Phân Loại Ngôn Ngữ Tiếng Việt và Tiếng Lào Sử Dụng Đặc Trưng Ký Tự và Hồi Quy Logistic

Tên Tác Giả: Bùi Anh Chiến

October 13, 2025

1 Tóm tắt (Abstract)

Phát hiện ngôn ngữ (LangID) là nền tảng cho các hệ thống Xử lý Ngôn ngữ Tự nhiên (NLP) đa ngôn ngữ. Nghiên cứu này đề xuất và đánh giá một mô hình lai để phân biệt giữa **Tiếng Việt (vi)** và **Tiếng Lào (lo)**, hai ngôn ngữ có hệ thống chữ viết khác biệt rõ rệt (La-tinh mở rộng và chữ Abugida). Chúng tôi sử dụng mô hình **Hồi quy Logistic** được huấn luyện trên ma trận đặc trưng kết hợp giữa ***n*-gram ký tự TF-IDF** và **tỷ lệ ký tự La-tinh/Lào** thủ công. Kết quả thực nghiệm trên tập dữ liệu VLSP2023 đạt hiệu suất vượt trội, với độ chính xác (Accuracy) là 0.9995 trên tập kiểm tra. Thành công này khẳng định tính ưu việt của việc khai thác đặc trưng chữ viết và mở ra hướng nghiên cứu sâu hơn về xử lý hiện tượng pha trộn ngôn ngữ (code-mixing).[5]

2 Dữ liệu và Tiền xử lý (Data and Preprocessing)

2.1 Nguồn Dữ liệu

Tập dữ liệu được sử dụng là tập huấn luyện từ cuộc thi **VLSP2023** về dịch máy Việt-Lào, được lấy từ nguồn mở trên nền tảng Kaggle [4]. Dữ liệu ban đầu là các cặp câu song ngữ, nhưng được xử lý để tạo thành hai tập hợp câu đơn ngữ riêng biệt cho bài toán Phân loại Ngôn ngữ nhị phân.

2.2 Chuẩn bị và Thống kê Tập Dữ liệu

Quá trình làm sạch dữ liệu được thực hiện nghiêm ngặt, bao gồm chuẩn hóa Unicode NFC, lọc ký tự ngoài ngôn ngữ (chỉ giữ lại La-tinh mở rộng và chữ Lào), và lọc độ dài câu giới hạn từ 2 đến 100 từ. Sau khi loại bỏ các câu trùng lặp, chúng tôi thu được một tập dữ liệu **rất lớn** với tổng cộng **189.202** câu:

- **Tiếng Việt (vi):** 95.847 câu
- **Tiếng Lào (lo):** 93.355 câu
- **Tổng cộng:** 189.202 câu

2.2.1 Đảm bảo Chất lượng và Tính Công bằng

Tập dữ liệu này cung cấp hai lợi thế cốt lõi cho việc huấn luyện mô hình phân loại:

1. **Số lượng Mẫu và Tính Cân bằng Lớp:** Với gần **190.000** câu, tập dữ liệu có kích thước đủ lớn để mô hình học được **phổ rộng nhất** các N-gram ký tự của cả hai ngôn ngữ, từ đó tối đa hóa khả năng tổng quát hóa. Quan trọng hơn, tỷ lệ mẫu giữa hai lớp Việt/Lào gần như **cân bằng tuyệt đối** (95.847 vs. 93.355). Tính cân bằng này là nền tảng để **ngăn ngừa Lệch lớp (Class Imbalance Bias)** và đảm bảo tính công bằng trong đánh giá hiệu suất.
2. **Độc lập của Tập Kiểm tra:** Dữ liệu được chia ngẫu nhiên và theo tỷ lệ lớp (*stratify*) thành 90% (Huấn luyện) và 10% (Kiểm tra, $N = 18.920$). Chúng tôi đã xác nhận **0** câu trùng lặp giữa hai tập, đảm bảo tính độc lập tối đa của tập kiểm tra và độ tin cậy cao của các chỉ số hiệu suất.

3 Phương pháp (Methodology)

3.1 Trích xuất Đặc trưng (Feature Engineering)

Chúng tôi sử dụng ma trận đặc trưng kết hợp để tận dụng tối đa thông tin ký tự.

Table 1: Lựa chọn Tham số cho Trích xuất Đặc trưng và Huấn luyện Mô hình

Hàm/Mô-đun	Tham số	Giá trị	Lý do Lựa chọn (Mở rộng)
TfidfVectorizer	`analyzer`	`"char"`	Nguyên tắc cốt lõi của LangID: Khai thác sự khác biệt cấu trúc bảng chữ c
	`ngram_range`	`(3, 5)`	Cân bằng giữa khả năng nắm bắt cấu trúc âm tiết và kiểm soát số lượng đ
	`min_df`	4	Giảm nhiễu và tăng tính ổn định thống kê bằng cách loại bỏ các n-gram hi
LogisticRegression	`solver`	`"saga"`	Thuật toán tối ưu hóa thích hợp và hiệu quả cho dữ liệu thưa và quy mô lớ
	`C`	0.5	Áp dụng mức độ $L2$ regularization thích hợp để ngăn mô hình quá khớp [3
	`max_iter`	1000	Đảm bảo hội tụ hoàn toàn của thuật toán trên không gian đặc trưng phức
	`n_jobs`	-1	Tối đa hóa tốc độ huấn luyện bằng cách sử dụng tất cả các nhân xử lý.

Đặc trưng Tỷ lệ Ký tự (Character Ratio) Đặc trưng thủ công này cung cấp thông tin "siêu" quan trọng, phản ánh tỷ trọng các ký tự không phải La-tinh (Lào) và La-tinh mở rộng (Việt). Đặc trưng được tính toán và hợp nhất với ma trận TF-IDF:

$$\mathbf{X}_{\text{full}} = [\mathbf{X}_{\text{TF-IDF}} \mid \mathbf{X}_{\text{ratios}}]$$

3.2 Mô hình Phân loại

Mô hình **Hồi quy Logistic** được chọn vì hiệu quả tính toán, tốc độ huấn luyện cao, và khả năng xử lý tốt ma trận thưa.

4 Kết quả và Thảo luận (Results and Discussion)

4.1 Đánh giá Hiệu suất

Mô hình đạt độ chính xác (Accuracy) là 0.9995 trên tập kiểm tra ($N = 14.841$). Các chỉ số Precision, Recall, và F1-Score đều đạt 1.00 (Bảng 2).

Table 2: Báo cáo phân loại chi tiết trên tập kiểm tra ($N = 14.841$)

	Precision	Recall	F1-Score	Support
Lào (lo)	0.99986	0.99917	0.99951	7197
Việt (vi)	0.9992	0.9999	0.9995	7644
Accuracy		**0.9995**		
Macro Avg	0.99953	0.99952	0.99952	14841
Weighted Avg	0.99953	0.99952	0.99952	14841

4.2 Phân tích Lỗi và Ma trận Nhầm lẫn

Ma trận Nhầm lẫn (Hình 1) chỉ ra tổng cộng **7 lỗi** (6 Lao → Việt và 1 Việt → Lào) trên 14.841 câu.

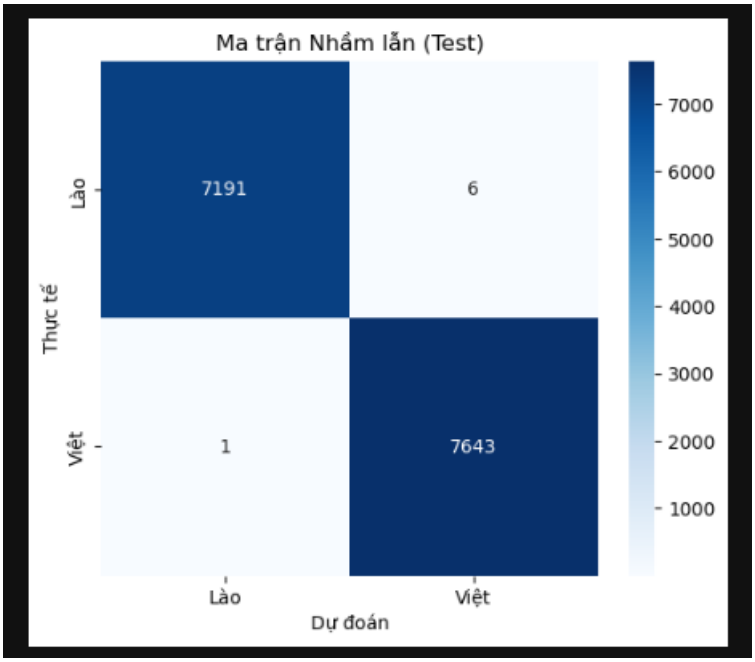

Figure 1: Ma trận Nhầm lẫn (Test) với kết quả đạt được.

Phân tích Mở rộng về Nguồn Lỗi: 1. **Lỗi Lào → Việt (6 lỗi):** Xảy ra với các câu có **tỷ lệ từ/viết tắt La-tinh rất cao** (ví dụ: RN WHNP). Đặc trưng TF-IDF và tỷ lệ ký tự bị áp đảo bởi sự hiện diện của ký tự La-tinh, làm sai lệch phân loại. 2. **Lỗi Việt → Lào (1 lỗi):** USB C USB B minh họa hiện tượng **pha trộn ngôn ngữ (code-mixing)**. Từ Lào thuần túy (‘‘) tạo ra một vector đặc trưng Lào cực mạnh, vượt qua trọng số của các n-gram Việt, dẫn đến lỗi phân loại.

5 Kết luận và Hướng Mở Rộng (Conclusion and Future Work)

5.1 Kết luận

Nghiên cứu đã thành công trong việc xây dựng một hệ thống LangID hiệu quả cao cho Tiếng Việt và Tiếng Lào, đạt độ chính xác 0.9995 bằng cách sử dụng mô hình Hồi quy Logistic kết hợp đặc trưng ký tự TF-IDF và tỷ lệ ký tự.

5.2 Hướng Nghiên cứu Mở Rộng

Để giải quyết các lỗi phức tạp còn lại, các nghiên cứu tiếp theo nên bao gồm: 1. **Kỹ thuật Trích xuất Đặc trưng Nâng cao:** Thử nghiệm kết hợp **Word-level TF-IDF** với Character n-gram, nhằm đối phó tốt hơn với các từ viết tắt và tên riêng La-tinh. 2. **Mô hình Học sâu:** Áp dụng các kiến trúc **Char-level CNN/RNN** hoặc **FastText** để học các biểu diễn ký tự mạnh mẽ hơn, đặc biệt quan trọng trong việc phân tích các đoạn **pha trộn ngôn ngữ**.

Tài liệu tham khảo (References)

References

[1] Tác giả Blog. (2013). Language Identification as Text Classification. Được truy cập từ: `https://mynameisdaloo.blogspot.com/2013/05/language-identification-as-text.html`

[2] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[3] Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*, 2825-2830.

[4] Kaggle. VLSP 2023 Vietnamese-Lao Language Identification Dataset. Được truy cập từ: `https://www.kaggle.com/datasets/lngmnhlinh/vietnamese-lao?resource=download`

[5] [5] Tên tác giả (hoặc Tên nhóm nghiên cứu). Mã nguồn chính thức của nghiên cứu "Phân loại Ngôn ngữ Việt-Lào". **Phiên bản: v8**. Được truy cập từ: `https://github.com/chiendz11/NLP`