồng thời chuẩn hoá khoảng trắng nhằm giảm sai lệch khi so khớp từ vựng. Ở bước xây dựng tri thức gán nhãn, nhóm tạo hệ thư viện gồm tiếng Anh mở rộng, Hán Việt, tên riêng và phiên âm nước ngoài. Các thư viện được tổng hợp từ nguồn công khai và từ kết quả trích xuất bằng mô hình ngôn ngữ ở giai đoạn chuẩn bị, sau đó được làm sạch chéo nhằm giảm chồng lấn giữa các nhóm nhãn.

Khâu trọng tâm của nghiên cứu là gán nhãn thành phần ngôn ngữ trong lời bài hát theo thứ tự ưu tiên để hạn chế nhầm lẫn. Pipeline gán nhãn được tổ chức theo trình tự: phiên âm nước ngoài, tên riêng, Hán Việt, tiếng Việt phổ thông và tiếng Anh.

Đối với phân tích theo thời gian, dữ liệu được gán giai đoạn dựa trên năm phát hành theo các khoảng 1990-2000, 2000-2010, 2010-2015, 2015-2020 và 2020-2025. Các bản ghi không có năm hợp lệ được tách khỏi phân tích theo giai đoạn để tránh tạo sai lệch. Kết quả thống kê cho thấy tập dữ liệu có năm hợp lệ phân bố không đồng đều giữa các giai đoạn; riêng giai đoạn 2020-2025 có số bản ghi lớn nhất. Trên cơ sở đó, nhóm tổng hợp các tỷ lệ ngôn ngữ theo giai đoạn và trực quan hoá để quan sát xu hướng.

Tổng hợp lại, nghiên cứu đóng góp một quy trình xử lý dữ liệu lời bài hát có thể tái lập, có cơ chế xây dựng và làm sạch thư viện gán nhãn, có bước gán nhãn theo thứ tự ưu tiên và có hệ đầu ra phục vụ tổng hợp theo thời gian và theo nhóm. Các kết quả định lượng và trực quan hoá cho phép mô tả xu hướng xuất hiện yếu tố tiếng Anh và phiên âm theo thời gian, cũng như sự khác biệt theo thể loại và tác giả, qua đó đáp ứng mục tiêu của đề tài theo hướng khách quan và kiểm chứng được. Bên cạnh đó, báo cáo cũng ghi nhận các hạn chế chính liên quan đến độ đầy đủ của trường năm, khả năng bao phủ của thư viện gán nhãn và phần token chưa gán nhãn, đồng thời đề xuất hướng phát triển như mở rộng dữ liệu, tinh chỉnh thư viện và bổ sung kiểm định thống kê cho các khác biệt quan sát được.

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

CHƯƠNG 1: GIỚI THIỆU

1.1 Lý do chọn đề tài

Âm nhạc có thể được xem là sự phản chiếu sinh động của đời sống văn hoá - xã hội. Trong bối cảnh toàn cầu hoá cùng sự bùng nổ của truyền thông số, các dòng chảy văn hoá quốc tế (US-UK, Hàn Quốc/K-pop, Trung Quốc, …) ngày càng để lại dấu ấn rõ rệt trên thị trường âm nhạc Việt Nam. Tác động này thể hiện trực diện trong lời bài hát, thông qua lớp từ vựng vay mượn hoặc phiên âm, hiện tượng trộn mã Việt-Anh (code switching), biến đổi cấu trúc câu, cũng như hệ thống chủ đề và ẩn dụ mang màu sắc ngoại lai. Tuy vậy, phần lớn các nhận định hiện nay vẫn nghiêng về cảm tính, dựa trên trải nghiệm cá nhân hoặc một số trường hợp tiêu biểu, vì thế chưa hình thành được một bức tranh tổng thể mang tính định lượng, có hệ thống và có thể theo dõi biến thiên theo thời gian.

Sự phát triển của khoa học dữ liệu (Data Science) và xử lý ngôn ngữ tự nhiên (NLP) tạo tiền đề để tiếp cận vấn đề theo hướng dữ liệu lớn, mô hình/thuật toán, và các thước đo có khả năng kiểm chứng. Cách tiếp cận này cho phép lượng hóa tương đối khách quan mức độ ảnh hưởng văn hóa thể hiện trong lời hát, theo dõi xu hướng biến đổi qua các giai đoạn, đồng thời đối chiếu mối tương quan giữa các nguồn tác động khác nhau. Tuy nhiên, bài toán vẫn đặt ra nhiều thách thức phương pháp luận và kỹ thuật. Thứ nhất, dữ liệu lời bài hát thường phân tán trên nhiều nền tảng, làm gia tăng chi phí chuẩn hóa và đối sánh. Thứ hai, quá trình thu thập chịu ảnh hưởng bởi ràng buộc bản quyền và các giới hạn truy cập theo robots.txt, dẫn đến nguy cơ thiếu hụt hoặc thiên lệch mẫu. Thứ ba, thông tin mốc thời gian (năm phát hành, thời điểm phổ biến, phiên bản phát hành lại) thường không đồng nhất, gây khó khăn cho phân tích theo chuỗi thời gian. Cuối cùng, tiếng Việt có các đặc trưng ngôn ngữ quan trọng-dấu thanh, tách từ, biến thể chính tả, và hiện tượng từ lóng/viết tắt-đòi hỏi quy trình tiền xử lý.

Đề tài **"Phân tích sự ảnh hưởng của văn hóa nước ngoài đến lời bài hát tiếng Việt"** được lựa chọn với kỳ vọng tạo ra đóng góp theo hai hướng. Về học thuật, nghiên cứu hướng đến việc đề xuất một khung đo lường cùng hệ chỉ số nhằm nhận diện và lượng hóa các biểu hiện ảnh hưởng văn hóa ngoại lai trong lời ca tiếng Việt. Về thực tiễn, kết quả nghiên cứu được kỳ vọng cung cấp cơ sở tham chiếu phục vụ hoạt động

sáng tác, sản xuất âm nhạc, cũng như hỗ trợ công tác truyền thông-giáo dục và các nghiên cứu liên quan đến văn hóa đại chúng.

1.2 Mục đích của đề tài

Đề tài hướng đến việc thiết lập một khung phân tích định lượng dựa trên dữ liệu lớn và các kỹ thuật NLP, nhằm mô tả và lý giải một cách khách quan mức độ cũng như hình thức ảnh hưởng của văn hoá nước ngoài lên lời bài hát tiếng Việt theo thời gian. Qua đó, nghiên cứu không chỉ dừng ở câu hỏi "ảnh hưởng nhiều hay ít", mà còn làm rõ ảnh hưởng thể hiện ở những vị trí nào trong văn bản lời hát, biểu hiện qua các dấu hiệu ngôn ngữ cụ thể nào, và được rút ra từ một bộ dữ liệu có cấu trúc minh bạch, có khả năng tái sử dụng cho các nghiên cứu tiếp theo.

Để hiện thực hoá mục tiêu tổng quát, nghiên cứu tập trung vào các mục tiêu cụ thể theo các nhóm nhiệm vụ sau:

Thứ hai, nghiên cứu đã tiến hành chuẩn hoá và làm sạch văn bản lời bài hát nhằm giảm nhiễu và gia tăng tính nhất quán của dữ liệu phục vụ phân tích. Các thao tác chính bao gồm: loại bỏ những ký tự, đoạn không liên quan đến nội dung lời hát; xử lý các biến thể thường gặp của từ lóng và viết tắt; chuẩn hoá chính tả ở mức hợp lý để hạn chế sai lệch mà vẫn giữ được đặc trưng biểu đạt; đồng thời tách từ và chuyển đổi dữ liệu sang định dạng phù hợp cho các bài toán NLP.

Thứ ba, nghiên cứu đã đề xuất và xây dựng một hệ chỉ số đo lường mức độ ảnh hưởng văn hoá theo nhiều lớp đặc trưng ngôn ngữ. Ở lớp từ vựng, hệ chỉ số tập trung vào việc thống kê tần suất và tỷ lệ xuất hiện của từ vay mượn, cụm từ ngoại lai, cũng như các đoạn trộn mã Việt-Anh trong lời bài hát. Ở lớp ngữ nghĩa, nghiên cứu đặt trọng tâm vào việc khai thác các mô hình nhúng từ (ví dụ Word2Vec hoặc FastText) để đối chiếu quan hệ ngữ nghĩa giữa từ tiếng Việt và các yếu tố ngoại lai, qua đó nhận diện cụm nghĩa và trường liên tưởng mang dấu ấn ảnh hưởng văn hoá.

Thứ tư, nghiên cứu đã triển khai phân tích xu hướng theo thời gian và theo nhóm bài hát để xác định sự biến thiên của từng dòng ảnh hưởng qua các giai đoạn. Trên cơ sở đó, phân tích cho phép so sánh khác biệt giữa nhóm bài hát thiên về "thuần Việt" và nhóm bài hát chuyển ngữ hoặc chịu ảnh hưởng mạnh từ nước ngoài; đồng thời đối chiếu

mối liên hệ giữa mức độ ảnh hưởng văn hoá với thể loại, thời kỳ phát hành và các đặc trưng ngôn ngữ thể hiện trong lời hát.

Cuối cùng, nghiên cứu đã trực quan hoá kết quả và tổng hợp hàm ý thông qua các biểu đồ xu hướng theo thời gian, thống kê tần suất và các biểu diễn ngữ nghĩa. Từ các kết quả này, nghiên cứu đã rút ra các kết luận bước đầu, đồng thời đề xuất gợi ý ứng dụng và xác lập hướng phát triển tiếp theo, trong đó bao gồm khả năng vận dụng học máy để suy đoán nguồn ảnh hưởng, năm sáng tác hoặc thể loại cho những bài hát còn thiếu thông tin mô tả.

1.3 Phát biểu đề tài

Đề tài được phát biểu dưới dạng một bài toán phân tích định lượng trên dữ liệu quy mô lớn như sau: Định lượng và phân tích các biểu hiện ảnh hưởng của văn hoá nước ngoài trong lời bài hát tiếng Việt theo các giai đoạn thời gian, dựa trên kho dữ liệu lời bài hát được thu thập từ nhiều nguồn bằng mã thu thập tự động (web scraper/script) và được xử lý bằng các kỹ thuật của khoa học dữ liệu và xử lý ngôn ngữ tự nhiên.

Theo hướng tiếp cận này, nghiên cứu không dừng ở các nhận xét mang tính cảm tính (ví dụ: “nhạc Việt gần đây giống K-pop/US-UK hơn”), mà chuyển các nhận xét đó thành các biến quan sát và thước đo có thể kiểm chứng. Cụ thể, dấu hiệu ảnh hưởng được mô hình hoá dưới dạng chỉ số định lượng và đặc trưng ngôn ngữ (từ vựng/cụm từ ngoại lai, hiện tượng trộn mã Việt-Anh, mô thức diễn đạt và các đặc trưng liên quan đến ngữ nghĩa…), sau đó được kiểm tra trên một tập dữ liệu có cấu trúc minh bạch và có khả năng tái sử dụng.

Về dữ liệu và kiểm soát chất lượng, phân tích được triển khai trên bộ dữ liệu đã được chuẩn hoá theo lược đồ thống nhất (title, composers, lyricists, year, genres, lyrics, urls, source, note). Trong đó, cột source được sử dụng như nhãn nguồn gốc/loại hình bài hát (ví dụ: gốc Việt, dịch/chuyển ngữ, cover hoặc trường hợp chịu ảnh hưởng trực tiếp theo nguồn ghi nhận), nhằm phục vụ các phép đối chiếu theo nhóm. Nhóm nghiên cứu không sử dụng AI để bổ sung trường năm phát hành (year); thay vào đó, thông tin năm được ưu tiên thu thập/đối soát trực tiếp từ các nguồn tin cậy thông qua mã, và các trường hợp thiếu hoặc chưa thống nhất được ghi nhận trong note để đảm bảo tính truy vết.

Từ phát biểu đề tài, nhóm nghiên cứu tập trung giải quyết bốn nhóm câu hỏi sau:

Thứ nhất, nhóm câu hỏi về mức độ ảnh hưởng. Nghiên cứu xem xét các yếu tố ngôn ngữ mang sắc thái văn hoá nước ngoài-bao gồm từ vựng/cụm từ ngoại lai, hiện tượng trộn mã Việt-Anh, và một số mô thức diễn đạt gợi phong cách (US-UK, K-pop, Trung Quốc…)-xuất hiện với tần suất và mức độ phổ biến như thế nào trong lời bài hát tiếng Việt.

Thứ hai, nhóm câu hỏi về biến đổi theo thời gian. Nghiên cứu phân tích mức độ và hình thức ảnh hưởng biến thiên theo các giai đoạn, đồng thời kiểm tra xem liệu có giai đoạn nào ghi nhận sự gia tăng rõ rệt của một dòng ảnh hưởng cụ thể (ví dụ K-pop hoặc US-UK).

Thứ ba, nhóm câu hỏi về khác biệt theo nguồn gốc và thể loại. Trên cơ sở nhãn source, nghiên cứu đối chiếu khác biệt giữa nhóm bài hát gốc Việt và các nhóm bài hát dịch/chuyển ngữ/cover hoặc chịu ảnh hưởng trực tiếp theo ghi nhận. Đồng thời, nghiên cứu xem xét các thể loại khác nhau (tiền chiến, bolero, nhạc trẻ, rap/hip-hop, ballad, OST…) có xu hướng sử dụng yếu tố ngoại lai theo những cách thức nào, và mức độ khác biệt đó có thể được kiểm định bằng các tiêu chí thống kê phù hợp.

Thứ tư, nhóm câu hỏi về mối liên hệ giữa đặc trưng ngôn ngữ và dòng ảnh hưởng văn hoá. Nghiên cứu tìm kiếm những đặc trưng ngôn ngữ có khả năng đóng vai trò chỉ báo cho từng dòng ảnh hưởng (US-UK, Hàn Quốc, Trung Quốc…), chẳng hạn nhóm từ khoá, kiểu trộn mã, cụm từ vay mượn, mô hình lặp/nhịp điệu câu, hoặc các quan hệ ngữ nghĩa rút trích từ biểu diễn vector (khi áp dụng mô hình nhúng từ).

1.4 Phạm vi chọn đề tài

Để bảo đảm đề tài đáp ứng yêu cầu học phần, phù hợp với thời lượng triển khai và nguồn lực của nhóm, đồng thời vẫn đủ cơ sở áp dụng các kỹ thuật khoa học dữ liệu và xử lý ngôn ngữ tự nhiên, phạm vi nghiên cứu được giới hạn theo hai khía cạnh: (i) phạm vi dữ liệu và (ii) phạm vi hệ thống-kỹ thuật triển khai. Việc giới hạn phạm vi giúp kiểm soát chất lượng dữ liệu, giảm rủi ro phát sinh trong quá trình thu thập và xử lý, và duy trì tính khả thi của nghiên cứu.

1.4.1 Phạm vi dữ liệu

Đối tượng dữ liệu. Nghiên cứu giới hạn phạm vi ở văn bản lời bài hát tiếng Việt thu thập từ các nguồn trực tuyến. Tập dữ liệu bao gồm hai nhóm: (1) các ca khúc gốc do tác giả Việt Nam sáng tác và (2) các ca khúc dịch/chuyển ngữ hoặc cover có xuất xứ

từ nước ngoài nhưng được phổ biến trong không gian âm nhạc Việt Nam. Đề tài không phân tích các thành tố âm nhạc (giai điệu, hoà âm, phối khí); toàn bộ xử lý được thực hiện trên phần lời ca dưới dạng văn bản, phù hợp với hướng tiếp cận bằng các kỹ thuật NLP.

Phạm vi thời gian. Nghiên cứu ưu tiên giai đoạn 1990-2025 nhằm phản ánh các làn sóng giao lưu văn hoá trong bối cảnh truyền thông số phát triển và phù hợp với mức độ sẵn có của dữ liệu lời bài hát trên web. Phạm vi thời gian có thể được điều chỉnh theo chất lượng metadata thu thập được để bảo đảm tính nhất quán của tập dữ liệu; các phân tích theo thời gian được tổ chức theo giai đoạn thay vì phụ thuộc hoàn toàn vào từng năm.

Tổ chức dữ liệu và nhãn nguồn gốc. Kho dữ liệu được chuẩn hoá theo cấu trúc bảng, với các trường thông tin chính phục vụ phân tích gồm: tên bài hát, nhạc sĩ, người viết lời, năm, thể loại, nội dung lời, đường dẫn tham chiếu, nhãn nguồn gốc và ghi chú. Trong đó, trường source được sử dụng làm nhãn nguồn gốc theo quy ước thống nhất: nếu bài hát do tác giả Việt Nam sáng tác thì ghi nhận Việt Nam; nếu thuộc nhóm dịch/chuyển ngữ/cover thì ghi nhận quốc gia/không gian văn hoá gốc khi có đủ căn cứ.

1.4.2 Phạm vi kỹ thuật và yêu cầu triển khai

Về phạm vi kỹ thuật, đề tài được triển khai theo hướng xử lý dữ liệu văn bản và phân tích định lượng trên máy tính cá nhân. Nhóm sử dụng Python làm ngôn ngữ chính, tổ chức quy trình theo dạng pipeline gồm các bước: chuẩn hoá và làm sạch dữ liệu, xây dựng thư viện phục vụ gán nhãn, gán nhãn thành phần ngôn ngữ trong lời bài hát, trích xuất đặc trưng và tổng hợp thống kê để trực quan hoá theo thời gian và theo nhóm.

Ở lớp xử lý dữ liệu, nhóm sử dụng các thư viện thao tác bảng và tính toán cơ bản như pandas và numpy, kết hợp biểu thức chính quy để chuẩn hoá văn bản, xử lý nhiễu và nhận diện một số mẫu ký tự đặc thù trong lyrics. Đối với bài toán gán nhãn từ vựng, nhóm triển khai cơ chế so khớp dựa trên hệ thư viện tự xây và các từ điển tham chiếu, đồng thời sử dụng công cụ từ điển tiếng Anh và tiếng Việt chuẩn ở mức phù hợp với dữ liệu. Nhóm có thử nghiệm mô hình chạy cục bộ nhưng không sử dụng cho quy trình chính do chất lượng đầu ra không đáp ứng yêu cầu.

Về trực quan hoá và báo cáo kết quả, nhóm sử dụng matplotlib để tạo các biểu đồ xu hướng theo giai đoạn, biểu đồ tần suất và các biểu đồ so sánh theo nhóm như

nguồn gốc, thể loại và nhạc sĩ. Các bảng tổng hợp được xuất ra định dạng CSV và được đóng gói thêm ở dạng Excel nhiều trang để thuận tiện kiểm tra, đối chiếu và phục vụ trình bày trong báo cáo.

1.5 Phương pháp nghiên cứu

Đề tài áp dụng phương pháp nghiên cứu thực nghiệm dựa trên dữ liệu, kết hợp thu thập dữ liệu từ web, tiền xử lý văn bản và các kỹ thuật xử lý ngôn ngữ tự nhiên để phân tích lời bài hát tiếng Việt. Quy trình nghiên cứu được thiết kế theo hướng có thể tái lập, gồm ba giai đoạn chính: (i) thu thập và tiền xử lý dữ liệu, (ii) phân tích dữ liệu, và (iii) đánh giá - trực quan hoá kết quả.

Trước hết, ở giai đoạn thu thập và tiền xử lý, nhóm xây dựng mã thu thập tự động để trích xuất lời bài hát và metadata từ các nguồn trực tuyến, sau đó chuẩn hoá dữ liệu theo cấu trúc bảng thống nhất gồm các trường: title, composers, lyricists, year, genres, lyrics, urls, source, note. Trong bước chuẩn hoá metadata, hai trường composers và lyricists được xử lý nhằm tăng tính nhất quán theo nguyên tắc bảo toàn thông tin: nếu một cột trống thì dùng giá trị của cột còn lại để điền; nếu cả hai cột đều có dữ liệu nhưng khác nhau thì gộp và ghi lại để giữ đầy đủ thông tin; nếu cả hai cột đều trống thì giữ nguyên. Đối với văn bản lời bài hát, nhóm thực hiện làm sạch bằng regex để xử lý dấu câu/ký tự và phát hiện các mẫu cần thiết cho các bước phân tích. Việc tách từ được triển khai ở mức đơn giản bằng phép tách theo khoảng trắng, phù hợp với đặc thù lời bài hát thường đã có khoảng trắng phân tách; nhóm không sử dụng các thư viện tách từ tiếng Việt như VnCoreNLP/underthesea/pyvi. Sau khi tách token, nhóm thực hiện đối sánh theo từ điển kết hợp regex để gán nhãn một số nhóm từ phục vụ phân tích, bao gồm tiếng Anh, từ Hán-Việt, tên riêng và một số dạng phiên âm.

Tiếp theo, ở giai đoạn phân tích dữ liệu, nhóm trích xuất và tính toán hệ chỉ số định lượng nhằm đo lường mức độ và hình thức ảnh hưởng thông qua các đặc trưng ngôn ngữ trong lời bài hát. Nhóm chỉ số trọng tâm về code-mixing gồm: Code-Mixing Index và Language Diversity. Bên cạnh đó, nhóm tính các chỉ số phân bố ngôn ngữ như tỷ lệ Vietnamese, Hán-Việt, English, ngoại ngữ khác và Proper Nouns. Trên cơ sở các chỉ số này, nhóm xây dựng thêm hai nhãn phân loại phục vụ so sánh và tổng hợp: Mixing Intensity (7 mức theo ngưỡng tỷ lệ tiếng Việt) và Style Category (5 nhóm kiểu phong cách).

Cuối cùng, ở giai đoạn đánh giá - trực quan hoá, các chỉ số được tổng hợp và so sánh theo chiều thời gian bằng cách phân nhóm năm phát hành thành 5 giai đoạn: 1990-2000, 2000-2010, 2010-2015, 2015-2020, 2020-2025. Kết quả được trình bày dưới dạng thống kê mô tả và các biểu đồ trực quan theo giai đoạn nhằm quan sát xu hướng biến đổi. Trong toàn bộ quy trình, nhóm thừa nhận hiện tượng thiếu và mâu thuẫn metadata là phổ biến khi thu thập từ web; các trường hợp này được giữ trống và ghi nhận trong note để phục vụ truy vết.

1.6 Kết cấu báo cáo

Báo cáo được tổ chức thành năm chương theo đúng tiến trình triển khai của đề tài, từ đặt vấn đề, xây dựng dữ liệu, thiết kế quy trình phân tích, đến trình bày kết quả và tổng kết. Cụ thể như sau.

Chương 1 giới thiệu đề tài, nêu lý do chọn đề tài, mục tiêu nghiên cứu và cách phát biểu bài toán theo hướng định lượng. Chương này cũng xác định phạm vi dữ liệu và phạm vi kỹ thuật, trình bày phương pháp nghiên cứu, đồng thời định hướng cách tiếp cận toàn bộ báo cáo.

Chương 2 trình bày giai đoạn xây dựng kho dữ liệu. Nội dung tập trung vào mục tiêu và vai trò của dữ liệu đầu vào, yêu cầu chất lượng của kho dữ liệu dùng chung, quy trình thu thập theo nguồn được phân công, và các bước chuẩn hoá - gộp dữ liệu. Chương này cũng mô tả cách xử lý dữ liệu thiếu hoặc mâu thuẫn metadata, cách bổ sung các trường quan trọng như năm và đường dẫn tham chiếu, tạo nền dữ liệu ổn định cho các bước phân tích.

Chương 3 mô tả chi tiết quy trình phân tích dữ liệu mà nhóm triển khai trên tập dữ liệu dùng chung. Chương này bao gồm tổ chức pipeline theo thư mục, xây dựng và làm sạch hệ thư viện gán nhãn, tiền xử lý lời bài hát, gán nhãn thành phần ngôn ngữ theo thứ tự ưu tiên, trích xuất đặc trưng và tạo các bảng tổng hợp theo thời gian và theo nhóm. Cuối chương là phần đóng gói đầu ra phục vụ kiểm tra và viết báo cáo.

Chương 4 trình bày kết quả và thảo luận dựa trên các bảng tổng hợp và biểu đồ trực quan hoá từ Chương 3. Nội dung gồm tổng quan dữ liệu phân tích, xu hướng theo thời gian của thành phần ngôn ngữ, thực nghiệm bổ sung theo năm về tỷ lệ bài có yếu tố tiếng Anh, so sánh theo thể loại và theo nhạc sĩ, và phân tích các nhóm phiên âm theo

ngôn ngữ. Các kết quả được diễn giải bám sát dữ liệu và phục vụ trực tiếp cho mục tiêu định lượng ảnh hưởng ngôn ngữ nước ngoài trong lời bài hát.

Chương 5 tổng kết nghiên cứu, khái quát các kết luận chính, nêu đóng góp của đề tài, chỉ ra các hạn chế trong dữ liệu và phương pháp, đồng thời đề xuất hướng phát triển để mở rộng và cải thiện nghiên cứu trong tương lai.

CHƯƠNG 2: XÂY DỰNG KHO DỮ LIỆU

2.1 Mục tiêu và vai trò của giai đoạn xây dựng dữ liệu

Trong các bài toán phân tích định lượng dựa trên khoa học dữ liệu và xử lý ngôn ngữ tự nhiên, tập dữ liệu đầu vào là nền tảng quyết định mức độ tin cậy của kết quả. Với đặc thù lời bài hát tiếng Việt được đăng tải phân tán trên nhiều website, thiếu chuẩn biểu diễn thống nhất, dễ phát sinh trùng lặp, đồng thời thường không đầy đủ siêu dữ liệu, giai đoạn xây dựng kho dữ liệu được tổ chức như một quy trình có kiểm soát, trong đó nhấn mạnh hai yêu cầu trọng tâm: truy vết nguồn và xử lý nhất quán.

Kho dữ liệu dùng chung của lớp được xây dựng trong tháng 10/2025 và sau khi gộp đạt quy mô xấp xỉ 60.000 bản ghi. Trong kho dữ liệu này, bản ghi là đơn vị thao tác chính: cùng một tiêu đề có thể xuất hiện ở nhiều biến thể khác nhau (remix, live, cover, chuyển ngữ), hoặc trùng tên nhưng khác tác giả và/hoặc khác năm. Vì vậy, kho dữ liệu không áp đặt cơ chế gộp cứng các trường hợp có khả năng là các phiên bản khác nhau, nhằm tránh mất thông tin và hạn chế rủi ro tạo sai lệch khi phân tích theo thời gian hoặc theo nguồn gốc.

Trong các bài toán phân tích định lượng dựa trên khoa học dữ liệu và xử lý ngôn ngữ tự nhiên, dữ liệu đầu vào là yếu tố quyết định trực tiếp đến độ tin cậy của kết quả. Với đặc thù lời bài hát tiếng Việt được đăng tải phân tán trên nhiều website, không có chuẩn biểu diễn thống nhất, dễ phát sinh trùng lặp và thường thiếu siêu dữ liệu, giai đoạn xây dựng kho dữ liệu cần được tổ chức như một quy trình có kiểm soát, ưu tiên các yêu cầu: thu thập được lời bài hát, lưu vết nguồn tham chiếu, và chuẩn hoá đầu ra để có thể gộp.

Trong khuôn khổ học phần, triển khai xây dựng kho dữ liệu dùng chung theo cơ chế phân công: mỗi nhóm được giao phụ trách một hoặc một số website để thu thập dữ liệu thô. Sau đó, dữ liệu từ các nhóm được đưa vào thảo luận chung để thống nhất lược đồ và gộp thành kho dùng chung. Cách tổ chức này giúp: (i) mở rộng độ phủ nguồn dữ liệu; (ii) giảm rủi ro mất mát; và (iii) tạo điều kiện đối chiếu, kiểm tra chéo giữa các nguồn. Kho dữ liệu dùng chung của lớp được xây dựng trong tháng 10/2025 và sau khi gộp đạt quy mô xấp xỉ 60.000 bản ghi.

Cần lưu ý rằng đơn vị thao tác chính là bản ghi: cùng một tiêu đề có thể tồn tại nhiều phiên bản (live/remix/cover/chuyển ngữ) hoặc trùng tên nhưng khác tác giả, khác

năm. Vì vậy, ở cấp kho dữ liệu, việc gộp không được đặt mục tiêu ép về một bản duy nhất, mà ưu tiên giữ nguyên thông tin để phục vụ các bài toán phân tích có xét đến khác biệt phiên bản và bối cảnh xuất bản.

2.2 Mục tiêu của kho dữ liệu dùng chung và yêu cầu chất lượng

Kho dữ liệu dùng chung được hình thành để tạo ra một nền dữ liệu thống nhất phục vụ nhiều đề tài trong cùng học phần. Cách tiếp cận này giúp giảm tình trạng mỗi nhóm thu thập theo một chuẩn riêng, từ đó hạn chế sai lệch do khác nguồn và khác quy tắc xử lý. Đồng thời, việc chia sẻ một kho dữ liệu thống nhất cũng tạo điều kiện để các kết quả phân tích giữa các nhóm có thể đối chiếu trên cùng nền dữ liệu.

Về yêu cầu chất lượng, kho dữ liệu được định hướng theo ba nguyên tắc: chuẩn xác, thống nhất và tái sử dụng.

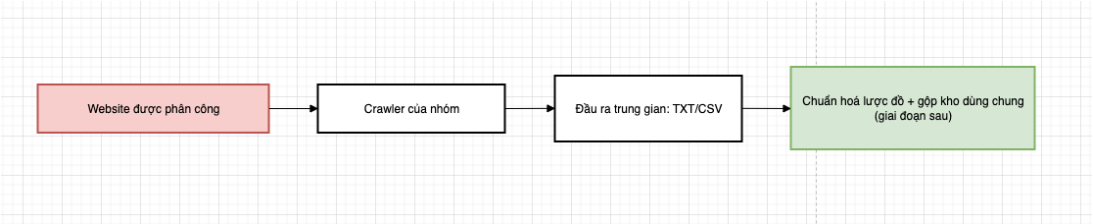
- Tính chuẩn xác theo nghĩa có thể kiểm chứng: dữ liệu cần gắn được với nguồn tham chiếu, đồng thời ghi nhận các trường hợp ngoại lệ (thiếu metadata, nghi vấn nội dung, trùng lặp kỹ thuật, nhiều phiên bản…) bằng cơ chế ghi chú để có thể truy vết và rà soát lại khi cần.
- Tính thống nhất: dù dữ liệu đến từ nhiều website, đầu ra cuối cùng phải được đưa về một lược đồ chung, bảo đảm khả năng gộp và khai thác đồng nhất ở các bước phân tích sau.
- Tính tái sử dụng: kho dữ liệu được xây dựng để có thể tiếp tục phục vụ nhiều mục đích và nhiều nhóm có chủ đề khác nhau trong học phần, không phụ thuộc vào một bài toán duy nhất. Điều này đòi hỏi dữ liệu được lưu trữ theo định dạng phổ biến, có cấu trúc trường rõ ràng, giữ được nguồn tham chiếu và có thể mở rộng bằng cách bổ sung cột đặc trưng mà không làm phá vỡ dữ liệu lõi.

2.3 Quy trình xây dựng dữ liệu theo cá nhân nhóm

Theo cơ chế phân công của lớp, mỗi nhóm phụ trách hai website và triển khai thu thập dữ liệu thô về dạng tệp trung gian; ở giai đoạn này, sản phẩm đầu ra được ưu tiên theo tiêu chí:

- Thu được dữ liệu thô (lời bài hát + metadata có sẵn).
- Có đầu ra trung gian (TXT/CSV) để bàn giao.
- Đảm bảo khả năng tiếp tục khi crawl dài ngày.

Dữ liệu ở giai đoạn này chưa bắt buộc khớp lược đồ kho dùng chung, vì bước chuẩn hoá, gộp sẽ diễn ra sau khi các nhóm hoàn tất việc cào dữ liệu và thống nhất cấu trúc chung cho dữ liệu tổng hợp cuối cùng.



Hình 1: Quy trình cào dữ liệu chung và kết quả mong muốn

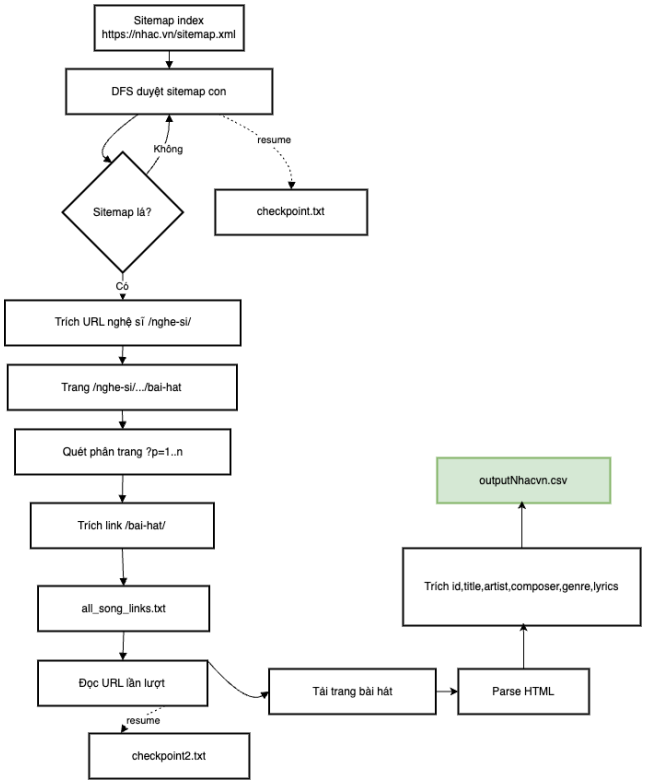
2.3.1 Nguồn nhac.vn

Như thể hiện trong sơ đồ, quy trình thu thập từ nhac.vn được tổ chức theo hai pha để kiểm soát tiến độ và đảm bảo có thể tiếp tục khi bị gián đoạn.

Ở pha thứ nhất, hệ thống bắt đầu từ sơ đồ chỉ mục của website và duyệt lần lượt các sơ đồ con theo cơ chế duyệt theo chiều sâu, đến khi gặp các nhánh cuối thì trích xuất danh sách trang nghệ sĩ và chuyển sang trang tổng hợp bài hát của từng nghệ sĩ. Từ đây, hệ thống quét lần lượt các trang phân trang để thu thập đầy đủ địa chỉ bài hát, sau đó ghi lại thành một danh sách liên kết dùng làm đầu vào cho pha tiếp theo. Trong toàn bộ pha này, trạng thái duyệt được lưu định kỳ để khi dừng giữa chừng có thể chạy tiếp đúng vị trí trước đó mà không phải quét lại từ đầu.

Ở pha thứ hai, hệ thống lần lượt truy cập từng địa chỉ bài hát trong danh sách đã thu được, tải nội dung trang và trích xuất các trường thông tin chính như mã định danh, tên bài, nghệ sĩ, nhạc sĩ, thể loại và phần lời. Phần lời được chuẩn hoá theo dạng văn bản nhiều dòng bằng cách chuyển các dấu ngắt dòng trên trang thành ký tự xuống dòng và loại bỏ các thành phần giao diện không thuộc nội dung lời, sau đó ghi nối tiếp ra tệp

dữ liệu tổng hợp. Tương tự pha một, pha hai cũng duy trì cơ chế lưu vị trí xử lý gần nhất để hỗ trợ tiếp tục khi tiến trình bị gián đoạn.



Hình 2: Quy trình cào dữ liệu chung và kết quả mong muốn

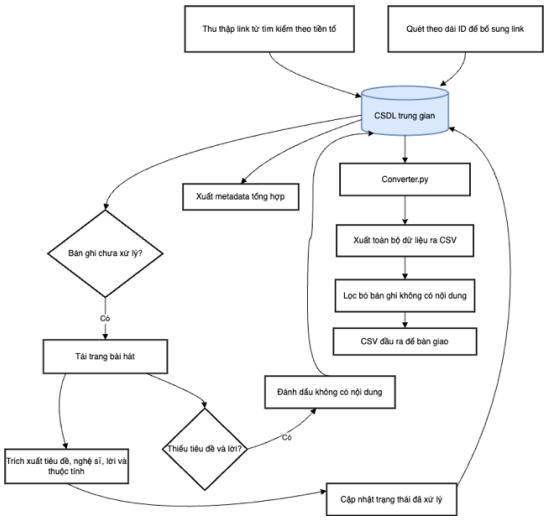
2.3.2 Nguồn lyric.tkaraoke.com

Theo sơ đồ, quy trình thu thập từ lyric.tkaraoke.com được thiết kế xoay quanh một cơ sở dữ liệu trung gian nhằm quản lý trạng thái thu thập và loại bỏ các trường hợp không có nội dung hợp lệ.

Trước hết, hệ thống thực hiện gom địa chỉ bài hát theo hai nhánh bổ trợ nhau. Nhánh thứ nhất khai thác trang kết quả tìm kiếm theo từ khoá và mở rộng theo tiền tố, đồng thời quét các trang phân trang để thu thập liên kết bài hát; nhánh thứ hai quét theo dải số định danh để bổ sung các địa chỉ ứng viên, giúp tăng độ bao phủ.

Tất cả địa chỉ thu được đều được lưu vào cơ sở dữ liệu kèm thông tin nguồn phát hiện và trạng thái xử lý, từ đó tránh việc thu thập trùng lặp và cho phép theo dõi tiến độ. Sau khi đã có tập địa chỉ, hệ thống chuyển sang bước thu thập nội dung chi tiết, lần lượt tải từng trang bài hát và trích xuất các thông tin như tiêu đề, nghệ sĩ và lời bài hát, đồng thời cập nhật trạng thái đã xử lý; trường hợp trang không có nội dung hợp lệ thì được đánh dấu tương ứng để loại khỏi dữ liệu khai thác.

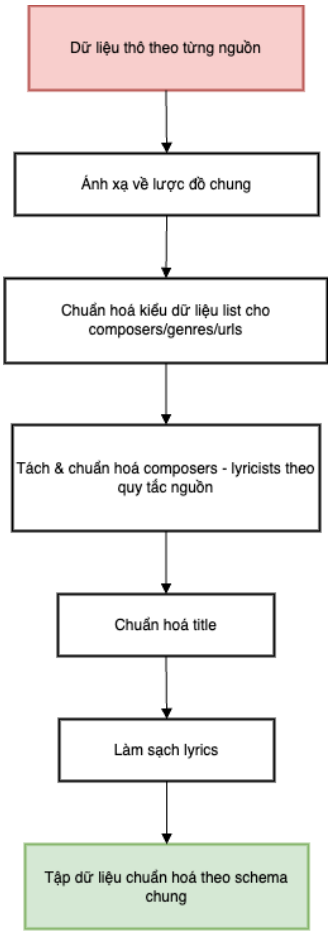
Khi cần bàn giao cho giai đoạn chuẩn hoá và gộp dữ liệu, hệ thống xuất toàn bộ dữ liệu từ cơ sở dữ liệu ra tệp dạng bảng và lọc bỏ các bản ghi đã bị đánh dấu không có nội dung, đồng thời lựa chọn mã hoá phù hợp để hiển thị tiếng Việt. Ở thời điểm kết thúc giai đoạn cá nhân nhóm, đầu ra vẫn phản ánh cấu trúc nội bộ của cơ sở dữ liệu trung gian; việc lựa chọn trường, đổi tên cột và đưa về lược đồ chuẩn dùng chung được thực hiện ở bước thảo luận và chuẩn hoá sau đó.



Hình 3:Quy trình thu thập dữ liệu từ lyric.tkaraoke.com theo mô hình cơ sở dữ liệu trung gian.

2.4 Chuẩn hoá dữ liệu theo quy ước lược đồ dùng chung.

Sau giai đoạn thu thập theo từng nguồn (mục 2.3), nhóm thực hiện bước chuẩn hóa nhằm đưa dữ liệu của mình về đúng quy ước lược đồ chung để có thể tham gia bước gộp dữ liệu ở cấp lớp. Trọng tâm của bước này là chuẩn hoá cấu trúc cột, thống nhất kiểu dữ liệu cho các trường dạng danh sách, và làm sạch các trường văn bản để giảm nhiễu khi so khớp và phân tích.



Hình 4: Quy trình chuẩn hóa dữ liệu thành cấu trúc chung đã thống nhất

Mô tả sơ đồ: dữ liệu thô được chuyển sang cấu trúc chung trước, sau đó lần lượt xử lý các trường "khó" (list fields, metadata tác giả, tiêu đề, lời bài hát) để tạo ra tập dữ liệu đã chuẩn hoá, sẵn sàng cho bước gộp dữ liệu giữa các nhóm.

Ở mức lược đồ, nhóm đưa dữ liệu về một cấu trúc thống nhất gồm các trường: title, composers, lyricists, year, genres, lyrics, urls, source, note; đối với các trường nguồn không có (ví dụ lyricists ở một số nguồn, year ở dữ liệu thu thập ban đầu), nhóm để rỗng hoặc null theo đúng trạng thái dữ liệu thay vì suy luận. Song song, nhóm chuẩn hoá các trường dạng danh sách bằng một hàm tiện ích để đọc được nhiều kiểu biểu diễn

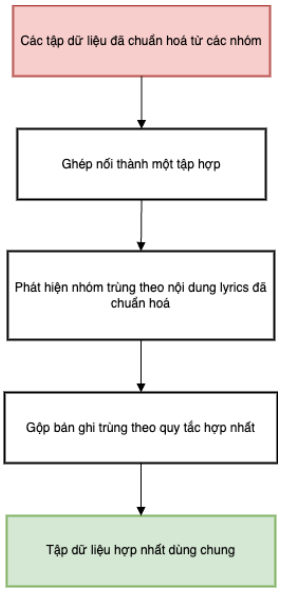
(list thật, chuỗi mô tả list, hoặc chuỗi phân tách bằng dấu phẩy), sau đó chuyển về cùng một chuẩn danh sách; bước này đặc biệt áp dụng cho composers, genres và urls nhằm tránh lỗi khi gộp và loại trùng về sau.

Với dữ liệu từ tkaraoke, nhóm thực hiện thêm một bước tách vai trò tác giả do nguồn thể hiện thông tin theo chuỗi có nhãn vai trò. Quy tắc tách được triển khai theo hướng bảo toàn thông tin: nhận diện các khối "Nhạc", "Lời", "Thơ", tách các tên theo các dấu phân tách thông dụng, sau đó ghi vào hai danh sách composers và lyricists; nếu không nhận diện được vai trò, toàn bộ được đưa vào composers để tránh mất thông tin. Với dữ liệu từ nhac.vn, do trường tác giả lời không được cung cấp trực tiếp trong dữ liệu thu thập ban đầu, lyricists được giữ rỗng ở giai đoạn chuẩn hoá lược đồ.

Sau khi chuẩn hoá lược đồ và metadata tác giả, nhóm chuẩn hoá trường tiêu đề theo các phép biến đổi nhằm ổn định so khớp giữa nguồn: đưa về chữ thường, thay "&" bằng "và", loại bỏ dấu câu không cần thiết, chuẩn hoá khoảng trắng. Tiếp theo, nhóm làm sạch trường lyrics theo bộ quy tắc thống nhất giữa các nguồn: loại ký tự ẩn, loại số thứ tự dạng "1.", loại các nhãn chỉ dẫn giọng/điệp khúc (bao gồm các biến thể viết tắt), loại nội dung trong ngoặc, chuẩn hoá "&" thành "và", đưa về chữ thường, giới hạn tập ký tự giữ lại để giảm nhiễu, và chuẩn hoá khoảng trắng. Kết quả của mục 2.4 là các tập dữ liệu đã được chuẩn hoá theo quy ước chung, bảo đảm có thể đưa vào bước gộp dữ liệu ở cấp lớp mà không cần can thiệp cấu trúc thêm.

2.5 Hợp nhất dữ liệu đa nguồn ở cấp lớp và giảm trùng lặp (Merge)

Sau khi từng nhóm hoàn tất bước chuẩn hoá theo quy ước chung, lớp triển khai bước gộp dữ liệu thành kho dùng chung. Việc gộp tập trung vào hai mục tiêu: (i) hợp nhất các tập dữ liệu về cùng một cấu trúc để tạo một kho thống nhất; và (ii) giảm trùng lặp kỹ thuật phát sinh khi nhiều nguồn đăng cùng một nội dung lời bài hát hoặc khi nhiều nhóm thu thập giao nhau.



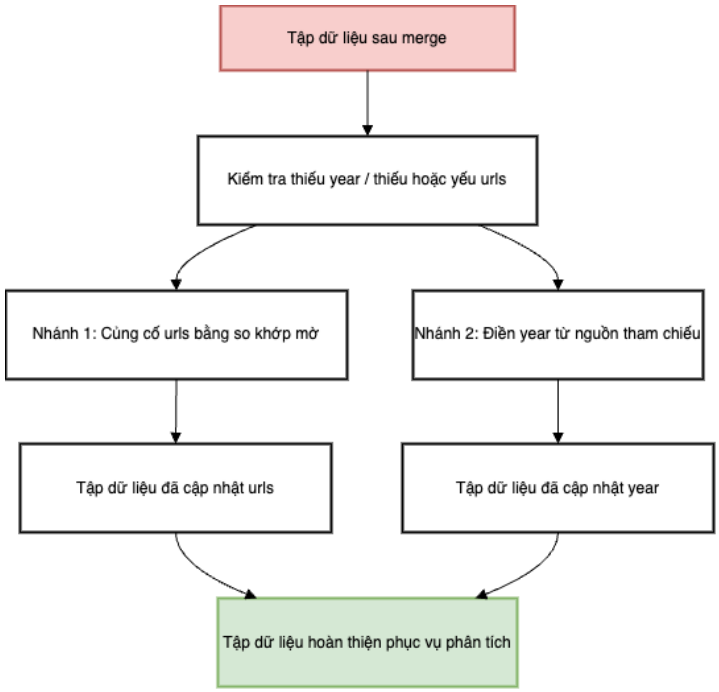
Hình 5: Quy cách gộp dữ liệu cào từ các nhóm thành tập dữ liệu chung

Mô tả sơ đồ: dữ liệu từ các nhóm được nối lại thành một tập hợp, sau đó triển khai phát hiện trùng lặp dựa trên nội dung lời đã chuẩn hoá; các bản ghi trùng được gộp theo quy tắc nhằm giữ lại thông tin tốt hơn và bảo toàn khả năng truy vết.

Ở bước phát hiện trùng lặp, nhóm chung sử dụng một biểu diễn lời bài hát đã được chuẩn hoá làm khoá nhóm, nhờ đó các bản ghi có lời giống nhau được gom vào cùng một nhóm trùng. Với mỗi nhóm trùng, bản ghi hợp nhất được tạo theo nguyên tắc ưu tiên dữ liệu đầy đủ hơn và bảo toàn thông tin: tiêu đề chọn theo tiêu chí đầy đủ hơn, các trường dạng danh sách như composers, lyricists, genres và urls được gộp bằng phép hợp nhất và loại trùng, trường year giữ giá trị hợp lệ nếu có, lyrics giữ nguyên nội dung, source và note được hợp nhất khi có nhiều giá trị khác nhau. Cách gộp này giúp giảm trùng lặp nhưng vẫn giữ được dấu vết nguồn, đặc biệt thông qua urls và note, để phục vụ kiểm chứng ở các chương phân tích sau.

2.6 Bổ sung trường thiếu sau gộp: củng cố URLs và điền cột year

Sau khi dữ liệu được gộp, tập dùng chung vẫn còn các trường có thể thiếu hoặc chưa đồng nhất, nổi bật là year và urls. Nhóm triển khai hai tuyến xử lý bổ sung tương ứng: (i) củng cố urls bằng cơ chế đối soát dựa trên chữ ký văn bản và so khớp mờ; và (ii) điền year bằng truy vấn tham chiếu từ các nguồn bên ngoài, có ghi nhận truy vết trong note và có cơ chế đánh dấu để đảm bảo tính ổn định khi xử lý quy mô lớn.



Hình 6: Kiểm tra thiếu sót dữ liệu và tiến hành bổ sung

Với việc điền cột urls, nhóm áp dụng quy trình theo hướng loại trước, đánh dấu, rồi bổ sung lại từ dữ liệu tham chiếu, để kiểm soát rõ các bản ghi bị tác động và hạn chế điền nhầm. Trước hết, nhóm chuẩn hoá cột đường dẫn về dạng danh sách, sau đó loại bỏ các đường dẫn thuộc miền nhac.vn trong từng bản ghi. Đồng thời, nhóm tạo thêm một cột đánh dấu để phân biệt hai trường hợp: bản ghi có đường dẫn nhac.vn đã bị loại bỏ và bản ghi không bị ảnh hưởng. Việc bổ sung lại đường dẫn chỉ được thực hiện cho nhóm bản ghi đã bị loại bỏ đường dẫn. Ở bước bổ sung, nhóm sử dụng một tệp tham chiếu giữ nguyên đường dẫn nhac.vn, xây dựng một khoá so khớp cho từng bài dựa trên tiêu đề và một đoạn ngắn ở phần đầu lời bài hát, sau đó chuẩn hoá khoá so khớp để giảm sai khác do cách viết. Nhóm thực hiện so khớp gần đúng để tìm bản ghi tham chiếu phù hợp nhất và chỉ chấp nhận kết quả khi mức tương đồng đạt ngưỡng đã đặt. Kết quả so khớp được xuất ra bảng rà soát, và việc cập nhật đường dẫn được thực hiện

theo đúng vị trí bản ghi trong tệp dữ liệu gốc; đường dẫn mới được thêm vào danh sách hiện có và loại trùng nhằm bảo toàn các tham chiếu hợp lệ khác.

Với việc điền cột year, nhóm xử lý trên tập dữ liệu sau khi gộp bằng cách lọc các bản ghi chưa có thông tin năm, sau đó truy vấn theo thứ tự ưu tiên từ các nguồn tham chiếu. Quy trình được triển khai theo nhiều tầng: ưu tiên truy vấn MusicBrainz theo tiêu đề và thông tin tác giả khi có, nếu không có kết quả thì chuyển sang Wikipedia, và cuối cùng thử iTunes. Năm được trích xuất từ các trường ngày phát hành hoặc phần mô tả của nguồn trả về. Mỗi trường hợp điền được năm đều được ghi rõ trong cột ghi chú kèm nguồn tham chiếu đã sử dụng; trường hợp không tìm thấy cũng được ghi nhận trạng thái để đảm bảo khả năng truy vết. Ngoài các nguồn truy vấn chính, nhóm có thêm một tệp tham chiếu chứa năm phát hành từ dữ liệu Spotify để đối chiếu theo tên bài hát trong các trường hợp cần bổ sung tham khảo.

2.7 Kết luận chương

Chương 2 đã trình bày toàn bộ quy trình xây dựng kho dữ liệu lời bài hát tiếng Việt trong khuôn khổ học phần theo hướng có kiểm soát và có thể truy vết. Từ cơ chế phân công thu thập theo nhóm, dữ liệu được thu thập đa nguồn và lưu thành các tệp trung gian nhằm bảo đảm tính liên tục của quá trình cào cũng như khả năng kiểm tra lại khi cần. Trên cơ sở dữ liệu đã thu thập, nhóm thực hiện bước chuẩn hoá theo lược đồ dùng chung, thống nhất cấu trúc cột và kiểu dữ liệu, đồng thời làm sạch các trường văn bản quan trọng như tiêu đề và lời bài hát để giảm nhiễu cho các bước so khớp và phân tích.

Sau khi các nhóm hoàn tất chuẩn hoá, dữ liệu được gộp ở cấp lớp để tạo thành kho dùng chung và triển khai giảm trùng lặp dựa trên nội dung lời bài hát đã được chuẩn hoá. Việc hợp nhất được thực hiện theo nguyên tắc bảo toàn thông tin: các trường dạng danh sách được gộp và loại trùng, dữ liệu cốt lõi về lời bài hát được giữ nguyên, đồng thời duy trì dấu vết nguồn thông qua trường đường dẫn và ghi chú.

Cuối cùng, chương 2 cũng mô tả các bước hoàn thiện dữ liệu sau gộp đối với những trường còn thiếu hoặc chưa đồng nhất. Cụ thể, nhóm triển khai quy trình củng cố đường dẫn tham chiếu bằng cơ chế đánh dấu các bản ghi bị ảnh hưởng và bổ sung lại từ dữ liệu tham chiếu theo phương pháp so khớp gần đúng có ngưỡng chấp nhận; đồng thời bổ sung trường năm phát hành bằng cách truy vấn theo thứ tự ưu tiên từ các

nguồn tham chiếu, kèm cơ chế ghi chú nguồn và trạng thái để phục vụ truy vết. Kết quả của chương này là một tập dữ liệu hợp nhất có cấu trúc thống nhất, chất lượng được kiểm soát ở mức dữ liệu, và đủ điều kiện làm đầu vào cho các bước trích xuất chỉ số và phân tích định lượng trong các chương tiếp theo.

CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU

3.1 Mục tiêu phân tích

3.1.1 Mục tiêu tổng quát của giai đoạn phân tích

Giai đoạn phân tích nhằm khai thác tập dữ liệu dùng chung final_dataset_cleaned để định lượng các biểu hiện ảnh hưởng ngôn ngữ trong lời bài hát, thông qua việc (i) chuẩn hoá lời bài hát theo các quy tắc đã thống nhất, (ii) gán nhãn thành phần từ vựng theo các nhóm ngôn ngữ/tín hiệu (phiên âm nước ngoài, tên riêng, Hán-Việt, tiếng Việt, tiếng Anh), và (iii) tính toán các chỉ số định lượng phục vụ phân tích theo giai đoạn thời gian và theo nhóm dữ liệu liên quan đến đề tài.

3.1.2 Đầu ra mong muốn của giai đoạn phân tích

Đầu ra của giai đoạn phân tích gồm ba nhóm chính:

- Thứ nhất là tập dữ liệu đã gán nhãn, trong đó lời bài hát được chuẩn hoá và mỗi bản ghi có thêm các trường đặc trưng phục vụ nhận diện tiếng Anh, phiên âm nước ngoài, tên riêng, Hán-Việt và tiếng Việt.
- Thứ hai là bộ chỉ số định lượng được tính trên từng bản ghi (bao gồm các chỉ số code-mixing, phân bố ngôn ngữ và chỉ số phong cách đã chốt trong đề tài).
- Thứ ba là các bảng tổng hợp theo giai đoạn thời gian, dùng để so sánh xu hướng biến đổi của các chỉ số qua các giai đoạn và làm dữ liệu đầu vào cho bước trực quan hoá, diễn giải ở các phần sau của báo cáo.

3.1.3 Nguyên tắc triển khai

Toàn bộ quy trình phân tích được triển khai theo nguyên tắc có thể tái lập bằng mã và truy vết được theo từng bước xử lý. Cụ thể, nhóm tổ chức quy trình xử lý theo các thư mục chức năng và xuất các tập tin trung gian theo từng bước (chuẩn hoá teencode, loại noise, gán nhãn theo thứ tự ưu tiên, trích xuất cột đặc trưng và tính chỉ số). Việc tách file theo bước giúp kiểm tra lỗi, so sánh trước-sau và đảm bảo rằng kết quả cuối cùng có thể được tái tạo từ dữ liệu đầu vào mà không phụ thuộc vào thao tác thủ công ngoài quy trình.

3.2 Dữ liệu đầu vào và tổ chức quy trình xử lý

3.2.1 Tập dữ liệu đầu từ kho dữ liệu lớp đã thống nhất.

Nhóm sử dụng final_dataset_cleaned_v3.csv làm tập dữ liệu đầu vào chính cho toàn bộ giai đoạn phân tích. Đây là tệp dữ liệu tổng hợp dùng chung của lớp, được hình thành sau quá trình thu thập đa nguồn, chuẩn hoá lược đồ, gộp dữ liệu và bổ sung một số trường còn thiếu theo quy ước thống nhất. Trên tập dữ liệu này, nhóm thực hiện các bước xử lý riêng phục vụ đề tài, bao gồm chuẩn hoá văn bản lời bài hát, xây dựng thư viện gán nhãn, gán nhãn thành phần từ vựng và tính toán các chỉ số định lượng.

3.2.2 Quy ước xử lý theo thư mục

Để thuận tiện cho việc theo dõi và kiểm soát chất lượng trong suốt quá trình phân tích, nhóm tổ chức toàn bộ pipeline theo các thư mục chức năng. Mỗi thư mục tương ứng với một nhóm công việc rõ ràng, giúp hạn chế nhầm lẫn giữa các bước và dễ kiểm tra lại khi phát sinh lỗi. Cụ thể, pipeline được tách thành ba phần chính: (1) tạo dữ liệu và kết quả trung gian để hình thành các thư viện phục vụ gán nhãn, (2) xây dựng và làm sạch các bộ từ điển dùng trong quá trình nhận diện từ vựng, và (3) gán nhãn trên toàn bộ tập dữ liệu và tính toán các chỉ số phục vụ phân tích. Với cách tổ chức này, các tệp đầu ra sau mỗi bước được lưu ngay trong thư mục tương ứng, qua đó đảm bảo có thể truy vết rõ dữ liệu đã thay đổi như thế nào theo từng công đoạn và thuận lợi khi tái lập quy trình.

3.2.3 Danh sách tập tin trung gian và ý nghĩa

Các tệp trung gian được tạo ra để đánh dấu dữ liệu sau từng bước xử lý quan trọng. Nhờ đó, nhóm có thể kiểm tra chất lượng, đối chiếu trước-sau, và chạy lại quy trình khi cần mà không phụ thuộc vào thao tác thủ công. Các nhóm tệp chính được tổ chức như sau:

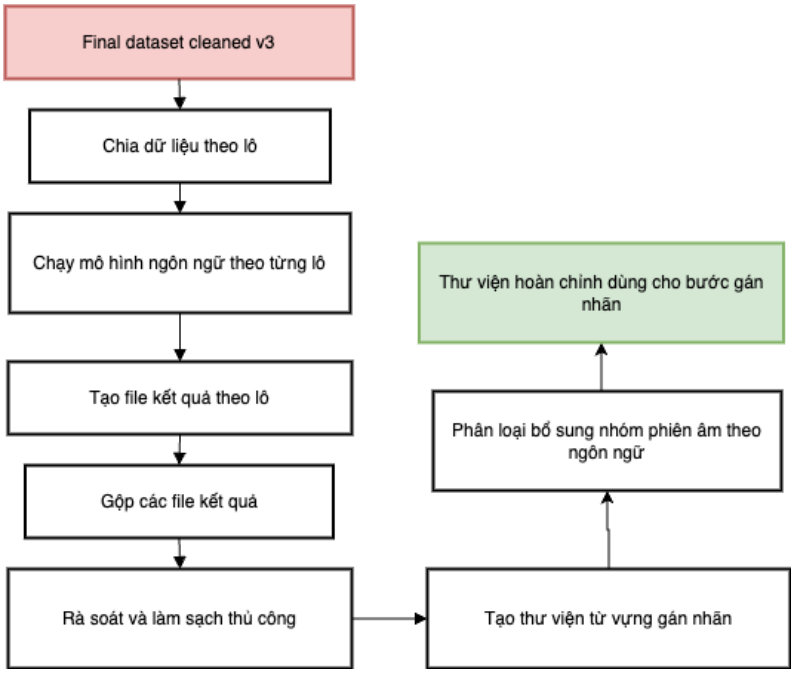
- Nhóm tách dữ liệu theo lô để tạo thư viện: từ tệp dữ liệu chung final_dataset_cleaned, nhóm chia thành nhiều phần nhỏ (part1 đến part4). Việc chia lô giúp xử lý ổn định hơn khi gọi mô hình ngôn ngữ và dễ tổng hợp kết quả theo từng đợt.
- Nhóm kết quả phân loại bằng mô hình ngôn ngữ: gồm các tệp đầu ra chứa chỉ mục gốc và các cột trích xuất như từ/cụm tiếng Anh, từ phiên âm nước ngoài và tên riêng. Các tệp này được sử dụng như nguồn dữ liệu đầu vào để nhóm tổng hợp, lọc và rà soát thủ công, từ đó hình thành và hoàn thiện các thư viện phục vụ bước gán nhãn.

- Nhóm thư viện gán nhãn: gồm các tệp từ điển phục vụ nhận diện trong lời bài hát, chẳng hạn teencode, danh sách nhiễu, danh sách tiếng Anh mở rộng, Hán-Việt, tên riêng và phiên âm. Các thư viện này được kiểm tra chéo và làm sạch để hạn chế trùng lặp giữa các nhóm nhãn.
- Nhóm dữ liệu sau chuẩn hoá văn bản: là các phiên bản dữ liệu sau những bước tiền xử lý chính trên lời bài hát, bao gồm bản sau chuẩn hoá teencode và bản sau khi loại nhiễu. Đây là dữ liệu nền được dùng trực tiếp cho bước gán nhãn.
- Nhóm dữ liệu gán nhãn theo thứ tự ưu tiên: sau mỗi bước gán nhãn theo trình tự ưu tiên, nhóm lưu một tệp trung gian tương ứng để phản ánh trạng thái dữ liệu tại thời điểm đó. Các tệp này thể hiện rõ nội dung đã được gán nhãn và các trường trích xuất phát sinh trong bước xử lý, qua đó hỗ trợ kiểm tra mức độ bao phủ nhãn và nhận diện các trường hợp chồng lấn giữa các bộ từ điển.
- Nhóm dữ liệu phục vụ tính toán chỉ số: là tệp dữ liệu tổng hợp cuối cùng sau khi hoàn tất gán nhãn và chuẩn bị đầy đủ các cột đặc trưng cần thiết. Tệp này được sử dụng làm đầu vào cho bước tính toán tỷ lệ phân bố ngôn ngữ, các chỉ số đo mức độ pha trộn ngôn ngữ và các chỉ số phong cách, phục vụ phân tích trong các mục tiếp theo.

3.3 Tạo dữ liệu huấn luyện và xây dựng thư viện gán nhãn bằng mô hình ngôn ngữ

3.3.1 Chia nhỏ dữ liệu để xử lý theo lô

Tập dữ liệu đầu vào final_dataset_cleaned có quy mô lớn, do đó nhóm chia dữ liệu thành nhiều phần nhỏ để xử lý theo lô. Việc chia lô giúp quá trình gọi mô hình ngôn ngữ ổn định hơn, dễ kiểm soát tiến độ và thuận tiện khi tổng hợp kết quả về sau. Các tệp sau khi chia được đặt theo quy ước thống nhất để đảm bảo có thể ghép lại đúng thứ tự và truy vết về chỉ mục gốc.



Hình 7: Làm sạch, phân loại và tạo bộ thư viện phục vụ gán nhãn

3.3.2 Phân loại sơ bộ thành phần trong lời bài hát bằng mô hình ngôn ngữ

Trên từng phần dữ liệu, nhóm sử dụng mô hình ngôn ngữ để trích xuất sơ bộ các thành phần từ vựng theo ba nhóm phục vụ xây thư viện: từ và cụm từ tiếng Anh, các từ nước ngoài được phiên âm, và tên riêng. Đầu ra của bước này được lưu theo cấu trúc bảng, trong đó có chỉ mục gốc của bản ghi, nội dung lyrics và ba cột kết quả trích xuất tương ứng. Các tệp đầu ra được xem như nguồn dữ liệu đầu vào để nhóm tổng hợp và xây dựng các bộ từ điển gán nhãn ở các bước tiếp theo.

3.3.3 Tổng hợp kết quả và rà soát thủ công

Sau khi hoàn tất xử lý theo lô, nhóm ghép các tệp kết quả thành một tập tổng hợp phục vụ xây dựng thư viện. Do kết quả trích xuất từ mô hình ngôn ngữ chưa đạt độ

chính xác tuyệt đối, nhóm thực hiện rà soát và lọc thủ công để loại bỏ các mục sai, mục trùng lặp hoặc các mục không đúng nhóm. Bước rà soát này nhằm nâng độ tin cậy của các thư viện trước khi đưa vào pipeline gán nhãn tự động, đồng thời giảm rủi ro lan truyền lỗi sang các bước tính toán chỉ số.

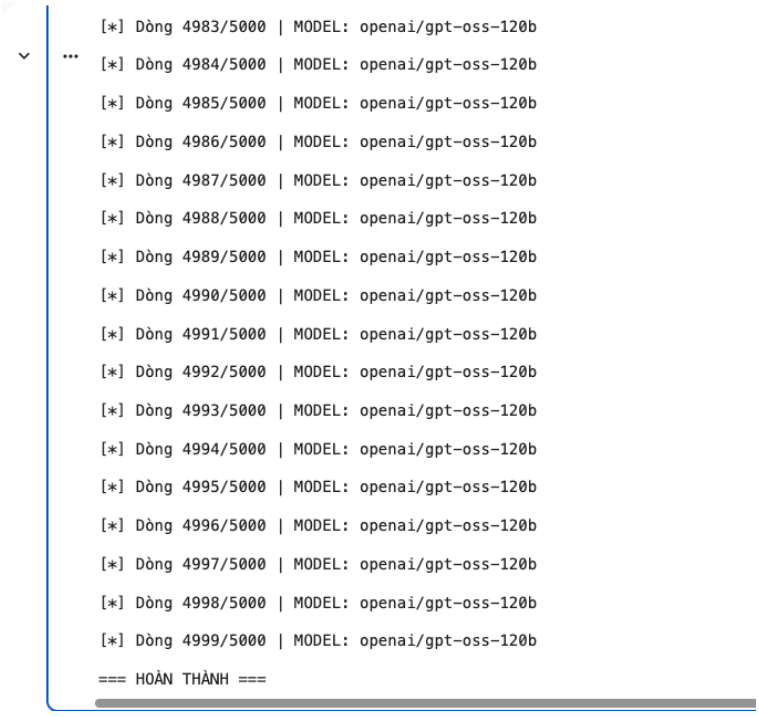
3.3.4 Phân loại bổ sung đối với nhóm phiên âm theo quốc gia và ngôn ngữ

Đối với nhóm từ phiên âm, sau khi có danh sách tổng hợp, nhóm thực hiện phân loại bổ sung theo quốc gia hoặc ngôn ngữ tương ứng. Mục tiêu của bước này là tổ chức thư viện phiên âm theo các nhóm rõ ràng, giúp bước gán nhãn sau đó có thể ghi nhận phiên âm theo từng nhóm ngôn ngữ, thay vì chỉ dừng ở nhãn chung. Việc phân loại bổ sung được thực hiện sau giai đoạn rà soát để hạn chế trường hợp phân loại dựa trên dữ liệu còn nhiễu hoặc sai lệch.

3.4 Thử nghiệm mô hình LLM chạy cục bộ

3.4.1 Mục tiêu thử nghiệm

Bên cạnh phương án sử dụng mô hình ngôn ngữ thông qua dịch vụ API, nhóm triển khai thử nghiệm một phương án thay thế là chạy mô hình LLM cục bộ trên môi trường Colab với mô hình được tải về và nạp theo cơ chế lượng tử hoá. Mục tiêu của thử nghiệm này là đối chiếu tính khả thi của hai hướng triển khai ở cùng một nhiệm vụ, cụ thể là trích xuất và phân loại từ vựng trong lời bài hát theo các nhóm phục vụ xây dựng thư viện gán nhãn. Việc so sánh tập trung vào khả năng áp dụng thực tế của mô hình cục bộ trong điều kiện tài nguyên hạn chế, mức độ ổn định khi xử lý dữ liệu lớn và chất lượng kết quả đầu ra so với phương án gọi API.



Hình 8: Quá trình chạy phân loại sử dụng mô hình ngôn ngữ lớn

3.4.2 Kết quả thử nghiệm và lý do không dùng trong pipeline chính

Kết quả thử nghiệm cho thấy mô hình chạy cục bộ có thể tạo đầu ra theo đúng định dạng yêu cầu, tuy nhiên chất lượng phân loại chưa đáp ứng tiêu chí của nhóm để đưa vào quy trình chính. Cụ thể, độ chính xác của kết quả trích xuất và phân loại không ổn định giữa các bản ghi và chưa đạt mức cần thiết để sử dụng làm nguồn xây dựng thư viện gán nhãn. Do đặc thù bài toán cần độ tin cậy cao ở cấp từ vựng để tránh lan truyền

sai lệch sang các bước gán nhãn tự động và tính toán chỉ số, nhóm quyết định không sử dụng phương án LLM cục bộ cho pipeline chính. Thay vào đó, nhóm sử dụng phương án gọi mô hình qua API để tạo danh sách ứng viên, sau đó kết hợp bước rà soát và làm sạch thủ công nhằm đảm bảo chất lượng thư viện.

index_goc	lyrics	eng	nuoc_ngoai_phien_am	ten_rieng
0	Mùa Giáng Sinh xưa anh hẹn anh sẽ về Ngày đó Noel bên hội sao trần thế Anh có nhớ không anh Em mặc màu áo xanh lam Xanh như liễu Đà Lạt Một chiều đông Giáng Sinh. Ngồi chờ Rê-vây- yông (Réveillon) Anh kể tích xưa rằng Vào một đêm giá lạnh Rợp trời hào quang thiên thần Và nơi hang Bê- Lem Thiên Chúa sinh Trên máng cỏ Là con Chúa trời. Ngày tháng trôi qua đi mau Mùa sao sáng năm nào Giờ cũng Noel Một mình em thẫn thờ Quỳ bên hang đá nguyện cầu Một người chân mây gió Được sống gần nhau. Tà áo năm xưa xanh màu thông Đà Lạt Nguyện đến năm nay khi cùng anh dạo phố Để nhớ Giáng Sinh xưa Kỷ niệm ngày Chúa ra đời Cho em sống lại màu xanh ái ân.	['Noel', 'Réveillon',]	['Réveillon', ':', 'Rê-vây- yông', 'Bê-Lem', 'Bê- Lem']	['Chúa trời']
1	đây miền quê hương em bao la thắm nương dâu , mênh mông sắc hương chè chúng em vui tung tăng bước chân vang trên đồi như đàn chim trong nắng đây miền quê thân thương khi sương sớm lung linh, có nắng mai nghiêng nghiêng em vui hát bên đồi em không quên công ơn biết bao người khó nhọc dành màu xanh cho em đây trường em thân yêu vui trong bóng cây xanh nhân giúp em bao điều chúng em như măng non có công ơn cô thầy nhắc nhở em năm tháng, cây phượng ơi thân thương khi sương sớm lung linh có nắng mai nghiêng nghiêng em vui bước đến trường em không quên công ơn biết bao người khó nhọc dành ngày mai cho em		[]	[]
2	khắp mọi miền đất nước bao la màu áo xanh qua vui hạnh phúc muôn nhà đem sức trẻ với tài ba dâng hiến tổ quốc thiêng liêng thanh niên việt nam vui lên đường hướng tới tương lai hợp sức chung tay tô đẹp nước non này yêu biết mấy với màu xanh thân ái màu áo thanh niên màu xanh tình nguyện ơi chiếc áo xanh chiếc áo xanh tình nguyện dấu có lên rừng hay đi xuống biển cũng chẳng ngại ngần những chàng trai hiên ngang vẫn luôn tươi cười những cô gái đảm đang	['Youth', 'Nation', 'Country', 'Future', 'Together']	[]	['Việt Nam']
3	Một ngày anh đến Nghe đời bỗng thấy reo vui Và rằng em biết Đã yêu nhau sẽ yêu dài lâu Em yêu anh như nắng yêu trời Em yêu anh không nói nên lời Em yêu anh như gió Yêu cơn mưa trong nắng hạ rơi. Kỷ niệm đó để đầu phai tàn Hạnh phúc đó đừng như khói sương Tan trong hư vô. Tình yêu xanh, màu xanh ước mơ Đêm từng đêm đời như giấc mơ Trên đường khuya cùng nhau sánh vai Trọn đời em luôn yêu anh mãi. Nguyện cầu cho tình yêu mãi xanh Như màu xanh tình em với anh Cho ngày mai mình luôn có nhau Mãi không rời xa.	['love', 'dream', 'forever']		
4	tiếng em ta cười vang trong bóng đêm mịt mù tiếng em ta cười vang để quên đau khổ hãy giữ lấy trái tim trẻ thơ thoát nơi chiến tranh, tối tăm chiến tranh gây niềm đau, gieo bao hận thù chiến tranh gây niềm đau, loài người phẫn ly nhìn vào sâu ánh mắt trẻ thơ một màu xanh nhưng nay còn đâu, hỡi em hãy cho em bình yên chỉ một phút thôi hãy cho em bình yên để được đến trường đừng gieo bao sầu đau đừng gây thêm bao niềm đau chia ly hát vang lên bài ca chung một tấm lòng hát vang lên bài ca xóa đi hận thù cùng nhau đem niềm bình yên cùng nhau đem niềm vui thần tiên một mai kia em xa cuộc đời vì hờn ghen của mỗi con người đã cướp mất trái tim trẻ thơ hỡi chiến thuật thu hồi cho em bình yên chỉ một phút thôi hãy cho em bình yên để	['trẻ thơ', 'trái tim trẻ thơ', 'chiến tranh', 'hận thù', 'sầu đau', 'niềm đau', 'bình yên', 'vui thần tiên']	[]	['Jay Cole', 'Sài Gòn']

Hình 9: Kết quả phân loại

3.4.3 Vai trò của phần thử nghiệm trong báo cáo

Phần thử nghiệm mô hình LLM chạy cục bộ được đưa vào báo cáo nhằm ghi nhận đầy đủ quá trình đánh giá phương án triển khai, thể hiện rằng nhóm đã xem xét các lựa chọn khác nhau trước khi chốt pipeline chính thức. Nội dung này đóng vai trò minh chứng cho cách tiếp cận có kiểm soát, trong đó nhóm không chỉ triển khai theo một hướng duy nhất mà có bước thử nghiệm, đối chiếu và loại bỏ phương án không đạt yêu cầu dựa trên kết quả thực tế.

3.5 Xây dựng và làm sạch hệ thư viện gán nhãn

Trong pipeline phân tích, nhóm sử dụng hệ thư viện gán nhãn như một lớp tri thức nền để nhận diện các nhóm từ vựng trong lời bài hát. Hệ thư viện này được xây dựng từ hai nguồn chính: dữ liệu trích xuất từ mô hình ngôn ngữ và các bộ từ điển công khai bên ngoài. Sau khi tổng hợp, nhóm thực hiện các bước làm sạch và kiểm tra chéo nhằm giảm chồng lấn giữa các nhóm nhãn và tăng độ ổn định khi gán nhãn trên toàn bộ tập dữ liệu.

3.5.1 Xây dựng thư viện Hán-Việt từ nguồn ngoài

Nhóm xây dựng thư viện Hán-Việt bằng cách tổng hợp từ nhiều tập dữ liệu bên ngoài và đưa về một cấu trúc thống nhất. Quy trình thực hiện gồm các bước chính:

- Đọc dữ liệu từ các nguồn Hán-Việt, giữ lại trường chứa từ Hán-Việt, loại các dòng rỗng, chuẩn hoá về chữ thường và chuẩn hoá khoảng trắng.
- Với nguồn có dữ liệu dạng danh sách, nhóm tách danh sách thành từng phần tử riêng để đưa về dạng một từ trên một dòng.
- Gắn nhãn nguồn cho từng dòng để phục vụ truy vết và hỗ trợ các bước lọc sau này.
- Gộp các nguồn lại thành một thư viện chung, loại trùng và xuất ra tệp từ điển Hán-Việt dùng trong pipeline gán nhãn.

Ngoài bước tổng hợp, nhóm triển khai một bước lọc bổ sung nhằm hạn chế đưa các từ phổ thông tiếng Việt vào nhóm Hán-Việt. Cụ thể, nhóm sử dụng một bộ từ điển tiếng Việt phổ thông để loại bỏ các mục Hán-Việt trùng với từ vựng phổ thông đối với một số nguồn nhất định, từ đó giữ lại các mục có tính đặc trưng cao hơn cho việc gán nhãn.

3.5.2 Xây dựng các thư viện phục vụ gán nhãn

Song song với thư viện Hán-Việt, nhóm xây dựng các thư viện còn lại để phục vụ các bước chuẩn hoá và gán nhãn trong pipeline, gồm:

- Thư viện teencode: chứa ánh xạ từ dạng viết tắt, biến thể mạng xã hội sang dạng chuẩn; được dùng trong bước chuẩn hoá lời bài hát để giảm nhiễu do cách viết không thống nhất.
- Thư viện noise: tập hợp các thành phần không mang giá trị ngôn ngữ trong lời bài hát, dùng để loại bỏ khỏi văn bản trước khi gán nhãn.
- Thư viện tiếng Anh mở rộng: ngoài từ điển tiếng Anh chuẩn, nhóm bổ sung danh sách từ và cụm từ thường gặp trong lời bài hát nhưng không phải lúc nào cũng có trong từ điển phổ thông; thư viện này phục vụ nhận diện tiếng Anh và tiếng lóng.
- Thư viện tên riêng: tổng hợp từ kết quả trích xuất bằng mô hình ngôn ngữ và quá trình rà soát, dùng để nhận diện địa danh, nghệ danh, thương hiệu hoặc tên người xuất hiện trong lời bài hát.
- Thư viện phiên âm nước ngoài: tổng hợp từ kết quả trích xuất và được tổ chức theo nhóm ngôn ngữ hoặc quốc gia; thư viện này phục vụ gán nhãn phiên âm theo nhóm thay vì gộp chung một nhãn.

Các thư viện trên được lưu theo định dạng bảng đơn giản, dễ đọc và dễ cập nhật, đồng thời được dùng trực tiếp trong các bước chuẩn hoá và gán nhãn của chương 3.

3.5.3 Làm sạch chéo giữa các thư viện để giảm chồng lấn

Do cùng một từ có thể xuất hiện trong nhiều nhóm, nhóm thực hiện làm sạch chéo giữa các thư viện để giảm trùng lặp và hạn chế gán nhãn sai. Các thao tác làm sạch chính gồm:

- Loại các từ tiếng Anh chuẩn khỏi thư viện tên riêng, nhằm tránh trường hợp từ tiếng Anh bị nhận diện nhầm thành tên riêng.
- Loại các mục phiên âm khỏi thư viện tên riêng, nhằm tách bạch rõ nhóm phiên âm và nhóm tên riêng.
- Loại các mục tên riêng khỏi thư viện tiếng Anh mở rộng để tránh chồng lấn giữa hai nhóm.
- Loại các mục phiên âm khỏi thư viện tiếng Anh mở rộng để hạn chế trường hợp phiên âm bị gán nhãn tiếng Anh.
- Lọc bỏ các từ phổ thông tiếng Việt khỏi thư viện tên riêng và một phần thư viện Hán-Việt, nhằm giảm nhiễu do các từ thông dụng không nên được xem là tín hiệu đặc thù.

Sau bước làm sạch chéo, nhóm duy trì các phiên bản thư viện đã lọc để sử dụng thống nhất trong pipeline gán nhãn, đồng thời giữ khả năng cập nhật khi phát sinh các trường hợp mới trong quá trình chạy trên dữ liệu lớn.

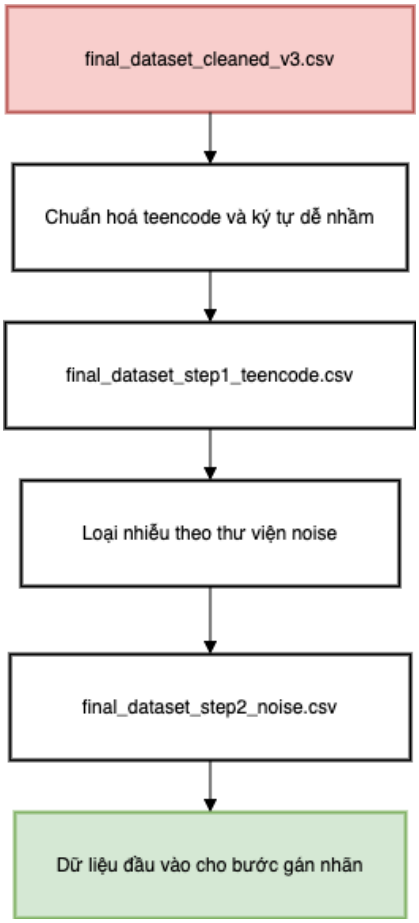
3.5.4 Kiểm tra độc lập giữa các nhóm thư viện và xử lý giao nhau

Sau khi làm sạch, nhóm thực hiện kiểm tra độc lập giữa các thư viện bằng cách chuẩn hoá toàn bộ từ vựng về cùng một dạng so sánh và tính các phần giao nhau giữa các cặp thư viện quan trọng, gồm tiếng Anh với phiên âm, tiếng Anh với tên riêng, và phiên âm với tên riêng. Mục tiêu của bước này là xác định các trường hợp còn trùng lặp và tiếp tục điều chỉnh thư viện nếu cần.

Bên cạnh kiểm tra bằng phần giao, trong quá trình xây dựng từ điển cụm từ để gán nhãn, nhóm duy trì cơ chế ghi nhận xung đột nhãn khi một mục xuất hiện đồng thời ở nhiều nhóm. Các trường hợp xung đột được xuất thành báo cáo riêng để rà soát và quyết định xử lý theo một trong hai hướng: loại bỏ khỏi một thư viện, hoặc giữ lại và giải quyết bằng thứ tự ưu tiên gán nhãn trong pipeline.

3.6 Tiền xử lý lời bài hát trước gán nhãn

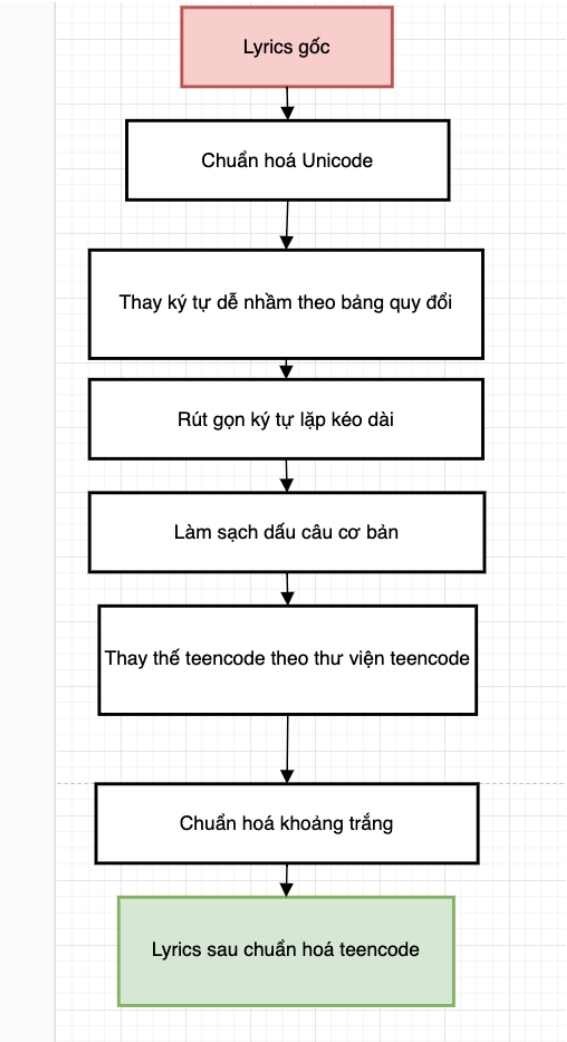
Trước khi thực hiện gán nhãn từ vựng, nhóm tiến hành tiền xử lý nội dung lời bài hát nhằm giảm nhiễu và đưa văn bản về dạng ổn định hơn cho các bước nhận diện theo thư viện. Các bước tiền xử lý được triển khai theo trình tự cố định và có tạo tệp trung gian để theo dõi ảnh hưởng của từng bước.



Hình 10: Chuẩn hóa các từ viết tắt, lệch chuẩn và xóa nhiễu trong bộ dữ liệu

3.6.1 Chuẩn hoá teencode và ký tự dễ nhầm theo thư viện teencode

Nhóm sử dụng thư viện teencode để chuyển các dạng viết tắt hoặc biến thể phổ biến trong môi trường mạng về dạng chuẩn. Việc chuẩn hoá được thực hiện theo cơ chế ánh xạ từ teencode sang từ chuẩn và ưu tiên thay thế các cụm dài trước để giảm sai lệch do thay thế chồng lấn. Bên cạnh teencode, nhóm xử lý thêm một số ký tự dễ nhầm do khác bảng mã hoặc ký tự đặc biệt, đưa về dạng tương đương trong bảng chữ cái thông thường. Ngoài ra, nhóm áp dụng quy tắc rút gọn chuỗi ký tự lặp kéo dài nhằm hạn chế trường hợp kéo dài âm tiết gây sai lệch thống kê token ở các bước sau.

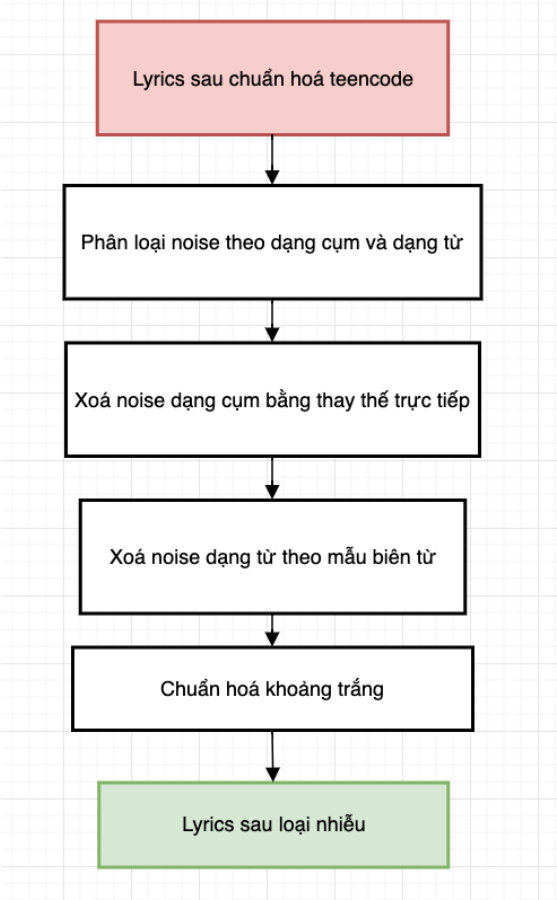


Hình 11:Quy trình xử lý chuẩn hóa các từ viết tắt, viết lệch chuẩn

3.6.2 Loại nhiễu theo thư viện noise và chuẩn hoá khoảng trắng

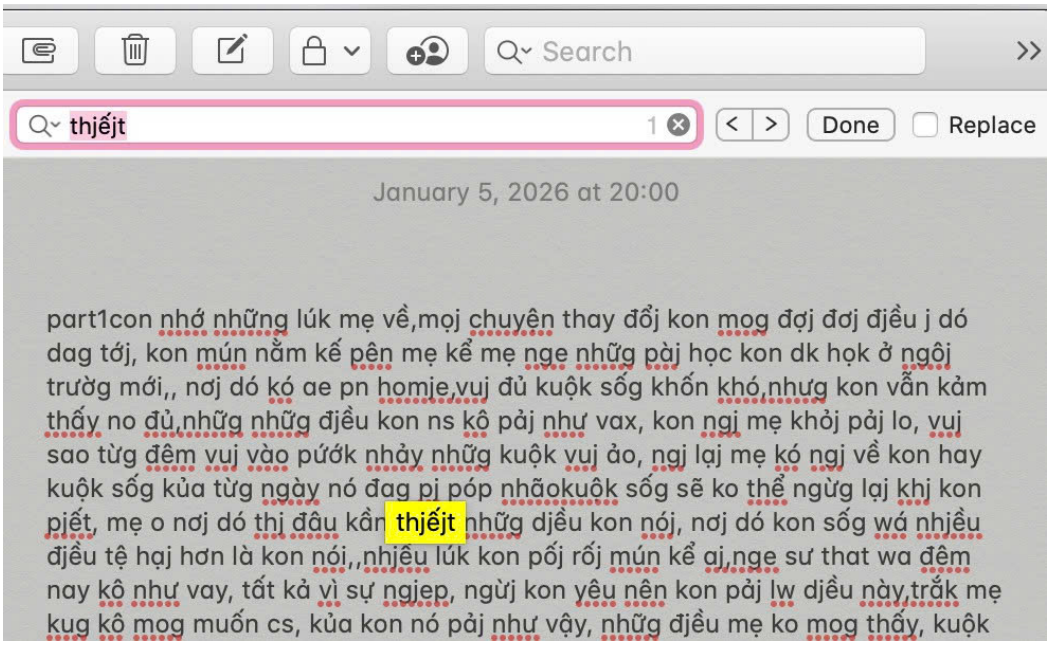
Sau bước chuẩn hoá teencode, nhóm loại bỏ nhiễu theo thư viện noise. Danh sách noise được tách theo hai dạng: noise là cụm dài và noise là từ đơn. Với noise dạng

cụm, nhóm loại theo cơ chế thay thế trực tiếp trong văn bản; với noise dạng từ đơn, nhóm áp dụng mẫu nhận diện để loại theo biên từ nhằm tránh xoá nhầm một phần của từ khác. Sau khi loại nhiễu, nhóm chuẩn hoá dấu câu cơ bản ở mức tối thiểu cần thiết cho gán nhãn và đưa khoảng trắng về dạng thống nhất để đảm bảo văn bản không chứa khoảng trắng thừa hoặc chuỗi dấu tách bất thường.



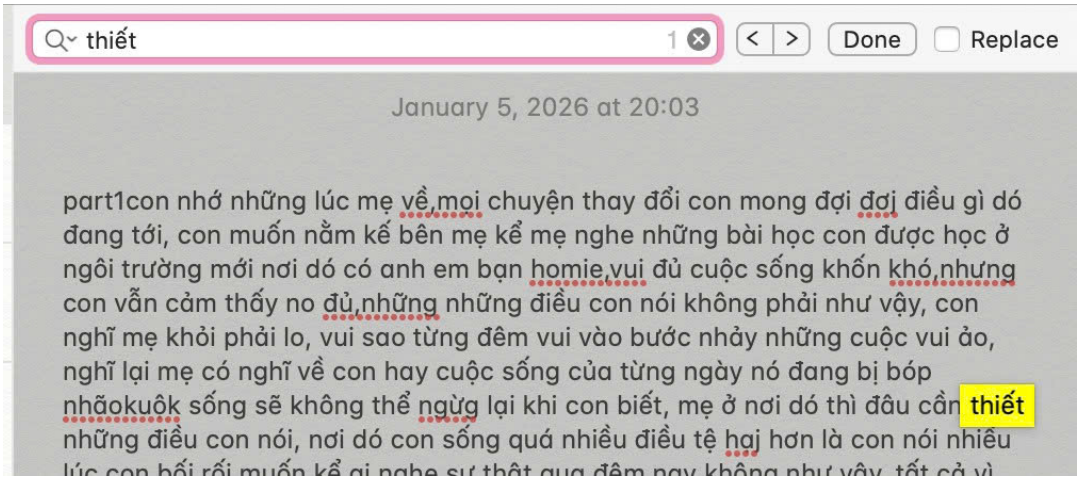
Hình 12: Quy trình loại bỏ các thành phần nhiễu trong tập dữ liệu

Hình 13 minh hoạ một đoạn lời bài hát ở trạng thái dữ liệu thô, trong đó xuất hiện các biến thể do gõ sai hoặc viết lệch chuẩn. Trường hợp từ khoá thiệt được thể hiện dưới dạng có ký tự thừa ở cuối (ví dụ thiệtj), làm cho việc tìm kiếm và đối sánh theo từ điển không ổn định.



Hình 13: Dữ liệu trước khi chuẩn hóa từ viết tắt, viết lệch chuẩn

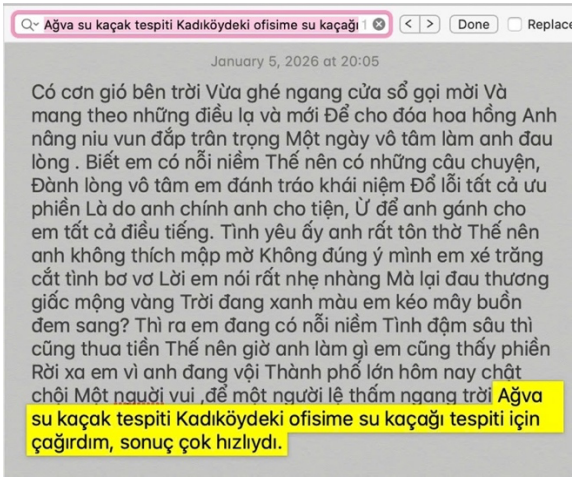
Hình 14 cho thấy cùng đoạn dữ liệu sau khi áp dụng bước chuẩn hoá văn bản. Các ký tự thừa và dạng viết lệch chuẩn đã được xử lý, đưa từ khoá về dạng thống nhất thiệt. Nhờ đó, từ vựng được biểu diễn nhất quán hơn, hỗ trợ tốt cho các bước nhận diện theo từ điển, giảm sai khác do chính tả không chuẩn và cải thiện khả năng truy vết, thống kê theo token trong các bước phân tích tiếp theo.



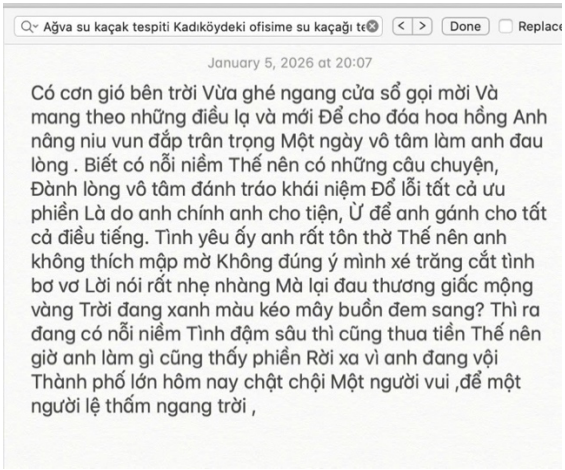
Hình 14: Kết quả dữ liệu sau khi chuẩn hóa từ viết tắt, viết lệch chuẩn

3.6.3 Tạo các phiên bản dữ liệu theo bước để truy vết

Để phục vụ kiểm tra và tái lập, nhóm lưu dữ liệu thành các phiên bản theo từng bước xử lý. Cụ thể, sau bước chuẩn hoá teencode, nhóm xuất một tệp trung gian phản ánh trạng thái lời bài hát đã được chuẩn hoá theo thư viện teencode. Tiếp đó, sau bước loại nhiễu theo thư viện noise, nhóm xuất thêm một tệp trung gian thứ hai phản ánh trạng thái lời bài hát sau khi đã loại nhiễu và chuẩn hoá khoảng trắng. Các phiên bản trung gian này giúp nhóm đối chiếu trước-sau khi cần đánh giá chất lượng tiền xử lý, đồng thời là cơ sở truy vết khi phát sinh sai lệch ở bước gán nhãn hoặc tính toán chỉ số.



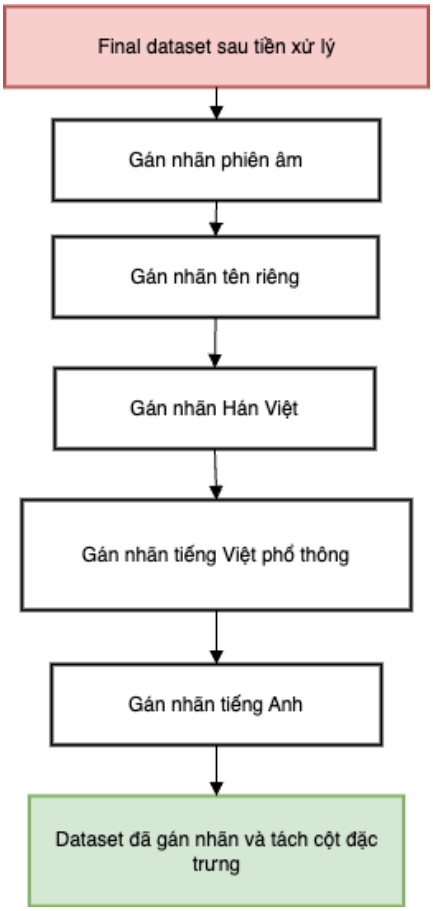
Hình 15: Dữ liệu trước khi xóa nhiễu



Hình 16: Dữ liệu sau khi xóa nhiễu

3.7 Gán nhãn thành phần ngôn ngữ trong lyrics

Giai đoạn gán nhãn được triển khai trên dữ liệu đã tiền xử lý ở mục 3.6. Mục tiêu là gán nhãn theo từng nhóm từ vựng để phục vụ hai việc: trích xuất các cột đặc trưng theo nhãn và làm đầu vào cho bước tính chỉ số định lượng ở các mục sau.



Hình 17: Quy trình gán nhãn theo thứ tự ưu tiên

3.7.1 Thiết kế thứ tự ưu tiên gán nhãn

Nhóm áp dụng thứ tự ưu tiên gán nhãn theo trình tự: phiên âm, tên riêng, Hán Việt, tiếng Việt, tiếng Anh. Thứ tự này được chọn để hạn chế chồng lấn giữa các nhóm từ vựng. Phiên âm được ưu tiên cao nhất vì thường có hình thức dễ bị nhầm với tiếng Anh hoặc tên riêng nếu không xử lý trước. Tên riêng được ưu tiên tiếp theo vì có tính đặc thù và thường cần nhận diện theo chuỗi nhiều từ. Hán Việt và tiếng Việt được xử lý trước tiếng Anh để tránh trường hợp một số từ bị gán nhãn tiếng Anh do hình thức ký tự, trong khi thực chất thuộc nhóm từ tiếng Việt hoặc Hán Việt theo thư viện của nhóm.

3.7.2 Gán nhãn phiên âm theo nhóm ngôn ngữ và xuất cột theo nhóm

Thư viện phiên âm được tổ chức theo các cột, mỗi cột đại diện cho một nhóm ngôn ngữ hoặc quốc gia. Nhóm xây dựng từ điển phiên âm bằng cách duyệt toàn bộ các cột trong tập dữ liệu phiên âm, chuẩn hoá về chữ thường và ánh xạ mỗi cụm phiên âm sang một nhãn tương ứng. Khi gán nhãn, hệ thống quét cụm từ theo độ dài giảm dần và nhận diện các cụm nhiều từ trước cụm một từ, nhằm tránh tách sai cụm phiên âm thành các phần nhỏ.

Đầu ra của bước này gồm hai phần:

- Phần thứ nhất là trường token đã gán nhãn, trong đó các cụm phiên âm được đánh dấu theo nhãn nhóm.
- Phần thứ hai là các cột trích xuất theo nhóm phiên âm, mỗi cột chứa danh sách các cụm phiên âm thuộc một nhóm ngôn ngữ tương ứng để thuận tiện tổng hợp thống kê.

3.7.3 Gán nhãn tên riêng theo nguyên tắc phân biệt hoa thường và xuất danh sách

Thư viện tên riêng được xây dựng dưới dạng danh sách và được làm sạch trước để loại bỏ các từ phổ thông. Khi gán nhãn, nhóm áp dụng nguyên tắc so khớp phân biệt hoa thường để hạn chế nhầm lẫn giữa tên riêng và từ thông thường. Bước này chỉ xét các token chưa được gán nhãn từ bước phiên âm. Tương tự bước phiên âm, hệ thống ưu tiên nhận diện cụm nhiều từ trước cụm một từ để giữ đúng cấu trúc tên người, địa danh hoặc tên tổ chức.

Kết quả được trích xuất thêm thành một cột riêng chứa danh sách tên riêng theo từng bản ghi, đồng thời cập nhật lại trường token đã gán nhãn để phục vụ các bước tiếp theo.

3.7.4 Gán nhãn Hán Việt dựa trên thư viện Hán Việt đã lọc

Thư viện Hán Việt được tạo từ nhiều nguồn và đã qua bước lọc để giảm lẫn với từ phổ thông. Khi gán nhãn, hệ thống kiểm tra từng token chưa được gán nhãn và đối chiếu theo dạng chữ thường. Nếu token nằm trong tập Hán Việt, token được gán nhãn Hán Việt và được trích xuất vào cột riêng phục vụ thống kê. Bước này được thực hiện sau phiên âm và tên riêng để tránh gán nhầm các cụm đặc thù vào nhóm Hán Việt.

3.7.5 Gán nhãn tiếng Việt phổ thông dựa trên từ điển tiếng Việt chuẩn

Ở bước này, nhóm sử dụng từ điển tiếng Việt chuẩn để gán nhãn cho các token còn lại chưa được nhận diện. Từ điển được nạp theo cơ chế tải về từ nguồn công khai, và có phương án thay thế bằng file nội bộ khi cần. Các token chưa có nhãn sau ba bước trước được đối chiếu với tập từ vựng tiếng Việt chuẩn theo dạng chữ thường. Kết quả được xuất thành cột riêng chứa danh sách từ tiếng Việt phổ thông, đồng thời giảm phần còn lại chưa được gán nhãn trước khi chuyển sang bước tiếng Anh.

3.7.6 Gán nhãn tiếng Anh dựa trên từ điển tiếng Anh chuẩn và thư viện mở rộng

Bước gán nhãn tiếng Anh được thực hiện cuối cùng để hạn chế chồng lấn với các nhóm từ trước đó. Nhóm kết hợp hai nguồn từ vựng: từ điển tiếng Anh chuẩn và thư viện tiếng Anh mở rộng do nhóm tổng hợp từ dữ liệu lời bài hát. Các token còn lại chưa được gán nhãn được đối chiếu theo dạng chữ thường, sau đó gán nhãn tiếng Anh nếu khớp. Kết quả được trích xuất thành cột riêng phục vụ bước tính tỷ lệ tiếng Anh và các chỉ số liên quan đến pha trộn ngôn ngữ.

3.7.7 Cơ chế phát hiện xung đột nhãn và cách xử lý

Trong quá trình xây dựng thư viện cụm từ, nhóm có cơ chế ghi nhận xung đột khi cùng một mục xuất hiện ở nhiều nhóm nhãn. Xung đột được kiểm tra theo các hướng chính: trùng giữa các nhóm phiên âm theo ngôn ngữ, trùng giữa tiếng Anh và tên riêng, trùng giữa tiếng Anh và phiên âm, và trùng giữa phiên âm và tên riêng. Các trường hợp này được xuất ra báo cáo riêng để rà soát.

Việc xử lý xung đột được thực hiện theo hai cách. Cách thứ nhất là làm sạch thư viện bằng cách loại mục trùng khỏi nhóm không phù hợp. Cách thứ hai là giữ mục trong thư viện nhưng giải quyết bằng thứ tự ưu tiên gán nhãn trong pipeline, bảo đảm mục đó chỉ nhận một nhãn duy nhất trong quá trình chạy thực tế.

3.8 Trích xuất đặc trưng từ kết quả gán nhãn và tạo dữ liệu đầu ra phục vụ phân tích

Phần này được triển khai trực tiếp trong thư mục Calculate_Analysis, trên dữ liệu đã qua hai bước tiền xử lý teencode và noise. Nhóm không tính chỉ số tổng hợp ở bước này, mà tập trung chuẩn hoá đầu ra gán nhãn và tách các cột đặc trưng theo từng nhóm nhãn để phục vụ thống kê và phân tích ở các mục sau.

3.8.1 Trường gán nhãn token theo từng bản ghi

Sau khi chạy các bước gán nhãn, mỗi bản ghi có một trường lưu danh sách token kèm nhãn. Trường này được dùng làm nguồn gốc để trích xuất các cột đặc trưng theo nhóm, đồng thời hỗ trợ kiểm tra mức độ bao phủ nhãn thông qua việc đếm nhãn chưa gán.

3.8.2 Trích xuất đặc trưng phiên âm theo nhóm ngôn ngữ

Trong bước gán nhãn phiên âm, dữ liệu đầu ra được tách thành các cột riêng theo từng nhóm ngôn ngữ. Mỗi cột chứa danh sách cụm phiên âm tương ứng nếu xuất hiện trong lời bài hát, và để trống nếu không có. Song song, nhóm tạo thêm một trường đếm số token chưa được gán nhãn để theo dõi mức độ bao phủ sau từng bước.

3.8.3 Trích xuất đặc trưng tên riêng, Hán Việt, tiếng Việt và tiếng Anh

Các bước gán nhãn tiếp theo cập nhật dần các token chưa gán nhãn và đồng thời xuất thêm các cột trích xuất theo từng nhóm. Cụ thể:

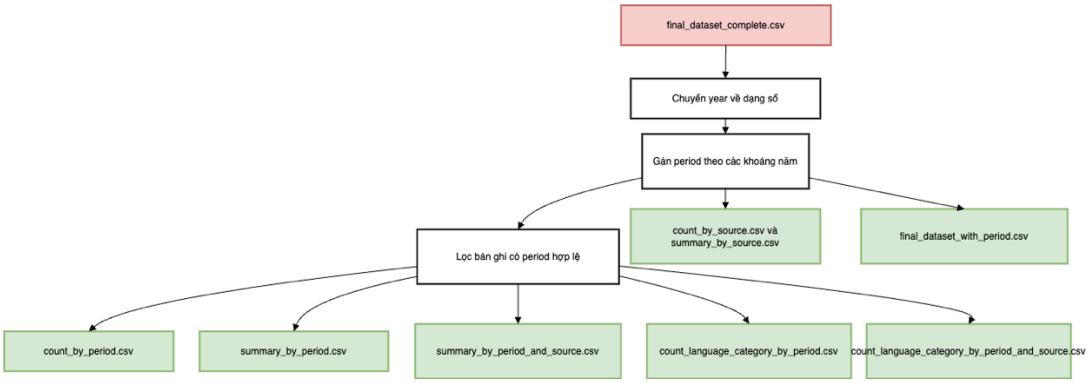
- Sau bước tên riêng, dữ liệu có thêm cột tổng hợp danh sách tên riêng theo bản ghi.
- Sau bước Hán Việt, dữ liệu có thêm cột tổng hợp danh sách từ Hán Việt theo bản ghi.
- Sau bước tiếng Việt phổ thông, dữ liệu có thêm cột tổng hợp danh sách từ tiếng Việt theo bản ghi.
- Sau bước tiếng Anh, dữ liệu có thêm cột tổng hợp danh sách từ tiếng Anh theo bản ghi.
- Các cột này được tạo theo đúng cấu trúc đã có trong notebook, nhằm tách rõ từng nhóm từ vựng phục vụ thống kê và so sánh.

3.8.4 Thống kê mức độ bao phủ nhãn và đầu ra theo từng bước

Sau mỗi bước gán nhãn, nhóm lưu một tệp dữ liệu trung gian để phản ánh trạng thái mới của dữ liệu. Đồng thời, nhóm thực hiện thống kê phân bố nhãn trên toàn bộ tập và thống kê mức độ xuất hiện theo từng nhóm, ví dụ số bản ghi có phiên âm theo từng ngôn ngữ, số bản ghi có tên riêng, số bản ghi có Hán Việt, tiếng Việt và tiếng Anh. Các thống kê này được dùng để kiểm tra nhanh chất lượng thư viện và mức độ phủ nhãn trước khi chuyển sang bước phân tích theo thời gian và theo nhóm.

3.9 Phân tích theo thời gian và theo nhóm

Phần này được thực hiện trên tệp dữ liệu tổng hợp sau gán nhãn final_dataset_complete.csv. Tệp này đã chứa các trường cần thiết cho phân tích theo thời gian và theo nhóm, gồm year, source, language_category và các cột thống kê. Mục tiêu của bước này là tạo biến giai đoạn thời gian từ year, sau đó tổng hợp dữ liệu theo giai đoạn và theo nhãn source, đồng thời xuất thêm các bảng phân bố theo language_category để phục vụ chương kết quả.



Hình 18: Quy trình tính toán, thống kê dựa trên dữ liệu đã gán nhãn

3.9.1 Tạo giai đoạn thời gian từ trường năm

Sau khi hoàn tất gán nhãn và tạo bộ đặc trưng, nhóm sử dụng trường năm phát hành để chuyển dữ liệu sang dạng phân tích theo giai đoạn. Trong xử lý, năm được chuyển về kiểu số; sau đó mỗi bản ghi được gán vào một trong năm khoảng thời gian: 1990-2000, 2000-2010, 2010-2015, 2015-2020 và 2020-2025. Những bản ghi không có năm hợp lệ không thể gán giai đoạn nên không được đưa vào các thống kê theo thời gian.

Kết quả đếm số bản ghi theo giai đoạn cho thấy dữ liệu có năm hợp lệ phân bố như sau: giai đoạn 1990-2000 có 4.271 bản ghi; 2000-2010 có 10.008 bản ghi; 2010-2015 có 5.678 bản ghi; 2015-2020 có 5.712 bản ghi; 2020-2025 có 14.935 bản ghi. Đây là tập dữ liệu nền để thực hiện các bước tổng hợp xu hướng theo thời gian ở phần tiếp theo.

3.9.2 Tổng hợp thống kê theo giai đoạn

Trên tập bản ghi đã gán giai đoạn, nhóm thực hiện tổng hợp theo từng giai đoạn bằng cách tính trung bình của các cột số đã có trong dữ liệu. Nhóm cột được đưa vào tổng hợp gồm các tỷ lệ theo nhãn ngôn ngữ như tỷ lệ tiếng Việt, Hán-Việt, tiếng Anh,

tỷ lệ chưa gán nhãn và tỷ lệ đã gán nhãn; đồng thời tổng hợp các cột đếm như số token theo từng nhóm và tổng số token của bài. Kết quả tổng hợp được xuất thành bảng theo giai đoạn để phục vụ trực quan hoá và so sánh xu hướng ở chương kết quả.

3.9.3 So sánh theo nguồn gốc và theo nhóm phân loại ngôn ngữ

Bên cạnh phân tích theo thời gian, nhóm tổng hợp dữ liệu theo trường nguồn gốc để quan sát sự khác biệt giữa các nhóm. Việc tổng hợp được thực hiện theo hai mức: tổng hợp theo nguồn gốc trên toàn bộ dữ liệu, và tổng hợp theo đồng thời giai đoạn và nguồn gốc để đối chiếu theo thời gian.

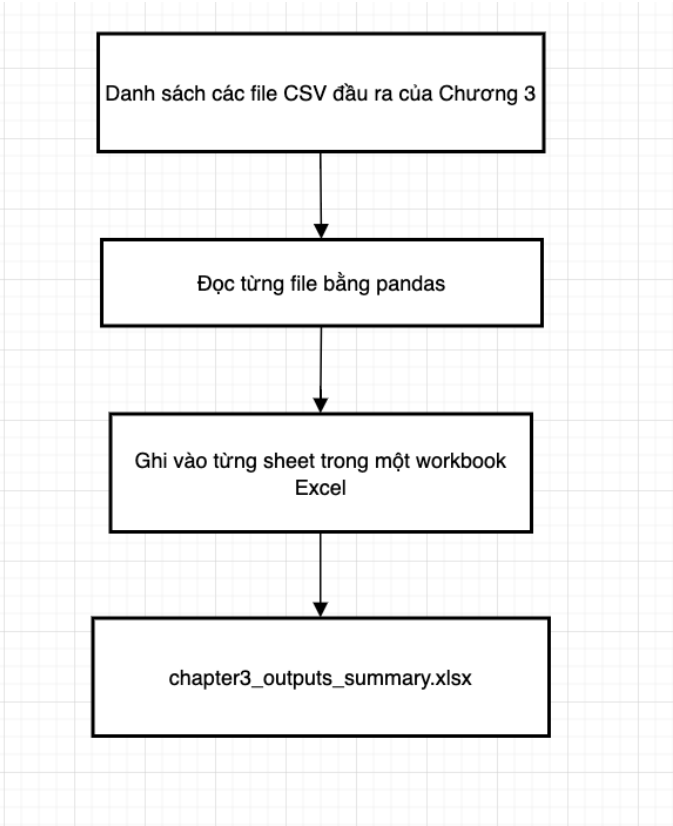
Ngoài ra, nhóm sử dụng trường phân loại mức độ ngôn ngữ đã có trong dữ liệu để tạo bảng phân bố số bản ghi theo giai đoạn, cũng như theo giai đoạn kết hợp với nguồn gốc. Trong bảng phân bố theo giai đoạn và nguồn gốc, có những tổ hợp cho kết quả bằng 0 do trong dữ liệu không phát sinh bản ghi thuộc tổ hợp đó. Ví dụ ở giai đoạn 1990-2000, một số nguồn có số bản ghi thuộc nhóm thuần Việt khác nhau, trong đó nguồn Ngoại ghi nhận 124 bản ghi thuộc nhóm thuần Việt.

3.9.4 Kết quả xuất ra và lưu ý khi chạy

Sau khi chạy, hệ thống tạo ra một bộ bảng kết quả gồm dữ liệu đã bổ sung giai đoạn thời gian, bảng đếm số bản ghi theo giai đoạn và theo nguồn gốc, bảng tổng hợp trung bình theo giai đoạn, bảng tổng hợp theo nguồn gốc, bảng tổng hợp theo giai đoạn kết hợp nguồn gốc, cùng các bảng phân bố theo nhóm phân loại ngôn ngữ. Trong quá trình đọc dữ liệu, có cảnh báo về kiểu dữ liệu không đồng nhất ở một số cột và cảnh báo về thay đổi mặc định của thao tác nhóm dữ liệu trong các phiên bản thư viện mới. Các cảnh báo này không làm thay đổi nội dung kết quả vì bước 3.9 chỉ sử dụng các cột năm, nguồn gốc, nhóm phân loại và các cột số để tổng hợp.

3.10 Tổng hợp và đóng gói đầu ra của Chương 3

Sau khi hoàn tất các bảng tổng hợp ở mục 3.9, nhóm thực hiện bước đóng gói kết quả theo dạng một tệp Excel để tiện theo dõi, kiểm tra và sử dụng khi viết báo cáo. Tệp Excel được tạo bằng cách đọc các bảng dữ liệu chính và ghi thành các trang riêng, bao gồm dữ liệu đầy đủ sau gán nhãn, dữ liệu đã có giai đoạn thời gian, các bảng tổng hợp theo giai đoạn, theo nguồn gốc, bảng tổng hợp theo giai đoạn kết hợp nguồn gốc, cùng các bảng đếm và bảng phân bố theo nhóm phân loại.



Hình 19: Quy trình tạo kết quả thống kê tổng quan, tiền đề cho phân tích và đưa ra kết luận

Kết quả đóng gói cho thấy tệp Excel được tạo thành công với đầy đủ các trang dữ liệu tương ứng, giúp người đọc có thể kiểm tra nhanh toàn bộ đầu ra của Chương 3 mà không cần mở nhiều tệp rời. Trong bước này cũng xuất hiện cảnh báo kiểu dữ liệu không đồng nhất ở một số cột khi đọc dữ liệu; cảnh báo không ảnh hưởng đến việc ghi tệp Excel vì thao tác thực hiện là đóng gói dữ liệu theo đúng trạng thái đã được xuất ở các bước trước, không tạo thêm biến đổi nội dung.

3.11 Kết luận chương

Chương 3 đã trình bày đầy đủ quy trình phân tích dữ liệu của nhóm trên tập dữ liệu dùng chung của lớp, bắt đầu từ chuẩn hoá văn bản lời bài hát, xây dựng hệ thư viện gán nhãn, đến gán nhãn thành phần ngôn ngữ theo thứ tự ưu tiên và trích xuất các cột đặc trưng phục vụ đo lường. Toàn bộ quy trình được triển khai bằng mã, có tổ chức theo thư mục chức năng và có các đầu ra trung gian theo từng bước nhằm bảo đảm khả năng tái lập và truy vết.

Trên cơ sở dữ liệu sau gán nhãn, nhóm thực hiện bước phân tích theo thời gian và theo nhóm bằng cách chuyển trường năm về dạng số, gán giai đoạn theo các khoảng 1990-2000, 2000-2010, 2010-2015, 2015-2020 và 2020-2025, sau đó tổng hợp thống kê theo giai đoạn, theo nguồn gốc và theo bảng chéo giai đoạn kết hợp nguồn gốc. Kết quả đếm cho thấy số bản ghi có năm hợp lệ được phân bố theo giai đoạn lần lượt là 4.271, 10.008, 5.678, 5.712 và 14.935 bản ghi, tạo nền dữ liệu rõ ràng cho việc quan sát xu hướng theo thời gian. Bên cạnh đó, nhóm xuất các bảng phân bố theo nhóm phân loại ngôn ngữ để mô tả cơ cấu dữ liệu theo giai đoạn và theo nguồn gốc, phục vụ trực tiếp cho phần trình bày kết quả.

Cuối chương, các bảng kết quả quan trọng được đóng gói vào một tệp Excel tổng hợp theo nhiều trang nhằm thuận tiện cho việc kiểm tra, đối chiếu và sử dụng trong báo cáo. Như vậy, Chương 3 đã hoàn thiện lớp dữ liệu phân tích và các bảng tổng hợp cần thiết, tạo tiền đề cho Chương 4 trình bày kết quả, trực quan hoá và thảo luận theo đúng mục tiêu nghiên cứu.

CHƯƠNG 4: KẾT QUẢ VÀ THẢO LUẬN

4.1 Tổng quan dữ liệu dùng cho phân tích kết quả

Phần kết quả và thảo luận của Chương 4 được xây dựng trên tập dữ liệu đã hoàn tất gán nhãn và trích xuất đặc trưng ở Chương 3. Dữ liệu ở thời điểm này đã có đầy đủ các trường phục vụ trực quan hóa và đánh giá, bao gồm năm phát hành, nguồn gốc, nhóm phân loại mức độ ngôn ngữ, các cột tỷ lệ theo nhóm ngôn ngữ, cùng các cột đếm token theo từng nhóm nhãn. Trên nền dữ liệu này, nhóm thực hiện tổng hợp theo giai đoạn thời gian và theo nhóm nguồn gốc để tạo các bảng thống kê và các biểu đồ phục vụ trình bày kết quả.

Trong phân tích theo thời gian, năm phát hành được chuyển về dạng số và dùng để gán giai đoạn theo năm khoảng: 1990 đến 2000, 2000 đến 2010, 2010 đến 2015, 2015 đến 2020 và 2020 đến 2025. Các bản ghi không có năm hợp lệ không được đưa vào các thống kê theo giai đoạn. Kết quả đếm số bản ghi theo giai đoạn cho thấy số lượng dữ liệu có năm hợp lệ phân bố như sau.

Giai đoạn	**Số bản ghi**
1990-2000	4.271
2000-2010	10.008
2010-2015	5.678
2015-2020	5.712
2020-2025	14.935

Bảng 4.1: Thống kê số bài hát theo từng giai đoạn từ tập dữ liệu

Tổng cộng, tập dữ liệu dùng cho các phân tích theo thời gian gồm 40.604 bản ghi có gán được giai đoạn. Bảng phân bố này là cơ sở để diễn giải mức độ đại diện dữ liệu theo từng giai đoạn, đồng thời là căn cứ quan trọng khi so sánh xu hướng giữa các giai đoạn, vì chênh lệch quy mô mẫu có thể ảnh hưởng đến độ ổn định của thống kê.

Bên cạnh thời gian, dữ liệu được tổng hợp theo trường nguồn gốc để phục vụ đối chiếu giữa các nhóm. Trường nguồn gốc được sử dụng như biến phân nhóm trong các bảng thống kê theo nhóm và trong bảng chéo giai đoạn kết hợp nguồn gốc. Ngoài ra, dữ liệu còn có trường phân loại mức độ ngôn ngữ dùng để mô tả cơ cấu các nhóm bài theo từng giai đoạn và theo nguồn gốc. Trong kết quả tổng hợp theo giai đoạn và nguồn

gốc, có những tổ hợp cho giá trị bằng không, phản ánh rằng trong dữ liệu không phát sinh bản ghi thuộc tổ hợp tương ứng. Trạng thái này là bình thường khi số lượng giá trị nguồn gốc lớn và khi dữ liệu ở một số nguồn có quy mô nhỏ.

Trong quá trình xử lý, hệ thống xuất hiện cảnh báo về kiểu dữ liệu không đồng nhất ở một số cột. Cảnh báo này phù hợp với cấu trúc dữ liệu hiện có, vì một số cột chứa danh sách hoặc chuỗi trích xuất theo nhãn, trong khi một số dòng để trống. Nhóm xử lý phần phân tích theo thời gian bằng cách chuyển trường năm về số và chỉ tổng hợp trên các cột số, nên các cảnh báo này không làm thay đổi các bảng đếm và bảng tổng hợp theo giai đoạn đã tạo ra.

4.2 Xu hướng theo thời gian của thành phần ngôn ngữ trong lời bài hát

4.2.1 Mục tiêu

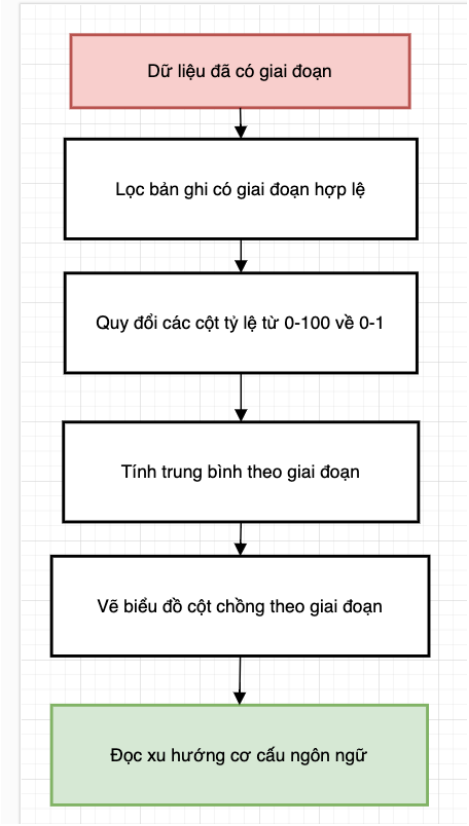
Mục tiêu của nhóm là mô tả sự thay đổi theo thời gian của thành phần ngôn ngữ trong lời bài hát, dựa trên các tỷ lệ đã được tạo sau bước gán nhãn. Kết quả được trình bày bằng trực quan hóa để người đọc có thể quan sát nhanh mức độ chiếm ưu thế của tiếng Việt, Hán Việt, tiếng Anh và phần chưa gán nhãn theo từng giai đoạn.

4.2.2 Dữ liệu và biến sử dụng

Phân tích sử dụng tập dữ liệu đã được gán giai đoạn thời gian. Mỗi bản ghi có trường giai đoạn và các cột tỷ lệ ngôn ngữ, gồm tỷ lệ tiếng Việt, tỷ lệ Hán Việt, tỷ lệ tiếng Anh và tỷ lệ chưa gán nhãn. Trong dữ liệu hiện tại, các tỷ lệ này được lưu theo thang phần trăm từ 0 đến 100, vì vậy khi vẽ biểu đồ theo định dạng phần trăm chuẩn, nhóm quy đổi các giá trị về thang 0 đến 1.

4.2.3 Quy trình tổng hợp và trực quan hóa

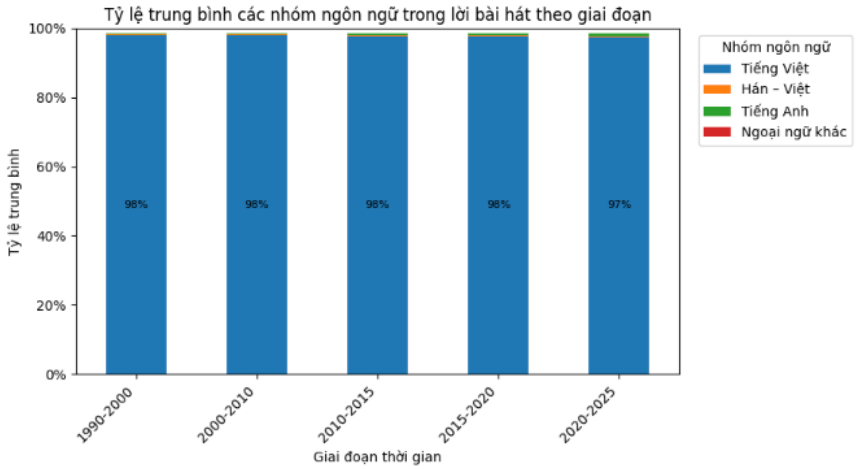
Nhóm thực hiện ba bước theo đúng trình tự: lọc các bản ghi có giai đoạn hợp lệ, quy đổi các cột tỷ lệ về cùng thang đo, và tính trung bình theo giai đoạn. Kết quả trung bình theo giai đoạn được đưa vào biểu đồ cột chồng để thể hiện cơ cấu tỷ lệ ngôn ngữ của từng giai đoạn.



Hình 20: Quy trình tổng hợp và trực quan hóa phân tích xu hướng theo thời gian của thành phần ngôn ngữ trong lời bài hát

Biểu đồ cột chồng được chọn vì thể hiện đồng thời nhiều thành phần trên cùng một trục thời gian. Để biểu đồ dễ đọc, nhãn phần trăm chỉ hiển thị với các phần có tỷ trọng đủ lớn; các phần quá nhỏ được ẩn nhãn để tránh rối.

4.2.4 Kết quả và hiện trực quan hoá

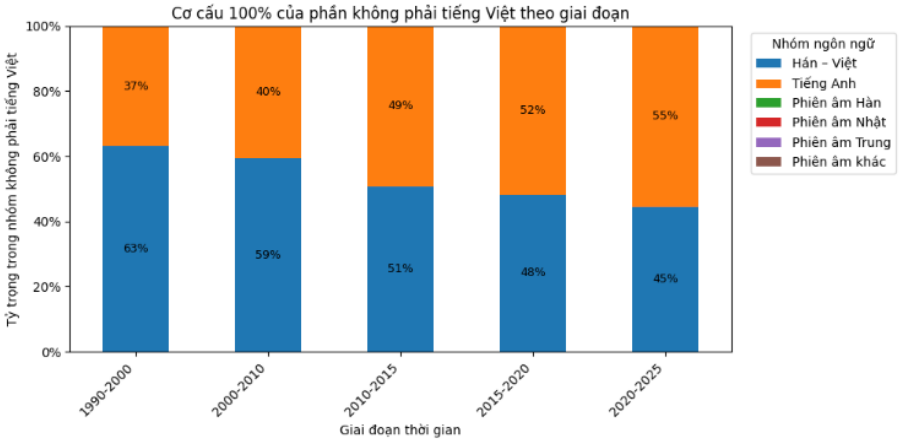


Hình 21: Tỷ lệ ngôn ngữ trung bình theo giai đoạn

Hình trên cho thấy tiếng Việt chiếm tỷ trọng áp đảo trong toàn bộ lời bài hát ở tất cả các giai đoạn. Cụ thể, tỷ lệ tiếng Việt trung bình xấp xỉ 98 phần trăm ở các giai đoạn từ 1990 đến 2020 và giảm nhẹ còn khoảng 97 phần trăm ở giai đoạn 2020-2025. Phần còn lại là các nhóm khác như Hán Việt, tiếng Anh và phần chưa gán nhãn, nhưng do tổng tỷ trọng rất nhỏ nên các lớp này gần như chỉ xuất hiện như một dải mỏng ở phía trên cột.

Ý nghĩa của hình này là xác nhận bức tranh tổng quan: xét trên toàn bộ token, lời bài hát vẫn chủ yếu là tiếng Việt; sự khác biệt theo thời gian nếu có sẽ nằm ở phần không phải tiếng Việt, vì phần này mới là nơi các yếu tố ngoại lai xuất hiện.

```
...  Tỷ lệ trung bình của phần không phải tiếng Việt trên toàn bộ lyrics theo giai đoạn:
        period  pct_non_vietnamese_on_total
0  1990–2000                     1.830817
1  2000–2010                     1.906121
2  2010–2015                     2.113123
3  2015–2020                     2.201145
4  2020–2025                     2.491261
```



Hình 22: Cơ cấu phần không phải tiếng Việt sau khi loại tiếng Việt

Hình trên được tạo để nhìn rõ hơn phần không phải tiếng Việt. Khi tiếng Việt chiếm khoảng 97-98 phần trăm như Hình 4.2, các thay đổi nhỏ của tiếng Anh, Hán Việt hoặc phần chưa gán nhãn rất khó quan sát nếu giữ nguyên thang đo toàn bộ. Vì vậy, việc loại tiếng Việt ra khỏi hình thứ hai là một thao tác hợp lý về trực quan hóa: thay vì nhìn toàn bộ cột bị đè bởi lớp tiếng Việt, hình này phóng đại phần còn lại để thấy rõ xu hướng.

Ba thành phần được thể hiện gồm Hán Việt, tiếng Anh và phần chưa gán nhãn. Kết quả thể hiện xu hướng rõ ràng theo thời gian đối với tiếng Anh: tỷ trọng tiếng Anh trong phần không phải tiếng Việt tăng dần từ khoảng 8 phần trăm ở giai đoạn 1990-2000, lên khoảng 10 phần trăm ở 2000-2010, 12 phần trăm ở 2010-2015, 15 phần trăm ở 2015-2020 và đạt khoảng 19 phần trăm ở 2020-2025. Trong khi đó, tỷ trọng Hán Việt gần như ổn định quanh mức 12-13 phần trăm qua các giai đoạn. Phần chưa gán nhãn giảm dần theo thời gian, từ khoảng 53 phần trăm xuống khoảng 46 phần trăm.

4.3 Thực nghiệm bổ sung theo thời gian: tỷ lệ bài hát có yếu tố tiếng Anh theo năm

4.3.1 Mục tiêu

Thực nghiệm này nhằm bổ sung một góc nhìn đơn giản nhưng trực quan về mức độ phổ biến của yếu tố tiếng Anh trong lời bài hát theo thời gian. Khác với phần 4.2 vốn nhìn theo cơ cấu token trung bình theo giai đoạn, mục 4.3 tập trung trả lời câu hỏi: trong từng năm, có bao nhiêu bài hát xuất hiện ít nhất một token được gán nhãn tiếng Anh.

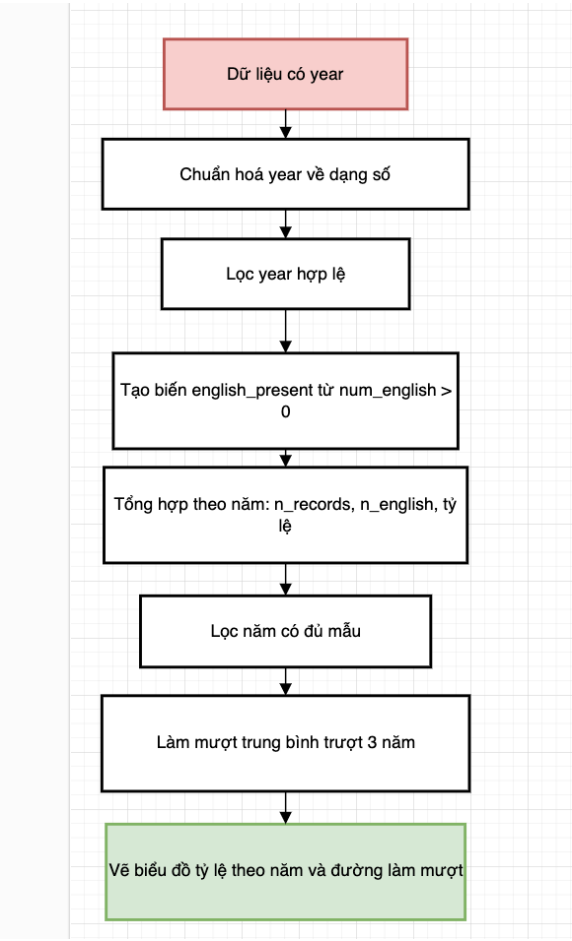
4.3.2 Dữ liệu và biến sử dụng

Dữ liệu được lấy từ tập đã hoàn tất gán nhãn và có trường năm. Trường year được chuẩn hoá về dạng số (year_num) và chỉ giữ các bản ghi có năm hợp lệ. Yếu tố tiếng Anh được định nghĩa theo biến nhị phân english_present, với điều kiện num_english > 0, nghĩa là trong lời bài hát có ít nhất một token được gán nhãn tiếng Anh. Cách định nghĩa này đo lường mức độ xuất hiện theo bài hát, không đo cường độ sử dụng trong từng bài.

4.3.3 Quy trình tổng hợp và trực quan hóa

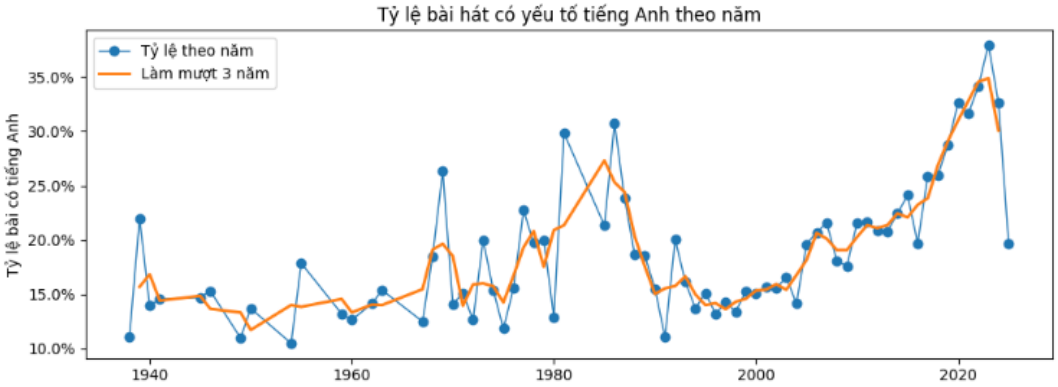
Quy trình xử lý gồm ba bước chính: chuẩn hoá năm và lọc dữ liệu có năm hợp lệ, tổng hợp theo từng năm, và trực quan hoá bằng biểu đồ đường. Khi tổng hợp theo

năm, nhóm tính n_records là tổng số bài có year hợp lệ trong năm đó, n_english là số bài thoả điều kiện english_present, và tỷ lệ pct_english_present = n_english / n_records. Để giảm nhiễu do các năm có số lượng bản ghi quá nhỏ, nhóm chỉ giữ những năm có tối thiểu 50 bản ghi. Sau đó nhóm vẽ hai đường: tỷ lệ theo năm và đường làm mượt theo trung bình trượt 3 năm để quan sát xu hướng ổn định hơn.



Hình 23: Quy trình trực quan hóa tỷ lệ bài hát có yếu tố tiếng Anh theo năm

4.3.4 Kết quả và hình trực quan hoá



Hình 24: Biểu đồ tỷ lệ bài hát có yếu tố tiếng Anh theo năm

Hình trên trình bày tỷ lệ bài hát có yếu tố tiếng Anh theo năm. Đường màu xanh thể hiện tỷ lệ quan sát theo từng năm, trong khi đường màu cam là trung bình trượt 3 năm giúp giảm dao động và làm rõ xu hướng chung. Một điểm cần lưu ý là dữ liệu theo năm có thể dao động mạnh ở một số giai đoạn do quy mô mẫu giữa các năm không đồng đều; vì vậy, nhóm sử dụng cả đường làm mượt để hỗ trợ diễn giải xu hướng tổng thể.

Từ hình vẽ có thể quan sát hai đặc điểm chính. Thứ nhất, tỷ lệ bài có yếu tố tiếng Anh biến động theo năm, phản ánh sự thay đổi về mức độ sử dụng theo từng thời điểm. Thứ hai, đường làm mượt cho thấy xu hướng tăng rõ hơn ở giai đoạn về sau, đặc biệt trong các năm gần đây, khi tỷ lệ đạt mức cao hơn so với nhiều giai đoạn trước. Kết quả này cung cấp một minh chứng định lượng đơn giản để bổ sung cho bức tranh tổng quan ở mục 4.2

Từ Hình trên, nhóm không chỉ ghi nhận xu hướng tăng của tỷ lệ bài hát có yếu tố tiếng Anh ở các năm gần đây, mà còn có thể đặt bối cảnh để diễn giải theo hướng thận trọng. Cụ thể, sự gia tăng này có thể liên hệ với quá trình hội nhập và giao lưu văn hoá ngày càng mạnh của Việt Nam trong giai đoạn sau Đổi mới, đặc biệt khi các kênh tiếp cận sản phẩm văn hoá quốc tế trở nên phổ biến hơn. Ở góc độ hạ tầng truyền thông, sự phát triển của Internet và nhạc số, sau đó là mạng xã hội và nền tảng video ngắn, làm tăng tần suất tiếp xúc với âm nhạc quốc tế và kéo theo xu hướng sử dụng tiếng Anh như một tín hiệu thẩm mỹ hoặc tín hiệu nhận diện phong cách trong lời bài hát.

Để gắn bối cảnh xã hội vào báo cáo một cách rõ ràng và có kiểm soát, nhóm có thể bổ sung các mốc tham chiếu trực tiếp lên trục thời gian của biểu đồ, theo hướng coi

đó là các điểm đánh dấu để so sánh trước và sau mốc. Ví dụ, nhóm có thể đánh dấu giai đoạn mở rộng hội nhập và giao thương quốc tế, giai đoạn phổ cập nhạc số và nền tảng trực tuyến, giai đoạn bùng nổ mạng xã hội và video ngắn, rồi quan sát xem tỷ lệ bài có yếu tố tiếng Anh thay đổi như thế nào quanh các mốc này. Nếu cần tăng độ chặt chẽ, có thể thực hiện thêm một kiểm định thống kê đơn giản theo nhóm giai đoạn trước và sau mốc, nhằm xác nhận sự khác biệt về tỷ lệ là đáng kể về mặt thống kê hay chỉ là dao động của mẫu.

Về ứng dụng thực tế, kết quả của hình trên có thể dùng như một chỉ báo định lượng cho nhà sáng tác và nhà sản kế mức độ chèn tiếng xuất khi thiết Anh phù hợp với bối cảnh thị trường và giai đoạn phát hành. Đối với truyền thông và phân phối, tỷ lệ bài có yếu tố tiếng Anh theo năm có thể được xem như một tín hiệu xu hướng để xây dựng thông điệp quảng bá, lựa chọn chiến lược nhắm mục tiêu theo nhóm khán giả và theo nền tảng. Ở góc độ giáo dục và nghiên cứu, kết quả này cung cấp một nền tham chiếu để thảo luận về mức độ tiếp xúc và vay mượn ngôn ngữ trong văn hoá đại chúng, đồng thời mở đường cho các phân tích sâu hơn, chẳng hạn so sánh theo thể loại, theo nguồn gốc, hoặc kết hợp với chỉ số cường độ để phân biệt giữa xuất hiện và mức độ sử dụng.

4.4 So sánh theo thể loại và tác giả

4.4.1 Mục tiêu

Mục tiêu của mục này là mở rộng quan sát theo hướng không gian đặc trưng của dữ liệu thay vì theo thời gian. Cụ thể, nhóm so sánh mức độ phổ biến của yếu tố tiếng Anh theo hai trục: thể loại và nhạc sĩ. Kết quả nhằm trả lời câu hỏi thực nghiệm: yếu tố tiếng Anh xuất hiện khác nhau như thế nào giữa các dòng nhạc, và giữa các nhóm nhạc sĩ có số lượng bài đủ lớn trong dữ liệu.

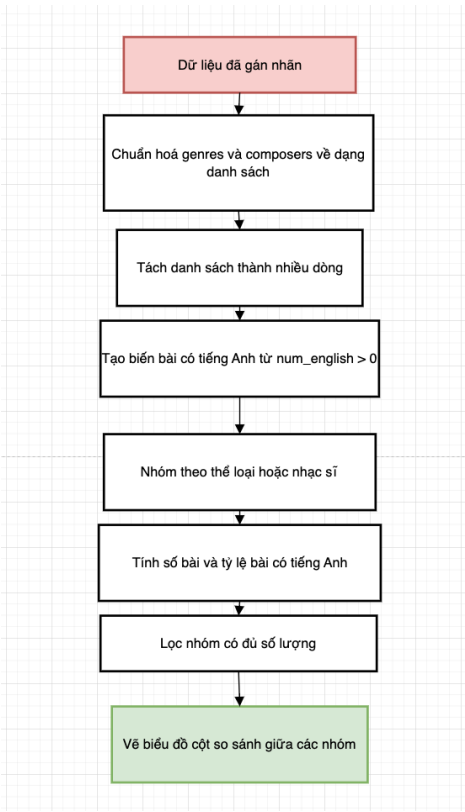
4.4.2 Dữ liệu và biến sử dụng

Phân tích sử dụng dữ liệu đã gán nhãn và có các trường genres, composers, số bài hát tiếng anh. Yếu tố tiếng Anh được đo bằng biến nhị phân, xác định bài hát có tiếng Anh khi bài hát tiếng anh lớn hơn 0. Cách đo này phản ánh mức độ xuất hiện theo bài, không phản ánh cường độ sử dụng trong từng bài.

Do genres và composers được lưu dưới dạng danh sách, một bài hát có thể thuộc nhiều thể loại hoặc có nhiều nhạc sĩ. Khi tổng hợp, nhóm tách danh sách này thành nhiều dòng để thống kê theo từng nhãn.

4.4.3 Quy trình tổng hợp và trực quan hóa

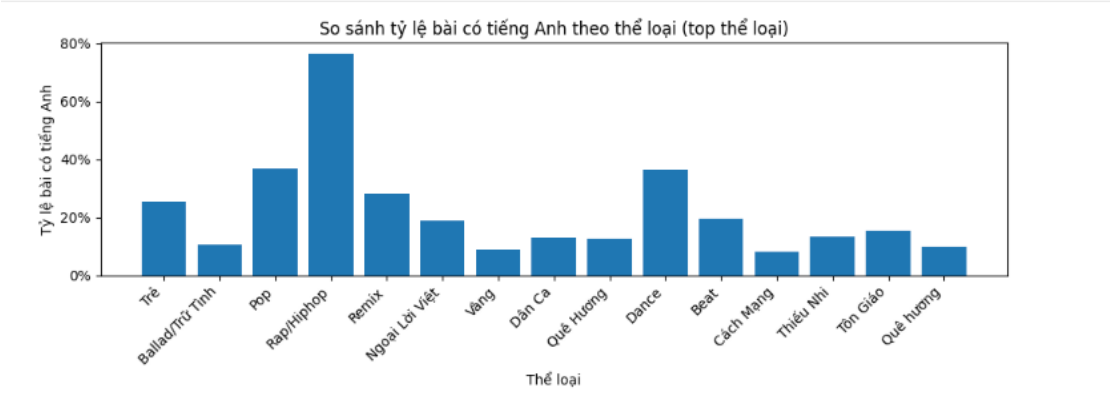
Quy trình xử lý được thực hiện theo cùng một khuôn cho cả hai trục thể loại và tác giả: chuẩn hoá trường danh sách, tách danh sách để nhóm dữ liệu, tính tỷ lệ bài có tiếng Anh theo từng nhóm, sau đó chỉ giữ các nhóm có đủ số lượng để biểu đồ ổn định.



Hình 25: Quy trình xử lý phục vụ so sánh tỷ lệ bài hát theo thể loại và tác giả

Trong bước lọc, nhóm đặt ngưỡng số lượng tối thiểu cho mỗi nhóm để hạn chế trường hợp một thể loại hoặc một nhạc sĩ có quá ít bài khiến tỷ lệ dao động mạnh và dễ gây hiểu nhầm.

4.4.4 Kết quả theo thể loại



Hình 26: So sánh tỷ lệ bài có tiếng Anh theo thể loại

Biểu đồ theo thể loại cho thấy mức độ xuất hiện tiếng Anh khác nhau rõ rệt giữa các dòng nhạc. Trong dữ liệu hiện tại, Rap/HipHop có tỷ lệ bài hát xuất hiện tiếng Anh cao nhất, tiếp theo là các nhóm như Pop và Dance; trong khi đó các thể loại mang tính truyền thống hơn hoặc thiên về trữ tình có xu hướng thấp hơn. Kết quả này gợi ý rằng việc chèn tiếng Anh không phải là hiện tượng phân bố đều trong toàn bộ thị trường âm nhạc, mà tập trung mạnh hơn ở những thể loại gắn với nhịp sống đô thị, văn hóa đại chúng và các kênh tiêu thụ nội dung hiện đại.

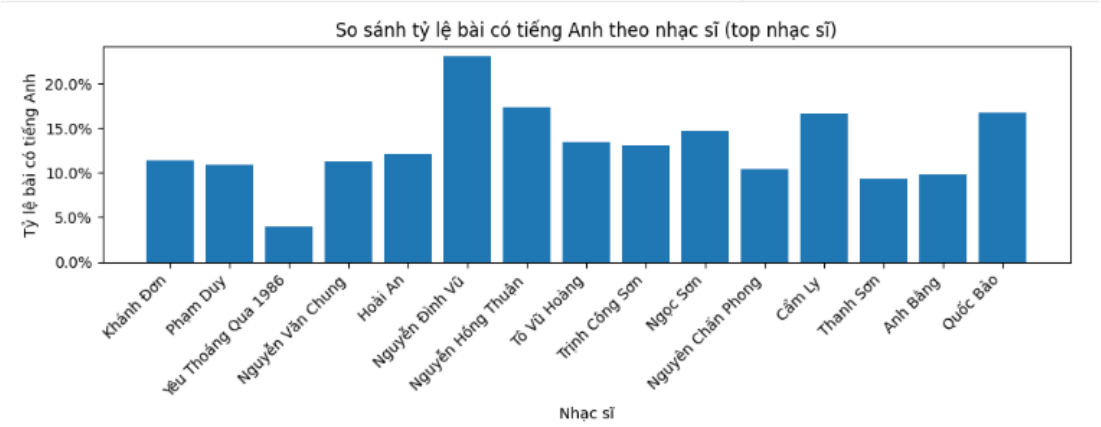
Ở góc độ xã hội và văn hoá, khác biệt theo thể loại có thể được hiểu như sự khác nhau về chức năng giao tiếp và định vị phong cách của lời bài hát. Với Rap/HipHop, tiếng Anh thường đóng vai trò như một tín hiệu nhận diện bản sắc thể loại và kết nối với văn hoá rap toàn cầu, nơi các thuật ngữ, khẩu ngữ và cách nói mang tính quốc tế xuất hiện thường xuyên trong thực hành sáng tác và biểu diễn. Với Pop và Dance, việc chèn tiếng Anh có thể gắn với mục tiêu tạo cảm giác hiện đại, bắt kịp xu hướng, tối ưu khả năng lan truyền trên nền tảng số và tăng mức độ "quốc tế hoá" của ca từ. Ngược lại, các thể loại truyền thống hoặc trữ tình thường ưu tiên tính kể chuyện, tính biểu cảm và tính gần gũi với tiếng Việt, do đó nhu cầu dùng tiếng Anh như một yếu tố thẩm mỹ hoặc biểu tượng có thể thấp hơn.

Ý nghĩa quan trọng của biểu đồ trên là làm rõ rằng ảnh hưởng ngôn ngữ nước ngoài cần được nhìn trong bối cảnh thể loại, thay vì chỉ quan sát theo thời gian. Cùng một giai đoạn, sự khác biệt giữa các thể loại có thể lớn hơn sự khác biệt giữa các năm, vì mỗi dòng nhạc có quy ước thẩm mỹ và cộng đồng người nghe khác nhau. Đồng thời, vì chỉ số đang dùng là xuất hiện tiếng Anh theo bài, hình phản ánh xu hướng "có chèn tiếng Anh hay không" trong từng dòng nhạc; để đánh giá sâu hơn về mức độ ảnh hưởng, có thể kết hợp thêm chỉ số cường độ như tỷ lệ tiếng Anh trên tổng token, nhằm phân biệt giữa việc chèn một vài từ khóa và việc sử dụng tiếng Anh với mật độ cao.

Về ứng dụng thực tiễn, kết quả này có giá trị tham chiếu cho ba nhóm đối tượng. Với người sáng tác và nhà sản xuất, đây là căn cứ định lượng để cân nhắc mức độ sử dụng yếu tố tiếng Anh phù hợp với chuẩn thẩm mỹ của thể loại và kỳ vọng của nhóm khán giả mục tiêu. Với truyền thông và phân phối nội dung, kết quả giúp xây dựng chiến lược định vị bài hát theo "tín hiệu phong cách" của lời ca, đặc biệt trong môi trường nền tảng số nơi tiêu đề và một vài cụm từ nổi bật có thể ảnh hưởng đến khả năng

tiếp cận. Với nghiên cứu văn hoá đại chúng, kết quả cung cấp một gợi ý rằng các làn sóng ảnh hưởng ngôn ngữ không diễn ra đồng loạt, mà có thể lan mạnh trong một số thể loại trước khi trở thành xu hướng rộng hơn, từ đó mở ra hướng phân tích tiếp theo theo nhóm thể loại, theo giai đoạn hoặc theo cấu trúc cộng đồng người nghe.

4.4.5 Kết quả theo nhạc sĩ



Hình 27: So sánh tỷ lệ bài có tiếng Anh theo nhạc sĩ

Biểu đồ theo nhạc sĩ được xây dựng trên nhóm nhạc sĩ có số lượng bài đủ lớn trong dữ liệu nhằm đảm bảo tỷ lệ ước lượng có độ ổn định nhất định. Kết quả cho thấy tỷ lệ bài hát có yếu tố tiếng Anh khác nhau giữa các nhạc sĩ, tuy nhiên mức chênh lệch nhìn chung không tạo ra khoảng cách lớn như khi so sánh theo thể loại. Điều này phù hợp với đặc điểm của dữ liệu: thể loại thường đại diện cách dùng ngôn ngữ tương đối ổn định, còn nhạc sĩ là một thực thể sáng tác có thể trải rộng nhiều giai đoạn và cộng tác với nhiều nghệ sĩ, khiến tỷ lệ ở cấp cá nhân thường bị trung hòa bởi độ đa dạng của tác phẩm.

Ở góc độ diễn giải, khác biệt theo nhạc sĩ gợi ý rằng việc chèn tiếng Anh không chỉ phụ thuộc vào đặc trưng thể loại mà còn liên quan đến phong cách sáng tác và bối cảnh sản xuất của từng tác giả. Một số nhạc sĩ có tỷ lệ cao hơn có thể phản ánh xu hướng ưu tiên ngôn ngữ hiện đại, khả năng tiếp cận nhóm nghệ sĩ hoặc thị trường có mức độ quốc tế hóa cao hơn, hoặc thường xuyên tham gia các dự án mang tính nhạc trẻ và nhạc đại chúng đương đại. Ngược lại, các nhạc sĩ có tỷ lệ thấp hơn thường phù hợp với nhóm tác phẩm thiên về cấu trúc lời truyền thống, tập trung vào tiếng Việt và tính biểu cảm thuần Việt.

Ý nghĩa của biểu đồ theo nhạc sĩ là bổ sung một lớp phân tích ở cấp tác giả, giúp nhận diện sự khác biệt về thói quen ngôn ngữ trong sáng tác. Khi kết hợp với phân tích theo thể loại, biểu đồ này hỗ trợ phân biệt hai tình huống: khác biệt do nhạc sĩ thực sự có phong cách riêng về ngôn ngữ, hay khác biệt chủ yếu do nhạc sĩ đó tập trung vào một số thể loại vốn có tỷ lệ tiếng Anh cao.

Về ứng dụng thực tiễn, kết quả theo nhạc sĩ có thể dùng làm tiêu chí chọn mẫu khi cần phân tích sâu theo hướng định tính hoặc phân tích trường hợp. Chẳng hạn, nhóm có thể chọn một số nhạc sĩ thuộc nhóm tỷ lệ cao và một số nhạc sĩ thuộc nhóm tỷ lệ thấp, sau đó đối chiếu cách chèn tiếng Anh trong cấu trúc lời, vị trí xuất hiện, hoặc mối liên hệ với chủ đề và phong cách biểu đạt.

4.5 Các ngôn ngữ nước ngoài khác và phân bố phiên âm theo nhóm ngôn ngữ

4.5.1 Mục tiêu

Mục tiêu của mục này là mở rộng phân tích sang các nhóm ngôn ngữ nước ngoài khác ngoài tiếng Anh và Hán Việt, dựa trên dữ liệu phiên âm đã được gán nhãn theo từng nhóm ngôn ngữ. Phần này trả lời hai câu hỏi thực nghiệm: những nhóm ngôn ngữ nào xuất hiện thường xuyên hơn trong tập dữ liệu, và mức độ xuất hiện đó thay đổi ra sao theo giai đoạn.

4.5.2 Dữ liệu và biến sử dụng

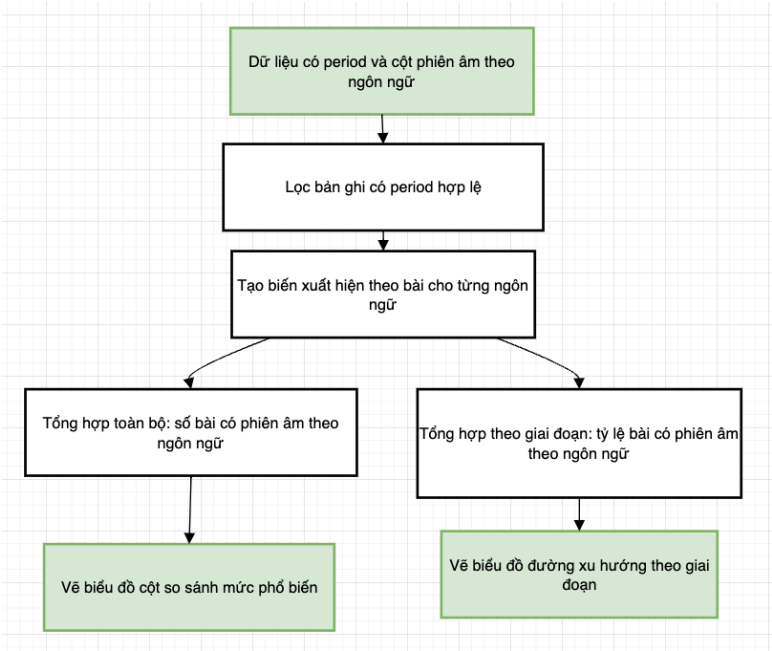
Phân tích sử dụng tập dữ liệu đã có giai đoạn thời gian và đã tách các cột phiên âm theo nhóm ngôn ngữ. Mỗi cột phiên âm biểu diễn danh sách cụm từ phiên âm của một nhóm ngôn ngữ trong một bài hát. Nếu bài không có phiên âm của nhóm đó, ô dữ liệu để trống.

Để so sánh ổn định theo nhóm, nhóm sử dụng biến xuất hiện theo bài. Một bài được xem là có phiên âm của một nhóm ngôn ngữ khi cột phiên âm tương ứng không rỗng. Cách đo này phản ánh mức độ phổ biến theo bài, không đo cường độ sử dụng trong một bài cụ thể.

4.5.3 Quy trình tổng hợp và trực quan hóa

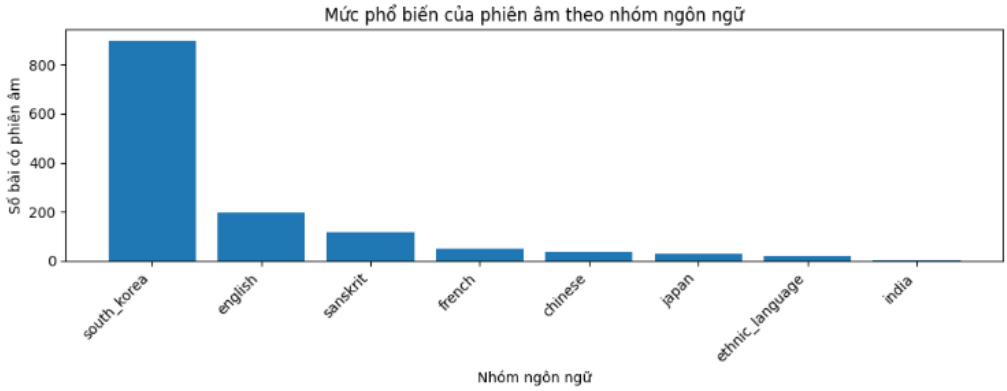
Quy trình gồm hai bước chính. Bước thứ nhất tổng hợp số lượng bài có phiên âm theo từng nhóm ngôn ngữ trên toàn bộ tập dữ liệu có giai đoạn. Bước thứ hai tổng hợp theo giai đoạn, tính tỷ lệ bài có phiên âm cho từng nhóm ngôn ngữ, từ đó vẽ xu

hướng theo thời gian. Vì số lượng nhóm ngôn ngữ có thể nhiều, nhóm chỉ trực quan hóa các nhóm xuất hiện nhiều nhất để biểu đồ dễ đọc.



Hình 28: Quy trình tổng hợp phục vụ phân tích tỉ trọng ngôn ngữ nước ngoài khác và phân bố phiên âm theo nhóm ngôn ngữ

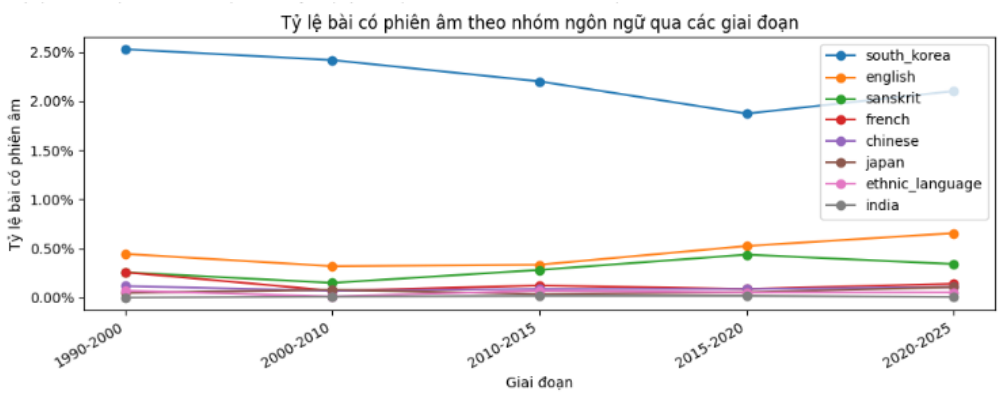
4.5.4 Kết quả và hình trực quan hoá



Hình 29: Biểu đồ cột thể hiện tỉ trọng ngôn ngữ nước ngoài khác và phân bố phiên âm theo nhóm ngôn ngữ

Hình trên trình bày mức phổ biến của phiên âm theo nhóm ngôn ngữ, đo bằng số lượng bài hát có xuất hiện ít nhất một cụm phiên âm thuộc nhóm đó. Kết quả cho thấy nhóm phiên âm tiếng Hàn chiếm ưu thế rõ rệt và cao hơn nhiều so với các nhóm còn lại. Nhóm phiên âm tiếng Anh đứng thứ hai, sau đó là Sanskrit, tiếng Pháp, tiếng Trung, tiếng Nhật, nhóm ngôn ngữ dân tộc và nhóm Ấn Độ. Chênh lệch lớn giữa nhóm tiếng Hàn và các nhóm còn lại cho thấy tín hiệu phiên âm trong tập dữ liệu hiện tại tập

trung mạnh vào một hướng ảnh hưởng chính, trong khi các hướng khác xuất hiện ở mức thấp hơn.



Hình 30: Biểu đồ tỷ lệ bài hát có phiên âm theo nhóm ngôn ngữ qua các giai đoạn

Hình trên thể hiện tỷ lệ bài có phiên âm theo nhóm ngôn ngữ qua các giai đoạn Trong đó tỷ lệ được tính theo phần trăm số bài trong từng giai đoạn. Kết quả cho thấy nhóm phiên âm tiếng Hàn duy trì mức cao nhất ở tất cả các giai đoạn và dao động quanh khoảng 2% đến 2,5%. Xu hướng của nhóm này giảm dần từ giai đoạn 1990-2000 đến 2015-2020, sau đó tăng trở lại nhẹ ở giai đoạn 2020-2025. Ngược lại, nhóm phiên âm tiếng Anh có xu hướng tăng dần về cuối chuỗi thời gian, đặc biệt rõ hơn từ giai đoạn 2015-2020 sang 2020-2025. Nhóm Sanskrit tăng lên đến giai đoạn 2015-2020 rồi giảm nhẹ ở giai đoạn 2020-2025. Các nhóm còn lại như tiếng Pháp, tiếng Trung, tiếng Nhật và nhóm ngôn ngữ dân tộc có tỷ lệ thấp và biến động nhỏ theo thời gian, thể hiện mức độ xuất hiện hạn chế trong tập dữ liệu.

Việc sử dụng hai hình là cần thiết vì mỗi hình trả lời một câu hỏi khác nhau. Hình 29 cho biết nhóm ngôn ngữ nào xuất hiện nhiều bài nhất trong toàn bộ dữ liệu, còn Hình 30 cho biết xu hướng thay đổi theo thời gian của từng nhóm. Kết hợp hai hình cho phép nhận diện đồng thời mức độ phổ biến và hướng biến động theo giai đoạn, qua đó làm rõ hình thức ảnh hưởng nước ngoài dưới dạng phiên âm trong lời bài hát.

Cần lưu ý rằng chỉ số đang dùng là xuất hiện theo bài, không phản ánh mức độ sử dụng nhiều hay ít trong từng bài. Đồng thời, các nhóm có tỷ lệ rất thấp có thể đến từ hai khả năng: hiện tượng thực tế xuất hiện ít, hoặc thư viện phiên âm của nhóm đó chưa bao phủ đầy đủ các biến thể trong dữ liệu. Vì vậy, khi cần đào sâu, nhóm có thể kết hợp

thêm chỉ số cường độ theo token và tiếp tục mở rộng thư viện phiên âm cho các nhóm còn mỏng dữ liệu.

4.6 Kết luận chương

Chương 4 đã trình bày các kết quả trực quan hoá và thảo luận dựa trên tập dữ liệu đã hoàn tất gán nhãn và trích xuất đặc trưng ở Chương 3. Trên cơ sở dữ liệu có năm phát hành hợp lệ, nhóm tiến hành phân tích theo giai đoạn thời gian và mở rộng một số thực nghiệm theo năm, theo thể loại, theo nhạc sĩ và theo nhóm phiên âm để làm rõ hơn hình thức và mức độ xuất hiện của các yếu tố ngôn ngữ nước ngoài trong lời bài hát tiếng Việt.

Kết quả tổng quan theo giai đoạn cho thấy tiếng Việt luôn chiếm tỷ trọng áp đảo trong toàn bộ lời bài hát ở tất cả các giai đoạn. Tuy nhiên, khi loại bỏ phần tiếng Việt để quan sát rõ hơn vùng không phải tiếng Việt, dữ liệu thể hiện xu hướng tăng của tỷ trọng tiếng Anh theo thời gian, trong khi tỷ trọng Hán Việt nhìn chung ổn định. Đây là cơ sở để kết luận rằng sự thay đổi quan sát được chủ yếu nằm ở phần ngôn ngữ ngoại lai và được phản ánh rõ hơn khi trực quan hoá theo cơ cấu không tính tiếng Việt.

Thực nghiệm bổ sung theo năm cho thấy tỷ lệ bài hát có xuất hiện ít nhất một token tiếng Anh biến động theo năm, nhưng xu hướng chung ở các năm gần đây là tăng rõ hơn khi nhìn bằng đường làm mượt. Kết quả này bổ sung cho phân tích theo giai đoạn bằng một thước đo khác ở cấp bài hát, phản ánh mức độ phổ biến của tiếng Anh trong toàn bộ tập bài theo thời gian, thay vì chỉ quan sát cơ cấu token trung bình.

Khi thay trục quan sát từ thời gian sang đặc trưng của dữ liệu, kết quả theo thể loại cho thấy yếu tố tiếng Anh phân bố không đồng đều giữa các dòng nhạc. Một số thể loại như Rap/HipHop, Pop và Dance có tỷ lệ bài xuất hiện tiếng Anh cao hơn đáng kể so với các thể loại truyền thống hoặc thiên về trữ tình. Kết quả theo nhạc sĩ tiếp tục cho thấy sự khác biệt ở cấp tác giả, dù mức chênh lệch nhìn chung không mạnh như theo thể loại, qua đó gợi ý rằng yếu tố tiếng Anh chịu ảnh hưởng đồng thời bởi quy ước thể loại và phong cách sáng tác.

Đối với các ngôn ngữ nước ngoài khác, phân tích phiên âm theo nhóm ngôn ngữ cho thấy nhóm phiên âm tiếng Hàn là nhóm nổi bật nhất về mức độ phổ biến và duy trì tỷ lệ cao nhất theo giai đoạn, trong khi các nhóm khác xuất hiện ở mức thấp hơn và biến động nhỏ hơn. Đồng thời, nhóm phiên âm tiếng Anh có xu hướng tăng dần về cuối

chuỗi thời gian. Hai hình trực quan hoá ở mục 4.6 giúp phân biệt rõ giữa mức độ phổ biến toàn bộ và xu hướng theo giai đoạn, qua đó bổ sung thêm một lớp bằng chứng về hướng ảnh hưởng ngôn ngữ nước ngoài trong lời bài hát.

Tổng hợp lại, Chương 4 đã hoàn thiện phần trình bày kết quả theo hướng bám sát dữ liệu và làm rõ được ba điểm chính: xu hướng theo thời gian của thành phần không phải tiếng Việt, khác biệt theo thể loại và theo nhạc sĩ đối với yếu tố tiếng Anh, và sự phân bố của phiên âm theo nhóm ngôn ngữ. Những kết quả này tạo nền để chuyển sang phần kết luận tổng quát và hạn chế nghiên cứu ở chương sau, đồng thời gợi mở các hướng mở rộng như phân tích cường độ sử dụng theo token, so sánh sâu theo thể loại hoặc kiểm định thống kê cho các khác biệt quan sát được.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận chính của nghiên cứu

Nghiên cứu đã xây dựng và khai thác một tập dữ liệu lời bài hát tiếng Việt theo hướng dữ liệu lớn, có lược đồ thống nhất và có cơ chế truy vết nguồn. Trên cơ sở tập dữ liệu dùng chung của lớp, nhóm triển khai pipeline tiền xử lý văn bản, xây dựng hệ thư viện gán nhãn và gán nhãn thành phần ngôn ngữ trong lời bài hát theo thứ tự ưu tiên. Kết quả gán nhãn được chuyển thành các cột đặc trưng và các bảng tổng hợp theo thời gian, theo nguồn gốc, theo thể loại và theo nhạc sĩ, tạo nền để trực quan hóa và diễn giải.

Từ kết quả trực quan hóa, có thể rút ra ba kết luận ở mức mô tả. Thứ nhất, tiếng Việt chiếm tỷ trọng chủ đạo trong lời bài hát ở tất cả các giai đoạn; các biến động đáng chú ý nằm ở phần không phải tiếng Việt. Thứ hai, tỷ trọng tiếng Anh trong phần không phải tiếng Việt tăng dần theo giai đoạn, và thực nghiệm theo năm cho thấy tỷ lệ bài hát có xuất hiện ít nhất một token tiếng Anh có xu hướng tăng rõ hơn ở giai đoạn gần đây khi quan sát bằng đường làm mượt. Thứ ba, yếu tố tiếng Anh phân bố không đồng đều theo thể loại và theo nhạc sĩ; ngoài tiếng Anh, nhóm phiên âm tiếng Hàn là nhóm xuất hiện nổi bật nhất trong dữ liệu phiên âm và duy trì tỷ lệ cao nhất theo giai đoạn.

5.2 Đóng góp của đề tài

Về mặt kỹ thuật và dữ liệu, đề tài đóng góp một quy trình xử lý có thể tái lập, gồm các bước chuẩn hoá văn bản, xây dựng thư viện gán nhãn, gán nhãn token và tạo bảng tổng hợp. Quy trình được tổ chức theo từng bước với các tệp trung gian, giúp kiểm tra trước-sau và truy vết khi cần.

Về mặt phân tích, đề tài cung cấp các kết quả định lượng theo nhiều trục quan sát gồm thời gian, nguồn gốc, thể loại, nhạc sĩ và nhóm phiên âm theo ngôn ngữ, tạo cơ sở tham chiếu cho các nghiên cứu tiếp theo về ngôn ngữ trong âm nhạc đại chúng.

5.3 Hạn chế

Thứ nhất, phân tích theo thời gian phụ thuộc vào độ đầy đủ và độ tin cậy của trường năm; các bản ghi thiếu năm không tham gia được vào tổng hợp theo giai đoạn và theo năm.

Thứ hai, hệ thư viện gán nhãn được xây dựng từ nhiều nguồn và có bước làm sạch chéo, tuy nhiên vẫn có khả năng còn thiếu bao phủ hoặc có trường hợp chồng lấn, đặc biệt ở nhóm phiên âm và tên riêng.

Thứ ba, một phần token vẫn thuộc nhóm chưa gán nhãn; tỷ lệ này ảnh hưởng đến độ đầy đủ của các thống kê, nhất là khi phân tích chi tiết các nhóm ngôn ngữ có tần suất thấp.

5.4 Hướng phát triển

Thứ nhất, có thể mở rộng dữ liệu và cải thiện metadata, ưu tiên tăng độ phủ và độ chính xác của trường năm, thể loại và nguồn gốc.

Thứ hai, có thể tiếp tục mở rộng và tinh chỉnh thư viện gán nhãn, đặc biệt với nhóm phiên âm theo ngôn ngữ và nhóm tiếng Anh mở rộng trong môi trường lời nhạc.

Thứ ba, có thể bổ sung các phân tích sâu hơn theo cường độ sử dụng, ví dụ kết hợp tỷ lệ token theo nhóm với các chỉ số phong cách, hoặc phân tích theo chủ đề khi có thêm lớp dữ liệu chủ đề.

Thứ tư, nếu cần tăng độ chặt chẽ, có thể bổ sung kiểm định thống kê cho một số khác biệt quan sát được theo giai đoạn hoặc theo nhóm, với điều kiện kiểm soát quy mô mẫu và giả định dữ liệu.

TÀI LIỆU THAM KHẢO

Tiếng Việt

1 Duyệt (n.d.), "Viet74K – Danh sách từ tiếng Việt chuẩn hóa", GitHub.

2 ph0ngp (n.d.), "Từ điển Hán–Việt phiên âm Pinyin, GitHub".

3 ryanphung (n.d.), "Tập hợp từ tương ứng Hán–Việt (Chinese–Hanviet Cognates)", GitHub

4 vndee (n.d.), "Tổng hợp tài nguyên Xử lý ngôn ngữ tự nhiên tiếng Việt (Awesome Vietnamese NLP)", GitHub.

5 Trường Đại học Tôn Đức Thắng (2023), Hướng dẫn thực hiện báo cáo đồ án học phần chuyên ngành Khoa học máy tính, Khoa CNTT, TP. Hồ Chí Minh.

6 Viện Công nghệ Thông tin (2023), "Báo cáo nghiên cứu ứng dụng xử lý ngôn ngữ tự nhiên tiếng Việt trong phân tích dữ liệu văn bản", Hà Nội.

Tiếng Anh

7 Groq. (2024) "Groq Cloud API Documentation", Groq Inc. Groq.com.

8 Bird, S., Klein, E., & Loper, E. (2009). "Natural Language Processing with Python." Reilly Media.

9 VanderPlas, J. (n.d.). "Python Data Science Handbook", GitHub repository.

10 Harris, C. R., Millman, K. J (2020). "Array Programming with NumPy." Nature.

11 Reitz, K. (n.d.). "Requests: HTTP for Humans".

12 Richardson, L. (n.d.). "Beautiful Soup Documentation." Crummy.com