NGUYEN VU HUY
CE180384
HUYNVCE180384@FPT.EDU.VN
AI1801 - SEG301 • Spring2025

COLLECT NEWS DATA AND BUILD NEWS CLASSIFICATION MODEL AND DEPLOY WEB APPLICATION

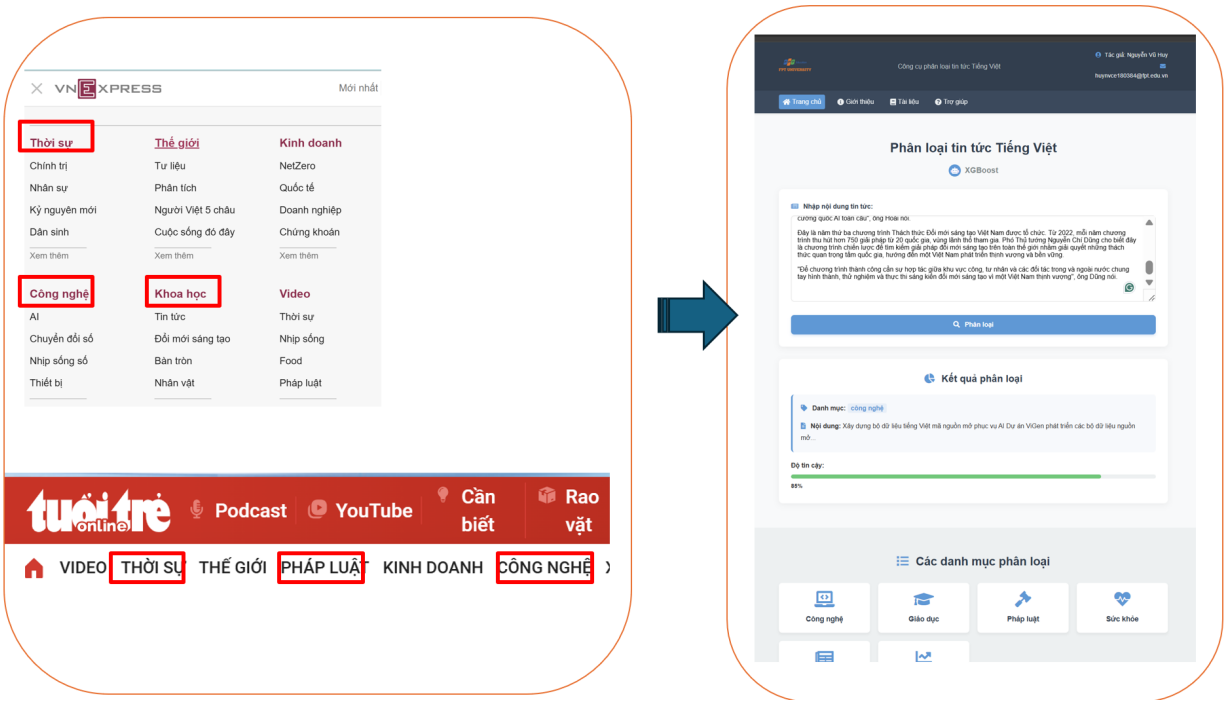

Figura 1: Project Overview

Introduction

This project focuses on collecting news data from vnexpress.net and tuoitre.vn, including fields such as URL, source, crawl time, category, title, description, publish time, content, and author. The collected data will be visualized to analyze trends and then used to fine-tune the XGBoost model for categorizing articles, with inputs combining title, description, and content. Finally, a Flask web application will be deployed to provide an interface for utilizing the model. The project aims to automate the classification of Vietnamese news, improving the efficiency of information organization and access.

Exploratory crawled data analysis for Vietnamese news classification

url	source	crawl_time	category	title	description	publish_time	content	author
https://tuoitre.vn/tong-bi-thu-to-lam-xay-dung...	tuoitre	2025-03-12 22:28:11	cong-nghe	Tổng Bí thư Tô Lâm: Xây dựng, ban hành danh mụ...	NaN	NaN	Tổng Bí thư Tô Lâm chủ trì phiên họp - Ảnh: TT...	THÀNH CHUNG
https://tuoitre.vn/cau-noi-viettel-xay-nen-gia...	tuoitre	2025-03-12 22:28:40	cong-nghe	'Cầu nối' Viettel xây nên giấc mơ giáo dục 'kh...	NaN	NaN	Trong gần 2 thập kỷ, Internet trường học đã gi...	VIETTEL - T.HÀ
https://tuoitre.vn/chi-dung-y-nghi-nguoi-dan-o...	tuoitre	2025-03-12 22:29:09	cong-nghe	Chỉ dùng ý nghĩ, người đàn ông bị liệt có thể ...	NaN	NaN	Giáo sư thần kinh học Karunesh Ganguly (giữa) ...	TTXVN
https://tuoitre.vn/bat-ngo-vi-lam-ly-lich-tu-p...	tuoitre	2025-03-12 22:29:40	cong-nghe	Bất ngờ vì làm lý lịch tư pháp qua VNeID thuận...	NaN	NaN	Một số người dân vẫn đến trực tiếp làm thủ tục...	ÁI NHÂN
https://tuoitre.vn/xac-thuc-dien-tu-qua-vneid-...	tuoitre	2025-03-12 22:30:11	cong-nghe	Xác thực điện tử qua VNeID tăng cường an ninh,...	NaN	NaN	Trung tâm RAR (Bộ Công an) và Sacombank ký kết...	DANH TRỌNG
...	...	...	...	...	...	...	...	...
https://vnexpress.net/chi-em-sinh-doi-gianh-ho...	vnexpress	2025-03-12 15:03:55	giao-duc	Chị em sinh đôi giành học bổng hơn 17 tỷ đồng ...	Quỳnh Anh và Quỳnh Hương, 19 tuổi, giành học b...	Thứ ba, 4/3/2025, 11:09 (GMT+7)	Tin trúng tuyển và học bổng đến vào sáng 22/2,...	NaN
https://vnexpress.net/chuyen-gia-giai-dap-co-n...	vnexpress	2025-03-12 15:04:34	giao-duc	Chuyên gia giải đáp 'có nên cho con du học THP...	Ông Hoàng Nam Tiến cùng các chuyên gia giáo dụ...	Thứ ba, 11/3/2025, 10:00 (GMT+7)	Sự kiện diễn ra vào 15h - 18h, thứ Bảy, ngày 2...	NaN
https://vnexpress.net/vung-dat-cua-dai-may-tra...	vnexpress	2025-03-12 15:05:14	giao-duc	'Vùng đất của dải mây trắng' là tên gọi nước nào?	Quốc gia này ở bán cầu Nam, được người Maori g...	Thứ sáu, 28/2/2025, 19:00 (GMT+7)	Bình Minh (Tổng hợp)\nNồng độ bụi mịn PM 2.5 v...	NaN

Figura 2: Sample data after crawling

Dataset Overview

The dataset I collected consisted of 1,752 articles, each assigned to a specific category. The dataset included fields such as title, content, and description, which were analyzed for word frequency and distribution patterns.

Category Distribution

The category distribution is visualized in Figure 3, showing that the dataset is imbalanced. The most represented category is "giao-duc"(education), followed by "phap-luat"(law) and "khoa-hoc"(science). The least represented categories are "kinh doanh"(business) and çong-nghe"(technology). This imbalance may affect model performance, requiring resampling techniques or weighted loss functions.

Word Frequency Analysis

Word Cloud Representation To analyze the most frequent words in different sections of the dataset, we generate word clouds for titles, content, and descriptions (Figure 4). The results show a dominance of high-frequency words, which are commonly found in Vietnamese text. These words might need to be filtered out as stopwords to enhance model performance.

Most Frequent Words in Titles Figure 5 presents a bar chart depicting the top 20 most frequently occurring words in article titles. These words provide valuable insights

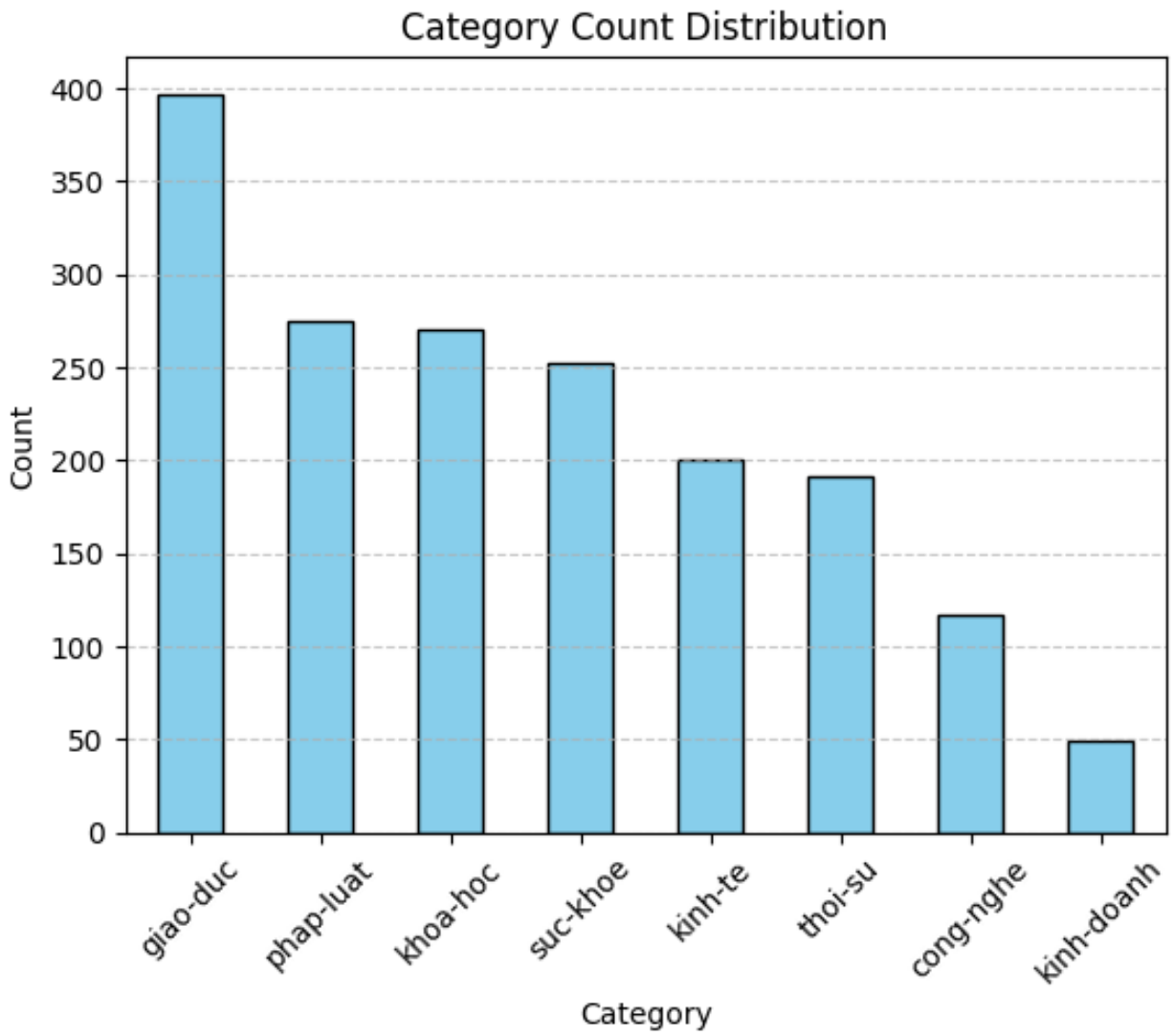

Figura 3: Category distribution

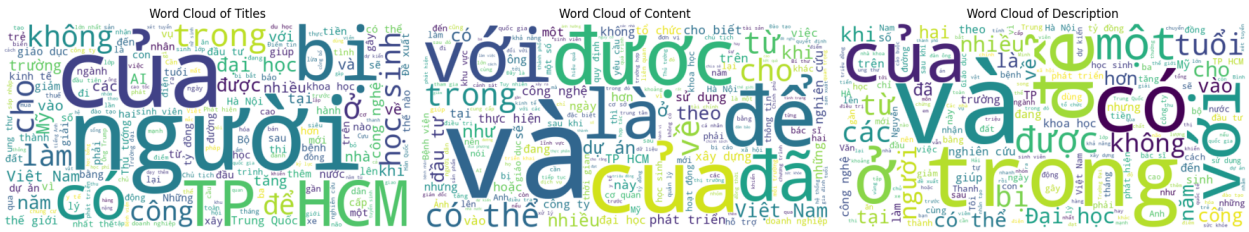

Figura 4: Word Cloud Representations for Titles, Content, and Descriptions

into the primary themes covered in the dataset. The most common terms include references to education, individuals, and societal issues, indicating prevalent topics in the collected data. Understanding these high-frequency words is essential for refining text preprocessing steps, such as stopword removal, and optimizing feature selection for downstream classification models. By analyzing title word distributions, we can enhance the representation of textual data and improve the overall performance of machine learning algorithms.

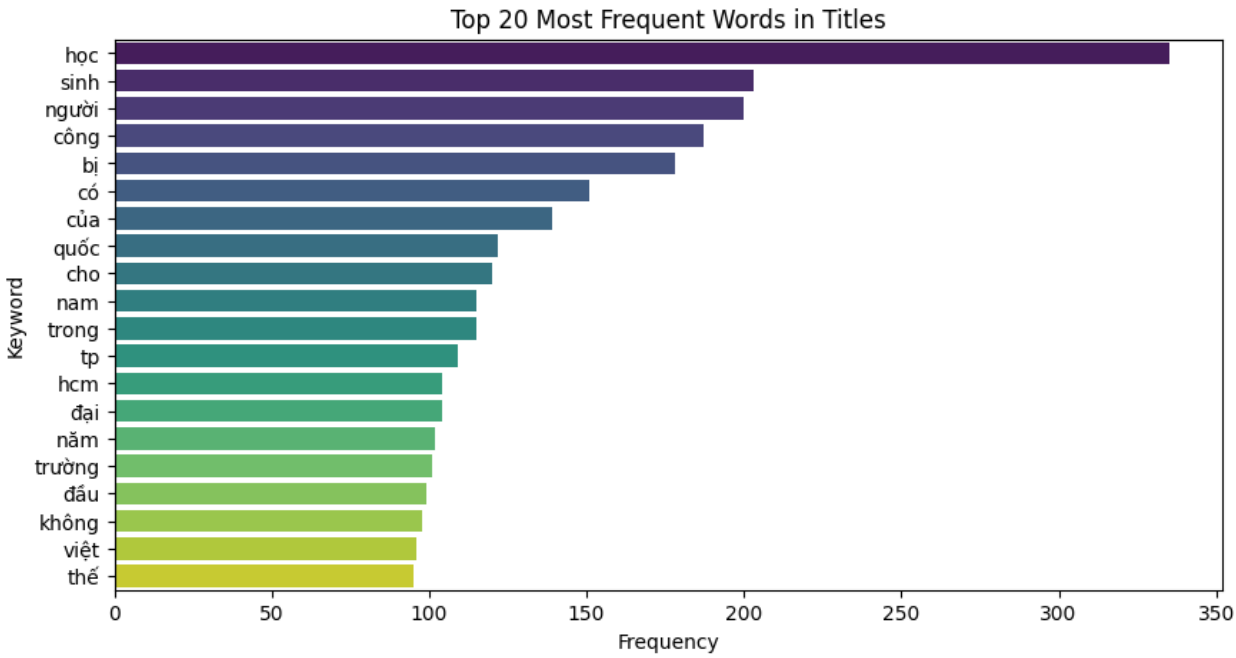

Figura 5: Top 20 Most Frequent Words in Titles

Pipeline and Model

XGBoost (eXtreme Gradient Boosting) is selected for its high performance and scalability in multi-class classification tasks. The model constructs an ensemble of decision trees sequentially, where each tree corrects the errors of the previous ones. The optimized hyperparameters include max-depth (10), learning-rate (0.1), n-estimators (300), objective ('multi:softmax'), num-class (determined by the number of unique labels in the training set), eval-metric ('mlogloss'), and random-state (42) to ensure reproducibility.

The model is trained using the XGBClassifier implementation with an evaluation set consisting of both training and test data. Performance is assessed using Accuracy, Precision, Recall, F1-score, and a Confusion Matrix. During training, the model provides updates at every 100 iterations to monitor progress and convergence.

The Vietnamese news classification system is built based on the XGBoost model with a complete data processing pipeline, from data collection to application deployment. The entire process consists of four main stages: data collection, preprocessing and exploratory data analysis (EDA), model training, and application deployment.

The data set is collected from reliable online news sources in Vietnam, including

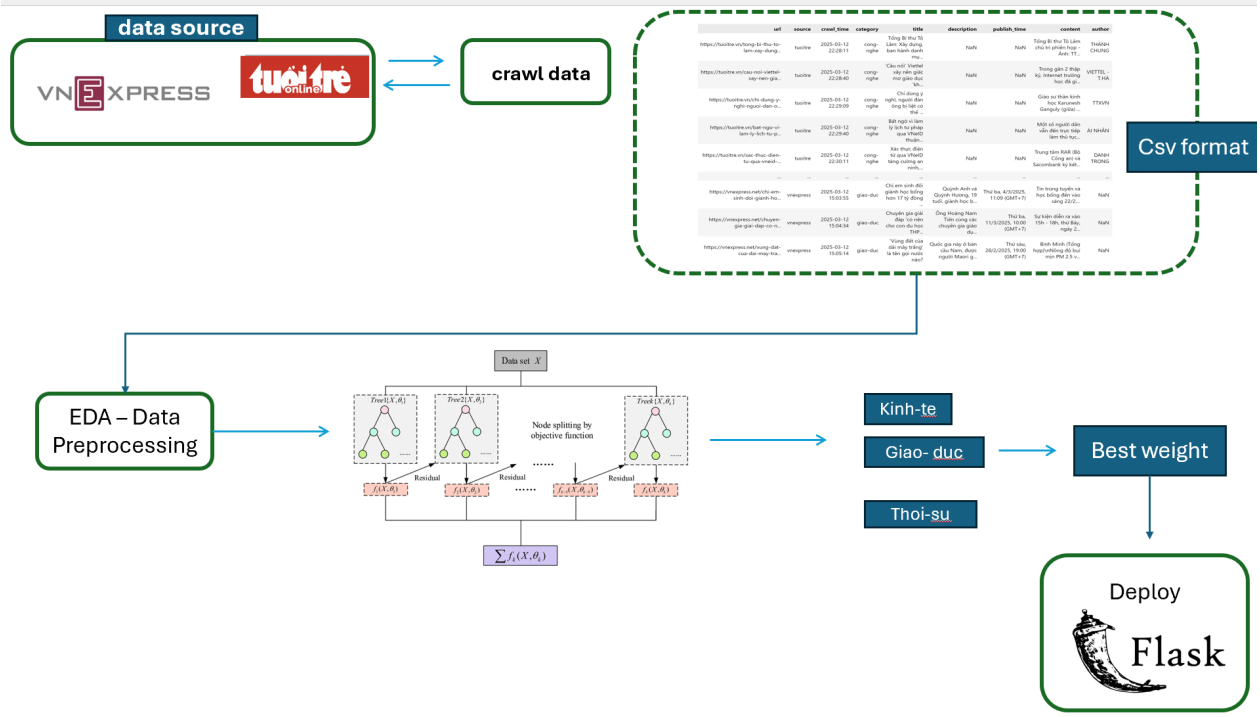

Figura 6: Complete processing pipeline

VnExpress and Tuoi Tre Online, using web crawling techniques. The collection process focuses on articles belonging to six main categories: Technology, Education, Law, Health, Current Affairs, Economy.

Data Collection News articles are collected from reputable Vietnamese sources such as VnExpress and Tuoi Tre Online using web crawling techniques. The data set consists of six categories: Technology, Education, Law, Health, Current Affairs and Economics. The collected data are stored in CSV format, including key attributes such as source URL, publication time, category, title, description, content, and author.

Data Preprocessing and EDA Text preprocessing involves removing special characters and unnecessary punctuation, normalizing Vietnamese diacritics, word segmentation using Vietnamese-specific libraries, removing stopwords, and transforming text into features using TF-IDF. EDA explores category distributions, average article length per category, frequent keywords, and textual feature correlations.

Results

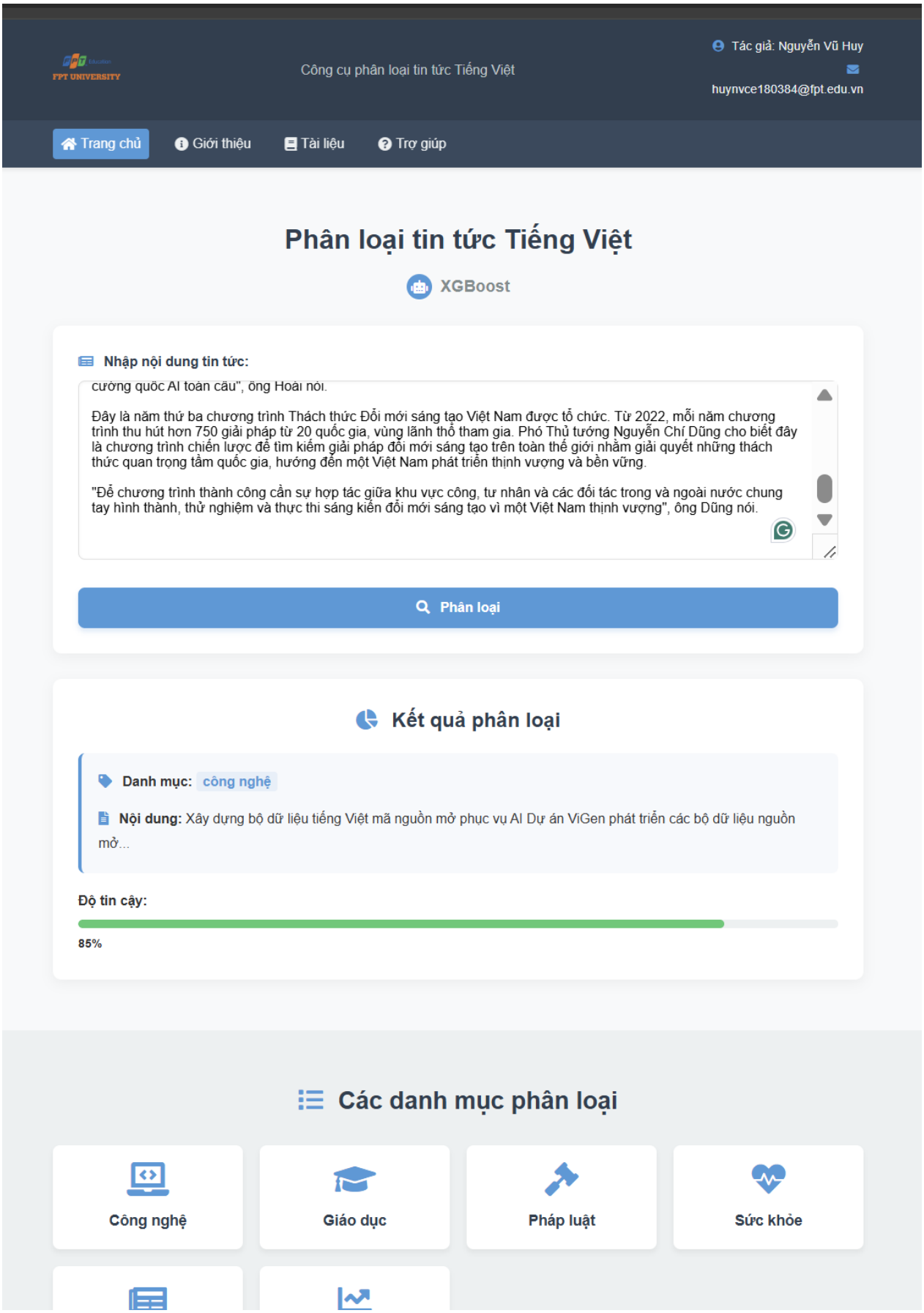

Figura 7: Web Application Interface

Cuadro 1: Classification Performance Metrics

Class	Precision	Recall	F1-score	Support
0	0.80	0.77	0.79	74
1	0.85	0.86	0.85	84
2	0.80	0.80	0.80	55
3	0.86	0.78	0.82	54
4	0.79	0.65	0.71	46
5	0.51	0.71	0.59	38
Accuracy		0.77		351
Macro avg	0.77	0.76	0.76	351
Weighted avg	0.79	0.77	0.78	351