Classification of Companies' Economic Activity: A Web Content and Machine Learning-Based Approach

Joan Fernández Navarro, Ana Debón, and Josep Domenech

Universitat Politècnica de València, Spain

Contents

1. Introduction
2. Methodology
3. Evaluation Strategies
4. Results
5. Conclusion

Introduction

- **Economic classification relevance**: Essential for economic analysis, policymaking, and market research.
- **Digitalization as an opportunity**: A significant percentage of Spanish companies have websites, enabling the potential for automatic classification based on web content.
- **Costly and time-consuming process:** Traditionally, the classification of companies has been carried out manually, and it is also prone to human errors and inconsistencies.
- **Proven feasibility of automated methods**: Previous research has shown that machine learning and text mining are viable tools for accurately classifying companies by economic sector.

Goal

Design and implement a methodology that allows the automatic classification of companies according to their economic activity, using as a main source the web content of the companies themselves.

2. Methodology

Main Scheme

The classification is carried out at the **division level**, which corresponds to level 2 of the CNAE classification system.

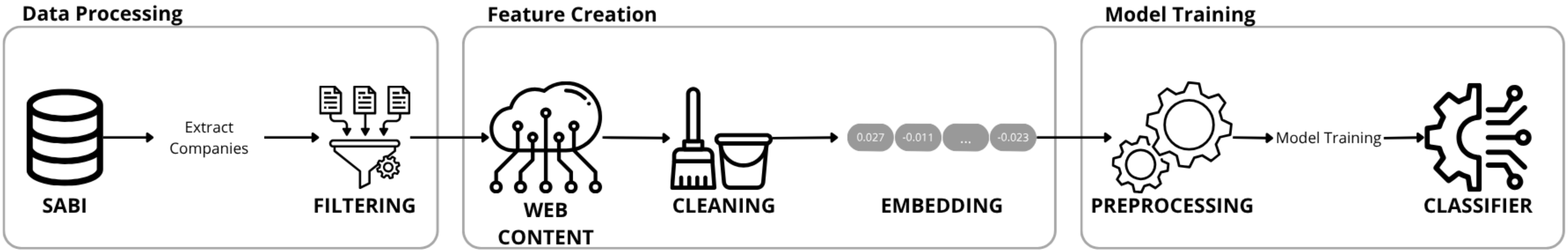
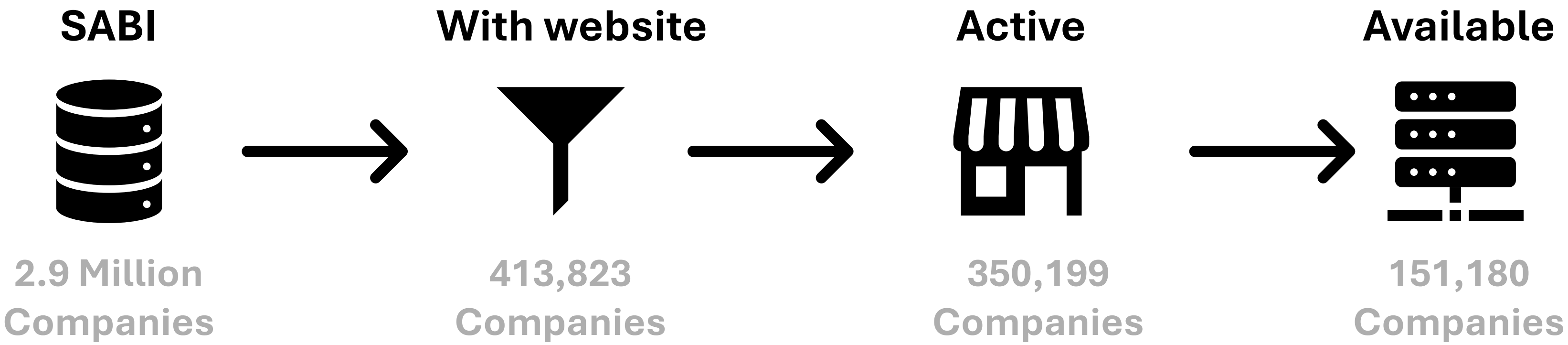


Figure 2. Pipeline from Data Extraction to Company Classification Using Web Content Embeddings.

Data

The data used in this project comes from **SABI** (Sistema de Análisis de Balances Ibéricos), a database that provides detailed **economic**, **financial**, and **corporate information** on companies operating in Spain and Portugal.



SABI

2.9 Million Companies

With website

413,823 Companies

Active

350,199 Companies

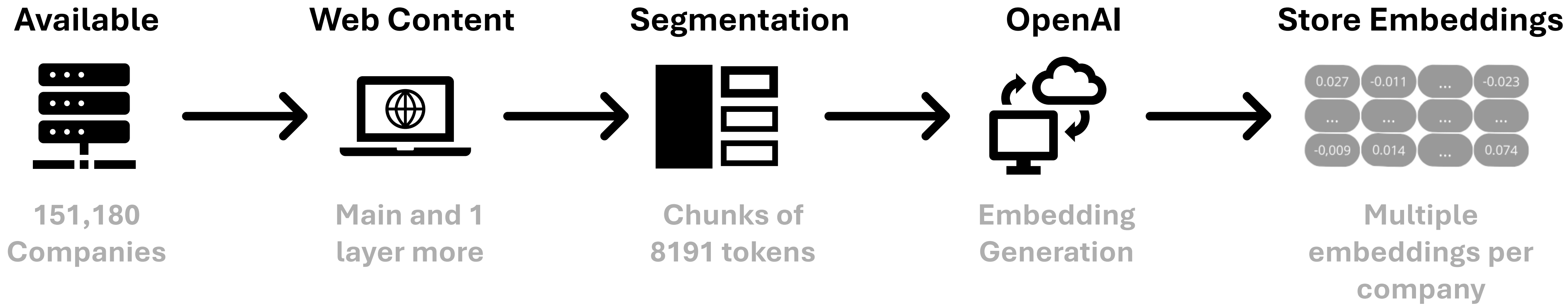
Available

151,180 Companies

Figure 3. Filtering process applied to the data extracted from SABI to select companies.

Feature Creation

To represent the web content of companies, the **multilingual** model ***text-embedding-3-small*** from OpenAI was used. This model transforms text into numerical vectors that capture its **semantic meaning.**



Available → **Web Content** → **Segmentation** → **OpenAI** → **Store Embeddings**

151,180 Companies

Main and 1 layer more

Chunks of 8191 tokens

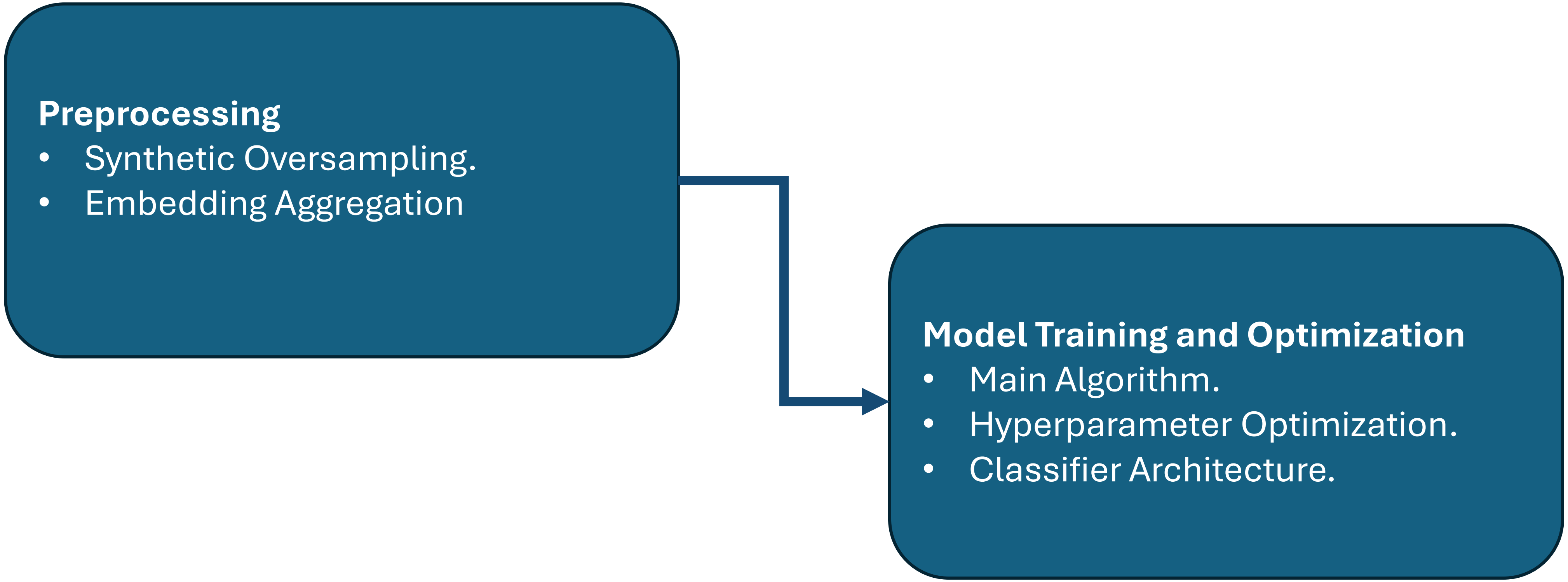
Embedding Generation

Multiple embeddings per company

0.027	-0.011	...	-0.023
...	...	...	...
-0,009	0.014	...	0.074

Figure 4. Process of the Embedding Generation.

Model Training



Preprocessing

- Synthetic Oversampling.
- Embedding Aggregation

Model Training and Optimization

- Main Algorithm.
- Hyperparameter Optimization.
- Classifier Architecture.

Preprocessing

Synthetic Oversampling

A	B	…	Z
1500	1500	…	783

↓

A	B	…	Z
1500	1500	…	1500

- SMOTE
- ADASYN
- Borderline SMOTE

The evaluation metric is the Jensen-Shannon Divergence.

Embedding Aggregation

Table 1. Description of the embedding aggregation techniques.

Technique	Operation
Truncate	$X' = X[:k]$
Max Pooling	$X'_i = \max(X_{1,i}, X_{2,i}, \dots, X_{n,i})$
Mean Pooling	$X'_i = \frac{1}{n}\sum_{j=1}^{n} X_{j,i}$
Sum Pooling	$X'_i = \sum_{j=1}^{n} X_{j,i}$

Hyperparameter Optimization

$$F1\ Generalization\ Score = \alpha \cdot F1_{test} - \beta \cdot \frac{|F1_{train} - F1_{test}|}{\sqrt{2}}$$

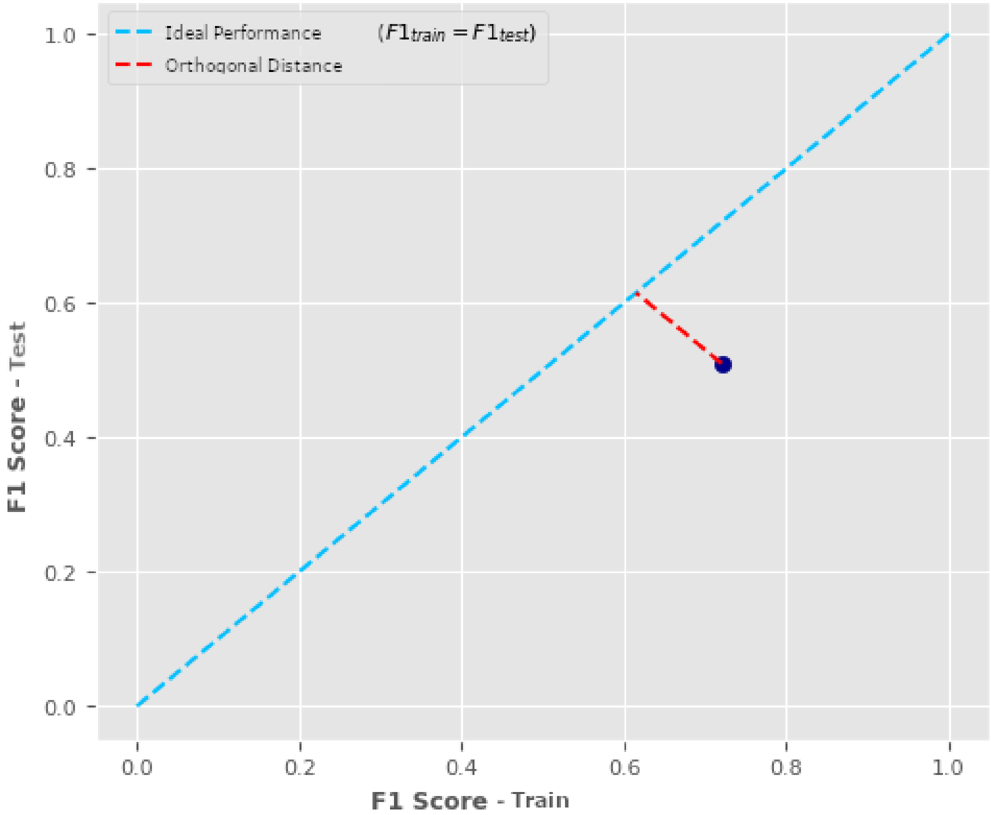


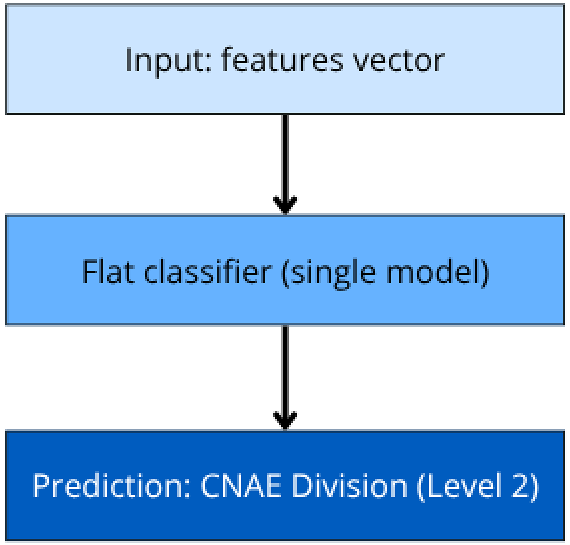
Figure 5. Example of calculation and operation of F1 Generalization Score.

where:

- α; weights the performance over the validation set.
- β; penalizes the degree of overfitting or underfitting with respect to the ideal model.

Classifier Architecture

Flat Classifier

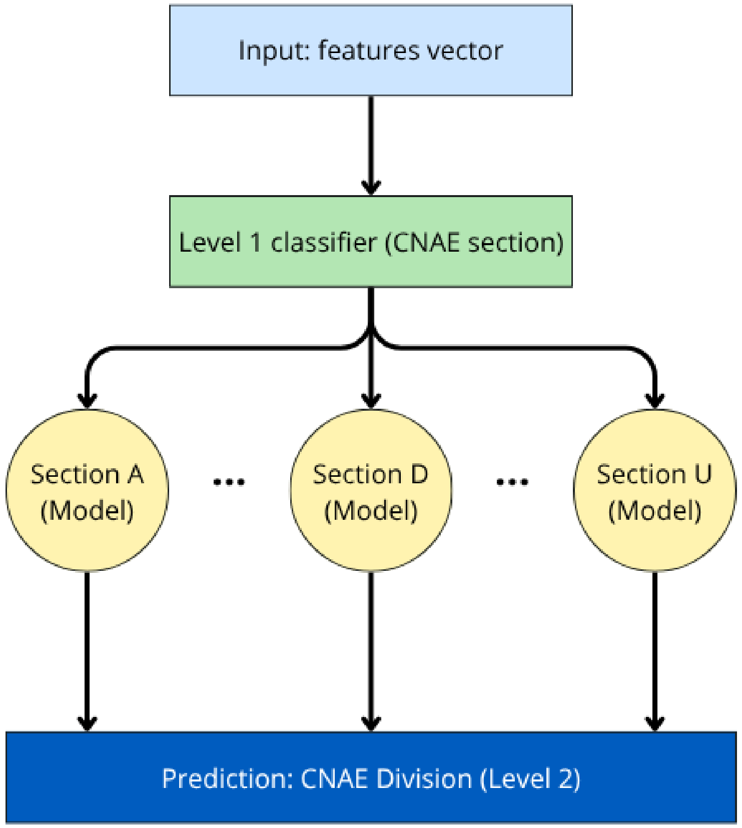


Input: features vector

Flat classifier (single model)

Prediction: CNAE Division (Level 2)

Hierarchical Classifier



Input: features vector

Level 1 classifier (CNAE section)

Section A (Model) ... Section D (Model) ... Section U (Model)

Prediction: CNAE Division (Level 2)

3. Evaluation Strategies

Proposed Strategies

Spanish Economy Distribution
To realistically assess the model, a subset of data is defined with a sectoral distribution that mirrors the structure of the national economy.

Advantages:

- Enables performance analysis in a more realistic setting.
- Helps identify underperforming categories.
- Allows estimation of model behavior under natural class imbalance.

Qualitative Error Analysis
A sample consisting of a subset of misclassified instances was selected to perform a manual inspection.

Types of errors:

- Ambiguity of web content.
- Non-representative web content.
- CNAE coding errors in SABI.
- Coincidence with secondary codes.
- Model error.
- Partially consistent prediction.

4. Results

Main Algorithm

The synthetic oversampling technique with better performance is **SMOTE**.

Table 2. Average F1 Score (± standard deviation) by classifier and aggregation technique.

	Truncate	Max Pooling	Mean Pooling	Sum Pooling
Baseline	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000
RF	0.372 ± 0.020	0.426 ± 0.012	0.472 ± 0.018	0.429 ± 0.009
NB	0.428 ± 0.031	0.434 ± 0.024	0.492 ± 0.023	0.098 ± 0.020
SVM	0.524 ± 0.022	0.558 ± 0.011	0.582 ± 0.026	0.237 ± 0.030
kNN	0.376 ± 0.013	0.448 ± 0.026	0.476 ± 0.026	0.409 ± 0.016
MLP	0.478 ± 0.023	0.505 ± 0.009	0.514 ± 0.015	0.517 ± 0.023

Aggregation Technique

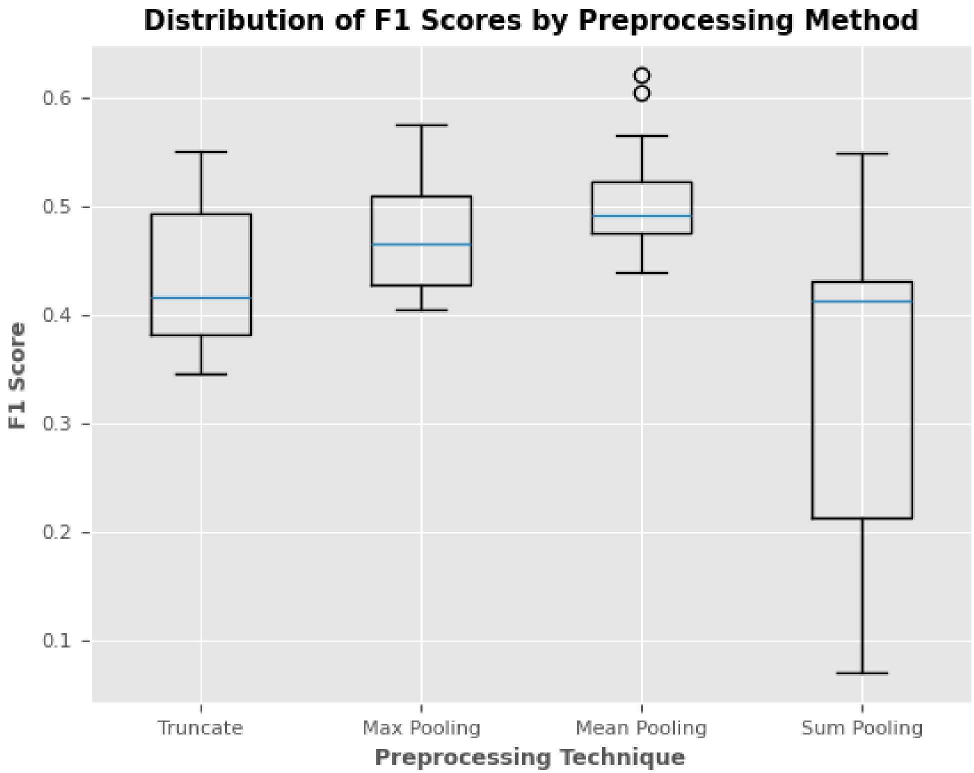


Figure 6. Distribution of F1 Score by aggregation technique.

Discarding Sum Pooling:

- The techniques follow a normal distribution, according to the Shapiro-Wilk test.
- The variances across the three groups are homogeneous, according to Levene's test.

Table 3. Results of ANOVA test.

Statistic	10.335
p-value	0.000113

Aggregation Technique

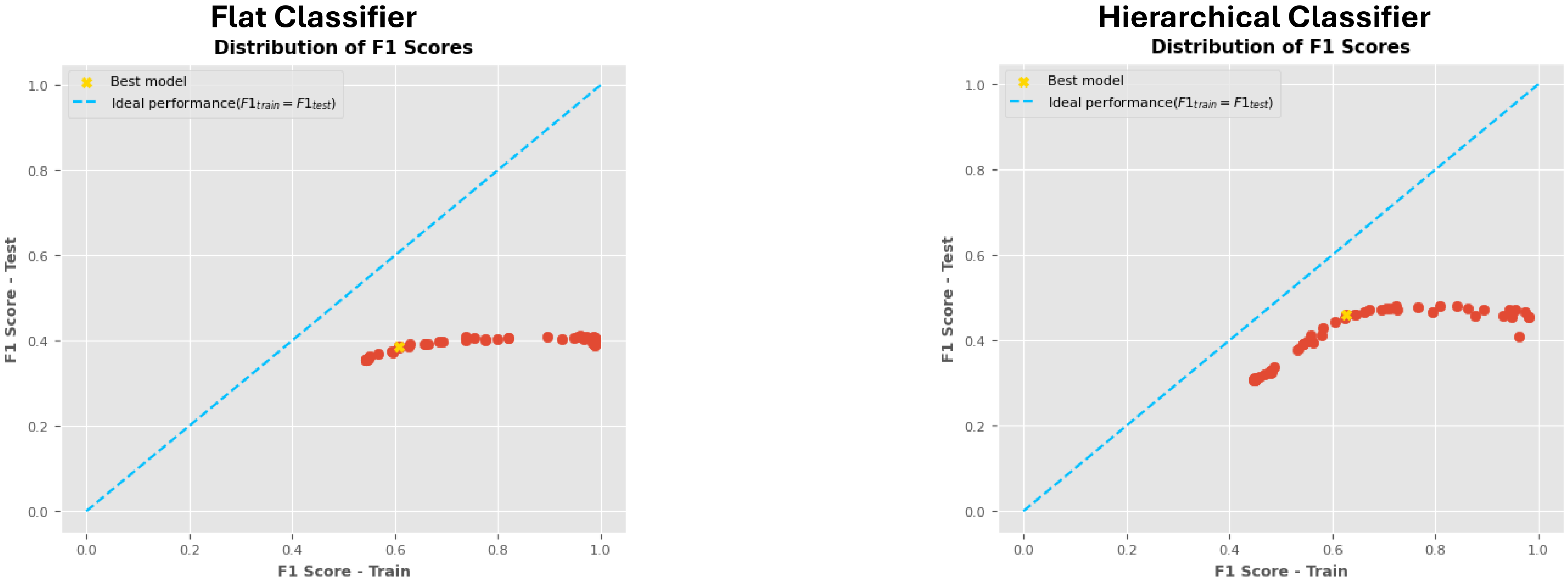
Table 4. Results of Tukey HSD test.

Group 1	Group 2	Mean Diff	p-value	Lower	Upper
Max Pooling	Mean Pooling	0.0331	0.0965	−0.0046	0.0707
Max Pooling	Truncate	−0.0384	0.0447	−0.076	−0.0007
Mean Pooling	Truncate	−0.0715	0.0001	−0.1091	−0.0338

Selected Technique

Mean pooling was chosen as the primary aggregation technique, as it yields a slightly higher average F1-score compared to max pooling.

Hyperparameter Optimization



Flat Classifier

Hierarchical Classifier

Figure 7. Distribution of F1 Scores for each Classifier Architecture. The yellow marker indicates the parameter combination that achieved the highest F1 Generalization Score within that architecture.

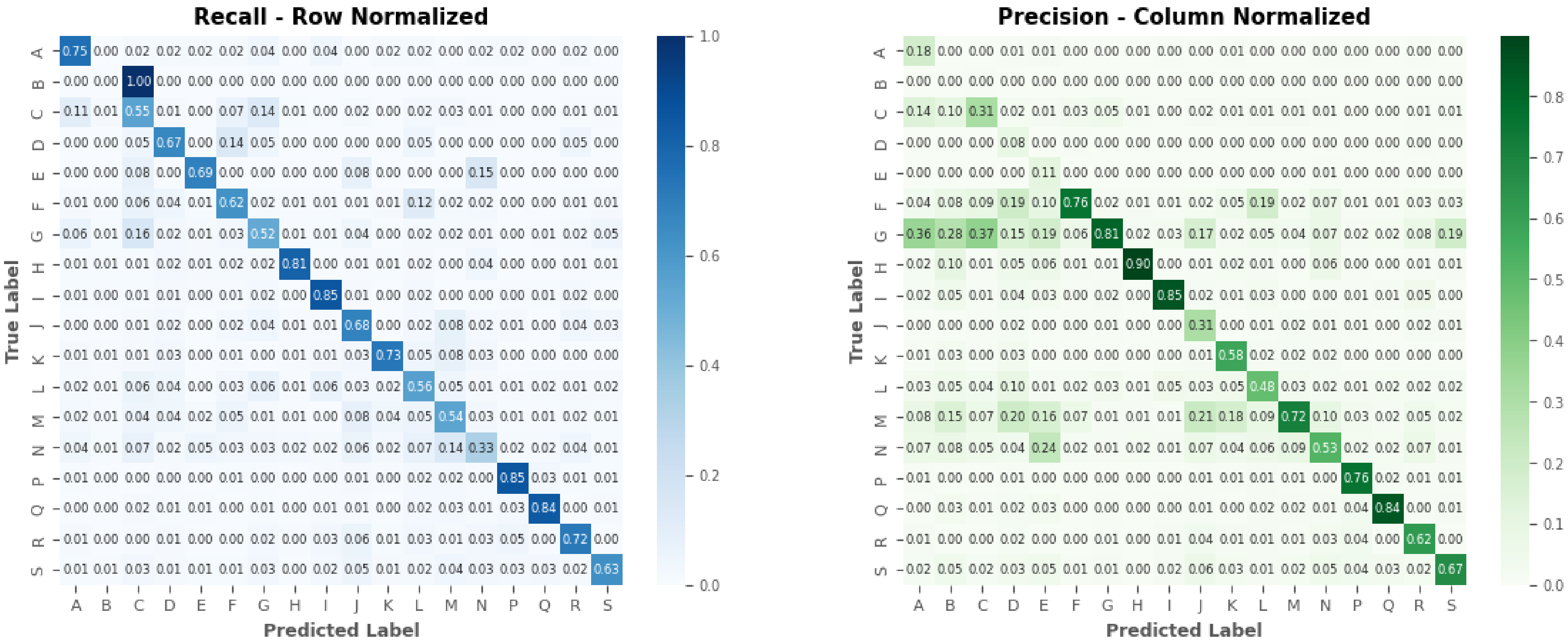
Hyperparameter Optimization

Table 5. Comparison of best performance model for each architecture.

	Flat Model	Hierarchical Model
Performance	0.3864	0.4593
Generalization	Large training-validation gap (~0.22)	Smaller gap (~0.17)
Robustness	More sensitive to overfitting	Less prone to severe overfitting
Computational Cost	~34.3s training time	~617.1s (approximately 15× slower)

Evaluation

The final model obtained an **F1 Score of 0.5608** on the validation sample of 5,400 companies.



Recall - Row Normalized

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S
A	0.75	0.00	0.02	0.02	0.02	0.02	0.04	0.00	0.04	0.00	0.02	0.02	0.00	0.02	0.02	0.00	0.02	0.00
B	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C	0.11	0.01	0.55	0.01	0.00	0.07	0.14	0.01	0.00	0.02	0.00	0.02	0.03	0.01	0.00	0.00	0.01	0.01
D	0.00	0.00	0.05	0.67	0.00	0.14	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00
E	0.00	0.00	0.08	0.00	0.69	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00
F	0.01	0.00	0.06	0.04	0.01	0.62	0.02	0.01	0.01	0.01	0.01	0.12	0.02	0.02	0.00	0.00	0.01	0.01
G	0.06	0.01	0.16	0.02	0.01	0.03	0.52	0.01	0.01	0.04	0.00	0.02	0.02	0.01	0.00	0.01	0.02	0.05
H	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.81	0.00	0.01	0.01	0.02	0.00	0.04	0.00	0.00	0.01	0.01
I	0.01	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.85	0.01	0.00	0.02	0.00	0.00	0.00	0.01	0.02	0.00
J	0.00	0.00	0.01	0.02	0.00	0.02	0.04	0.01	0.01	0.68	0.00	0.02	0.08	0.02	0.01	0.00	0.04	0.03
K	0.01	0.01	0.01	0.03	0.00	0.01	0.00	0.01	0.01	0.03	0.73	0.05	0.08	0.03	0.00	0.00	0.00	0.00
L	0.02	0.01	0.06	0.04	0.00	0.03	0.06	0.01	0.06	0.03	0.02	0.56	0.05	0.01	0.01	0.02	0.01	0.02
M	0.02	0.01	0.04	0.04	0.02	0.05	0.01	0.01	0.00	0.08	0.04	0.05	0.54	0.03	0.01	0.01	0.02	0.01
N	0.04	0.01	0.07	0.02	0.05	0.03	0.03	0.02	0.02	0.06	0.02	0.07	0.14	0.33	0.02	0.02	0.04	0.01
P	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.00	0.85	0.03	0.01	0.01
Q	0.00	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.02	0.03	0.01	0.03	0.84	0.00	0.01
R	0.01	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.03	0.06	0.01	0.03	0.01	0.03	0.05	0.00	0.72	0.00
S	0.01	0.01	0.03	0.01	0.01	0.01	0.03	0.00	0.02	0.05	0.01	0.02	0.04	0.03	0.03	0.03	0.02	0.63

Precision - Column Normalized

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S
A	0.18	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C	0.14	0.10	0.31	0.02	0.01	0.03	0.05	0.01	0.00	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01
D	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
F	0.04	0.08	0.09	0.19	0.10	0.76	0.02	0.01	0.01	0.02	0.05	0.19	0.02	0.07	0.01	0.01	0.03	0.03
G	0.36	0.28	0.37	0.15	0.19	0.06	0.81	0.02	0.03	0.17	0.02	0.05	0.04	0.07	0.02	0.02	0.08	0.19
H	0.02	0.10	0.01	0.05	0.06	0.01	0.01	0.90	0.00	0.01	0.02	0.01	0.00	0.06	0.00	0.00	0.01	0.01
I	0.02	0.05	0.01	0.04	0.03	0.00	0.02	0.00	0.85	0.02	0.01	0.03	0.00	0.00	0.01	0.01	0.05	0.00
J	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.31	0.00	0.01	0.02	0.01	0.01	0.00	0.02	0.01
K	0.01	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.58	0.02	0.02	0.02	0.00	0.00	0.00	0.00
L	0.03	0.05	0.04	0.10	0.01	0.02	0.03	0.01	0.05	0.03	0.05	0.48	0.03	0.02	0.01	0.02	0.02	0.02
M	0.08	0.15	0.07	0.20	0.16	0.07	0.01	0.01	0.01	0.21	0.18	0.09	0.72	0.10	0.03	0.02	0.05	0.02
N	0.07	0.08	0.05	0.04	0.24	0.02	0.01	0.02	0.01	0.07	0.04	0.06	0.09	0.53	0.02	0.02	0.07	0.01
P	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.76	0.02	0.01	0.01
Q	0.00	0.03	0.01	0.02	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.01	0.04	0.84	0.00	0.01
R	0.01	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.01	0.04	0.01	0.01	0.01	0.03	0.04	0.00	0.62	0.00
S	0.02	0.05	0.02	0.03	0.05	0.01	0.01	0.00	0.02	0.06	0.03	0.01	0.02	0.05	0.04	0.03	0.02	0.67

Figure 8. Confusion matrices normalized by row (left) and by column (right), representing class-wise recall and precision, respectively.

Evaluation

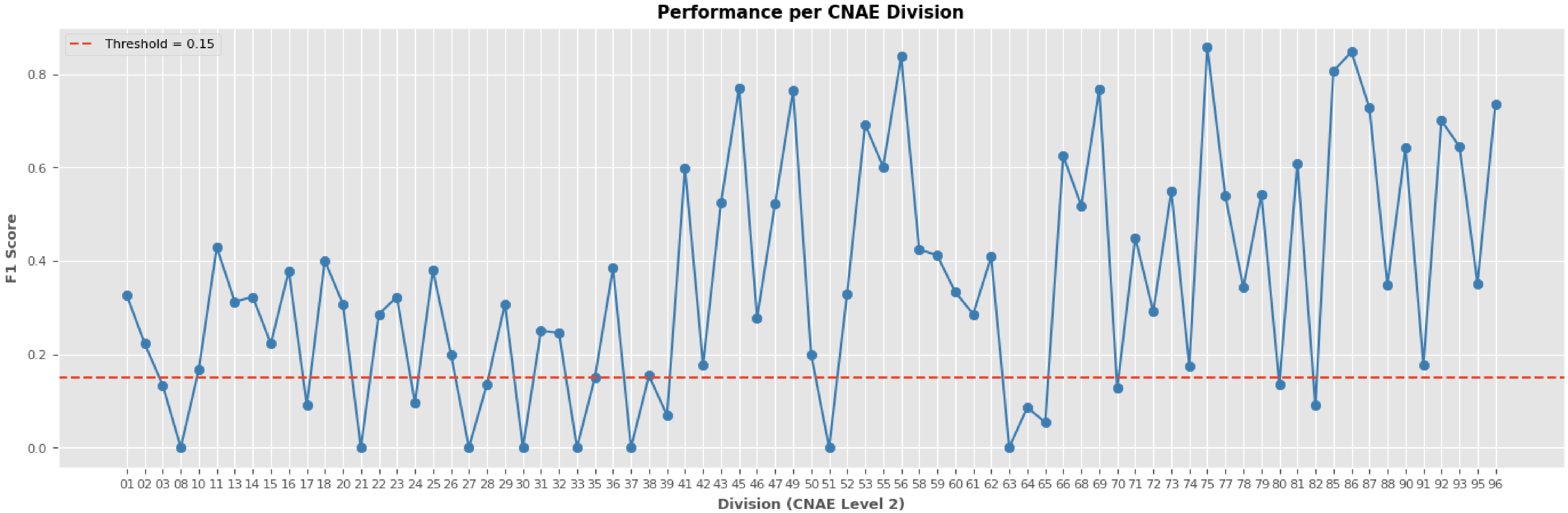


Figure 9. F1 Score per CNAE division (level 2) for the best-performing classifier. Each point represents the F1 score for a specific division. The red dashed line denotes a performance threshold set at 0.15, below which predictions are considered unreliable.

Qualitative Error Analysis

The analyzed sample consists of 36 randomly selected examples — 2 from each CNAE section — that were misclassified.

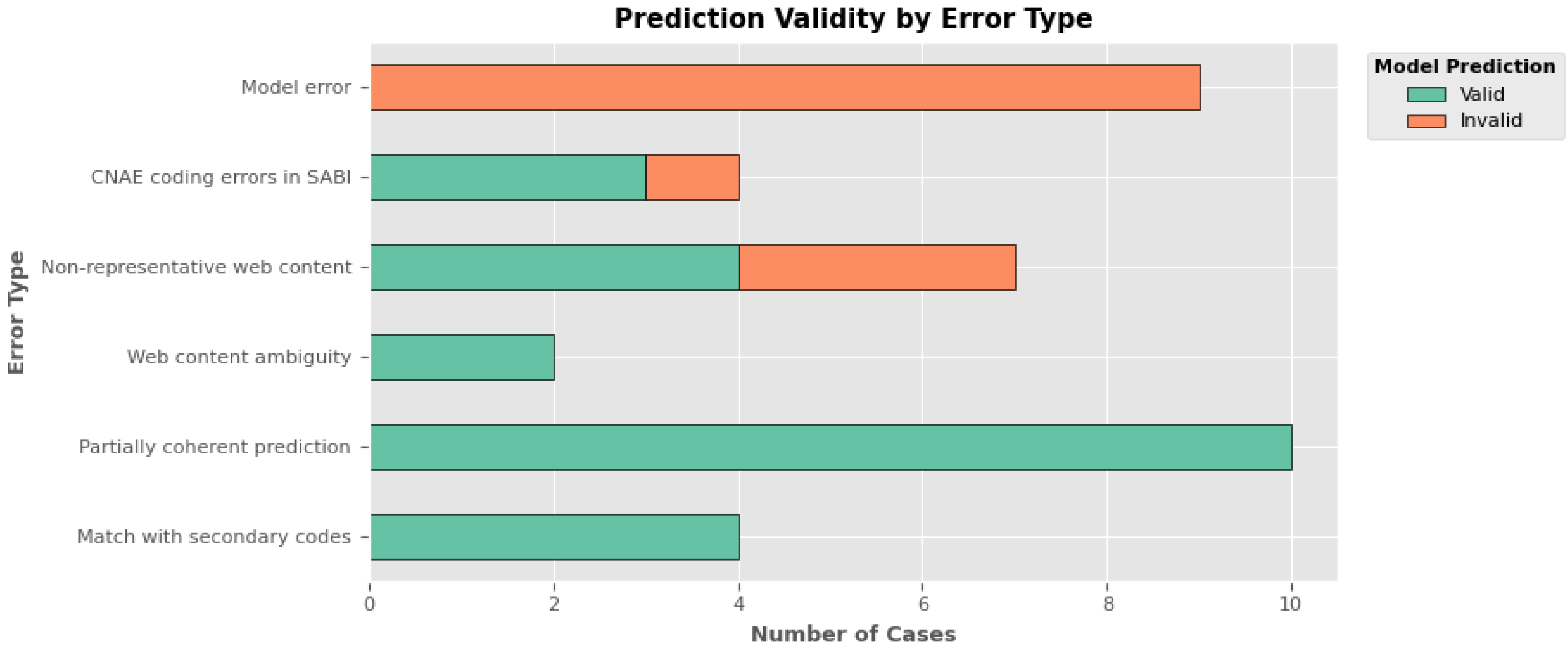


Figure 10. Distribution of prediction errors by type and validity. The analyzed sample consists of 36 randomly selected misclassified examples. Each bar represents the number of cases associated with a specific source of error, segmented by whether the model's prediction was valid (green) or invalid (orange).

5. Conclusions

Conclusions

- Despite errors in the original database (companies misclassified under economic activities), the model correctly classifies most cases.

- The large training corpus used reduces the impact of outliers (original misclassifications) and prevents them from skewing the model.

- The model’s ability to correct misclassifications suggests a potential use in **error detection in classification**.

Conclusions

- Automated classification based on web content is both **feasible** and **scalable**.

- The **aggregation** of embeddings is **essential** for preserving information and ensuring a coherent representation.

- **SMOTE** effectively addresses class imbalance, but it introduces **distortions** that may impact the model's generalization ability.

Future Work

- Expanding and diversifying the data sources by incorporating web content from other countries.
- Exploring more advanced oversampling techniques specifically designed for textual data.
- Generating synthetic web content using large language models (LLMs).
- Investigating more complex models based on deep learning architectures.

Thank you!

Any question?

References

Nahoomi, N. (2018). *Automatically coding occupation titles to a standard occupation classification* [Master's Thesis, The University of Guelph]. https://atrium.lib.uoguelph.ca/server/api/core/bitstreams/fd2005cf-97b4-430a-836e-ead2bee7c386/content

Cuffe, J., Bhattacharjee, S., Etudo, U., Smith, J. C., Basdeo, N., Burbank, N., & Roberts, S. R. (2019). Using public data to generate industrial classification codes. En *Big data for twenty-first-century economic statistics* (pp. 229–246). National Bureau of Economic Research. https://ideas.repec.org/h/nbr/nberch/14278.html

Roelands, M., van Delden, A., & Windmeijer, D. (2018, November). *Classifying businesses by economic activity using web-based text mining*. https://www.cbs.nl/-/media/_pdf/2017/47/topsectoren.pdf

Stateva, G., Dabrowski, D., Lasslop, G., Munter, P., Cierpial-Wolan, M., van Delden, A., Phelps, S., & ESSnet Trusted Smart Statistics - Web Intelligence Network. (2022). *WP3 1st interim technical report (Deliverable 3.1)* [Technical report]. European Union.

Smalbil, J. (2020). *Web-based economic activity classification* [Master's thesis, Delft University of Technology]. https://repository.tudelft.nl/record/uuid:f5f96ef9-8665-4c93-a932-34b8441976b0