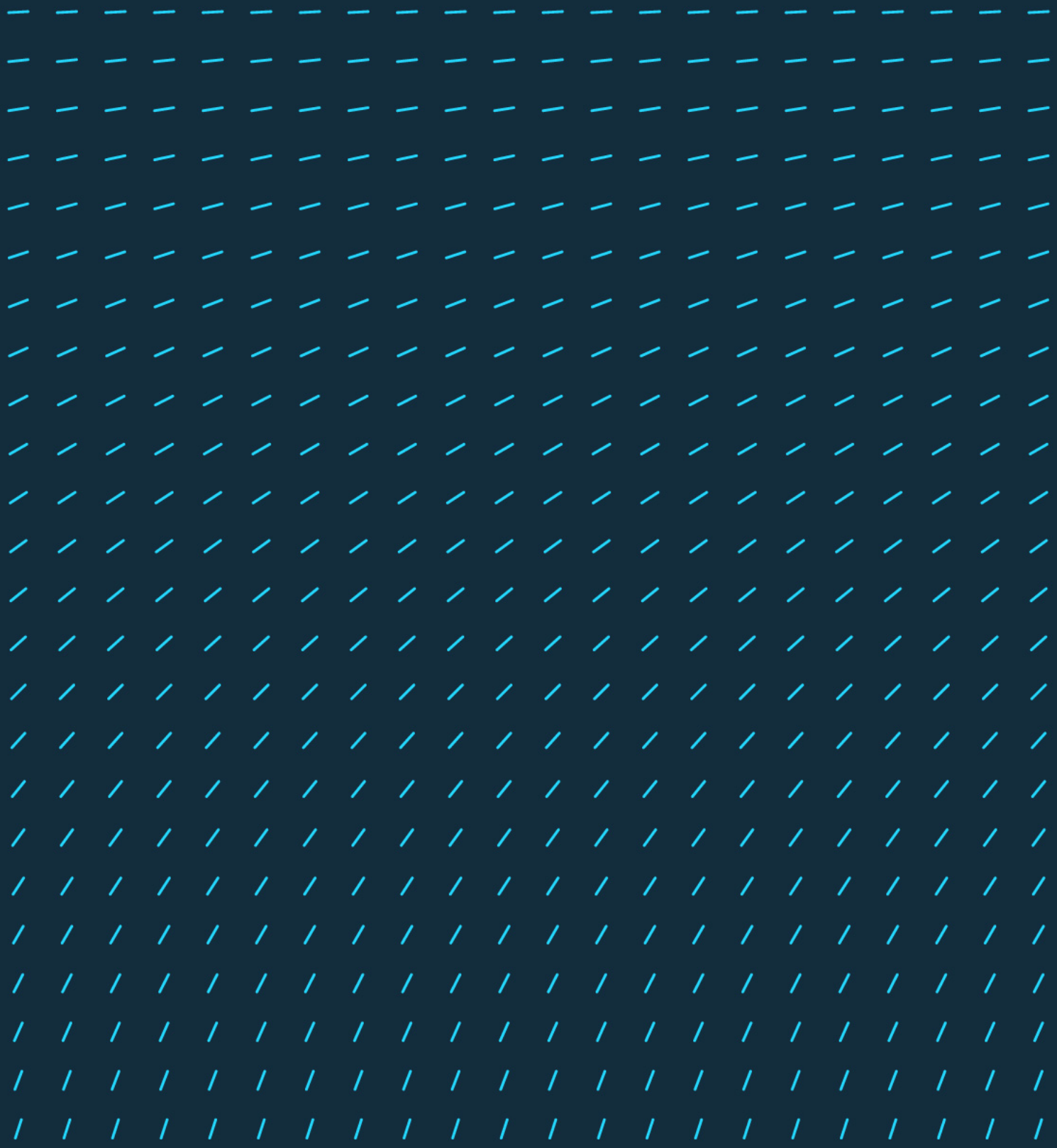

August 2025

SiFT

130 Person-Years of Rule-Writing Powering the Fastest, Most Accurate Event Extraction

Commetric

Table of Contents

Overview

Since 2005, Commetric has developed and deployed almost every major AI/ML and NLP:

- LLM-based Services & Models: OpenAI, Gemini, Deepseek; on-premises Llama, Qwen, Gemma, Mistral, Granite
- Embedding & Retrieval: Semantic Embedding, FAISS, Retrieval Augmented Generation (RAG), Clustering, Translation, Topic-modelling
- Traditional ML: Gradient Boosted Trees (XGBoost), SVMs, Regression, Ensembles, Heuristics
- Finite State Transducers (FSTs)

Event Extraction with SiFT

This page concerns one specific and critical competency: automatic Event Extraction from corporate news. We outline Commetric SiFT (Siera Finite-state Transducer), comparing our implementation with cheaper-to-develop AI methods.

What is SiFT?

SiFT serves as the pivotal link between unstructured news and the actionable intelligence generated by multi-agent AI systems. As outlined in the sections below, SiFT's architecture consistently delivers the highest performance across all KPIs—achieving superior recall, unmatched precision, blazing-fast processing speeds, and operational cost efficiency—though it does come with one trade-off: a high development cost (130 man-years).

SiFT is Commetric's proprietary event-harvesting platform. It recognizes over 440 unique event types across a two-tier event taxonomy, such as DISRUPTION_Explosion, FINANCIAL_margindecline, OPERATIONS_maintenanceshutdown, LITIGATION_securitiesclassactionfiled, CEO_bonus, and CSR_racialdiscrimination, among others.

Each extracted event is mapped to the relevant corporate entity (via company ID) and stored in our advanced knowledge base, SieraPlex, enabling precise and contextual insights. SiFT is a rule-based, fully deterministic system, relying on compiled C-language executables to run Commetric's NLP (Natural Language Processing) rules on incoming textual data.

This architecture ensures SiFT has broad access to corporate news, even in environments with restrictive publisher policies, making it highly efficient in extracting and processing critical data from diverse sources.

Pre-SiFT Lexicographic Filtering

Purpose: Reduce noise and improve system efficiency by only submitting articles with relevant company mentions.

- Custom dictionaries for over 5,000 companies, each containing:

- Distinctive name forms (e.g., ExxonMobil, Exxon)
- Ambiguous/short forms (e.g., Apple, BP, GM), which require corroboration
- Associated officers (e.g., Darren Woods, Mary Barra)
- Subsidiary names (e.g., Texaco under Chevron)
- Not classified as AI processing—permissible even under strict publisher terms.

On-premises Operations

SiFT can process 1 million articles in about 6 minutes. A five-year run (1.8 billion articles) takes approximately 2,300 core-days. Standard processing scope is up to 1,000 bytes per article (title + lead text).

Suggested hardware:

- Two Dell PowerEdge R7625, each with 192 physical cores (2 × AMD EPYC 9654) and 1TB RAM.
- One Dell PowerEdge R7625, with 64 physical cores (AMD EPYC 9554) and 1TB RAM.

Operational Cost @ 24/7: $103/day ($37,595/year): Server depreciation $68, Power $15, Cooling $6.

Alternatives to Finite State Transducers

Recent AI advances have lowered development effort but fall short in several areas:

- Low recall: >30% missed events
- Low precision: ≥35% mis-tagging
- High cost: 10×–50× slower → expensive GPU/cloud
- Publisher bans: Narrower content reach
- Poor semantic mapping: 90% mis-tagging for fine-grained events

Example: For 1,000 articles tagged FINANCIAL_margindecline, semantic similarity may cluster on irrelevant dimensions like sector, macro, or PR spin, not the actual event.

Why SiFT Wins

- Highest Accuracy: Thousands of handcrafted rules = near-perfect recall/precision
- Blazing Throughput: Linear C-compiled processing @ sub-millisecond/article
- Minimal Cost: Commodity servers, no GPU, no cloud
- Full Coverage: Not classed as AI/ML = unrestricted access
- On-premises processing

SiFT Output → SieraPlex Knowledge Base

- 440 event types
- Disambiguated company links (ID, ticker)
- 480M+ extractions across 10 years
- Equity prices linked to news for impact analysis
- No restriction on post-SiFT AI use
- Includes: frequency, elevation, price/volume impact, dates, sources, named entities
- Best practice tools and data for enterprise risk and crisis analytics

---

"No one will ever match our 130 person-year, multi-generational investment in event extraction expertise."

Human Language Translation

SiFT processes English. For other languages, pre-stage machine translation (MT) works well. However, MT may be subject to publisher terms.

Stability is key—changes in MT quality may create false trends. For major updates, rerun SiFT on archived content.

Translation speed:

- Mac Ultra: 0.83/sec → 71,000/day
- A100 GPU: ~15/sec → daily coverage for all major EU languages
- Running cost: ~$17/day (server amortisation) + $3.50 (power/cooling)
- Google Translate: likely thousands/day
- Post-lexicographic filter insertion → 5× capacity gain
- A100 rental: ~$1,000/month per GPU

Where SiFT Powers Commetric Solutions

- Cogent: Hybrid media analysis platform (1M+ usage hours)
- Internal IRIS: Auto-configured corporate newsletters (job-function-specific)
- External EventDaaS: Issues & Risks enterprise reports

- XSELL & PINT: Sales intelligence tools
- ComVix: Equity feed with corporate event context

Client Deployments

- A prominent provider of innovative payment solutions
- A major global player in the automotive industry
- A renowned Swiss-based pharmaceutical corporation
- A well-established American biopharmaceutical enterprise
- One of Switzerland’s top insurance firms
- A globally recognised name in biotechnology
- A key American company in food processing and manufacturing
- A world-renowned leader in energy and chemical sectors
- A respected British multinational in financial technology
- A distinguished American consultancy specializing in brand strategy

—— Ready to integrate the most accurate, highest speed event extraction into your analytics workflow? Reach out to Commetric for a demo of SiFT today.

Terminology / Primer

What is a Transducer?

Computational model that maps input sequences (e.g. text) to output sequences (e.g. annotations).

Finite State Transducer (FST)

- Outputs symbols, not raw text
- Memory-efficient: only state is stored
- Deterministic: high performance
- Linear time complexity (unlike exponential alternatives). Processing time linear growth with volume increase.

FST Use Cases

- Tokenizer: 'She didn’t go.' → [She, did, n’t, go]
- Morphological Analyser: Unhappiness → un- + happy + -ness
- Tagger:
    - POS: 'The cat sat on the mat.' → [The/DET, cat/NOUN, sat/VERB, ...]
    - NER: 'Donald Trump was not born in Hawaii.' → [Donald Trump/PER, Hawaii/LOC]

- Noun Phrase ID: 'the powerful but disgraced CEO' → NP → e.g. Jeffrey Skilling (Enron)

SiFT Specifics: C-based FST network. Thousands of compiled rules. Sub-millisecond per article. Stateless and thread-safe. Outperforms neural nets on long-tail/fine-grained events. Fully auditable and deterministic. On-premise processing, but cloud compatible.