THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://youtu.be/4jPJpdLyezU
- Link slides (dạng .pdf đặt trên Github của nhóm):
  https://github.com/tuananhdev1102/CS2205.CH201/blob/main/Anh%20Nguy%E1%BB%85n%20Tr%E1%BB%8Dng%20Tu%E1%BA%A5n%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Nguyễn Trọng Tuấn Anh
- MSSV: 250101004

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 2
- Link Github:
  https://github.com/tuananhdev1102/CS2205.CH201

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HỆ THỐNG PHÁT HIỆN ĐẠO VĂN VĂN BẢN DỰA TRÊN BIỂU DIỄN NGỮ NGHĨA VÀ MACHINE LEARNING

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

A MACHINE LEARNING–BASED SEMANTIC TEXT PLAGIARISM DETECTION SYSTEM

TÓM TẮT *(Tối đa 400 từ)*

Trong lĩnh vực giáo dục hiện đại, sự phát triển của công nghệ số dẫn đến tần suất nộp bài luận, báo cáo dưới dạng điện tử ngày một nhiều, và đi kèm với đó là vấn nạn sao chép tài liệu thiếu trung thực. Những công cụ phát hiện đạo văn truyền thống chủ yếu dựa vào kỹ thuật so sánh từ khóa hoặc chuỗi ký tự giống nhau, nên chỉ có thể xác định được các trường hợp copy nguyên văn. Đối với những hình thức tinh vi hơn như viết lại câu, thay đổi trật tự từ hay dùng từ đồng nghĩa thì các phương pháp cũ tỏ ra kém hiệu quả.

Gần đây, sự tiến bộ của Học máy và Xử lý ngôn ngữ tự nhiên đã đưa ra một hướng tiếp cận khác cho bài toán này. Thay vì chỉ xem xét mặt chữ, văn bản được chuyển thành các vector biểu diễn ngữ nghĩa, giúp nắm bắt được ý chính và nội dung sâu bên trong. Cách làm này cho phép đánh giá sự tương đồng về mặt ý nghĩa giữa các đoạn văn, từ đó nhận diện cả những trường hợp nội dung đã bị chỉnh sửa kỹ lưỡng.

Nghiên cứu này nhằm phát triển một hệ thống nhận diện văn bản sao chép sử dụng biểu diễn ngữ nghĩa kết hợp với Học máy. Hệ thống sẽ tiếp nhận văn bản cần kiểm tra cùng với kho tài liệu tham khảo, thực hiện các bước làm sạch và chuẩn hóa, rồi chuyển đổi chúng thành vector ngữ nghĩa thông qua các mô hình Học máy có sẵn. Những vector này sau đó được lưu trữ và đem ra so sánh để tính toán mức độ giống

nhau về mặt ngữ nghĩa, cuối cùng chỉ ra các phần văn bản có dấu hiệu sao chép.
Thay vì phải xây dựng và huấn luyện mô hình từ đầu, nghiên cứu tận dụng những mô hình biểu diễn ngữ nghĩa hiện có để đảm bảo tính thực tiễn. Kết quả đầu ra dự kiến là một báo cáo chi tiết thể hiện điểm số tương đồng, các đoạn văn khả nghi cùng hình ảnh minh họa trực quan. Hướng tiếp cận này kỳ vọng sẽ nâng cao hiệu quả đấu tranh với gian lận học thuật và mở đường cho những công cụ kiểm tra đạo văn thông minh hơn trong tương lai.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Ở bậc đại học và các chương trình đào tạo trực tuyến, việc đánh giá chính xác năng lực của người học thông qua bài tập dạng văn bản là rất quan trọng. Tuy nhiên, cùng với sự bùng nổ của Internet và các công cụ hỗ trợ biên tập, hành vi sao chép tài liệu đã trở nên phổ biến và khó kiểm soát. Hành vi này không chỉ làm giảm chất lượng giáo dục mà còn phá vỡ sự công bằng và các chuẩn mực trong môi trường học thuật.
Đa phần các công cụ phát hiện đạo văn hiện nay hoạt động dựa trên cơ chế so khớp từ khóa, đối chiếu chuỗi n-gram hoặc so sánh các đoạn ký tự trùng lặp. Dù có hiệu quả với việc sao chép y nguyên, chúng thường bỏ qua những hình thức đạo văn ở tầng ngữ nghĩa, chẳng hạn như diễn đạt lại ý, thay đổi cấu trúc câu hay dùng lối hành văn khác. Vì thế, vấn đề nhận diện đạo văn cần được giải quyết ở khía cạnh nội dung và ý nghĩa, chứ không dừng lại ở hình thức bề ngoài của văn bản.
Học máy và Xử lý ngôn ngữ tự nhiên cung cấp những công cụ giúp biểu diễn văn bản dưới dạng vector chứa đựng thông tin ngữ nghĩa. Thông qua các biểu diễn này, chúng ta có thể đo lường mức độ tương đồng giữa các văn bản dựa trên ý nghĩa nội tại, từ đó phát hiện được những trường hợp sao chép tinh vi. Đây được xem là hướng đi phù hợp với xu hướng phát triển của các hệ thống phân tích văn bản thông minh ngày nay.
Xuất phát từ nhu cầu thực tế trong giáo dục và những hạn chế của phương pháp truyền thống, nghiên cứu này đề xuất xây dựng một hệ thống phát hiện đạo văn vận dụng kỹ thuật Học máy. Đầu vào của hệ thống bao gồm văn bản cần đánh giá và bộ

tài liệu tham khảo, trong khi đầu ra là báo cáo thể hiện mức độ sao chép thông qua điểm số tương đồng và những đoạn văn nghi ngờ. Hệ thống được thiết kế theo một quy trình rõ ràng, nhằm đảm bảo khả năng triển khai và mở rộng trong thực tế.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Thiết lập một quy trình xử lý văn bản chuyên biệt cho nhiệm vụ phát hiện sao chép, bao gồm giai đoạn tiền xử lý, phân tách đoạn văn và chuyển đổi văn bản thành vector ngữ nghĩa bằng các mô hình Học máy.

2. Thiết kế và phát triển hệ thống có khả năng so sánh mức độ tương đồng ngữ nghĩa giữa văn bản đầu vào và văn bản tham chiếu, từ đó nhận diện được cả hành vi sao chép nguyên bản lẫn hành vi diễn đạt lại nội dung.

3. Kiểm tra hiệu suất của hệ thống thông qua các số liệu định lượng về độ tương đồng và khả năng phát hiện, đồng thời đánh giá tính khả thi khi áp dụng vào môi trường giáo dục thực tế.

NỘI DUNG VÀ PHƯƠNG PHÁP

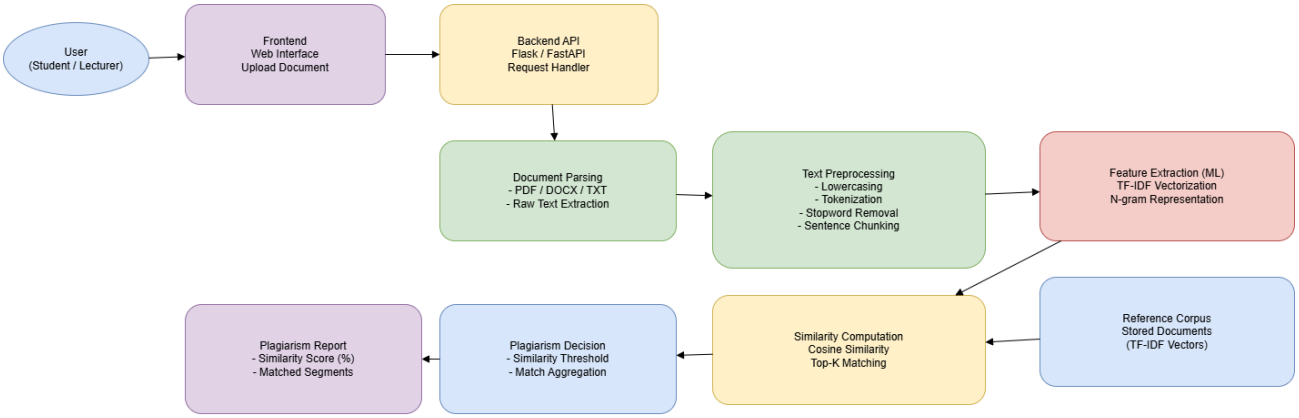
Nội dung

Nghiên cứu tập trung vào bài toán phát hiện văn bản sao chép dựa trên biểu diễn ngữ nghĩa và Học máy. Nhu cầu nghiên cứu bắt nguồn từ thực trạng gian lận học thuật ngày càng phức tạp và những điểm yếu của phương pháp truyền thống chỉ dựa trên so khớp từ khóa. Khoảng trống cần lấp đầy nằm ở việc phát hiện các hình thức đạo văn tinh vi, khi nội dung được viết lại bằng ngôn từ khác nhưng vẫn giữ nguyên ý tưởng gốc.

Nghiên cứu hướng đến việc xây dựng một hệ thống hoàn chỉnh, từ bước chuẩn bị dữ liệu văn bản cho đến khi đưa ra kết quả và hình ảnh hóa. Văn bản sẽ được chia nhỏ thành các đoạn phù hợp để nâng cao độ chính xác khi so sánh. Các biểu diễn vector

ngữ nghĩa sẽ đóng vai trò phản ánh nội dung từng đoạn, tạo nền tảng cho việc đối chiếu và tìm ra dấu vết sao chép ở cấp độ ý nghĩa.

Phương pháp



Hình 1. Sơ đồ pipeline tổng quát của hệ thống phát hiện đạo văn văn bản dựa trên Machine Learning.

Nghiên cứu áp dụng phương pháp thực nghiệm kết hợp giữa Xử lý ngôn ngữ tự nhiên và Học máy. Quy trình tổng thể của hệ thống bao gồm nhiều bước xử lý nối tiếp, mỗi bước đảm nhận một chức năng cụ thể để biến đổi dữ liệu văn bản đầu vào thành kết luận về khả năng đạo văn.

Đầu tiên, văn bản cần kiểm tra và tập văn bản tham khảo sẽ được làm sạch và chuẩn hóa để loại bỏ các yếu tố gây nhiễu. Sau đó, văn bản được phân đoạn nhằm đảm bảo việc so sánh diễn ra ở mức độ chi tiết thích hợp. Tiếp theo, mỗi đoạn văn sẽ được chuyển thành vector ngữ nghĩa nhờ các mô hình Học máy đã được huấn luyện trước. Việc sử dụng biểu diễn vector giúp mô hình hóa ý nghĩa của văn bản, thay vì chỉ phụ thuộc vào từ ngữ cụ thể.

Các vector ngữ nghĩa này sau đó được lưu trữ và đem so sánh trong không gian vector thông qua những phép đo độ tương đồng. Những đoạn văn có vector có mức độ tương đồng cao sẽ được xem là có dấu hiệu đáng nghi. Cuối cùng, hệ thống tổng hợp kết quả so sánh để tạo báo cáo, bao gồm điểm số tương đồng và danh sách các đoạn văn cần xem xét. Kết quả sẽ được trình bày một cách trực quan để hỗ trợ người dùng

trong công tác đánh giá.

KẾT QUẢ MONG ĐỢI

Luận văn kỳ vọng sẽ xây dựng được một hệ thống phát hiện văn bản sao chép dựa trên Học máy với quy trình xử lý minh bạch và có tính ứng dụng cao. Hệ thống không chỉ nhận diện được các trường hợp copy nguyên văn mà còn phát hiện được cả những hình thức đạo văn thông qua diễn đạt lại.

Về mặt kết quả cụ thể, hệ thống dự kiến cung cấp các chỉ số định lượng về mức độ tương đồng ngữ nghĩa cùng báo cáo trực quan chỉ ra các đoạn văn khả nghi. Việc kiểm tra đánh giá kết quả sẽ giúp phân tích hiệu quả của từng thành phần trong quy trình, đồng thời làm rõ tiềm năng áp dụng hệ thống vào thực tế giáo dục và đào tạo.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Christopher D. Manning, Hinrich Schütze:
Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[2]. Dan Jurafsky, James H. Martin:
Speech and Language Processing. Pearson, 3rd Edition, 2025.

[3]. Sebastian Raschka, Vahid Mirjalili:
Python Machine Learning. Packt Publishing, 3rd Edition, 2019.

[4]. Wikipedia:
Plagiarism Detection. URL: https://en.wikipedia.org/wiki/Plagiarism_detection
(truy cập 30-12-2025).