Đề tài: **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**

Bài báo "**Artificial Text Detection via Examining the Topology of Attention Maps**" (2021) [1]

Link: <https://aclanthology.org/2021.emnlp-main.50v2.pdf>

Bài báo khảo sát: “**Unveiling Topological Structures in Text: A Comprehensive Survey of Topological Data Analysis Applications in NLP**” [2] (2024) <https://arxiv.org/pdf/2411.10298>

Source code [1]: <https://github.com/danchern97/tda4atd>

# Mục tiêu

* Tối ưu hóa hiệu suất phân lớp văn bản tiếng Việt thông qua việc tích hợp đặc trưng hình học từ phân tích dữ liệu tô-pô (Topological Data Analysis - TDA) vào biểu diễn ngữ nghĩa của PhoBERT [3].
* Kết hợp kỹ thuật tăng cường dữ liệu với PhoBERT để cải thiện hiệu suất mô hình trong các tình huống dữ liệu hạn chế hoặc không đồng đều.

# Động lực

* PhoBERT đã chứng minh hiệu quả vượt trội trong xử lý ngôn ngữ tiếng Việt nhờ khả năng biểu diễn ngữ nghĩa mạnh mẽ. Tuy nhiên, PhoBERT vẫn bị hạn chế trong việc nắm bắt các thông tin hình học sâu hơn về cấu trúc dữ liệu.
* Phân tích tô-pô có khả năng trích xuất các đặc trưng hình học liên quan đến mối quan hệ giữa các thành phần dữ liệu, bổ sung hiệu quả cho các đặc trưng ngữ nghĩa do PhoBERT cung cấp.
* Việc kết hợp các đặc trưng tô-pô, tăng cường dữ liệu, và PhoBERT sẽ tạo ra một phương pháp toàn diện hơn cho bài toán phân lớp văn bản.

# Phương pháp nghiên cứu

## Trích xuất đặc trưng tô-pô (TDA)

* Sử dụng TDA để phân tích cấu trúc hình học của dữ liệu văn bản, tạo ra các đặc trưng như độ liên thông, chu trình bậc cao từ bản đồ attention của PhoBERT hoặc các biểu diễn ngữ nghĩa khác.
* Kết hợp đặc trưng tô-pô này với biểu diễn ngữ nghĩa từ PhoBERT.

## Huấn luyện mô hình

* Tích hợp đặc trưng tô-pô vào tầng biểu diễn cuối cùng của PhoBERT (hoặc sử dụng như một đầu vào bổ sung).
* Áp dụng phương pháp fine-tuning PhoBERT trên các tập dữ liệu cụ thể, sử dụng thêm các đặc trưng hình học.

## Đánh giá và phân tích

* Sử dụng các chỉ số như Accuracy, F1-score, Precision, và Recall để đánh giá hiệu suất.
* So sánh kết quả với các mô hình cơ bản như PhoBERT không tích hợp TDA, hoặc các mô hình học sâu khác.

# References

|  |  |
| --- | --- |
| [1] | Kushnareva, L., Cherniavskii, D., Mikhailov, V., Artemova, E., Barannikov, S., Bernstein, A., ... & Burnaev, E., "Artificial text detection via examining the topology of attention maps," in *EMNLP*, 2021. |
| [2] | Uchendu, A., & Le, T., "Unveiling Topological Structures in Text: A Comprehensive Survey of Topological Data Analysis Applications in NLP," *arXiv preprint arXiv:2411.10298,* 2024. |
| [3] | Nguyen, D. Q., & Nguyen, A. T., "PhoBERT: Pre-trained language models for Vietnamese," in *EMNLP*, 2020. |