

Measuring the Market Impact of Financial News Using Lightweight NLP Models

Baptiste PRAS, Martin LEIVA, Vladimir HERRERA-NATIVI, Javier PEÑA-CASTAÑO

February 13, 2026

Project Objective

- Financial markets react to textual disclosures.
- Challenge: transform raw financial news into structured market events.
- Constraint: limited computational budget.
- Goal: build an end-to-end lightweight NLP pipeline.

Final representation:

(date, ticker, impact summary)

Bridge between text and market returns.

Global Pipeline

- ① Text cleaning and normalization
- ② Impact summarization (hierarchical)
- ③ Evaluation (ROUGE + LLM-as-a-judge)
- ④ Entity extraction and ticker linking
- ⑤ Market data retrieval

Why Text Cleaning Matters

Financial disclosures contain:

- Legal boilerplate
- Safe harbor statements
- Formatting noise
- Metadata headers

If not removed:

- Model focuses on irrelevant text
- Numeric distortion
- Entity confusion

Preprocessing Steps

1. Language Filtering

- FastText language identification
- English-only corpus

2. Boilerplate Removal

- Regex patterns for safe harbor
- Removal of structural blocks

3. Minimum Length Constraints

- Remove too-short articles
- Remove degenerate summaries

Summarization Challenge

Financial documents are:

- Long
- Numerically dense
- Multi-entity

Risk:

- Truncation
- Hallucination
- Numeric errors

Hierarchical Map–Reduce Strategy

Step 1: Chunking

- Overlapping token chunks
- Preserve context continuity

Step 2: Map Stage

- Each chunk → factual bullet notes

Step 3: Reduce Stage

- Consolidate notes
- Final 2–4 sentence impact summary

Training Design

- Supervised fine-tuning
- Same global target for each chunk
- Higher loss weight on first chunk

Prompt constraints:

- First sentence: “The [Issuer] . . . ”
- Preserve all numerical values
- No filler language

1. ROUGE-L

- Lexical overlap
- Useful baseline

Limitations:

- No hallucination detection
- No numeric verification
- No entity grounding check

LLM-as-a-Judge Framework

Six evaluation dimensions (0–5 scale):

- Accuracy
- Issuer grounding
- Numeric fidelity
- Coverage
- Conciseness
- Anti-filler professionalism

Focus on financial reliability.

Evaluation Results

Average scores:

- Accuracy: 3.00
- Issuer grounding: 3.17
- Numeric fidelity: 3.01
- Coverage: 3.07
- Conciseness: 2.08
- Anti-filler: 2.38

2 GPU-hours training → competitive performance.

Why Entity Extraction?

To link text to markets, we need:

(date, ticker)

NER → company names Linking → valid tradable ticker Merge → price data

Named Entity Recognition

- Pre-trained BERT NER model
- Extract ORG entities
- Inference only, no fine-tuning

Lightweight and efficient.

Ticker Linking

- Use yfinance metadata
- Normalize entity names
- Similarity-based matching

Coverage filtering:

- Keep reliable tickers
- 30% dataset retained

Final Structured Event

Each article becomes:

(date, ticker, impact summary)

Then:

- Retrieve daily prices
- Compute short-horizon returns
- Prepare regression dataset

Conclusion

What is completed:

- Robust text normalization
- Hierarchical summarization
- Finance-oriented evaluation
- Entity extraction and ticker linking
- Market data retrieval

Ongoing work:

- Regression modeling
- Abnormal return analysis
- Statistical significance testing

Lightweight models + structured integration = Practical financial NLP pipeline.