TabSketchFM
Sketch based Tabular Representation Learning for Data Discovery over Data Lakes

Aamod Khatiwada, Harsha Kokel, Ibrahim Abdelaziz, Subhajit Chaudhury, Julian Dolby, Oktie Hassanzadeh, Zhenhan Huang, Tejaswini Pedapati, Horst Samulowitz, **Kavitha Srinivas**
IBM Research, Northeastern University, RPI

Table representation for data discovery

Data discovery often hinges on column similarity.

Building	Age
Chrysler	96
Trinity	178

People	Age
John Smith	23
Mary Taylor	45

Pyramids	Age
Khufu	4600
Khafre	4550

Column semantics is governed by:

- The larger table context
- Values in the column

Transformer architectures are great for capturing context BUT

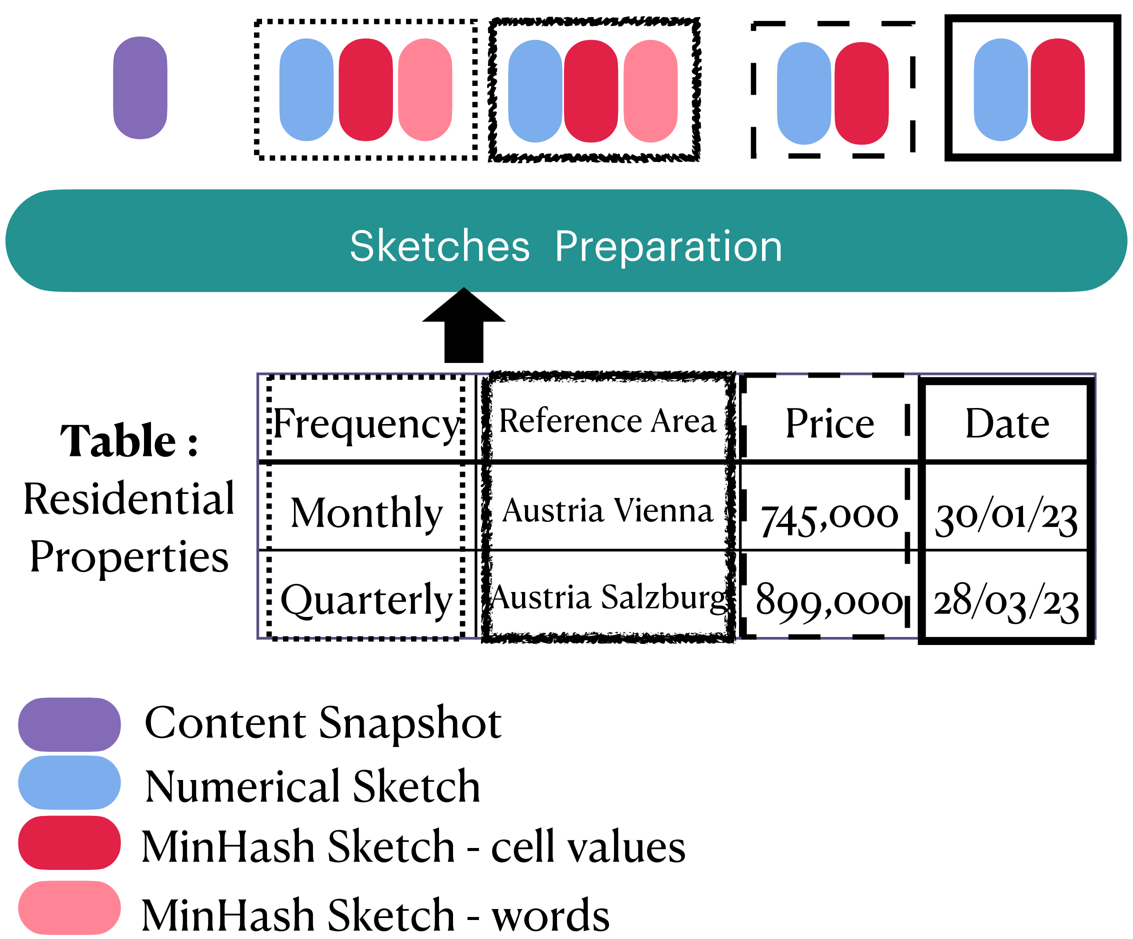
✘ Context Length

✘ Numeric values representation

Sketch based inputs for tables

MinHash: A locality sensitive hash that captures Jaccard similarity between sets of values.

Numerical: Number of unique values, NaNs, and for numbers: statistics such as mean, percentiles, max, min etc.

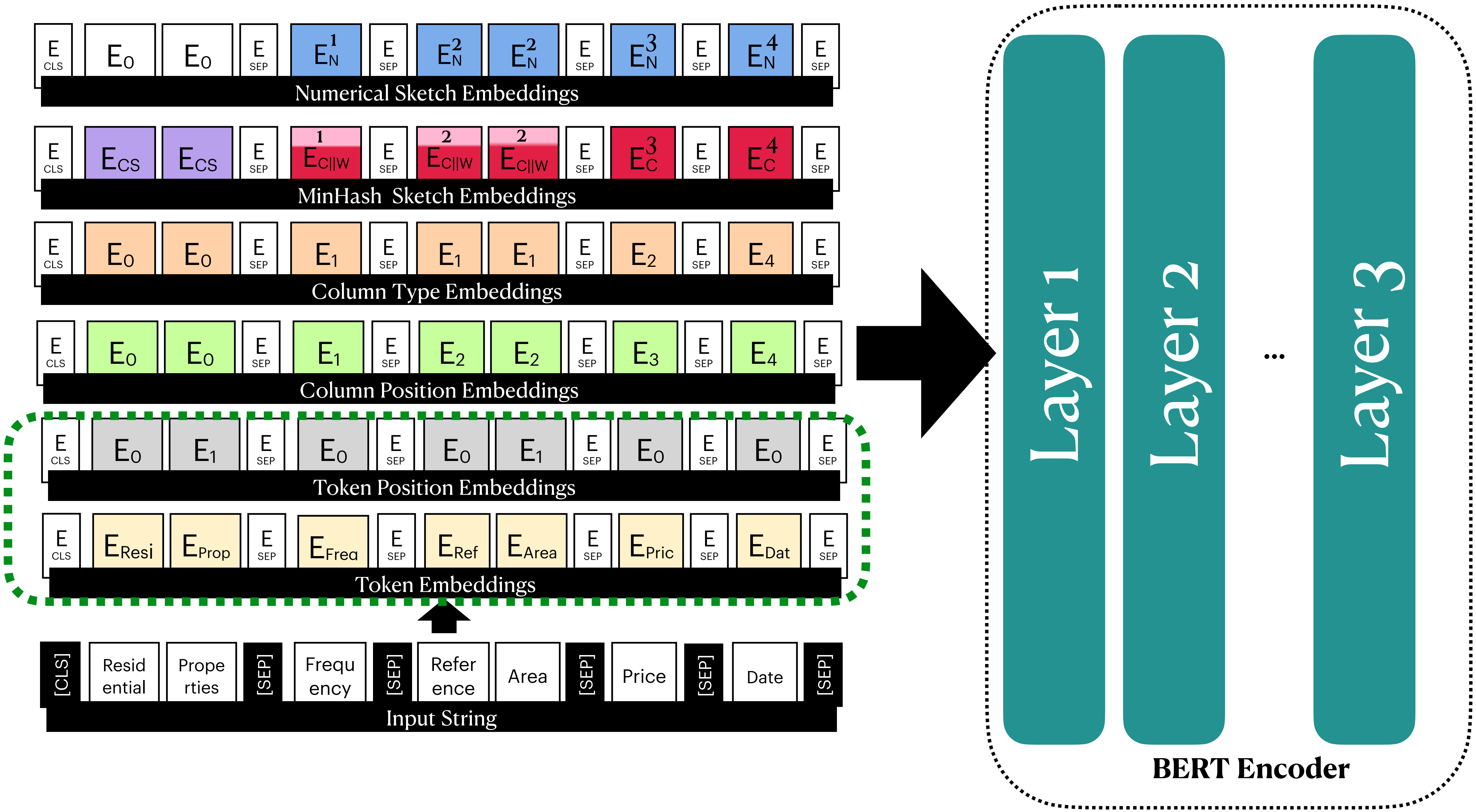


Content snapshot: Serialize each row as string, take hash.

String columns: Compute MinHash on cell values and words because these can often contain semantics of column (e.g. avenue/road)

TabSketchFM - Architecture

Overall approach - treat all sketches as inputs into embedding or linear layers. Sum after all vectors are of the same dimensionality.



Pretraining

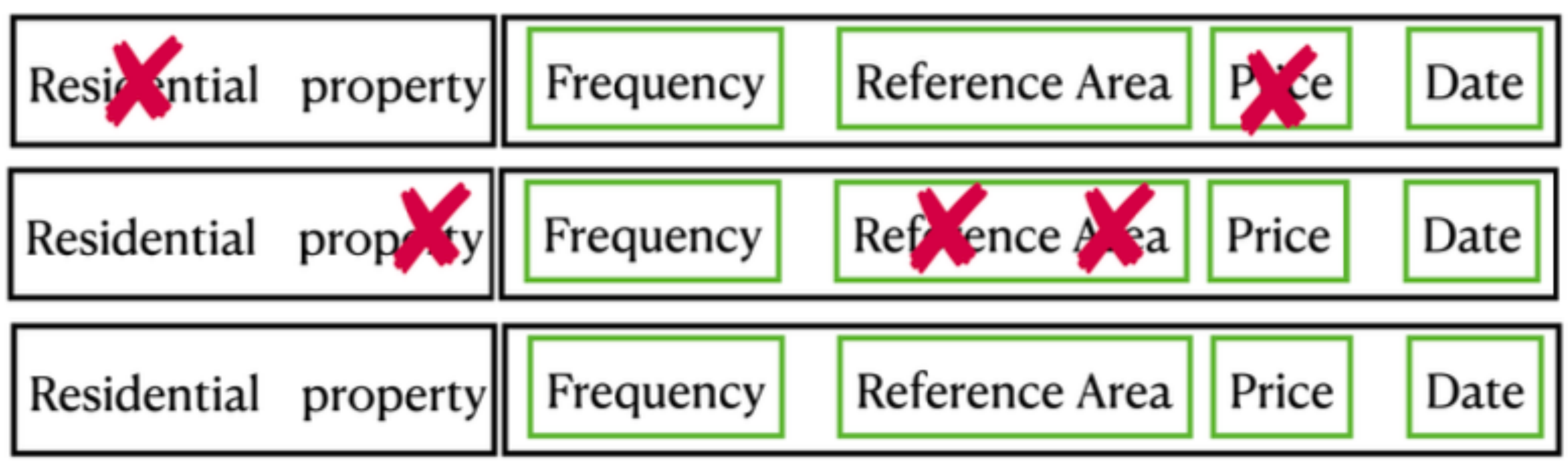
Dataset

197,254 CSVs from CKAN and Socrata

Average of 2234.5 rows, 35.8 columns

Data augmentation (change order of columns, per table)

MLM loss on column names or table description with whole column masking.



Finetuning

Lakebench - a dataset for tabular representation learning (**https://ibm.biz/LakeBench**)

Train

TABLE I: Cardinality of all the datasets in LakeBench, as well as search benchmarks in this paper.

	Benchmark	Task	# Tables	Avg. Rows	Avg. Cols	# Table Pairs Train	Test	Valid	Data type distribution (%) String	Int.	Float	Date
Union	TUS-SANTOS	Binary Classification	1127	4285.17	13.04	16592	3566	3552	77.94	8.62	7.51	5.93
	Wiki Union	Binary Classification	40752	51.05	2.66	301108	37638	37638	57.97	14.38	24.25	3.4
	ECB Union	Regression	4226	292.47	36.3	15344	1910	1906	47.72	14.05	36.31	1.92
Join	Wiki Jaccard	Regression	8489	47.3	2.8	12703	1412	1572	57.5	15.66	19.76	7.07
	Wiki Containment	Regression	10318	47.15	2.79	21007	2343	2593	57.26	15.26	20.58	6.9
	Spider-OpenData	Binary Classification	10730	1208.87	9.09	5146	742	1474	42.22	18.51	32.62	6.66
	ECB Join	Multi-label Clasification	74	8772.24	34.47	1780	222	223	52.14	7.8	37.79	2.27
Subsets	CKAN Subset	Binary Classification	36545	1832.58	25.37	24562	2993	3010	31.75	17.53	46.14	4.58
	Eurostat Subset	Search	38904	2157	10.46				64.63	9.03	7.83	18.50
	Wikijoin	Search	46521	49.64	2.68				58.13	13.35	25.0	3.50

Search at test only

Result: Fine tuned models for embedding tables for related table search. Join/Union/Subset discovery.

Join search

TABLE V: F1, Precision & Recall for Wiki-join search.

Baseline	**Mean F1**	**P@10**	**R@10**
TaBERT-FT	5.88	0.43	0.04
LSH-Forest	10.48	0.8	0.08
Josie	19.56	**0.99**	0.12
DeepJoin	18.88	0.96	0.11
WarpGate	18.58	0.95	0.11
SBERT	83.67	0.95	0.89
TabSketchFM	89.09	0.97	0.94
TabSketchFM-SBERT	**92.81**	0.98	**0.99**

SBERT: A baseline that takes a standard sentence encoder and encodes comma separated column values as a sentence.

TabSketchFM-SBERT: A combination of TabSketchFM embeddings and SBERT embeddings

Subset search

TABLE VIII: F1, Precision & Recall for Eurostat subset search.

Baseline	**Mean F1**	**P@10**	**R@10**
TABERT-FT	4.03	0.05	0.05
TUTA-FT	9.82	0.13	0.12
SBERT	43.12	0.56	0.51
TabSketch	**49.96**	**0.59**	**0.53**
TabSketch-SBERT	47.54	0.58	0.52

Union search

TABLE VII: F1, Precision & Recall for TUS union search.

Baseline	**Mean F1**	**P@60**	**R@60**
TaBERT-FT	26.66	0.90	0.30
TUTA-FT	27.27	0.89	0.31
Starmie	27.48	0.96	0.32
D3L	18.98	0.75	0.21
SANTOS	20.83	0.81	0.23
SBERT	**31.13**	**0.99**	**0.36**
TabSketchFM	30.43	0.97	0.35
TabSketchFM-SBERT	30.72	**0.99**	0.35

TABLE VI: F1, Precision & Recall for SANTOS union search.

Baseline	**Mean F1**	**P@10**	**R@10**
TaBERT-FT	36.64	0.63	0.46
TUTA-FT	25.34	0.43	0.3
Starmie	**54.08**	**0.97**	0.72
D3L	26.44	0.54	0.4
SANTOS	50.36	0.89	0.67
SBERT	53.86	**0.97**	**0.73**
TabSketchFM	51.38	0.92	0.69
TabSketchFM-SBERT	**54.09**	**0.97**	**0.73**

Conclusions

- Across a wide variety of tasks, TabSketchFM models did better than other systems both neural and traditional.
- Very good generalization across tasks and datasets.
- Important to include strong LLM baselines (e.g. SBERT)
- Artifacts:
  - Paper: https://ibm.biz/tabsketchfm_arxiv
  - Code: https://github.com/ibm/TabSketchFM
  - Models: coming soon on hugging face.