

TIẾP CẬN PHÁT HIỆN TIN GIẢ BẰNG TRÍ TUỆ NHÂN TẠO CÓ GIẢI THÍCH

Sự lan truyền nhanh chóng của tin giả trên mạng xã hội đặt ra thách thức lớn đối với độ tin cậy thông tin và nhận thức cộng đồng, đặc biệt trong bối cảnh tiếng Việt khi nghiên cứu còn hạn chế. Nghiên cứu này so sánh hai mô hình học sâu BiLSTM và PhoBERT trong phát hiện tin giả tiếng Việt, đồng thời tích hợp LIME thuộc nhóm Explainable AI (xAI) để giải thích cơ chế quyết định của mô hình. Thí nghiệm trên bộ dữ liệu ReINTEL 2020 cho thấy cả hai mô hình đạt hiệu quả cao, trong đó PhoBERT vượt trội với độ chính xác 93.6% và điểm F1 đạt 80.0%. Việc áp dụng LIME giúp trực quan hóa mức độ đóng góp của từng từ, tăng tính minh bạch và độ tin cậy, qua đó góp phần xây dựng hệ thống phát hiện tin giả vừa hiệu quả vừa dễ giải thích.

Từ khóa

Phát hiện tin giả, Xử lý ngôn ngữ tự nhiên tiếng Việt, PhoBERT, BiLSTM, Trí tuệ nhân tạo có giải thích

Thông tin các tác giả

1/Đào Xuân Kỳ, ThS., đang công tác trường Đại học Công Thương Tp. Hồ Chí Minh, email: kydx@huit.edu.vn;
2/Đinh Minh Hoà, ThS., đang công tác Trường Đại học Ngoại Ngữ Tin Học Tp. Hồ Chí Minh, email: hoadm@huflit.edu.vn, NCS tại Đại học Công nghệ Tp. Hồ Chí Minh; Email: dmhoa24nct@hutech.edu.vn; 3/ Tiểu Phùng Mai Sương, ThS., đang công tác Trường Đại học Ngoại Ngữ Tin Học Tp. Hồ Chí Minh, email: suongtpm@huflit.edu.vn, NCS tại Đại học Công nghệ Tp. Hồ Chí Minh; Email: tpmsuong24nct@hutech.edu.vn

Authors: Mr ĐINH, Hoà (Trường Đại học Ngoại Ngữ-Tin Học Tp. Hồ Chí Minh); Mr ĐÀO, Kỳ (Trường Đại học Công Thương Thành phố Hồ Chí Minh)

Co-author: MAI, Sương (Trường Đại học Ngoại Ngữ-Tin Học Tp. Hồ Chí Minh)

Track Classification: Tiểu ban 3: Trí tuệ Nhân tạo và Ngôn ngữ học Tính toán trong phát triển Kinh tế, Văn hóa và Xã hội