**Proposal**

1. Thông tin nhóm: Nhóm 14

- Huỳnh Thị Hương Ly 18133029

- Huỳnh Thiên Long 18133025

- Trần Gia Nguyên 18133036

- Phan Thành Trung 18133059

2. Giới thiệu sơ lược về project

- Tên đề tài: Nhận biết tin giả

- Tập dữ liệu:

<https://www.kaggle.com/vovanthuong/pyspark-nlp-detect-fake-news-97-acc-99-auc/data>

Tập dữ liệu được thu thập từ nhiều nguồn khác nhau; các bài báo trong file true.csv có được bằng cách thu thập thông tin các bài báo từ Reuters.com. Đối với các bài báo giả mạo, chúng được thu thập từ nhiều nguồn khác nhau từ các trang web không đáng tin cậy đã bị Politifact gắn cờ (kiểm tra xác thực tổ chức ở Hoa Kỳ) và Wikipedia. Tập dữ liệu chứa các loại bài báo khác nhau trên các chủ đề, tuy nhiên, phần lớn các bài báo tập trung vào các chủ đề chính trị và tin tức thế giới.

Dữ liệu bài toán được chia thành 2 file là true.csv và fake.csv.

- Mô tả bài toán:

Sự lan truyền thông tin sai lệch là một vấn đề ngày càng gia tăng. Việc phân loại tin tức theo cách thủ công thì tốn thời gian. Bài toán phân loại hi vọng có thể giúp phân biệt đâu là tin tức đáng tin hay tin giả.

- Mô tả tập dữ liệu:

Tập dữ liệu gồm 4 thuộc tính:

+ subject : loại của bài báo

+ title : tên bài báo

+ text : văn bản bài báo

+ date : ngày bài báo được xuất bản

+ label : nhãn cho biết tin thật hay giả

Tập dữ liệu được chia thành hai tập train và test với tỉ lệ là 80:20

- Mô tả về độ đo dùng để đánh giá kết quả: ConfusionMatrix, Accuracy và areaUnderROC vì muốn so sánh thuật toán nào nhận biết tin giả thành tin thật ít nhất. Chọn độ đo này có thể giúp đánh giá thuật toán nào có tỉ lệ bỏ sót tin giả thấp nhất. Vì trong bài toán xác định tin giả, việc nhầm tin giả thành tin thật là 1 sai sót nghiêm trọng.

3. Giải pháp

- Các phương pháp, thuật toán để giải quyết bài toán: decision tree, gradient boost, random forest, naive bayes

- Lý do lựa chọn phương pháp, thuật toán trên: hợp với bài toán phân loại.

4. Thực nghiệm

- Tiền xử lý : Gộp 2 tập dữ liệu, gán nhãn, trích lấy 1 phần từ tập dữ liệu gốc, kiểm tra và xử lý các giá trị NA, xử lý dữ liệu text... Sắp xếp dữ liệu và chia dữ liệu thành 2 tập train và test.

- Sử dụng thuật toán để giải quyết bài toán: decision tree, gradient boost, random forest, naive bayes.

- Tinh chỉnh các siêu tham số bằng cross validation.

- Dựa trên độ đo đánh giá các thuật toán trên, xác định thuật toán phù hợp với bài toán này nhất.

5. Phác thảo kế hoạch làm việc:

- Tuần 11: Tiền xử lí dữ liệu và EDA

- Tuần 12: Giải quyết bài toán dựa trên thuật toán đã chia cho các thành viên:

Cả nhóm: Tiền xử lý dữ liệu

Trung: Giải quyết bài toán bằng thuật toán Random Forest.

Nguyên: Giải quyết bài toán bằng thuật toán Naive Bayes. Tổng kết, so sánh các thuật toán. Tiền xử lý dữ liệu

Long: Giải quyết bài toán bằng thuật toán Decision tree.

Ly: Giải quyết bài toán bằng thuật toán Gradient boost. Tiền xử lý dữ liệu

- Tuần 13-14: Dựa trên độ đo so sánh thuật toán, tìm thuật toán phù hợp, tinh chỉnh các siêu tham số: Cả nhóm

- Tuần 15: Tổng hợp, viết báo cáo cuối cùng: Cả nhóm