Việc sao chép và đạo văn hiện tại đang được biết tới như những vấn nạn lớn trong môi trường học thuật. Vấn đề trên có thể được diễn ra ở bất cứ đâu, từ những bài báo, những bài viết học thuật hoặc những quyển luận văn trong bất cứ trường đại học nào cho tới những bài viết và tác phẩm phẩm không được công bố.Việc sử dụng các phần mềm chống đạo văn thường không thể kiểm tra bao quát được nhiều trường hợp.

Trong bài báo của chúng tôi, chúng tôi đưa ra giải pháp đo độ tương đồng trên câu dựa trên thuật toán Levenshtein và dùng mô hình học sâu Siamese để cải thiện cho tiếng việt.

Tiền xử lý dữ liệu tách văn bản thành các câu và lưu trữ vào database của chương trình. Khi tiến hành so sánh thì sẽ tìm tập các văn bản có khả năng cao xảy ra tương đồng từ database. Sau đó dùng thuật toán Levenshtein để đo độ tương đồng giữa các câu được tách ra từ văn bản. Chúng tôi cải tiến bằng cách sử dụng mô hình học sâu để huấn luyện bộ dữ liệu tiếng việt để tang khả năng nhận biết ngữ nghĩa của các câu.

Mô hình có vai trò rất lớn trong việc đo lường độ tương đồng giữa hai văn bản. So với các phương pháp truyền thống như thuật toán Levenshtein, thì thuật toán này không thể cho độ chính xác cao nếu hai câu giống nhau về mặt ngữ nghĩa nhưng khác nhau hoàn toàn về mặt ký tự. Nhưng với mô hình Siamese Recurrent Neural Network sẽ giúp giải quyết vấn đề đó.