1. Tiền xử lý dữ liệu

+ Sử dụng thư viện nltk

+ Loại bỏ tất cả các ký tự đơn

+ Loại bỏ tất cả các ký tự đơn từ bắt đầu

+ Thay thế nhiều khoảng trắng vào thành một khoảng trắng

+ Tách từ

+ Chuyển sang dạng chữ thường

1. Trộn dữ liệu

+ Tránh trường hợp mất cân bằng label cho dữ liệu train

1. Bóc tách feature

+ Sử dụng phương pháp TF-IDF : Mức độ quan trọng của từ trong một văn bản , bản thân văn bản đang xét nằm trong một tập các văn bản.

1. Chia train , test , validation

+ Chia theo tỉ lệ 60 20 20

1. Lựa chọn thuật toán

+ Sử dụng thuật toán xác suất **Multinomial Naive Bayes**

1. Turing tham số mô hình

+ Turning tham số mô hình cho phương pháp TF-IDF

+ Các tham số đã turning : n\_gram , max\_features , stropwords

1. Đánh giá mô hình