Graph-Based Classification of Facebook Pages Using Traditional Machine Learning Algorithms

Haleemah Amisu,
Department of Computer and Information Sciences,
Towson University, Maryland, USA
hamisu1@students.towson.edu

***Abstract*—This project applies traditional machine learning models to classify Facebook pages using structural features derived from a real-world social network. Using the Facebook Page-Page Network dataset, an undirected graph was built where nodes represent verified pages and edges represent mutual "likes." Graph-based features such as degree, clustering coefficient, pagerank, and degree centrality were extracted using NetworkX. Three models—Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (KNN)—were trained to predict page categories. Random Forest achieved the most balanced performance with a macro F1 score of 0.49. SVM improved significantly after scaling. These results show that structural graph features alone can provide meaningful insights for node classification.**

I. Introduction

Graph-structured data is common across many domains, including social networks, recommendation systems, and biological networks. Node classification is a foundational task in this space, and while deep learning approaches like graph neural networks (GNNs) have gained popularity, traditional models remain valuable for their simplicity and interpretability. This project investigates how graph-derived features can be used with classical machine learning models to classify Facebook pages by type.

II. Related Work

Previous research has shown that traditional machine learning models can perform reasonably well on graph classification tasks when paired with handcrafted features such as degree, clustering, and centrality. Random Forest and SVM have served as strong baselines for tasks in social network analysis. This project builds on those ideas and provides a baseline for later comparisons with deep learning methods.

III. Dataset Description

The Facebook Page-Page Network dataset, obtained from the Stanford SNAP repository[1], consists of 22,470 nodes representing verified Facebook pages, and 171,002 undirected edges representing mutual page "likes." Each page is labeled with one of four categories: *company*, *government*, *politician*, or *tvshow*. Only the edge list and the page labels were used; feature metadata in the JSON file was not utilized in this project.

[1]https://snap.stanford.edu/data/facebook-large-page-page-network.html

IV. Methodology

A. Graph Construction

Using NetworkX, an undirected graph was created from the page-page edge list. Each node corresponds to a Facebook page, and each edge reflects a mutual "like" connection between two pages.

B. Feature Engineering

NetworkX was used to compute the following structural features for each node:

- Degree
- Clustering Coefficient
- Pagerank
- Degree Centrality

The resulting feature matrix (X) was paired with the encoded page types (y).

C. Classifiers

Three traditional models were trained and evaluated:

- Random Forest
- Support Vector Machine (SVM), with standardized features
- k-Nearest Neighbors (KNN)

D. Evaluation

The data was split into training and test sets using a 70/30 stratified split. Each model was evaluated using accuracy, precision, recall, F1 score, and confusion matrices.

V. Experiments and Results

TABLE I
Model Performance Summary

Model	Accuracy	Macro F1	Notes
Random Forest	52%	0.49	Most balanced
SVM (scaled)	52%	0.41	High precision, weak recall
KNN	48%	0.42	Weakest performance

Random Forest performed best in terms of balance and generalization. **SVM** required feature scaling to avoid collapsing into majority class predictions. After scaling, it performed well on majority classes but poorly on the minority class *tvshow*. **KNN** underperformed overall, showing limitations in capturing global graph structure.

VI. Discussion

Results show that traditional classifiers can effectively use structural graph features for multiclass node classification. Random Forest offered the most consistent performance. SVM improved significantly with proper preprocessing, while KNN lacked robustness in this graph context. These results offer a baseline for future comparisons with graph neural networks.

VII. Conclusion

This project demonstrated the feasibility of classifying Facebook pages using only graph structure. The best-performing model, Random Forest, used structural features like degree and pagerank to achieve a macro F1 score of 0.49. These results establish a practical benchmark for node classification and support further exploration of deep learning techniques like GCNs and GATs.

References

- SNAP: Facebook Page-Page Network Dataset. https://snap.stanford.edu/data/facebook-large-page-page-network.html
- Colab Notebook: https://colab.research.google.com/drive/1mC8IjbfGm-5XTez96WHH37R367Or9qmx?usp=sharing
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.