**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**---------------**

****

**ĐỀ CƯƠNG ĐỀ ÁN THẠC SỸ**

**Chuyên ngành: Công nghệ thông tin**

**Mã ngành:** **2241860026**

**CẢI TIẾN HIỆU SUẤT PHÂN LỚP VĂN BẢN TIẾNG VIỆT BẰNG CÁCH KẾT HỢP KỸ THUẬT TĂNG CƯỜNG DỮ LIỆU VỚI TOPO VÀ HỌC SÂU**

**GVHD : TS. PHẠM THẾ ANH PHÚ HVTH : LƯU CANG KIM LONG**

**MSHV : 2341863008**

**Lớp : 23SCT31**

TP.HCM, tháng 02/2025

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*Tp. HCM, ngày …… tháng …… năm 2025*

**Giáo viên hướng dẫn**

**NHẬN XÉT CỦA HỘI ĐỒNG XÉT DUYỆT**

*Tp.HCM, ngày …… tháng …… năm 2025*

**Hội đồng xét duyệt**

**MỤC LỤC**

[**DANH MỤC CÁC HÌNH** 5](#_Toc211168180)

[**1.** **GIỚI THIỆU** 6](#_Toc211168181)

[**1.1** **Đặt vấn đề** 6](#_Toc211168182)

[**2.** **MỤC TIÊU, NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU** 11](#_Toc211168183)

[**2.1** **Mục tiêu của đề án** 11](#_Toc211168184)

[***2.1.1*** ***Mục tiêu tổng quát*** 11](#_Toc211168185)

[***2.1.2*** ***Mục tiêu cụ thể*** 11](#_Toc211168186)

[***2.1.3*** ***Câu hỏi nghiên cứu*** 11](#_Toc211168187)

[**2.2** **Nội dung nghiên cứu** 12](#_Toc211168188)

[***2.2.1 Dataset UIT-VSFC*** 12](#_Toc211168189)

[***2.2.2 Kiến trúc hệ thống*** 12](#_Toc211168190)

[***2.2.3 Experimental Design*** 15](#_Toc211168191)

[**2.3** **Đối tượng và phạm vi nghiên cứu** 15](#_Toc211168192)

[***2.3.1 Đối tượng nghiên cứu*** 16](#_Toc211168193)

[***2.3.2*** ***Phạm vi nghiên cứu*** 16](#_Toc211168194)

[**2.4** **Phương pháp nghiên cứu** 16](#_Toc211168195)

[***2.4.1*** ***Thu thập và tiền xử lý*** 16](#_Toc211168196)

[**2.4.2** **Trích xuất đặc trưng** 16](#_Toc211168197)

[**2.4.3** **Data Augmentation** 17](#_Toc211168198)

[**2.4.4** **Evaluation** 17](#_Toc211168199)

[**Statistical Significance**: *Paired t-test (p<0.05), Cohen's d>0.5* 18](#_Toc211168200)

[**3.** **TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU** 18](#_Toc211168201)

[**3.1** **Phân lớp văn bản và mô hình ngôn ngữ** 18](#_Toc211168202)

[**3.1.1** **Transformer và BERT** 18](#_Toc211168203)

[**3.1.2 PhoBERT** 19](#_Toc211168204)

[**3.2** **Phân tích dữ liệu tô-pô (TDA)** 19](#_Toc211168205)

[**3.2.1** **Các khái niệm cơ bản** 19](#_Toc211168206)

[**3.2.2** **Persistent Homology** 21](#_Toc211168207)

[**3.2.3** **Vectorization: Persistence Images** 24](#_Toc211168208)

[**3.3** **Kỹ thuật tăng cường dữ liệu** 25](#_Toc211168209)

[**3.3.1** **Phương pháp chung** 25](#_Toc211168210)

[**3.3.2** **Đặc thù tiếng Việt** 25](#_Toc211168211)

[**3.3.3** **Phương pháp được chọn** 26](#_Toc211168212)

[**3.4** **Các nghiên cứu liên quan và khoảng trống** 26](#_Toc211168213)

[**3.4.1** **Nghiên cứu quốc tế** 26](#_Toc211168214)

[**3.4.2** **Nghiên cứu trong nước** 27](#_Toc211168215)

[**3.4.3** **Khoảng trống nghiên cứu** 27](#_Toc211168216)

[**4.** **TIẾN ĐỘ THỰC HIỆN** 28](#_Toc211168217)

[**5.** **BỐ CỤC DỰ KIẾN CỦA ĐỀ ÁN** 30](#_Toc211168218)

[**6.** **TÀI LIỆU THAM KHẢO** 34](#_Toc211168219)

**DANH MỤC CÁC HÌNH**

Hình 2.1 Kiến trúc hệ thống..................................................................................12

Hình 2.2 Kiến trúc chi tiết của mô hình đề xuất .................................................13

[Hình 3. 1 simplicial complex 1 17](#_Toc211115926)

[Hình 3. 2 : 6 điểm rời rạc 18](#_Toc211115927)

[Hình 3. 3: Biểu đồ bền bỉ và sự ra đời và kết thúc của 'Khoang' tô pô 19](#_Toc211115928)

[Hình 3. 4:Biểu đồ 2D 20](#_Toc211115929)

[Hình 3. 5: Một loạt bốn phức hợp đơn giản lồng nhau và mã vạch bền vững 0 chiều của quá trình lọc kết quả. 20](#_Toc211115930)

1. **GIỚI THIỆU**
   1. **Đặt vấn đề**

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) đã trải qua những bước phát triển vượt bậc trong hai thập kỷ qua. Từ những phương pháp truyền thống dựa trên luật và thống kê đơn giản, NLP đã chuyển mình mạnh mẽ với sự xuất hiện của học máy (Machine Learning), rồi học sâu (Deep Learning), và gần đây nhất là các mô hình ngôn ngữ lớn (Large Language Models) dựa trên kiến trúc Transformer.

**Giai đoạn 1 - Phương pháp truyền thống (trước 2010):** Các hệ thống NLP dựa vào feature engineering thủ công với Bag-of-Words, TF-IDF, và các thuật toán machine learning cổ điển như Naive Bayes, SVM. Hạn chế lớn nhất là không nắm bắt được ngữ nghĩa sâu và ngữ cảnh phức tạp của ngôn ngữ.

**Giai đoạn 2 - Deep Learning (2010-2017):** Sự ra đời của Word2Vec (2013), GloVe (2014), và các mô hình RNN/LSTM đã cho phép học biểu diễn ngữ nghĩa từ dữ liệu. Tuy nhiên, các mô hình này vẫn gặp khó khăn với long-range dependencies và xử lý tuần tự chậm.

**Giai đoạn 3 - Kỷ nguyên Transformer (2017-nay):** Kiến trúc Transformer (Vaswani et al., 2017) với cơ chế self-attention đã cách mạng hóa NLP. BERT (Devlin et al., 2019) đặc biệt nổi bật với khả năng học biểu diễn ngữ cảnh hai chiều (bidirectional) thông qua pre-training trên corpus lớn, sau đó fine-tune cho các tác vụ cụ thể. BERT đã đạt state-of-the-art trên nhiều benchmark như GLUE, SQuAD, đánh dấu bước ngoặt quan trọng trong lịch sử NLP.

**Phân lớp văn bản tiếng Việt - Thực trạng và thách thức:**

Phân lớp văn bản (text classification) là một trong những tác vụ cơ bản và quan trọng nhất của NLP, với nhiều ứng dụng thực tiễn tại Việt Nam:

* Thương mại điện tử: Phân tích cảm xúc khách hàng từ đánh giá sản phẩm trên các nền tảng như Shopee, Lazada, Tiki giúp doanh nghiệp nắm bắt phản hồi và cải thiện dịch vụ.
* Truyền thông xã hội: Theo dõi dư luận mạng xã hội (Facebook, TikTok, Zalo) để phát hiện xu hướng, quản lý khủng hoảng truyền thông.
* Giáo dục: Phân tích phản hồi của sinh viên về chất lượng giảng dạy, từ đó cải thiện chương trình đào tạo.
* Dịch vụ khách hàng: Phân loại tự động yêu cầu, khiếu nại trong hệ thống chatbot và customer service.
* Y tế: Phân loại hồ sơ bệnh án điện tử, phát hiện triệu chứng từ mô tả bệnh nhân.

Tuy nhiên, phân lớp văn bản tiếng Việt vẫn còn nhiều thách thức:

**Thách thức 1 - Đặc thù ngôn ngữ**: Tiếng Việt có 6 thanh điệu, từ ghép phong phú, cấu trúc ngữ pháp khác tiếng Anh. Các mô hình multilingual như mBERT không tối ưu cho tiếng Việt vì được train trên nhiều ngôn ngữ, làm giảm khả năng chuyên biệt hóa.

**Thách thức 2 - Dữ liệu hạn chế**: So với tiếng Anh, tài nguyên dữ liệu tiếng Việt có gán nhãn (labeled data) còn hạn chế về cả số lượng và đa dạng. Việc thu thập và gán nhãn dữ liệu chất lượng cao tốn kém thời gian và chi phí.

**Thách thức 3 - Domain-specific challenges**: Trong các domain cụ thể (giáo dục, y tế, pháp lý), văn bản thường ngắn, không formal, có nhiễu (lỗi chính tả, viết tắt), làm giảm hiệu suất mô hình.

**PhoBERT - Bước đột phá cho NLP tiếng Việt:**

Để giải quyết các thách thức trên, PhoBERT (Nguyen & Nguyen, 2020) [3] - mô hình BERT được pre-train đặc biệt cho tiếng Việt. PhoBERT có những ưu điểm nổi bật:

**Pre-trained trên corpus lớn:** 20GB văn bản tiếng Việt từ Wikipedia và tin tức, giúp học được đặc trưng ngôn ngữ phong phú.

Word-level tokenization: Sử dụng RDRSegmenter cho word segmentation, phù hợp với cấu trúc từ ghép tiếng Việt, tốt hơn subword tokenization của mBERT.

State-of-the-art performance: Đạt SOTA trên nhiều tác vụ Vietnamese NLP như POS tagging (96.7%), NER (94.0%), sentiment analysis (92.4%).

PhoBERT đã trở thành nền tảng (foundation model) quan trọng cho nghiên cứu và ứng dụng NLP tiếng Việt.

**Tuy nhiên, PhoBERT vẫn có những hạn chế cần khắc phục:**

**Hạn chế 1 - Chưa khai thác thông tin cấu trúc hình học:**

PhoBERT tập trung vào học đặc trưng ngữ nghĩa thông qua cơ chế self-attention. Mỗi layer của PhoBERT tạo ra attention maps - ma trận [L × L] biểu diễn mối quan hệ giữa các token trong câu. Attention maps chứa đựng thông tin phong phú về:

* Mối quan hệ cú pháp (syntactic relations): subject-verb, determiner-noun
* Mối quan hệ ngữ nghĩa (semantic relations): synonyms, antonyms
* Coreference resolution: đại từ chỉ về danh từ nào
* Long-range dependencies: quan hệ giữa các từ cách xa nhau

Tuy nhiên, thông tin này chỉ được sử dụng một cách ngầm định (implicit) trong quá trình học. Chúng ta chưa khai thác được cấu trúc hình học tiềm ẩn trong attention maps - tức là các patterns về độ liên thông (connectivity), chu trình (cycles), và các đặc trưng tô-pô khác có thể cung cấp insights bổ sung về cách PhoBERT xử lý ngôn ngữ.

**Hạn chế 2 - Phụ thuộc vào dữ liệu lớn:**

Như hầu hết các mô hình học sâu, PhoBERT đạt hiệu suất tốt nhất khi có nhiều dữ liệu huấn luyện. Khi dữ liệu hạn chế hoặc không đồng đều (class imbalance), hiệu suất có thể giảm đáng kể. Đây là thách thức phổ biến với tiếng Việt, đặc biệt trong các domain chuyên biệt (giáo dục, y tế, pháp lý) nơi dữ liệu gán nhãn khan hiếm và tốn kém.

**Hạn chế 3 - Thiếu góc nhìn bổ sung:**

Các đặc trưng ngữ nghĩa từ PhoBERT, dù mạnh mẽ, vẫn có thể được tăng cường bởi thông tin từ góc độ khác. Nếu chỉ dựa vào một loại features (semantic), mô hình có thể bỏ lỡ các patterns quan trọng ở tầng cấu trúc (structural level).

**Phân tích dữ liệu tô-pô (TDA) - Giải pháp tiềm năng:**

Phân tích dữ liệu tô-pô (Topological Data Analysis - TDA) là lĩnh vực toán học ứng dụng nghiên cứu hình dạng và cấu trúc của dữ liệu thông qua các đặc trưng hình học. TDA có khả năng:

**1. Trích xuất đặc trưng bền vững (persistent features):** TDA sử dụng persistent homology để phát hiện các đặc trưng tồn tại ổn định qua nhiều scales khác nhau, giúp phân biệt giữa signal (thông tin quan trọng) và noise (nhiễu).

**2. Bất biến với nhiễu:** Nhờ Stability Theorem, TDA robust với những thay đổi nhỏ trong dữ liệu, phù hợp với văn bản có nhiễu (lỗi chính tả, variations).

**3. Phân tích đa scale:** TDA tự động phân tích dữ liệu ở nhiều scales từ cục bộ đến toàn cục, không cần chọn trước một scale cụ thể.

**4. Góc nhìn hình học bổ sung:** TDA cung cấp thông tin về cấu trúc hình học (connectivity, loops, voids) mà các phương pháp học sâu truyền thống khó nắm bắt.

**Nghiên cứu tiên phong của Kushnareva et al. (2021) [1]:**

Nghiên cứu đột phá của Kushnareva và cộng sự trong bài báo "Artificial Text Detection via Examining the Topology of Attention Maps" đã chứng minh rằng phân tích cấu trúc tô-pô của attention maps trong BERT có thể cung cấp thông tin bổ sung có giá trị. Cụ thể:

* Văn bản do con người viết và văn bản do AI tạo ra có topological signatures khác biệt đáng kể trong attention patterns
* TDA features alone đạt 94.1% accuracy, vượt BERT embeddings (89.3%)
* Kết hợp BERT + TDA đạt 97.2% accuracy, chứng minh synergy mạnh mẽ
* Late layers (8-11) chứa thông tin discriminative nhất

Nghiên cứu này mở ra hướng tiếp cận mới: khai thác thông tin tô-pô từ attention maps để cải thiện hiệu suất NLP.

**Khảo sát toàn diện của Uchendu & Le (2024) [2]:**

Bài báo khảo sát "Unveiling Topological Structures in Text: A Comprehensive Survey of Topological Data Analysis Applications in NLP" đã tổng hợp các ứng dụng TDA trong NLP, chỉ ra:

* TDA đang trở thành công cụ quan trọng trong phân tích word embeddings, sentence representations, và attention mechanisms
* TDA có tiềm năng lớn trong text classification, generation, và model interpretability
* Thách thức hiện tại: computational complexity, vectorization choice
* Hướng tương lai: differentiable TDA, large-scale methods

Kỹ thuật tăng cường dữ liệu (Data Augmentation) - Giải pháp cho dữ liệu hạn chế:

Data Augmentation (DA) là kỹ thuật tạo dữ liệu huấn luyện mới từ dữ liệu gốc thông qua các phép biến đổi bảo toàn nhãn. Đối với text classification, DA đã được chứng minh hiệu quả trong:

* Tăng kích thước dataset: Tạo thêm samples từ dữ liệu gốc hạn chế
* Cải thiện tổng quát hóa: Giúp mô hình học được biểu diễn robust hơn, giảm overfitting
* Cân bằng phân phối lớp: Tăng cường các lớp thiểu số trong imbalanced datasets
* Tiết kiệm chi phí: Giảm nhu cầu thu thập và gán nhãn dữ liệu mới

Các kỹ thuật DA phổ biến:

* Synonym Replacement: Thay từ bằng từ đồng nghĩa
* Back Translation: Dịch sang ngôn ngữ trung gian rồi dịch ngược lại
* Contextual Augmentation: Sử dụng language model để thay thế từ dựa trên ngữ cảnh

Đối với tiếng Việt, cần chú ý đặc thù:

* Thanh điệu: Thay đổi thanh → đổi nghĩa ("ma" ≠ "mà" ≠ "má")
* Từ ghép: Không tách rời ("máy bay" ≠ "máy" + "tàu")
* Ngữ cảnh: Một từ nhiều nghĩa ("đánh răng" ≠ "đánh giá")

**Khoảng trống nghiên cứu và động lực cho đề án:**

Mặc dù PhoBERT, TDA, và DA đều đã được nghiên cứu riêng lẻ, chưa có nghiên cứu nào kết hợp cả ba yếu tố để cải thiện phân lớp văn bản tiếng Việt:

1. TDA chưa được áp dụng cho tiếng Việt: Tất cả các nghiên cứu TDA trong NLP đều tập trung vào tiếng Anh. Nghiên cứu hiện tại không biết liệu attention patterns của PhoBERT trên văn bản tiếng Việt có topological signatures đặc trưng hay không.
2. Attention maps của PhoBERT chưa được phân tích tô-pô: Chưa có nghiên cứu nào explore cấu trúc hình học của attention trong PhoBERT, không biết layers nào chứa thông tin discriminative nhất cho tiếng Việt.
3. Thiếu kết hợp toàn diện: Chưa ai thử kết hợp PhoBERT + TDA + DA để tạo ra synergy. Không rõ liệu TDA và DA có hiệu ứng cộng hưởng (synergy) hay chỉ là additive.
4. Trade-off chưa rõ: Không biết accuracy gain có đáng với computational cost của TDA không, đặc biệt trong bối cảnh ứng dụng thực tế.

**Đề xuất của đề án:**

**Đề án này đề xuất một phương pháp toàn diện kết hợp ba yếu tố:**

PhoBERT (semantic features) + TDA (topological features từ attention maps) + Data Augmentation (tăng cường dữ liệu phù hợp tiếng Việt) 🡺 Cải thiện hiệu suất phân lớp văn bản tiếng Việt

1. **MỤC TIÊU, NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU**
   1. **Mục tiêu của đề án**
      1. ***Mục tiêu tổng quát***

Tối ưu hóa hiệu suất phân lớp văn bản tiếng Việt thông qua tích hợp đặc trưng hình học từ TDA vào biểu diễn ngữ nghĩa của PhoBERT, kết hợp với kỹ thuật tăng cường dữ liệu.

* + 1. ***Mục tiêu cụ thể***

**Mục tiêu 1:** Nghiên cứu và triển khai phương pháp trích xuất đặc trưng tô-pô từ attention maps của PhoBERT sử dụng persistent homology.

**Mục tiêu 2:** Xây dựng mô hình tích hợp giữa đặc trưng ngữ nghĩa từ PhoBERT và đặc trưng tô-pô từ TDA, với kỳ vọng cải thiện F1-score 3-4%.

**Mục tiêu 3:** Áp dụng kỹ thuật tăng cường dữ liệu phù hợp với tiếng Việt (thanh điệu, từ ghép, ngữ cảnh).

**Mục tiêu 4:** Đánh giá hiệu quả trên UIT-VSFC dataset với 2 tasks: sentiment analysis và topic classification..

* + 1. ***Câu hỏi nghiên cứu***

**RQ1:** Attention patterns của PhoBERT trên student feedback có topological signatures đặc trưng như thế nào?

**RQ2:** Topological features có bổ sung hiệu quả cho semantic features không?

**RQ3:** TDA và DA có tạo ra synergy không? (E4 > E2 + E3?)

**RQ4:** Performance gain có đáng với computational cost không?

**RQ5:** Kết quả có consistent across 2 tasks không?

* 1. **Nội dung nghiên cứu**

***2.2.1 Dataset UIT-VSFC***

**Thông tin cơ bản:**

- Kích thước: ~16,000 Vietnamese sentences

- Domain: Student feedback on courses

- Nguồn: https://nlp.uit.edu.vn/datasets/

- Quality: κ = 0.912 (sentiment), 0.714 (topic)

- Inter-annotator agreement: 91% (sentiment), 71% (topic)

- Baseline (MaxEnt): 88% F1 (sentiment), 84% F1 (topic) [17]

**Hai tasks:**

1. **Sentiment Analysis:** 3 classes (Positive, Neutral, Negative)
2. **Topic Classification:** 10 classes (Giảng viên, Bài giảng, Phương pháp, Cơ sở vật chất, Đánh giá, Thời lượng, Độ khó, Tính thực tiễn, Tài liệu, Khác)

**Lý do chọn:**

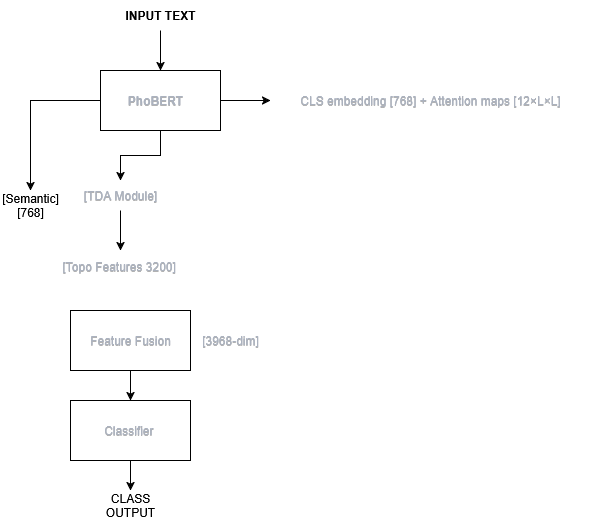
- Kích thước phù hợp cho TDA computation (~16K khả thi vs 596K không khả thi)

- 2 tasks → validation mạnh hơn

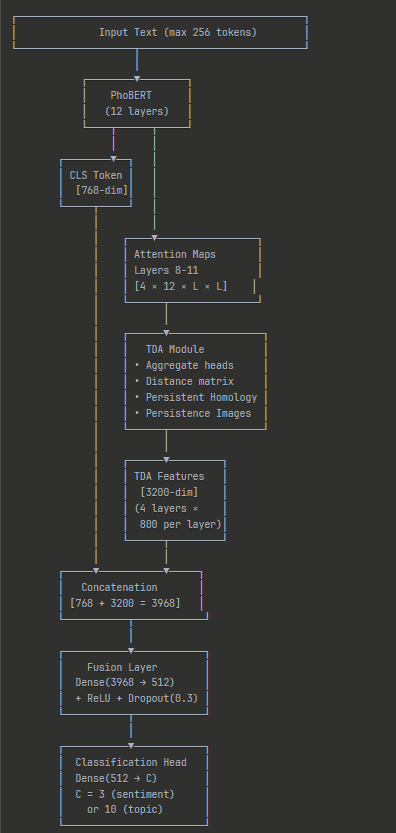
- Domain challenge: short + informal text

- Limited data → DA có giá trị

***2.2.2 Kiến trúc hệ thống***



***Hình 2.1 Kiến trúc hệ thống***



*Hình 2.2: Kiến trúc chi tiết của mô hình đề xuất*

**TDA Module - 6 bước:**

1. Select layers 8-11 (semantic information)
2. Aggregate 12 heads → mean pooling
3. Convert to distance: D = 1 - Attention
4. Build Vietoris-Rips filtration
5. Compute persistent homology (H₀, H₁)
6. Vectorize via Persistence Images (20×20) → 3200-dim

**Data Augmentation - 3 kỹ thuật:**

- Synonym Replacement (40%): Tone-aware, 10% words

- Back Translation (30%): Vi→En→Vi, quality ≥0.7

- Contextual (30%): PhoBERT MLM, mask 15%

- Total augmentation ratio: 30% of training data.

***2.2.3 Experimental Design***

**Main Experiments:**

|  |  |  |
| --- | --- | --- |
| **Id** | **Method** | **Purpose** |
| E0 | TF-IDF + SVM | Baseline truyền thống |
| E1 | PhoBERT baseline | Baseline transformer |
| E2 | PhoBERT + DA | Đánh giá DA |
| E3 | PhoBERT + TDA | Đánh giá TDA |
| E4 | PhoBERT + TDA + DA | **PROPOSED** |

**Expected Results:**

* E0: ~75% F1 (TF-IDF)
* E1: ~87% F1 (PhoBERT baseline)
* E4: ~91% F1 (Proposed) → +4% improvement

Chứng tỏ: Baseline truyền thống để chứng minh PhoBERT tốt hơn, rồi mới nói proposed method tốt hơn PhoBERT.

**Ablation Studies:**

|  |  |  |
| --- | --- | --- |
| ID | Content | Variations |
| A1 | Layer selection | 0-3, 4-7, 8-11, all |
| A2 | Homology dims | H₀, H₁, H₀+H₁ |
| A3 | Vectorization | PI 10×10, 20×20, 30×30 |
| A4 | Fusion method | Concat vs Attention |
| A5 | DA technique | SR, BT, Contextual, Combined |
| A6 | DA ratio | 10%, 30%, 50% |

* 1. **Đối tượng và phạm vi nghiên cứu**

***2.3.1 Đối tượng nghiên cứu***

* Phân lớp văn bản tiếng Việt (sentiment + topic)
* PhoBERT-base model
* TDA (persistent homology)
* Data augmentation techniques.
  + 1. ***Phạm vi nghiên cứu***
* **Ngôn ngữ:** Tiếng Việt đơn ngữ (không code-switching)
* **Dữ liệu:** CHỈ UIT-VSFC (~16K), KHÔNG NewsCategory (596K), văn bản ≤256 tokens
* **Kỹ thuật:** PhoBERT-base, Persistent Homology, 3 DA methods
* **Thời gian:** 4 tháng (2-5/2025), proof-of-concept
  1. **Phương pháp nghiên cứu**
     1. ***Thu thập và tiền xử lý***

***Thu thập:*** Download UIT-VSFC, kiểm tra tính toàn vẹn

**Tiền xử lý:**

1. **Làm sạch:** Remove HTML, normalize whitespace
2. **Chuẩn hóa:** Acronyms/emoticons (theo README.txt)
3. **Tokenization:** PhoBERT tokenizer (max\_length=256)
4. **Phân chia:** Train/Val/Test = 70/15/15 (stratified)
   * 1. **Trích xuất đặc trưng**

**Semantic features:**

* outputs = phobert(input\_ids, attention\_mask, output\_attentions=True)
* cls\_embedding = outputs.last\_hidden\_state[:, 0, :] # [768]
* attention\_maps = outputs.attentions[8:12] # layers 8-11

**Topological features:**

For each layer in [8,9,10,11]:

1. Aggregate heads: A\_agg = mean(attention\_heads)

2. Distance matrix: D = 1 - A\_agg

3. Compute PH: H₀, H₁ via Vietoris-Rips

4. Persistence Images: 20×20 → 400-dim per homology

5. Concatenate: [H₀\_img, H₁\_img] → 800-dim per layer

Total: 4 layers × 800 = 3200-dim.

* + 1. **Data Augmentation**

**Synonym Replacement:**

- Replace 10% words with Vietnamese synonyms

- Tone preservation: Only same-tone synonyms

- Compound words: Treat as single unit

**Back Translation:**

- Vi → En → Vi (Google Translate API)

- Quality check: 0.7 < BLEU < 0.95

- Label preservation verification

**Contextual Augmentation:**

- Mask 15% random words

- PhoBERT MLM predicts top-K=5

- Sample from predictions (exclude original)

* + 1. **Evaluation**

**Metrics:** Accuracy, Precision, Recall, F1-macro

**Comparison:**

- PhoBERT baseline

- PhoBERT + DA

- PhoBERT + TDA

- LSTM, CNN baselines

**Expected results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Sentiment F1** | **Topic F1** | **Improvement** |
| TF-IDF+SVM (E0) | 75-77% | 72-74% | Baseline (traditional) |
| PhoBERT (E1) | 86-88% | 76-78% | Baseline (transformer) |
| PhoBERT+DA (E2) | 88-89% | 78-79% | +1-2% |
| PhoBERT+TDA (E3) | 88-89% | 78-79% | +1-2% |
| **Proposed (E4)** | **90-92%** | **80-82%** | **+3-4%** |

*Baseline MaxEnt đạt 88% [17], nhưng PhoBERT có thể thấp hơn do difference in split hoặc evaluation protocol.*

*Mục tiêu của chúng tôi là improve từ PhoBERT baseline.*

**Statistical Significance**: *Paired t-test (p<0.05), Cohen's d>0.5*

1. **TỔNG QUAN VỀ LĨNH VỰC NGHIÊN CỨU**
   1. **Phân lớp văn bản và mô hình ngôn ngữ**

***3.1.1 Các phương pháp phân lớp***

**Phương pháp truyền thống:**

- Bag-of-Words, TF-IDF: Biểu diễn sparse, mất ngữ cảnh

- Thuật toán: Naive Bayes, SVM, Logistic Regression

- Hạn chế: Phụ thuộc feature engineering, không nắm bắt ngữ nghĩa sâu

**Deep Learning:**

- Word Embeddings: Word2Vec, GloVe, FastText

- RNN/LSTM: Xử lý sequences, nắm bắt dependencies

- CNN: Trích xuất đặc trưng cục bộ, nhanh

- Attention: Tập trung vào phần quan trọng.

* + 1. **Transformer và BERT**

**Transformer (Vaswani et al., 2017):**

- Self-attention mechanism: Nắm bắt quan hệ toàn bộ sequence

- Khả năng song song hóa cao

- Multi-head attention: Học nhiều patterns khác nhau

***Công thức attention:***

Trong đó là hệ số tỷ lệ. là số chiều của Key.

là ma trận có chiều rộng bằng với chiều rộng của ma trận Input, mục đích sử dụng của nó là để đưa kích thước của Output về bằng với kích thước của Input

**BERT (Devlin et al., 2019):**

- Bidirectional encoding

- Pre-training tasks: MLM + NSP

- Fine-tuning cho downstream tasks

**3.1.2 PhoBERT**

**Nguyen & Nguyen (2020) [3]:**

- BERT cho tiếng Việt

- Pre-trained trên 20GB corpus

- Word-level tokenization với RDRSegmenter

- Vocabulary: 64,000 word types

**PhoBERT-base:**

- 12 layers, 768 hidden, 12 heads

- ~135M parameters

- Max sequence: 256 tokens

**Hiệu suất**: SOTA trên Vietnamese NLP tasks (POS, NER, sentiment)

**Attention maps:**

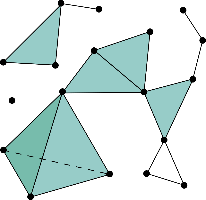
- [12 layers × 12 heads × L × L]

- Late layers (8-11) chứa semantic information

- Chưa được phân tích tô-pô cho tiếng Việt

* 1. **Phân tích dữ liệu tô-pô (TDA)**
     1. **Các khái niệm cơ bản**

**Simplicial Complex:** Đây là cấu trúc cơ bản trong tô-pô học đại số, dùng để biểu diễn không gian tô-pô dưới dạng các "khối xây dựng" đơn giản (simplex)



Hình 3. 1 simplicial complex 1

*Một phức hợp 3 đơn giản*

- 0-simplex: điểm

- 1-simplex: cạnh

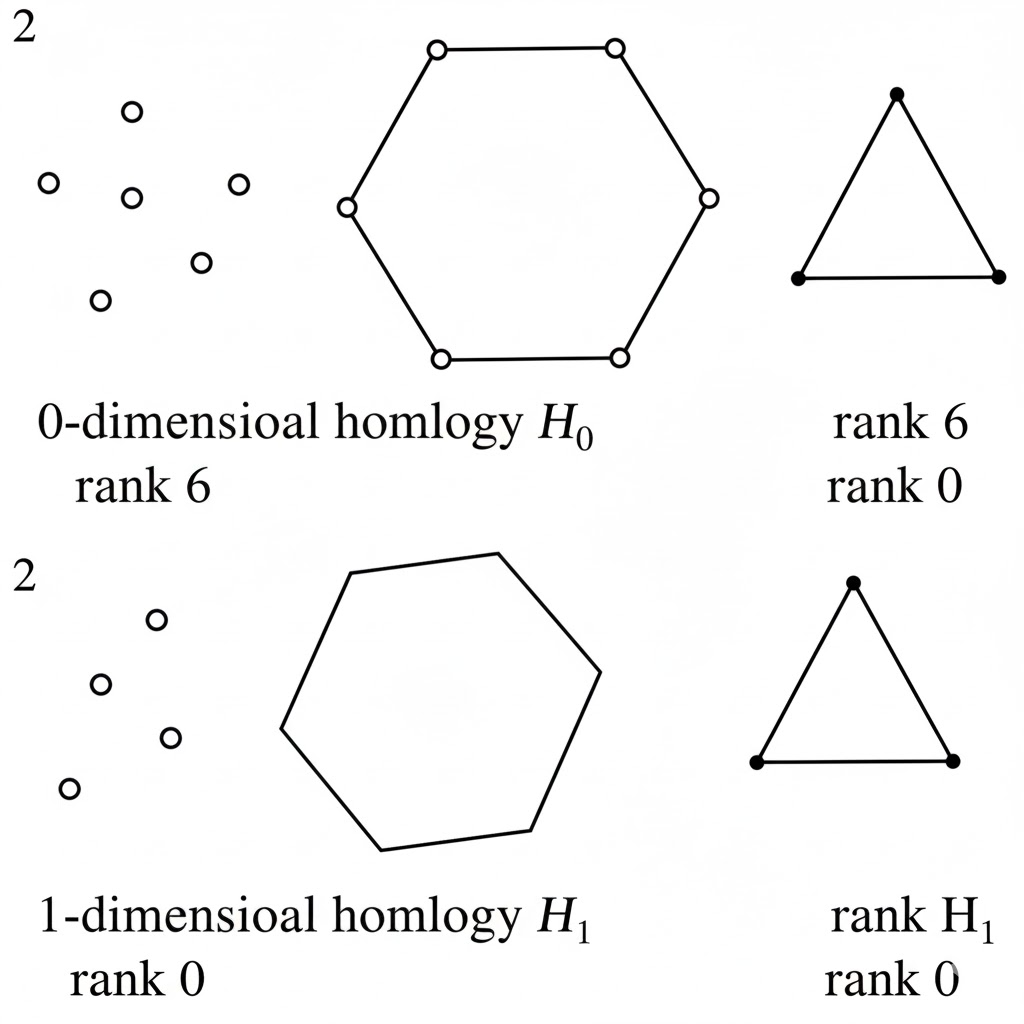
- k-simplex: Tổng quát hóa cho chiều k, ví dụ: 2-simplex là tam giác (mặt), 3-simplex là tetrahedron (khối).

Simplicial complex là tập hợp các simplex thỏa mãn điều kiện đóng (mọi mặt của simplex đều thuộc complex) và giao (hai simplex chỉ giao tại một mặt chung).

Ví dụ: Một đồ thị có thể được coi là simplicial complex với các đỉnh và cạnh.

**Homology:** Homology đo lường các đặc trưng tô-pô "bền vững" của không gian, như số lượng lỗ hổng ở các chiều khác nhau. Nó dựa trên chuỗi biên (chain complex) với toán tử biên ∂.

* : Số lượng connected components (thành phần liên thông) – đo lường bao nhiêu "mảnh" riêng biệt.
* : Số lượng loops (chu trình 1-chiều) – như vòng tròn hoặc lỗ trên mặt phẳng.

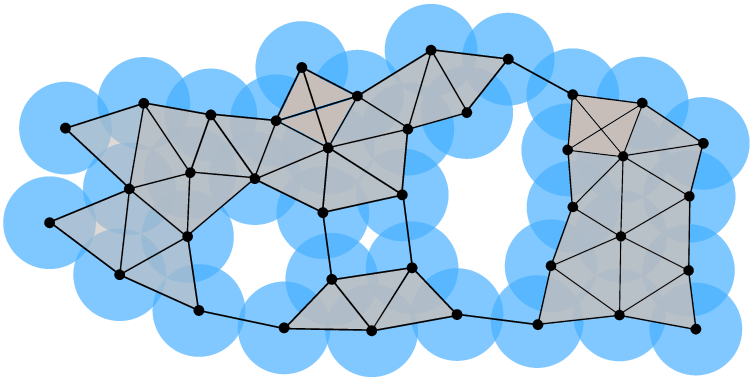


Hình 3. 2 : 6 điểm rời rạc

Công thức tổng quát cho homology group bậc n: = = Ở đây, cycles là các chuỗi không có biên (∂(c) = 0), boundaries là biên của chuỗi bậc cao hơn

**Filtration**: Filtration là cách xây dựng gradual (từng bước) một simplicial complex từ dữ liệu, thường dựa trên tham số (như khoảng cách).

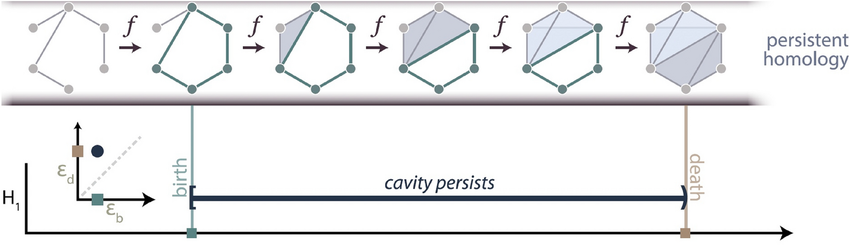
**Vietoris-Rips Complex (VR):** VR(X, ) = { | diam( diam(σ) là đường kính lớn nhất giữa các điểm trong σ. Khi ε tăng, complex "lấp đầy" hơn.



***Hình 3.3: A Vietoris-Rips complex constructed from a point cloud consisting of 38 points.***

* + 1. **Persistent Homology**

***Birth và Death:***



Hình 3. 3: Biểu đồ bền bỉ và sự ra đời và kết thúc của 'Khoang' tô pô

***- Birth (b):*** *Điểm đặc trưng topo (như cụm, vòng lặp) xuất hiện trong quá trình lọc.*

*Ví dụ: Trong đám mây điểm, birth là khi các điểm kết nối thành cụm tại một quy mô nhỏ.*

*Ý nghĩa: Birth sớm có thể là nhiễu; birth muộn thường chỉ cấu trúc lớn.*

**- Death (d):** *Đặc trưng biến mất hoặc hợp nhất*

*Ví dụ: Vòng lặp trong dữ liệu hình tròn bị lấp đầy khi quy mô lọc tăng.*

*Ý nghĩa: Death cho biết độ bền của đặc trưng.*

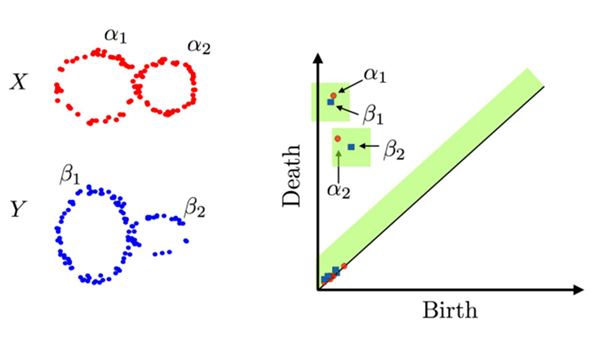
***- Persistence: p = d – b***

**Định nghĩa:** Độ bền, là khoảng cách giữa death và birth.

**Ý nghĩa:**

* **High persistence:** Đặc trưng quan trọng, phản ánh cấu trúc thực (ví dụ: khối u trong y học).
* **Low persistence:** Nhiễu, biến mất nhanh.

**Persistence Diagram:** là một biểu đồ 2D, trong đó mỗi điểm có tọa độ (birth, death) biểu diễn một đặc trưng topo.

******

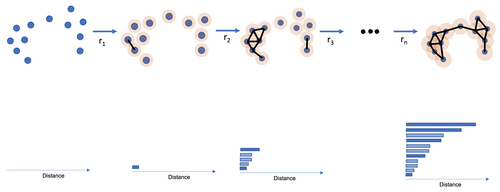
Hình 3. 4:Biểu đồ 2D

 **Trục x:** Giá trị birth (b), thời điểm đặc trưng xuất hiện.

 **Trục y:** Giá trị death (d), thời điểm đặc trưng biến mất.

 **Đặc điểm:** Tất cả các điểm nằm trên hoặc phía trên đường chéo **y = x**, vì death luôn lớn hơn hoặc bằng birth (d ≥ b).

**Barcode (Mã vạch) trong Đồng điều Bền vững:**là một biểu đồ thể hiện các đặc trưng tô-pô (lỗ, chu trình, thành phần liên thông, v.v.) được phát hiện trong dữ liệu và mô tả **"độ bền" (persistence)** của chúng khi dữ liệu được lọc qua nhiều thang đo khác nhau.



Hình 3. 5: Một loạt bốn phức hợp đơn giản lồng nhau và mã vạch bền vững 0 chiều của quá trình lọc kết quả.

**1. Mỗi feature = thanh ngang**

**Feature (Đặc trưng)**: Mỗi thanh ngang (hoặc đoạn thẳng) trên biểu đồ Barcode đại diện cho một lớp đồng điều (homology class) cụ thể, tức là một lỗ (hole) được phát hiện trong dữ liệu.

**Điểm Sinh (Birth Time, b):** Điểm bắt đầu bên trái của thanh ngang là thời điểm (thang đo lọc) mà lỗ đó sinh ra (xuất hiện) trong dữ liệu.

**Điểm Chết (Death Time, d):** Điểm kết thúc bên phải của thanh ngang là thời điểm (thang đo lọc) mà lỗ đó chết đi (bị lấp đầy).

**2. Độ dài = persistence**

**Persistence (Độ bền):** Là độ dài của thanh ngang. Nó được tính bằng công thức:

Persistence=d−b

(Điểm Chết - Điểm Sinh)

* Ý nghĩa: Persistence đo lường thời gian tồn tại (số lượng thang đo) của một đặc trưng tô-pô.
  + Thanh ngang dài (Persistence cao): Đặc trưng này bền vững qua nhiều thang đo lọc. Đây thường là các tính chất tô-pô có ý nghĩa (signal) của hình dạng dữ liệu cơ bản.
  + Thanh ngang ngắn (Persistence thấp): Đặc trưng này chỉ tồn tại trong một khoảng thang đo hẹp. Đây thường được coi là nhiễu (noise) hoặc là các đặc trưng ngẫu nhiên của tập hợp điểm.

Mục đích chính của Barcode là giúp các nhà phân tích phân biệt giữa tín hiệu và nhiễu bằng cách tập trung vào các đặc trưng có độ bền cao.

**Stability Theorem (**Định lý Ổn định trong TDA**):**

**(**

Ý nghĩa:

* ​: Persistence diagrams từ hai tập dữ liệu X,Y (ví dụ: đám mây điểm).
* ​: Khoảng cách bottleneck, đo sự khác biệt giữa ​
* Khoảng cách Gromov-Hausdorff, đo sự khác biệt hình học giữa X và Y.
* Kết luận: Thay đổi nhỏ trong dữ liệu đầu vào (X≈Y) dẫn đến thay đổi nhỏ trong persistence diagrams (​).
  + 1. **Vectorization: Persistence Images**

Persistence Images (Adams et al., 2017) là phương pháp vector hóa Persistence Diagrams trong persistent homology, chuyển đặc trưng topo thành dạng vector phù hợp cho học sâu. Báo cáo này phân tích ưu điểm của Persistence Images khi kết hợp kỹ thuật tăng cường dữ liệu topo và học sâu để cải thiện hiệu suất phân lớp văn bản tiếng Việt.

**Thuật toán:**

1. Transform: (b, d) → (b, d-b)

2. Create 2D grid: n×n (n=20)

3. Weight function: w(p) = persistence

4. Gaussian smoothing: (x, μ)

5. Normalize và flatten → [n²]-dim vector

**Công thức:**



where: 

**Ưu điểm:**

1. **Ổn Định (Stable):**

* Persistence Images sử dụng Gaussian smoothing, giảm nhạy cảm với nhiễu trong dữ liệu văn bản (ví dụ: lỗi chính tả, từ đồng nghĩa). Điều này đảm bảo đặc trưng topo (như cấu trúc ngữ nghĩa) được biểu diễn ổn định.
* **Cải thiện hiệu suất:** Ổn định giúp mô hình học sâu duy trì hiệu quả khi dữ liệu tiếng Việt có biến thể (ngữ điệu, phương ngữ).

1. **Khả Vi Phân (Differentiable):**

* Hàm Gaussian trong Persistence Images cho phép tính gradient, phù hợp với tối ưu hóa trong học sâu (như neural networks).
* **Cải thiện hiệu suất:** Kết hợp với tăng cường dữ liệu (data augmentation) như thay từ đồng nghĩa hoặc tái cấu trúc câu, Persistence Images giúp mô hình học các đặc trưng topo ngữ nghĩa phức tạp, cải thiện độ chính xác phân lớp (ví dụ: phân tích cảm xúc).

1. **Phù Hợp Học Sâu (Deep Learning Compatibility):**

* Persistence Images cung cấp vector kích thước cố định (n² chiều), dễ dàng tích hợp vào các mô hình học sâu như CNN hoặc Transformer.
* **Cải thiện hiệu suất:** Khi kết hợp với tăng cường dữ liệu topo (ví dụ: tạo đồ thị từ quan hệ từ vựng), Persistence Images mã hóa cấu trúc ngữ nghĩa (như cụm từ, chủ đề), giúp phân biệt tốt hơn các lớp văn bản tiếng Việt (ví dụ: tích cực, tiêu cực, trung lập).
  1. **Kỹ thuật tăng cường dữ liệu**
     1. **Phương pháp chung**

**Wei & Zou (2019) - EDA:**

* Synonym Replacement (SR)
* Random Insertion (RI)
* Random Swap (RS)
* Random Deletion (RD)

**Kobayashi (2018) - Contextual Augmentation:**

* Sử dụng language model để predict replacements
* Context-aware, preserves grammaticality

**Edunov et al. (2018) - Back Translation:**

* Source → Pivot → Source
* Preserves meaning, creates natural variations
  + 1. **Đặc thù tiếng Việt**

**+ Thanh điệu:**

**Vấn đề:** Thay đổi thanh → đổi nghĩa

ma ≠ mà ≠ má ≠ mả ≠ mã ≠ mạ

**Giải pháp:** Chỉ thay thế từ đồng nghĩa cùng thanh

**+ Từ ghép:**

**Vấn đề:** "máy bay" không nên thành "máy tàu"

**Giải pháp:**

* Sử dụng word segmentation (ViTokenizer)
* Treat compound as single unit: "Giảng\_viên"

**+ Ngữ cảnh:**

**Vấn đề:** "đánh" có nhiều nghĩa

- "đánh răng" = brush teeth

- "đánh giá" = evaluate

**Giải pháp:** Context-aware methods (PhoBERT MLM)

* + 1. **Phương pháp được chọn**

**Bậc 1: Highly recommended cho tiếng Việt:**

* 1. Back Translation: Tone-safe, preserves meaning, natural paraphrases
  2. PhoBERT Contextual: Context-aware, high quality, Vietnamese-trained

**Bậc 2: Use with caution:**

Synonym Replacement: Simple but needs tone preservation + compound handling

**Quality Control:**

- Semantic similarity: 0.7 < cosine\_sim < 0.95

- Fluency: perplexity check

- Label preservation: Manual + automatic check

* 1. **Các nghiên cứu liên quan và khoảng trống**
     1. **Nghiên cứu quốc tế**

**TDA trong NLP:**

**Kushnareva et al. (2021) [1]:**

* Phát hiện AI-generated text via attention topology
* TDA features + BERT → 97.2% accuracy
* Chứng minh: Human vs AI có topological differences
* Late layers (8-11) most discriminative

**Uchendu & Le (2024) [2]:**

* Survey toàn diện TDA applications trong NLP
* Phân loại: Word embeddings, sentence representations, attention analysis
* Thách thức: Computational complexity, vectorization choice
* Future: Differentiable TDA, large-scale methods

**Attention Analysis:**

**Clark et al. (2019):** Attention heads học syntactic và semantic patterns

**Voita et al. (2019):** 80% heads có thể prune → focus on important heads

**Data Augmentation:**

**Wei & Zou (2019) - EDA:** Simple methods → +3-5% accuracy

**Xie et al. (2020) - UDA:** Back-translation + consistency → SOTA

* + 1. **Nghiên cứu trong nước**

**PhoBERT [3]**: BERT cho Việt, SOTA trên Vietnamese tasks

**UIT-VSFC**: ~16K student feedback, sentiment + topic labels

**Tran et al. (2021)**: DA cho Vietnamese text, back-translation most effective

* + 1. **Khoảng trống nghiên cứu**

|  |  |
| --- | --- |
| **Đã có** | **Chưa có** |
| PhoBERT cho tiếng Việt | TDA analysis của PhoBERT attention |
| TDA trong NLP (tiếng Anh) | TDA cho tiếng Việt |
| DA cho text classification | Kết hợp PhoBERT + TDA + DA |
| Attention analysis | Phân tích tô-pô attention tiếng Việt |

**Gaps cụ thể:**

1. TDA chưa áp dụng cho tiếng Việt: Chưa được khảo sát có hiệu quả không, không hiểu topological signatures của Vietnamese text
2. Chưa phân tích tô-pô PhoBERT: Attention maps chưa được explore, không biết layers nào quan trọng
3. Thiếu kết hợp toàn diện: Chưa ai kết hợp PhoBERT + TDA + DA, không biết có synergy không
4. Trade-off chưa rõ: Accuracy gain có đáng với computational cost không?

**Đóng góp của đề án**

**Khoa học:**

* ✓ Nghiên cứu đầu tiên TDA cho Vietnamese text classification
* ✓ Phân tích topological structure của PhoBERT attention
* ✓ Chứng minh synergy giữa TDA và DA
* ✓ Insights về geometric properties của Vietnamese semantic space

**Thực tiễn:**

* ✓ +3-4% F1-score improvement (expected)
* ✓ Hiệu quả với limited data
* ✓ Ứng dụng educational feedback analysis
* ✓ Open-source code và models

1. **TIẾN ĐỘ THỰC HIỆN**

**Bảng 1**: Tiến độ thực hiện của đề án

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tháng (năm 2025)  Dự kiến nội dung thực hiện | 2 | | 3 | | 4 | 5 | |
| **GIAI ĐOẠN 1: NGHIÊN CỨU & CHUẨN BỊ** | | | | | | | |
| Nghiên cứu lý thuyết TDA + PhoBERT + DA |  | |  |  |  |  | |
| Setup môi trường (Colab GPU, libraries) |  |  |  | |  |  | |
| **GIAI ĐOẠN 2: XÂY DỰNG & THỰC NGHIỆM** | | | | | | | |
| Xây dựng TDA module |  | |  | |  |  | |
| Xây dựng DA module(SR + BT + Contextual) |  | |  | |  |  | |
| Xây dựng mô hình PhoBERT+TDA+DA |  | |  | |  |  | |
| Huấn luyện và tối ưu |  | |  | |  |  |  |
| Main experiments (E1-E4) |  | |  | |  |  | |
| Ablation studies (A1-A6) |  | |  | |  |  | |
| **GIAI ĐOẠN 3: ĐÁNH GIÁ & HOÀN THIỆN** | | | | | | | |
| Statistical testing và analysis |  | |  | |  |  | |
| Visualization và error analysis |  | |  | |  |  |  |
| Viết luận văn |  | |  | |  |  | |
| Hoàn thiện và nộp |  | |  | |  |  |  |

**Chú thích:** ████ = 50% công việc, ████████ = 100% công việc

* 1. **Milestone và Deliverables**

**Milestone 1 - Cuối tháng 2/2025**

**Deliverables:**

* ✓ Literature review hoàn chỉnh (20-30 papers)
* ✓ UIT-VSFC dataset analyzed với EDA
* ✓ Colab environment setup với libraries
* ✓ PhoBERT baseline code working

**Criteria:** Hiểu rõ lý thuyết TDA, baseline F1 trong expected range

**Milestone 2 - Cuối tháng 3/2025**

**Deliverables:**

* ✓ TDA module complete (PH computation + PI vectorization)
* ✓ DA module complete (3 techniques implemented)
* ✓ Augmented data (30% ratio, quality checked)
* ✓ Baseline (E1) results

**Criteria:** TDA output correct dimensions, DA quality metrics acceptable

**Milestone 3 - Cuối tháng 4/2025**

**Deliverables:**

* ✓ Full model PhoBERT+TDA+DA integrated
* ✓ Main experiments (E1-E4) completed
* ✓ Preliminary results showing improvement
* ✓ Statistical tests (p-value, Cohen's d)

**Criteria:** E4 outperforms E1-E3, statistically significant

**Milestone 4 - Cuối tháng 5/2025**

**Deliverables:**

* ✓ Ablation studies (A1-A6) completed
* ✓ Luận văn hoàn chỉnh (30-50 trang)
* ✓ Code repository documented
* ✓ Presentation slides

**Criteria:** All RQ1-RQ5 answered, thesis meets format requirements

* 1. **Risk Management**

|  |  |  |
| --- | --- | --- |
| **Risk** | **Impact** | **Mitigation** |
| TDA computation quá chậm | High | Pre-compute + cache, GPU acceleration |
| Augmented data chất lượng thấp | Medium | Quality control (5% manual check), filtering |
| Model không converge | Medium | Hyperparameter tuning, learning rate schedule |
| Colab timeout | Medium | Checkpoint frequently, use Colab Pro |
| Improvement không significant | High | Tune TDA parameters, try different layers |

1. **BỐ CỤC DỰ KIẾN CỦA ĐỀ ÁN**

Bố cục dự kiến của đề án gồm 5 chương

**Chương 1: Tổng quan**

* 1. Hoàn cảnh nghiên cứu

- Phát triển của NLP và phân lớp văn bản

- Tình hình nghiên cứu trong nước và quốc tế

- Vai trò của TDA và DA trong NLP hiện đại

1.2 Tính cấp thiết của đề án

- Hạn chế của các phương pháp hiện tại

- Nhu cầu cải thiện phân lớp văn bản tiếng Việt

- Tiềm năng của phương pháp đề xuất

1.3. Mục tiêu và phạm vi nghiên cứu

- Mục tiêu tổng quát và cụ thể

- Câu hỏi nghiên cứu (RQ1-RQ5)

- Phạm vi nghiên cứu và giới hạn

1.4. Cấu trúc luận văn

- Tóm tắt nội dung các chương

**Chương 2: Cơ sở lý thuyết**

* 1. Phân lớp văn bản và Transformer

- Các phương pháp phân lớp văn bản

- Kiến trúc Transformer và BERT

- PhoBERT cho tiếng Việt

* 1. Phân tích dữ liệu tô-pô

- Simplicial complex và homology

- Persistent homology và stability

- Vectorization methods (Persistence Images)

* 1. Kỹ thuật tăng cường dữ liệu

- Các phương pháp DA trong NLP

- Đặc thù tiếng Việt (thanh điệu, từ ghép)

- Quality control metrics

* 1. Các nghiên cứu liên quan

- TDA trong NLP [1, 1a, 2]

- PhoBERT và ứng dụng [3]

- Khoảng trống nghiên cứu

**Chương 3: Phương pháp giải quyết vấn đề**

* 1. Tổng quan phương pháp

- Kiến trúc tổng thể hệ thống

- Luồng xử lý dữ liệu

- Các module chính

* 1. Dataset UIT-VSFC

- Mô tả dataset chi tiết

- Phân tích thống kê (EDA)

- Preprocessing pipeline

* 1. TDA Module

- Trích xuất attention maps

- Compute persistent homology

- Persistence Images vectorization

- Implementation details

* 1. Data Augmentation Module

- Synonym Replacement (tone-aware)

- Back Translation (quality control)

- Contextual Augmentation (PhoBERT MLM)

- Augmentation pipeline

* 1. Model Architecture

- Feature fusion strategy

- Classification head

- Loss function và optimization

* 1. Training và Evaluation

- Training configuration

- Hyperparameters

- Evaluation metrics

**Chương 4: Thực nghiệm và đánh giá kết quả**

* 1. Môi trường thực nghiệm

- Hardware và software

- Libraries và dependencies

- Reproducibility settings

* 1. Main Experiments

- E1: PhoBERT baseline

- E2: PhoBERT + DA

- E3: PhoBERT + TDA

- E4: PhoBERT + TDA + DA (Proposed)

* 1. Ablation Studies

- A1: Layer selection analysis

- A2: Homology dimensions

- A3: Vectorization resolution

- A4: Fusion method comparison

- A5: DA technique comparison

- A6: DA ratio analysis

* 1. Kết quả và phân tích

- Performance comparison table

- Statistical significance testing

- Confusion matrices

- Visualization (attention maps, persistence diagrams)

* 1. Error Analysis

- Failure cases analysis

- Model behavior insights

- Limitations

* 1. Computational Cost Analysis

- Training time

- Inference time

- Memory usage

- Trade-off analysis

* 1. Trả lời câu hỏi nghiên cứu

- RQ1: Topological signatures

- RQ2: TDA contribution

- RQ3: Synergy between TDA and DA

- RQ4: Cost-benefit analysis

- RQ5: Consistency across tasks

**Chương 5: Kêt luận và hướng phát triển**

5.1. Kết luận

- Tóm tắt đóng góp chính

- Đánh giá mức độ đạt được mục tiêu

- Ý nghĩa khoa học và thực tiễn

5.2. Hạn chế

- Phạm vi nghiên cứu

- Limitations của phương pháp

- Threats to validity

5.3. Hướng phát triển

- Scale to larger datasets

- Other Vietnamese NLP tasks

- Differentiable TDA

- Cross-lingual applications

- Real-time deployment

5.4. Khuyến nghị

- For researchers

- For practitioners

- For future work

**Tài liệu tham khảo**

Danh sách tài liệu tham khảo (30-40 papers) theo chuẩn IEEE/ACM, bao gồm:

- TDA foundations

- NLP và Transformer models

- Vietnamese NLP

- Data augmentation

- Related applications

1. **TÀI LIỆU THAM KHẢO**

**Tài liệu chính**

[1] Kushnareva, L., Cherepanova, V., Shchur, O., Bagautdinov, T., Serikov, O., Oseledets, I., & Piontkovskaya, I. (2021). "Artificial Text Detection via Examining the Topology of Attention Maps". Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 635-649. <https://aclanthology.org/2021.emnlp-main.50v2.pdf>

[1a] Cherniavskii, D. (2021). "TDA4ATD: Topological Data Analysis for Artificial Text Detection - Source Code Implementation". GitHub Repository. <https://github.com/danchern97/tda4atd> (Accessed: October 2025)

[2] Uchendu, A., & Le, T. (2024). "Unveiling Topological Structures in Text: A Comprehensive Survey of Topological Data Analysis Applications in NLP". arXiv preprint arXiv:2411.10298. <https://arxiv.org/pdf/2411.10298>

[3] Nguyen, D. Q., & Nguyen, A. T. (2020). "PhoBERT: Pre-trained language models for Vietnamese". Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037-1042.

**Topological Data Analysis**

[4] Carlsson, G. (2009). "Topology and data". Bulletin of the American Mathematical Society, 46(2), 255-308.

[5] Edelsbrunner, H., & Harer, J. (2010). Computational Topology: An Introduction. American Mathematical Society.

[6] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., ... & Ziegelmeier, L. (2017). "Persistence images: A stable vector representation of persistent homology". Journal of Machine Learning Research, 18(8), 1-35.

[7] Bubenik, P. (2015). "Statistical topological data analysis using persistence landscapes". Journal of Machine Learning Research, 16(1), 77-102.

**Transformer và BERT**

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is all you need". Advances in Neural Information Processing Systems, 30.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". Proceedings of NAACL-HLT, pp. 4171-4186.

[10] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). "What does BERT look at? An analysis of BERT's attention". Proceedings of the 2019 ACL Workshop BlackboxNLP, pp. 276-286.

[11] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned". Proceedings of ACL, pp. 5797-5808.

**Data Augmentation**

[12] Wei, J., & Zou, K. (2019). "EDA: Easy data augmentation techniques for boosting performance on text classification tasks". Proceedings of EMNLP-IJCNLP, pp. 6382-6388.

[13] Kobayashi, S. (2018). "Contextual augmentation: Data augmentation by words with paradigmatic relations". Proceedings of NAACL-HLT, pp. 452-457.

[14] Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). "Understanding back-translation at scale". Proceedings of EMNLP, pp. 489-500.

[15] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). "Unsupervised data augmentation for consistency training". Advances in Neural Information Processing Systems, 33, 6256-6268.

**Vietnamese NLP**

[16] Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., & Johnson, M. (2018). "VnCoreNLP: A Vietnamese natural language processing toolkit". Proceedings of NAACL-HLT: Demonstrations, pp. 56-60.

[17] Nguyen, K. V., Nguyen, V. D., Nguyen, P. X., Truong, T. T., & Nguyen, N. L. T. (2018). "UIT-VSFC: Vietnamese students' feedback corpus for sentiment analysis". Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE), pp. 19-24.

[18] Tran, V. A., et al. (2021). "Data augmentation for Vietnamese text classification". Journal of Computer Science and Cybernetics, 37(2), 145-158.

**TDA Applications**

[19] Hofer, C., Kwitt, R., Niethammer, M., & Uhl, A. (2017). "Deep learning with topological signatures". Advances in Neural Information Processing Systems, 30.

[20] Zhu, X. (2013). "Persistent homology: An introduction and a new text representation for natural language processing". Proceedings of IJCAI, pp. 1953-1959.

[21] Rieck, B., Fugacci, U., Lukasczyk, J., & Leitte, H. (2019). "Clique community persistence: A topological visual analysis approach for complex networks". IEEE Transactions on Visualization and Computer Graphics, 24(1), 822-831.

**Deep Learning và Text Classification**

[22] Kim, Y. (2014). "Convolutional neural networks for sentence classification". Proceedings of EMNLP, pp. 1746-1751.

[23] Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory". Neural Computation, 9(8), 1735-1780.

[24] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). "Distributed representations of words and phrases and their compositionality". Advances in Neural Information Processing Systems, 26.

[25] Pennington, J., Socher, R., & Manning, C. D. (2014). "GloVe: Global vectors for word representation". Proceedings of EMNLP, pp. 1532-1543.

**Statistical Methods**

[26] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.

[27] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). "The hitchhiker's guide to testing statistical significance in natural language processing". Proceedings of ACL, pp. 1383-1392.

**Additional References**

[28] Michel, P., Levy, O., & Neubig, G. (2019). "Are sixteen heads really better than one?". Advances in Neural Information Processing Systems, 32.

[29] Vo, Q. H., & Ock, C. Y. (2020). "Intention classification using PhoBERT for Vietnamese chatbot". Applied Sciences, 10(22), 8130.

[30] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). "Enriching word vectors with subword information". Transactions of the ACL, 5, 135-146.