|  |
| --- |
| **TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**  **KHOA CÔNG NGHỆ THÔNG TIN**  A blue circle with white text  AI-generated content may be incorrect.  **Xử lý ngôn ngữ tự nhiên – 7080124**  **Đề tài**: Phát hiện & sửa lỗi chính tả tiếng việt – Vietnamese spelling detection & correction  **Sinh viên thực hiện:** *1. Vũ Hoài Nam - 2221050493*  *2. Mai Thành Khải – 2221050074*  *3. Phạm Xuân Tiến - 2221050418*  **Giảng Viên Hướng Dẫn: TS. Nguyễn Thế Lộc**    **Hà Nội - 2025** |

Mục Lục

[Lời Mở Đầu 3](#_Toc199987238)

[I. Mô tả đề tài 4](#_Toc199987239)

[1. Bài toán thực tế 4](#_Toc199987240)

[2. Khó khăn chính 4](#_Toc199987241)

[a) Đặc thù tiếng Việt 4](#_Toc199987242)

[b) Thiếu dữ liệu huấn luyện 4](#_Toc199987243)

[c) Đánh đổi giữa hiệu suất và độ chính xác 4](#_Toc199987244)

[3. Ví dụ minh họa 5](#_Toc199987245)

[4. Tại sao cần giải quyết? 5](#_Toc199987246)

[II. Phân công công việc 5](#_Toc199987247)

[III. Cách tiếp cận / giải pháp 5](#_Toc199987248)

[1. Kiến Trúc Tổng Thể Hệ Thống 5](#_Toc199987249)

[2. Giải pháp kỹ thuật 5](#_Toc199987250)

[2.1 Giải Pháp Kỹ Thuật Chi Tiết 5](#_Toc199987251)

[2.1.1 Module Tiền Xử Lý 5](#_Toc199987252)

[2.1.2 Module Phát Hiện Lỗi 6](#_Toc199987253)

[2.1.3 Module Sinh Ứng Viên Sửa Lỗi 6](#_Toc199987254)

[2.2 Quy Trình Huấn Luyện 6](#_Toc199987255)

[2.2.1 Chuẩn bị dữ liệu: 6](#_Toc199987256)

[2.2.2 Huấn luyện mô hình Seq2seq: 6](#_Toc199987257)

[IV. Dự kiến kết quả đạt được 7](#_Toc199987258)

[**1.** Bảng Tóm Tắt Dự Kiến Kết Quả 7](#_Toc199987259)

[**2.** Hiệu Suất Theo Loại Lỗ 7](#_Toc199987260)

[**3.** So Sánh Với Các Giải Pháp Hiện Có 7](#_Toc199987261)

[V. Tài liệu tham khảo 8](#_Toc199987262)

# Lời Mở Đầu

Trong thời đại công nghệ thông tin phát triển mạnh mẽ như hiện nay, khối lượng dữ liệu ngôn ngữ tự nhiên do con người tạo ra ngày càng trở nên khổng lồ và đa dạng, từ văn bản, email, bình luận trên mạng xã hội cho đến văn nói trong các cuộc đối thoại hằng ngày. Việc khai thác, hiểu và xử lý hiệu quả nguồn dữ liệu này trở thành một thách thức lớn, đồng thời cũng mở ra nhiều cơ hội ứng dụng rộng rãi trong các lĩnh vực như trí tuệ nhân tạo, dịch máy, trợ lý ảo, tìm kiếm thông tin, phân tích cảm xúc, kiểm tra chính tả và nhiều hơn nữa.

**Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)** là một lĩnh vực liên ngành kết hợp giữa **khoa học máy tính, trí tuệ nhân tạo** và **ngôn ngữ học**, nhằm mục tiêu giúp máy tính hiểu, diễn giải, tạo ra và phản hồi ngôn ngữ của con người theo cách có ý nghĩa.

Thông qua môn học này, sinh viên được tiếp cận với các kiến thức nền tảng như biểu diễn văn bản, phân tích cú pháp, gán nhãn từ loại, trích xuất thực thể, mô hình ngôn ngữ, cũng như các kỹ thuật hiện đại dựa trên học máy và học sâu như RNN, Transformer, BERT,... Ngoài lý thuyết, môn học còn chú trọng thực hành và ứng dụng, giúp sinh viên xây dựng được các hệ thống xử lý văn bản thực tế như phát hiện lỗi chính tả, chatbot, phân loại văn bản, hoặc dịch máy.

Việc học và nghiên cứu môn Xử lý ngôn ngữ tự nhiên không chỉ giúp sinh viên rèn luyện tư duy thuật toán, tư duy ngôn ngữ mà còn mở ra hướng phát triển nghề nghiệp trong lĩnh vực trí tuệ nhân tạo – một xu hướng đang ngày càng đóng vai trò quan trọng trong thời đại số.

Xuất phát từ tầm quan trọng của việc xử lý và hiểu ngôn ngữ tự nhiên trong các hệ thống thông minh hiện đại, nhóm chúng em đã lựa chọn đề tài **“PHÁT HIỆN VÀ SỬA LỖI CHÍNH TẢ TIẾNG VIỆT – VIETNAMESE SPELLING DETECTION AND CORRECTION”** để nghiên cứu và triển khai.

Trong khuôn khổ báo cáo này, chúng em sẽ trình bày tổng quan về cơ sở lý thuyết liên quan, các phương pháp xử lý dữ liệu văn bản, mô hình học máy/học sâu được áp dụng, cùng với quy trình xây dựng hệ thống và đánh giá hiệu quả thực nghiệm.

Thông qua quá trình thực hiện báo cáo, chúng em không chỉ củng cố kiến thức lý thuyết đã học trên lớp mà còn có cơ hội tiếp cận thực tiễn ứng dụng NLP trong các bài toán cụ thể. Chúng em hy vọng rằng báo cáo này sẽ phần nào thể hiện được sự nỗ lực, tinh thần học hỏi và khả năng vận dụng kiến thức vào thực tế.

Mã nguồn và tài liệu dự án được có thể được tìm thấy tại GitHub: <https://github.com/devk1ngz/Vietnamese-Spelling-Correction>

# Mô tả đề tài

## Bài toán thực tế

Lỗi chính tả là một vấn đề phổ biến trong quá trình soạn thảo văn bản, đặc biệt trong tiếng Việt do sự phức tạp về:

* Phát âm gần giống (vd: "d" vs "gi" vs "r", "n" vs "l").
* Từ ghép và từ địa phương (vd: "trời" vs "giời", "vô" vs "zo").

Việc phát hiện và sửa lỗi tự động giúp:

* Nâng cao chất lượng văn bản (bài báo, luận văn, email).
* Hỗ trợ công cụ soạn thảo (Google Docs, Microsoft Word).
* Cải thiện hiệu suất các hệ thống NLP (dịch máy, chatbot, tìm kiếm).

## Khó khăn chính

### Đặc thù tiếng Việt

* Từ đồng âm, dị nghĩa:

Ví dụ: "bàn" (bàn ghế) và "bàn" (bàn bạc).

* Khó phân biệt nếu không có ngữ cảnh.
* Biến thể ngôn ngữ:
* Cách viết khác nhau giữa miền Bắc & Nam (vd: "dành" và "giành").
* Từ địa phương không có trong từ điển chuẩn.

### Thiếu dữ liệu huấn luyện

* Không có dataset công khai chứa cặp "câu sai → câu đúng" cho tiếng Việt.
* Cần tự sinh dữ liệu bằng cách:
* Tạo lỗi giả từ văn bản sạch (vd: thêm/xóa dấu, hoán đổi phụ âm).
* Thu thập từ các trang báo (VnExpess, comment website) – nhưng cần gán nhãn thủ công.

### Đánh đổi giữa hiệu suất và độ chính xác

* Rule-based (từ điển + thuật toán sửa lỗi) → Nhanh nhưng kém linh hoạt.
* Deep Learning (BERT, Transformer) → Chính xác hơn nhưng tốn tài nguyên.

## Ví dụ minh họa

|  |  |  |
| --- | --- | --- |
| Câu có lỗi | Nguyên nhân lỗi | Câu đúng |
| Tôi đánh rơi ny nước. | Nhầm "l" và "n"  (phát âm gần) | Tôi đánh rơi ly nước. |
| Tôi đáh rơi ly nước | Thiếu ký tự  (n → đánh) | Tôi đánh rơi ly nước. |
| Tôi đanh rơi ly nước. | Thiếu dấu  (dấu “sắc” 🡪 đánh) | Tôi đánh rơi ly nước. |
| Tôi đành rơi ly nước. | Sai dấu  (đành 🡪 đánh) | Tôi đánh rơi ly nước. |
| Tôi đáhh rơi ly nước. | Sai thứ tự chữ cái  (đàhn 🡪 đánh) | Tôi đánh rơi ly nước. |

## 4. Tại sao cần giải quyết?

- Giao tiếp chuyên nghiệp: Văn bản không lỗi tăng độ tin cậy.

- Hỗ trợ người học tiếng Việt: Sửa lỗi tự động giúp người nước ngoài viết đúng.

- Ứng dụng trong AI: Chatbot, voice assistant hoạt động chính xác hơn.

# Phân công công việc

|  |  |
| --- | --- |
| Họ tên | Nhiệm vụ |
| Phạm Xuân Tiến | Thu thập dữ liệu  Tiền xử lý dữ liệu  Viết báo cáo |
| Vũ Hoài Nam | Xây dựng mô hình n-grams  Xây dựng mô hình RNN  Huấn luyện mô hình |
| Mai Thành Khải | Triển khai và thử nghiệm hệ thống sửa lỗi chính tả (CorrectSpelling.ipynb), đánh giá kết quả |

# Cách tiếp cận / giải pháp

## Kiến Trúc Tổng Thể Hệ Thống

Hệ thống được thiết kế theo mô hình pipeline 5 tầng:

[Input Text] → [Tiền xử lý] → [Phát hiện lỗi] → [Sinh ứng viên] → [Lựa chọn] → [Output]

## Giải pháp kỹ thuật

### Giải Pháp Kỹ Thuật Chi Tiết

### Module Tiền Xử Lý

Các bước thực hiện:

* Chuẩn hóa Unicode:

Chuyển đổi các dạng mã hóa về chuẩn Unicode dấu riêng

* Xử lý viết hoa/viết thường:

Áp dụng lowercase cho các từ không phải danh từ riêng

* Tách từ tiếng Việt

Sử dụng Underthesea với cấu hình tối ưu

### Module Phát Hiện Lỗi

Phương pháp kết hợp 3 lớp:

1. Lớp từ điển:

* Sử dụng bộ từ điển mở rộng 150.000 từ
* Bao gồm cả từ chuyên ngành và phương ngữ

1. Lớp thống kê:

* Mô hình 5-gram được huấn luyện trên kho ngữ liệu 10GB

1. **Lớp học sâu**:

* Seq2seq kết hợp Bi LSTM

### Module Sinh Ứng Viên Sửa Lỗi

**Thuật toán cải tiến:**

1. **SymSpell tối ưu cho tiếng Việt**:
   * Edit distance tối đa = 2
   * Từ điển nén sử dụng bloom filter
2. **Phương pháp lai**:

**Tạo ra cá từ sai gần giống với từ đúng theo các trường hợp**:

* + **Thiếu chữ cái**
  + **Thừa chữ cái**
  + **Sai dấu**
  + **Thiếu dấu**
  + **Từ đồng âm**
  + **…**

### Quy Trình Huấn Luyện

### Chuẩn bị dữ liệu:

* Tập dữ liệu 50,000 cặp câu lỗi-đúng
* Cân bằng 5 loại lỗi chính:
  + **Thiếu chữ cái**
  + **Thừa chữ cái**
  + **Sai dấu**
  + **Thiếu dấu**
  + **Từ đồng âm**

### Huấn luyện mô hình Seq2seq:

* Mô hình học sâu sử dụng kiến trúc Sequence-to-sequence (Seq2seq) với hai thành phần chính:
  + Encoder: mạng Bi-LSTM 2 tầng giúp mã hoá thông tin ngữ cảnh hai chiều (trước và sau)
  + Decoder: LSTM để sinh ra chuỗi văn bản được sửa lỗi
* Dữ liệu huấn luyện gồm 50,000 cặp câu, được chia thành tập train (80%), validation (10%) và test (10%).
* Mô hình được huấn luyện với loss function là categorical cross-entropy, tối ưu bằng thuật toán Adam
* Sau huấn luyện, mô hình được lưu dưới định dạng .h5

# Dự kiến kết quả đạt được

## Bảng Tóm Tắt Dự Kiến Kết Quả

|  |  |  |
| --- | --- | --- |
| Tiêu Chí Đánh Giá | Chỉ Số Mục Tiêu | Phương Pháp Đo Lường |
| Độ chính xác | ≥ 90% | Tỷ lệ câu được sửa đúng hoàn toàn |
| Độ phủ (Recall) | ≥ 87% | Tỷ lệ phát hiện đúng tổng số lỗi |
| F1-score | ≥ 88% | Cân bằng giữa Precision và Recall |
| Thời gian xử lý | < 80ms/câu | Đo trên 1000 câu test |

## Hiệu Suất Theo Loại Lỗ

|  |  |  |
| --- | --- | --- |
| Loại Lỗi | Độ Chính Xác Dự Kiến | Mức Độ Khó |
| Thiếu dấu | 94% | Dễ |
| Thiếu chữ | 93% | Dễ |
| Thừa chữ | 91% | Trung Bình |
| Sai dấu | 92% | Dễ |
| Đồng âm khác nghĩa | 80% | Rất Khó |

## So Sánh Với Các Giải Pháp Hiện Có

|  |  |  |  |
| --- | --- | --- | --- |
| Tính năng | Mô hình Seq2seq + BiLSTM | Google Docs | VSpell |
| Độ phủ tiếng Việt | 95% | 82% | 88% |
| Sửa được từ đồng âm | Có | Không | Hạn chế |
| Tốc độ xử lý (từ/giây) | 100 | 150 | Hạn Chế |
| Hỗ trợ phương ngữ | Có | Không | Một phần |

**Kết Luận**

Hệ thống dự kiến đạt:

* Độ chính xác tổng thể **>90%**
* Xử lý **đa dạng loại lỗi** đặc thù tiếng Việt
* Tốc độ **nhanh** (<80ms/câu), phù hợp ứng dụng thực tế
* Vượt trội so với các công cụ hiện có về **khả năng sửa từ đồng âm và hỗ trợ phương ngữ**

# Tài liệu tham khảo

[1] P. Norvig, "How to Write a Spelling Corrector", 2016. [Online]. Available: https://norvig.com/spell-correct.html  
[2] Đ. Điền, "Xử lý tiếng Việt trong NLP", ĐH KHXH&NV TP.HCM, 2019.  
[3] N. V. Vui, "Ứng dụng mô hình ngôn ngữ trong sửa lỗi chính tả tiếng Việt", Tạp chí Khoa học Công nghệ, 2020.  
[4] GitHub, "SymSpell: Fast spelling correction", [Online]. Available: <https://github.com/wolfgarbe/SymSpell>  
[5] F. Chollet, "Sequence-to-sequence learning with Keras", Keras documentation, 2021. [Online]. Available: https://keras.io/examples/nlp/lstm\_seq2seq/