Data Preprocessing

Feature Extraction

Predictive Model

Expected Label

New

Document

**Phân loại**

Training Text/

Document

Data Preprocessing

Feature Extraction

Supervised Learning Algorithm/

Classification Algorithm

Labels

**Huấn luyện**

Mô hình tổng quát hệ thống

Mô hình tổng quan cách làm cho bài toán

Raw data

Văn bản có cấu trúc

Feature Vector

Result

Predictive Model

Data preprocessing

Feature extraction

Training

Evaluation

tiền xử lý dữ liệu

Văn bản thô

Xóa code HTML

Loại bỏ từ dừng

Tách từ và đưa về văn bản thường

Chuẩn hóa dấu từ

Chuẩn hóa dấu câu tiếng Việt

Xóa khoảng trắng, ký tự đặc biệt

Chuẩn hóa bảng mã Unicode

Văn bản có cấu trúc

Mô hình tách từ

Chuẩn hóa văn bản

Chuyển về văn bản thường

Phân tích cú pháp

Danh sách các từ

Tách từ