Ngày nay cùng với sự phát triển mạnh mẽ của công nghệ mọi nguồn dữ liệu trong các lĩnh vực dần được số hóa và mang lại cho chúng ta một nguồn tài nguyên phong phú để có thể tận dụng khai thác. Vì vậy những năm gần đây việc phân tích khai thác dữ liệu ngày được chú trọng và phát triển hơn, các đề tài nghiên cứu ra đời ngày càng nhiều. Tuy nhiên số lượng các nghiên cứu tại Việt Nam vẫn còn hạn chế đặc biệt là ở lĩnh vực giáo dục, vì vậy tôi quyết định chọn hướng nghiên cứu này cho luận văn của mình.

Luận văn này nghiên cứu về kỹ thuật khai thác dữ liệu bằng các phương pháp khác nhau và ứng dụng vào bài toán thực tế trong lĩnh vực giáo dục. Cụ thể tôi đã tập trung nghiên cứu về các khái niệm xử lý văn bản, tóm tắt văn bản, phân loại văn bản đồng thời tìm hiểu những phương pháp phân tích văn bản, phân lớp văn bản khác nhau để áp dụng vào cụ thể bài toán của luận văn. Đặc biệt là tập trung nghiên cứu về kỹ thuật phân tích xử lý văn bản và phân lớp dữ liệu.

Phương pháp tiếp cận bài toán phân tích xử lý văn bản và phân lớp dữ liệu tôi trải qua các bước sau:

Bước 1: Thực hiện việc tách câu từ từ các tập dữ liệu.

Bước 2: Mô hình hóa các câu từ thành vector.

Bước 3: Chạy huấn luyện và phân lớp dữ liệu qua các phương pháp phân lớp khác nhau.

Với cách tiếp cận trên của tôi và áp dụng trực tiếp vào bài toán phân tích ý kiến khảo sát đánh giá giảng viên tại trường Đại học Công nghệ TP.HCM tôi đã thu về một số kết quả nhất định. Sau quá trình thực nghiệm bài toán tôi đã xây dựng một bộ phân lớp dữ liệu đánh giá giảng viên đáng tin cậy với độ chính xác lên đến ~75% và rút ra được những nhận xét về ưu nhược điểm của các phương pháp phân lớp.

Kết quả này luận văn đã đóng góp một bộ phân lớp dữ liệu mới trong lĩnh vực giáo dục cụ thể là bộ phân lớp ý kiến đánh giá giảng viên và có thể áp dụng vào các ứng dụng thực tế. Bên cạnh đó luận văn còn là nguồn tài liệu tham khảo về ưu nhược điểm của một số phương pháp phân lớp trên cùng tập dữ liệu qua đó quyết định lựa chọn phương pháp phù hợp.

Luận văn này bao gồm 5 chương – trình bày chi tiết các ý tưởng, phương thức thực hiện, các thực nghiệm và kết luận cũng như hướng phát triển tiếp theo cho đề tài.