

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, University College Dublin, 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.
- CASE STUDY: 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 stan-



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.**

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

Findings:

- Longer horizons (30–90 days) reduce MSE (more stable trends).

