From Rules to Flexibility: A Resource and Method for SEC Item Extraction in Post-2021 10-K Filings

Xiao Li, Changhong Jin, Ruihai Dong

Insight Research Ireland Centre for Data Analytics
School of Computer Science, Univerisity College Dublin, Ireland

Insight
RESEARCH IRELAND CENTRE FOR DATA ANALYTICS

Taighde Éireann Research Ireland

1 Problem & Motivation

- Since **2021**, the SEC mandates **iXBRL** for 10-K filings. Traditional **RegEx / rule-based** parsers (built for plain HTML) often break under hybrid, deeply nested layouts.
- Researcher need **reliable**, **scalable item-level segmentation** on **recent filings** (2021–2024) to study today's markets.

🎯 ***GOAL:*** Deliver a **layout-robust** extraction method and a **standardized recent-years dataset** that works across heterogeneous iXBRL structures.

2 Contributions

Method: A **DOM-flattening** + **fuzzy-matching** framework that detects Item titles robustly, independent of brittle tree positions or exact phrasing.

Resource: A **Standardization item-segmented corpus** for **2021–2024** filings, and compatibility with previous 10-K corpus **EDGAR-CORPUS**.

Evaluation: An **automated validation protocol** assessing **coverage** and **ordering consistency**, plus expert spot-checks.

Outcome: **87.8% average extraction accuracy,** consistently outperforming RegEx and tree-based baselines.

Use Case: A **volatility forecasting** case study using **Item 1A** text demonstrates downstream utility .

3 Method: Layout-Robust Item Extraction

🔶 *Key Idea:*

Avoid brittle tree assumptions. **Flatten** visible content blocks, **normalize** text, then **fuzzy-match** against standardized Item title templates (e.g., "item 1 business").

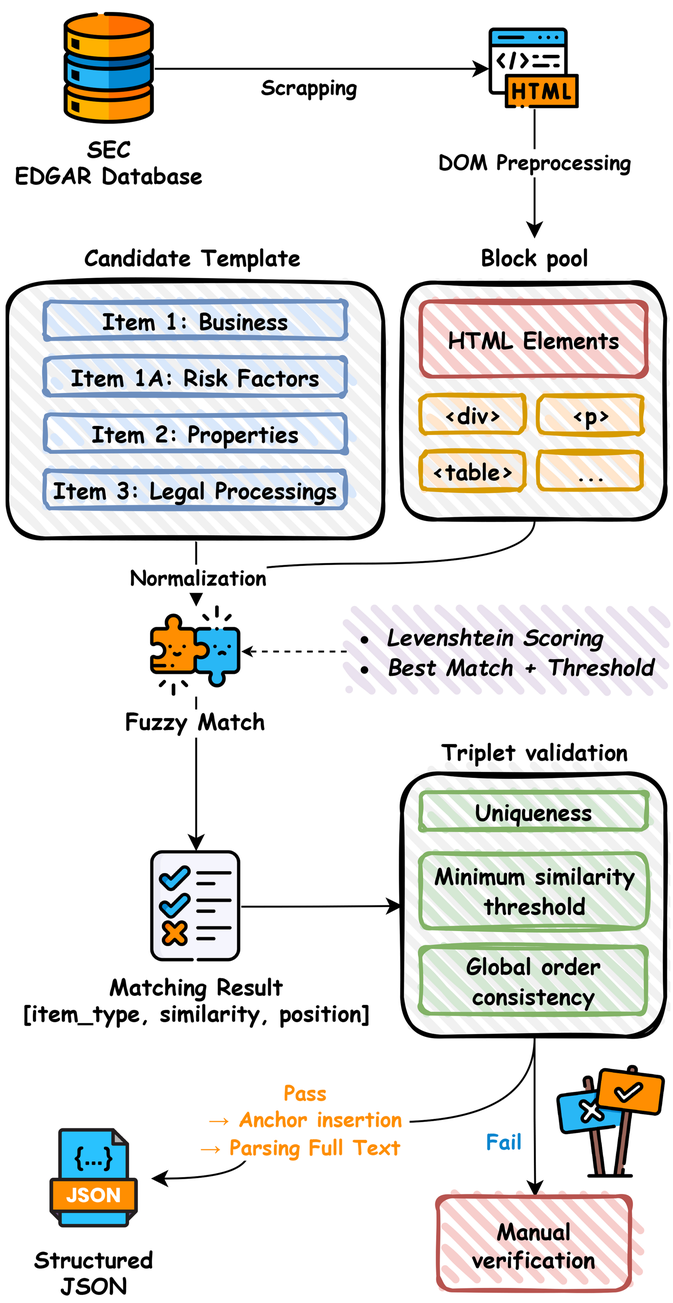
🔶 *Validation Steps:*

1. **Uniqueness** (one best match per Item)
2. **Minimum similarity threshold**
3. **Global order consistency** (monotonic titles)

Output: Annotated anchors → item-level **JSON** with titles, content, and positions.

🔶 *Dataset & Evaluation*

- **Scope:** S&P 500 universe, **~490 filings per year**, **2021–2024** (total **1968**).
- **Success criterion:** All standard Items correctly identified and validated; any missing/-misaligned Item →failure.
- **Human verification: 50 filings × 4 years** (200 total) manually checked—**no duplicated/misaligned** Items observed among the auto-successful set.



4 Results

Unified automated validation on 2021–2024 filings shows our method at **87.8%** average accuracy, outperforming **RegEx 72.9%** and **tree-based 67.0%**, with a 2024 **peak of 91.7%**.

Year	Regex (%)	Tree-based (%)	**Ours** (%)	Sample Size
2021	71.4	68.1	85.7	489
2022	72.7	69.5	87.4	491
2023	74.6	67.7	86.6	493
2024	72.9	67.3	91.7	495
Avg.	72.9	67.0	**87.8**	Total: 1968

5 Case study: Volatility Forecasting

🔶 *Task:*

Predict **post-filing n-day volatility** using **Item 1A**.

🔶 *Data:*

- **2005–2020** Item 1A from **EDGAR-CORPUS.**
- **2021–2024** from **our corpus** → mixed 2005–2024 & recent-only 2021–2024 settings.

🔶 *Findings:*

- Longer horizons (30–90 days) **reduce MSE** (more stable trends).
- Recent-only vs mixed: compatibility observed; **transformers may not always benefit** from longer history.
- Demonstrates **practical utility** of robust item segmentation for downstream finance tasks.

Table: Forecasting errors (MSE) across models and horizons.

Dataset	Models	Forecasting Horizon					
		n=3	n=7	n=15	n=30	n=60	n=90
Our Data (2021–2024)	XGBoost	0.068	0.041	0.028	0.016	0.015	0.014
	FinBERT	0.054	0.036	0.028	0.020	0.023	0.019
	RoBERTa	0.055	0.040	0.028	0.019	0.020	0.020
	Longformer	0.054	0.036	0.029	0.019	0.020	0.019
Mixed Data (2005–2024)	XGBoost	0.062	0.037	0.025	0.014	0.015	0.013
	FinBERT	0.053	0.034	0.036	0.037	0.025	0.022
	RoBERTa	0.052	0.034	0.032	0.018	0.024	0.017
	Longformer	0.054	0.036	0.035	0.027	0.020	0.017

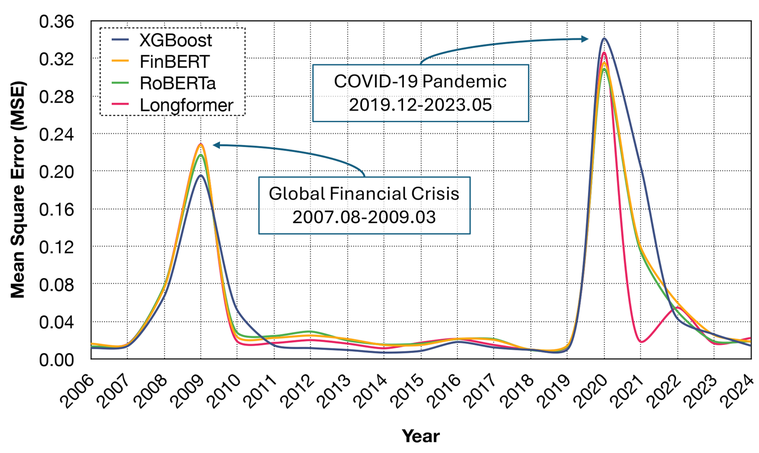


Figure: Model Error Trends in Forecasting Volatility

6 Error Analysis & Mitigations

Failures mainly from **non-standard headings** ("Item 1 Our Business") and **omitted Items** (e.g., 6 or 16).

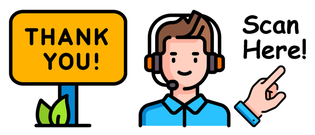
Mitigations: Title normalization, tolerance for non-mandatory Items, strict 7 vs 7A disambiguation, order constraints.

7 Acknowledgments & Contact

This work received support from Research Ireland [12/RC/2289_P2] and the China Scholarship Council (CSC), for which the authors are grateful.

Contact: xiao.li@ucdconnect.ie

GitHub: johnny-xiao-li/Flex_10K



Github

HOST INSTITUTIONS: DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University; Ollscoil na Gaillimhe University of Galway; UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh; University College Dublin An Coláiste Ollscoile, Baile Átha Cliath

PARTNER INSTITUTIONS: Maynooth University National University of Ireland Maynooth; Tyndall National Institute; Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin; University of Limerick Ollscoil Luimnigh

FUNDED BY: Taighde Éireann Research Ireland

CIKM SEOUL 2025
November 10 - 14