Logistics Research Papers Classification

By Iffan Kepan

Table of contents

01 **Introduction**

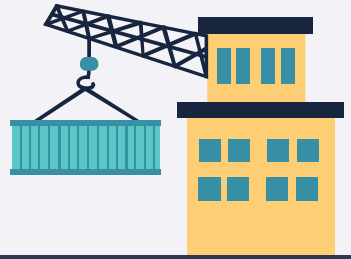
Overview of the problem and the project's objectives.

02 **Methodology**

Processes and techniques used to conduct the research.

03 **Summary**

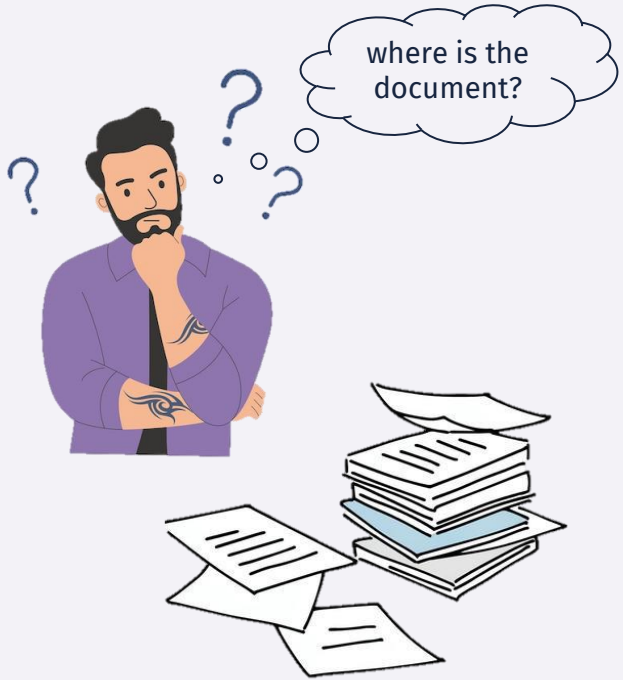
Key findings from models and future development.



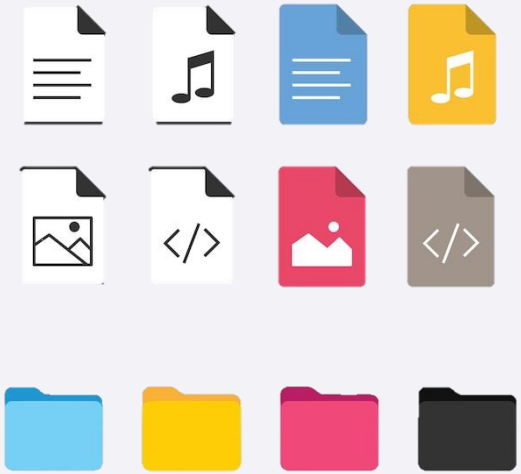
01

Introduction

Do you ever struggle with finding documents?



Life is better when...



Problem Statement

> Students face significant difficulty in finding literature reviews or research papers from their seniors, as there is no categorization of research work, and all papers are stored in a single folder.

What is the criteria to classify ?

Demand Planning

Procurement and Sourcing

Manufacturing and Production

Warehouse and Inventory Management

Distribution and Transportation

Import - Export

After Sales Services

02

Methodology

Data Collection

Students

The data comes from logistics bachelor's degree students.

Language

All documents are written in Thai.

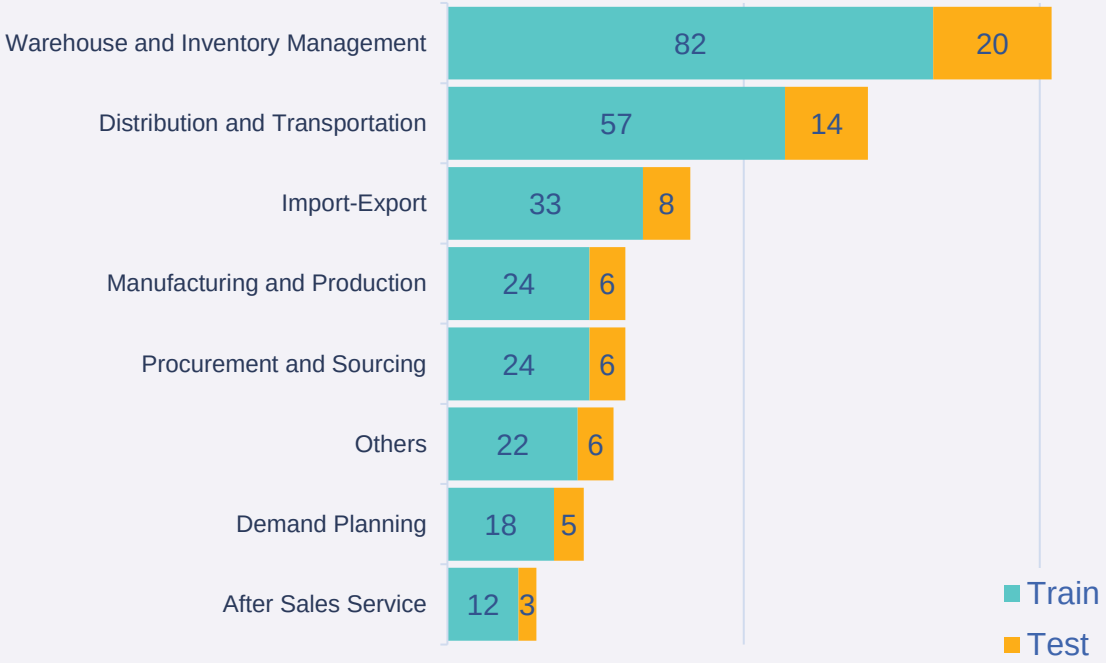
Dataset

The dataset includes a total of 340 files.

Size

Each file contains an average of 6 pages.

Category Distribution



- **Warehouse and Inventory Management** is the most topic with **102 papers** (30% of the total 340 files).
- While **After Sales Service** has the fewest papers with only 15.
- Train/Test with **70%/30%**

Flow of Process

Raw Data

Plain text from
Research paper

Tokenization

Used two Thai tokenizers:

Newmm, WangchanBERTa

TF-IDF Vectorization

Tested different parameters:

max features: 5000, 7000, None

max df: 0.9, 0.8

Model Training

Compared 4 models:

Neural Network, SVM,

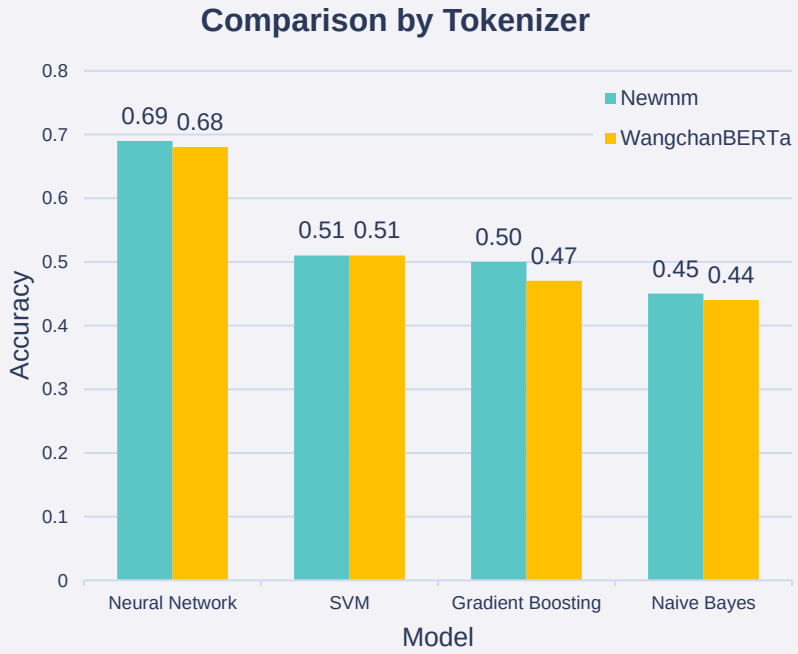
Gradient Boosting, Naive Bayes

Evaluation

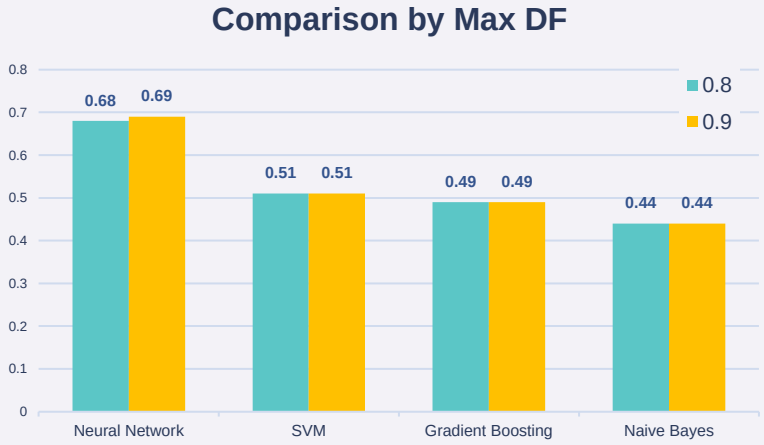
Measured performance
using Accuracy

Performance Comparison

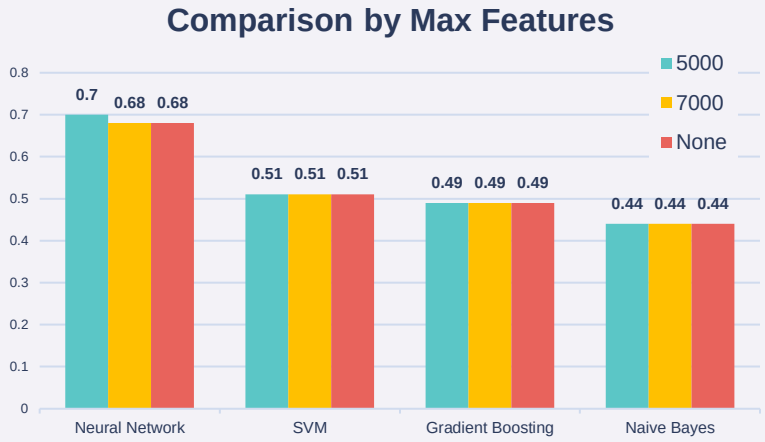
Newmm performs better than **WangchanBERTa** in many models except SVM where both have the same score.



Performance Comparison (2)



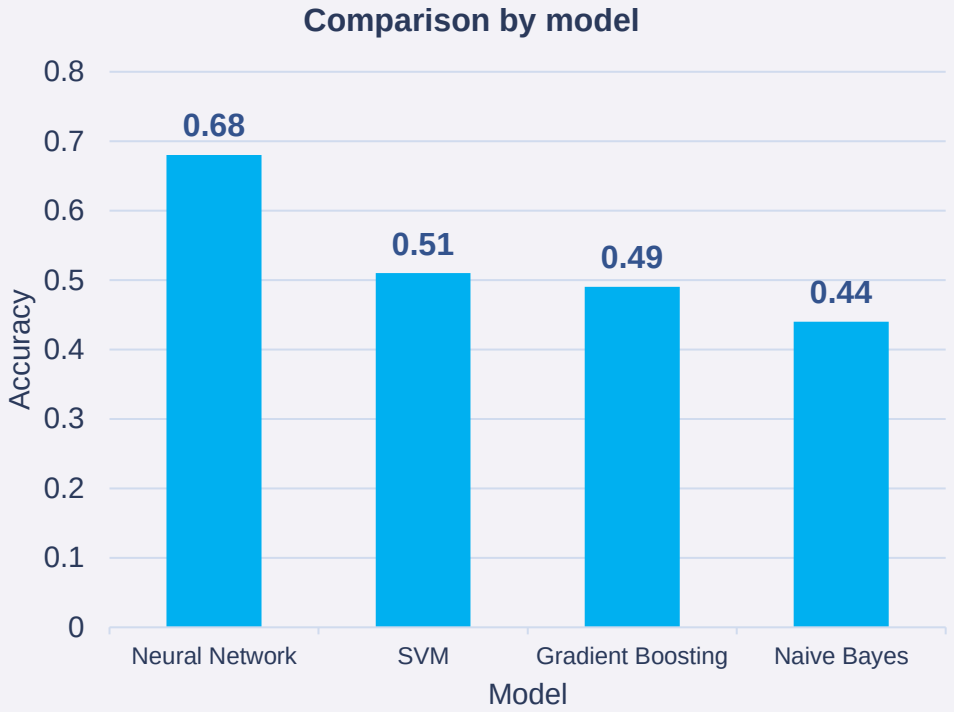
Constraining **Max DF** to **0.9** marginally improves **Neural Network** performance, but has no significant impact on other models.



Neural Network benefits most from constraining **Max Features** to **5000**, while other models remain unaffected.

Model Evaluation

- **Neural Network** performs the best, with an accuracy of **0.68**
- Other models like SVM (0.51), Gradient Boosting (0.49), and Naive Bayes (0.44) show lower accuracies.



03

Summary

Key Findings

Accuracy

71.57 %

Best Performance

- Import-Export (92%)
- Procurement and Sourcing (89%)
- Warehouse and Inventory Management (84%)

Best Combination

- Model : Neural Network
- Tokenizer : Newmm Tokenizer
- max_features = 5000
- max_df = 0.9

Challenges

- After Sales Service (20%)
- Others (38%)
- Manufacturing and Production (44%)

Future Development and Recommendations

End-to-end pipeline automates the entire process, from data collection to final output:

- Categorization
- Summarization
- Analysis
- Recommender System

Streamlit

Q & A

Thank you