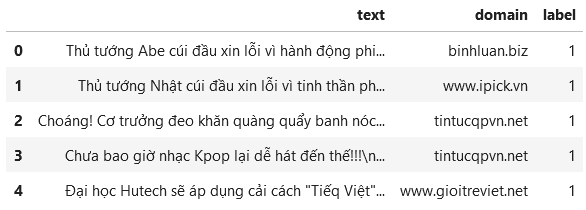
CHƯƠNG 2: DỮ LIỆU

2.1 Thu thập dữ liệu

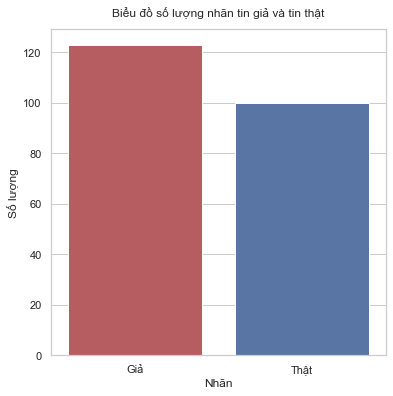
Dữ liệu vn\_news\_223\_tdlfr.csv chứa các thông tin về các bài báo và được lấy từ trang github vfnd vietnamese fake news dataset. Dữ liệu bao gồm 3 cột: text, domain và label. Cột text chứa văn bản bài báo như tiêu đề và nội dung; cột domain chứa tên miền của bài báo; cột label chứa nhãn 1 hoặc 0 để xác định bài báo là tin giả hoặc tin thật. Các chủ đề của tin tức là: Thể thao, Văn hóa, Xã hội, Kinh tế, Pháp luật.



Hình 2.1: Dữ liệu các bài báo

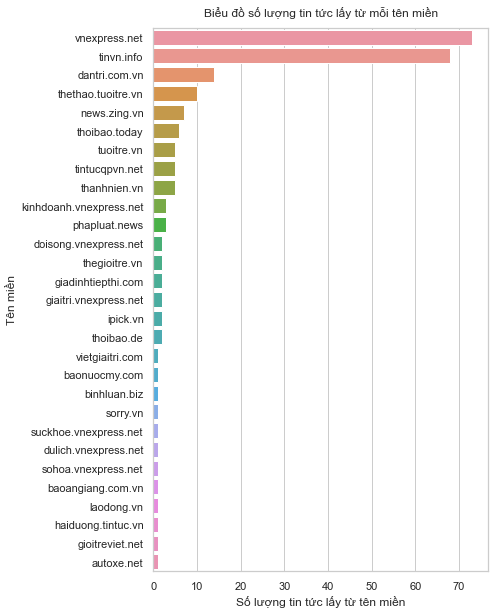
2.1.1 Chi tiết dữ liệu

Dữ liệu gồm 223 dòng, trong đó có 100 bài tin thật và 123 bài tin giả



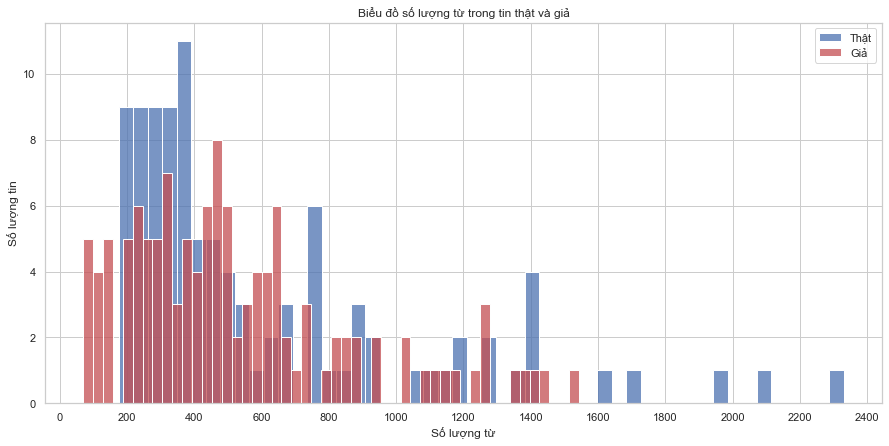
Hình 2.2: Số lượng nhãn tin giả và tin thật

Dữ liệu có 29 tên miền khác nhau, phần lớn tin tức được lấy từ trang vnexpress.net và tinvn.info.



Hình 2.3: Số lượng tin tức lấy từ mỗi tên miền

Các bài tin tức trong dữ liệu thường có khoảng từ 200 đến 400 từ.



Hình 2.4: Số lượng từ trong tin thật và giả

2.2 Xử lí dữ liệu

**1.3.2.3 Logistic regression (LR)**

Logistic regression được sử dụng khi phân loại văn bản dựa trên một bộ đặc tính lớn với kết quả ra kiểu nhị phân(Ture/False hoặc tin thật/giả). Hàm giả thuyết của Logistic regression là:

Logistic regression sử dụng hàm sigmoid để chuyển kết quả thành một giá trị xác suất; mục tiêu là giảm thiểu hàm cost, hàm được tính như sau:

**1.3.2.4 K-Nearest Neighbor (KNN)**

KNN là mô hình học không giám sát sử dụng dữ liệu huấn luyện để quyết định một điểm dữ liệu mới thuộc nhóm điểm dữ liệu nào. Mô hình KNN đo khoảng cách giữa điểm dữ liệu mới với các điểm gần nhất để xác định giá trị của K; nếu giá trị K=1 thì điểm dữ liệu mới chung nhãn với điểm gần nhất. Các phép tính đo khoảng cách là:

**1.3.2.5 Linear Support Vector Machine**

Support vector machine là mô hình học có giám sát phổ biến dùng để phân loại dữ liệu bằng cách tìm một siêu phẳng (hyperplane) sao cho nó chia bộ dữ liệu thành hai nhóm có đặc tính riêng.

**1.3.2.6 Decision trees (DT)**

Decision trees là một mô hình học có giám sát thường được sự dụng để giải bài toán phân loại. Decision trees có cấu trúc cây với mỗi điểm thể hiện một đặc tính dữ liệu, mỗi nhánh thể hiện một quy luật và mỗi lá thể hiện kết quả.

**1.3.2.7 Random forest (RF)**

Random forest là dạng cải tiến hơn của decision trees (DT) và cũng là loại mô hình học có giám sát. Random forest gồm nhiều cây decision trees hoạt dộng độc lập để dự đoán nhãn dựa trên kết quả của các cây.

**XGBoost**

XGBoost là thư viện cho gradient boosting học bằng giải thuật theo kiểu cây, được thiết kế để có hiệu suất cao.

**LightGBM**

LightGBM là thư viện cho gradient boosting học bằng giải thuật theo kiểu cây. LightGBM phát triển cây dựa trên leaf-wise, trong khi các giải thuật khác dựa trên level-wise. Leaf-wise giúp giảm nhiều loss hơn giải thật level-wise.