Detecting AI-Generated Content Workshop – COLING 2025

Cross-table Synthetic Tabular Data Detection

G. Charbel N. Kindji [1,2], Lina M. Rojas-Barahona [1], Elisa Fromont [2], Tanguy Urvoy [1]

1 2

Introduction. (1/2)

- Generative AI: more effective data generation models (text, image, audio, video)
- Risk: data forgery
- Important to develop detectors
- Challenging task → Distribution shift (detectors struggle with new content)

- Tabular Data Generation → Hot topic
- High quality tabular data generators (TabSyn [1], TabDDPM [2], TVAE [3], CTGAN [3])
- Data forgery → E.g. fake accounting results
- Specific table issue → structure shift

1- Zhang et al., *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space*, ICLR 2024
2- Kotelnikov et al., *TabDDPM: Modelling Tabular Data with Diffusion Models*, ICML 2023
3- Xu et al., *Modeling Tabular data using Conditional GAN*, NeurIPS 2019

Introduction. (2/2)

- Synthetic tabular data detection → Classification problem
- Can use the same table structure:

Product ID	Price	Rating	Label
P001	19.99	4.5	Real
P265	29.99	3.0	Real
P4565	199.99	5.0	Synthetic
P018	39.99	4.2	Real
P107	100.00	8.5	Synthetic

- We focus on **cross-table** (cross-structure) and **table-agnostic** detectors:

Table Rows	Source
Product ID | Price | Rating P001 | 19.99 | 4.5	Real
Fruit | Quantity Apple | 54,80	Synthetic
Employee | Department | Salary | Level E2535 | Sales | 75000 | Senior	Real
Model | Brand | Year | RAM | Price XPS 13 | Dell | 2022 | 16GB | 1299.99	Real
Name | Major | GPA Alice S. | Biology | 15.5	Synthetic
Fruit | Quantity Banana | 15	Real

Synthetic Tabular Data Detection – No Shift.

- Training and testing on the same set of tables
- Two approaches:
  - Single Generator vs Real
  - All Generators vs Real

Training Data

Table Rows	Source
Product ID | Price | Rating P001 | 19.99 | 4.5	Real
Fruit | Quantity Apple | 54,80	TVAE
Employee | Department | Salary | Level E2535 | Sales | 75000 | Senior	Real
Model | Brand | Year | RAM | Price XPS 13 | Dell | 2022 | 16GB | 1299.99	Real
Name | Major | GPA Alice S. | Biology | 15.5	TabSyn

Testing Data

Table Rows	Source
Employee | Department | Salary | Level E0458 | Marketing | 35000 | Junior	TabSyn
Name | Major | GPA John D. | History | 3.9	TabDDPM
Fruit | Quantity Mango | 3,14	Real
Product ID | Price | Rating P6659 | 100.00 | 4.5	Real
Model | Brand | Year | RAM | Price ThinkPad X1 | Lenovo | 2023 | 16GB | 2265,98	Real

Synthetic Tabular Data Detection – Cross-table Shift.

Training Data

Training and testing on different table structures (cross-structure detectors)

Table Rows	Source
Product ID | Price | Rating P001 | 19.99 | 4.5	Real
Fruit | Quantity Apple | 54,80	TVAE
Employee | Department | Salary | Level E2535 | Sales | 75000 | Senior	Real
Model | Brand | Year | RAM | Price XPS 13 | Dell | 2022 | 16GB | 1299.99	Real
Name | Major | GPA Alice S. | Biology | 15.5	TabSyn

Testing Data

Table Rows	Source
Country | Population Canada | 409,19	CTGAN
Event ID | Name | Date | Location | Attendees 001 | COLING | 2025 | Abu Dhabi | 14678	Real
Course ID | Instructor | Credits CS4A A | Jack S. | -75	TabSyn
Brand | Model | Year Toyota | Camry | 1256	TVAE
Month | Sales | Region | Growth January | 450000 | South | 15%	Real

Text-based Encodings.

- Row linearization:

Table Rows	Source
Product ID | Price | Rating P001 | 19.99 | 4.5	Real
Fruit | Quantity Apple | 54,80	Synthetic
Employee | Department | Salary | Level E2535 | Sales | 75000 | Senior	Real
Model | Brand | Year | RAM | Price XPS 13 | Dell | 2022 | 16GB | 1299.99	Real
Name | Major | GPA Alice S. | Biology | 15.5	Synthetic

Table Rows	Source
‘’Product ID:P001,Price:19.99,Rating:4.5’’	Real
‘’Fruit:Apple,Quantity:54,80’’	Synthetic
‘’Employee:E2535,Department:Sales,Salary:75000,Level:Senior’’	Real
‘’Model:XPS 13,Brand:Dell,Year:2022,RAM:16GB,Price:1299.99’’	Real
‘’Name:Alice S.,Major:Biology,GPA:15.5’’	Synthetic

- Logistic Regression → characters trigrams:
- “Pro”, “duc”, “t l”, “D:P”

- Text-based Transformer → sequence of character embedded + positionnal embedding

Table-based Encoding.

- Data is kept in its tabular form

Step 1: Separate numerical and categorical features

Num. Features		Cat. Features			Source
Price	Rating	Product ID			Real
19.99	4.5	P001			
Quantity		Fruit			Synthetic
54,80		Apple			
Salary		Employee	Department	Level	Real
75000		E2535	Sales	Senior	
Year	Price	Model	Brand	RAM	Real
2022	1299.99	XPS 13	Dell	16GB	
GPA		Name	Major		Synthetic
15.5		Alice S.	Biology		

Step 2: Apply padding

Num. Features		Cat. Features			Source
Price	Rating	Product ID	X	X	Real
19.99	4.5	P001	X	X	
Quantity	X	Fruit	X	X	Synthetic
54,80	X	Apple	X	X	
Salary	X	Employee	Department	Level	Real
75000	X	E2535	Sales	Senior	
Year	Price	Model	Brand	RAM	Real
2022	1299.99	XPS 13	Dell	16GB	
GPA	X	Name	Major	X	Synthetic
15.5	X	Alice S.	Biology	X	

Experiments.

14 Tables

Name	Size	#Num	#Cat
Abalone	4177	7	2
Adult	48842	6	9
Bank Marketing	45211	7	10
Black Friday	166821	6	4
Bike Sharing	17379	9	4
Cardio	70000	11	1
Churn Modelling	4999	8	4
Diamonds	26970	7	3
HELOC	5229	23	1
Higgs	98050	28	1
House 16H	22784	17	0
Insurance	1338	4	3
King	21613	19	1
MiniBooNE	130064	50	1

4 Generators [1]

3 Detectors

- Logistic Regression on character trigrams
- Two Transformer Architectures:
  - text-based and
  - table-based

3 Setups

- No-shift setup: training and testing on the same set of tables
  - Single Generator
  - All Generators
- Cross-table shift: testing on a distinct set of tables

1- Kindji et al., *Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning*, ArXiv preprint 2024
2- Xu et al., *Modeling Tabular data using Conditional GAN*, NeurIPS 2019
3- Kotelnikov et al., *TabDDPM: Modelling Tabular Data with Diffusion Models*, ICML 2023
4- Zhang et al., *Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space*, ICLR 2024

Results – No Shift Setup.

Single Generator (TVAE) vs Real

- Decent performance of detectors (both AUC and Accuracy)
- AUC = 0.92 with table-based Transformer
- Kindji et al. [1] → Avg. AUC for table-wise TVAE's sample detection = 0.81

All Generators vs Real

- Little drop of performance
- Transformer-based approaches delivers best results

Cross-structure Setup	Model	Metrics	
		AUC	Accuracy
Single Generator (TVAE) vs Real	3grm-LReg.	0.72 ± 0.00	0.65 ± 0.00
	Text-Transf.	0.76 ± 0.01	0.67 ± 0.01
	Table-Transf.	$\mathbf{0.92 \pm 0.00}$	$\mathbf{0.83 \pm 0.00}$
All Generators vs Real	3grm-LReg.	0.64 ± 0.00	0.59 ± 0.00
	Text-Transf.	0.73 ± 0.04	0.66 ± 0.05
	Table-Transf.	$\mathbf{0.77 \pm 0.00}$	$\mathbf{0.69 \pm 0.00}$

1- Kindji et al., *Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning,* ArXiv preprint 2024

Results – Cross-table Shift.

- Cross-table shift is a challenging task
- Table-based model struggle to generalize (AUC=0.51)
- Text-based approaches seems to generalize better (AUC=0.60)

Model	Metrics	
	AUC	Accuracy
3grm-LReg.	**0.60 ± 0.05**	**0.52 ± 0.03**
Text-Transf.	**0.60 ± 0.07**	**0.52 ± 0.01**
Table-Transf.	0.51 ± 0.00	0.50 ± 0.00

Conclusion and Future Work.

- No-shift → reasonable performance
- Cross-table synthetic tabular data detection→ challenging task
- Data encoding is key for detector's performance (table-based vs text-based Transformer)

- Further investigation on Transformer-based detectors
  - Incorporating metadata in the model (column names / description, …)
  - Using pretrained models like TaBERT [1]
  - Explore additional distribution shift setups (cross-generator shift)

1- Yin et al., *TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data,* ACL 2020

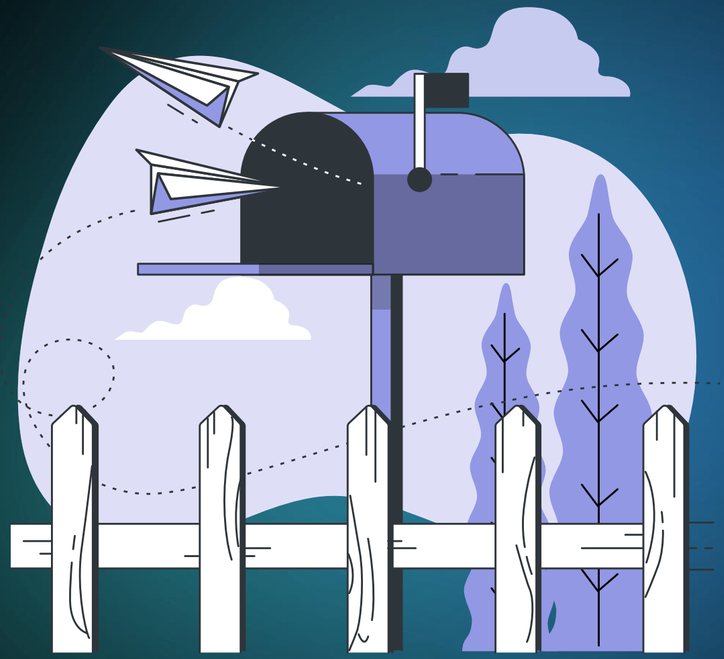
Cross-Table Synthetic Tabular Data Detection

G. Charbel N. Kindji, Lina M. Rojas-Barahona, Elisa Fromont, Tanguy Urvoy

Thank you

charbel.kindji@orange.com

charbel.kindji.orange@gmail.com



Illustrations from https://storyset.com/