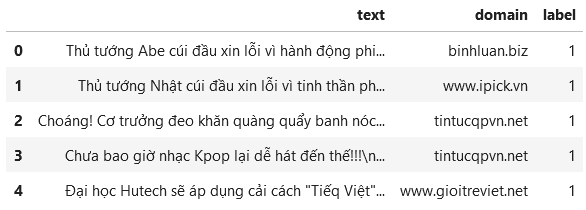
CHƯƠNG 2: DỮ LIỆU

2.1 Thu thập dữ liệu

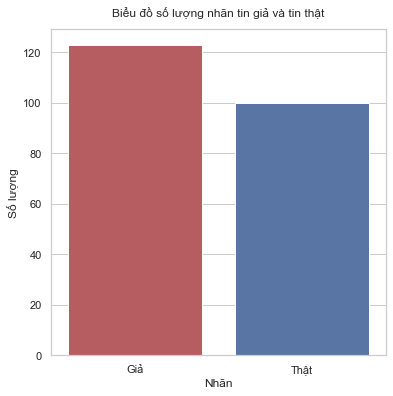
Dữ liệu vn\_news\_223\_tdlfr.csv chứa các thông tin về các bài báo và được lấy từ trang github vfnd vietnamese fake news dataset. Dữ liệu bao gồm 3 cột: text, domain và label. Cột text chứa văn bản bài báo như tiêu đề và nội dung; cột domain chứa tên miền của bài báo; cột label chứa nhãn 1 hoặc 0 để 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

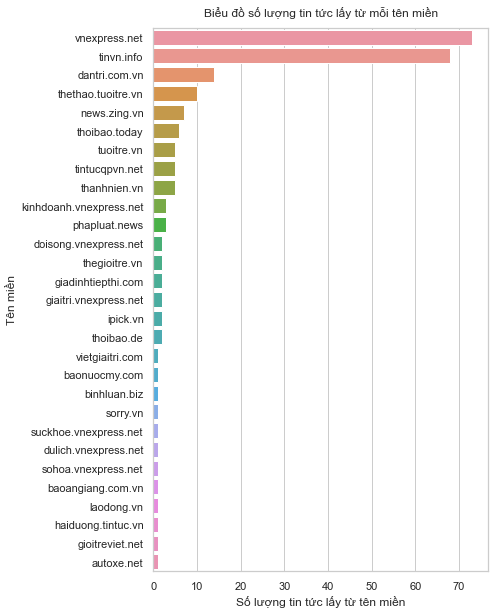
2.1.1 Chi tiết dữ liệu

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



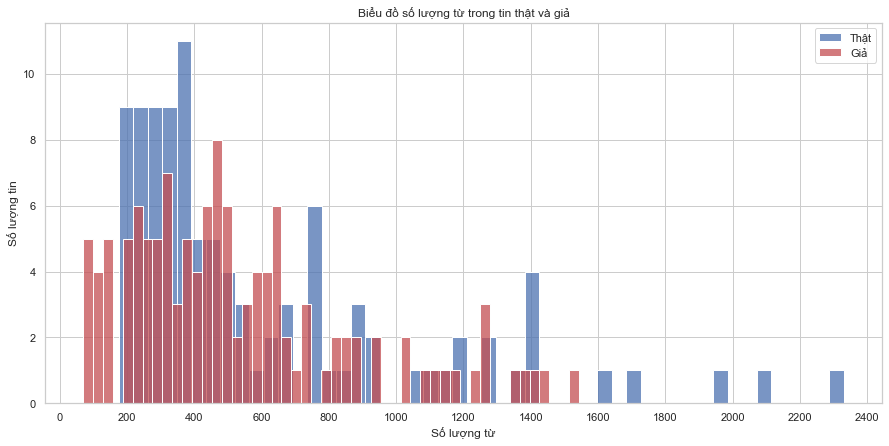
Hình 2.2: Số lượng nhãn tin giả và tin thật

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