FinSight AI

Directional Sentiment Pressure (DSP) Engine for Equity Markets

White Paper – November 2025

Abstract

FinSight AI is a lightweight, production-oriented sentiment intelligence system that converts raw global news into actionable institutional signals. The engine ingests financial news, regulatory filings, and market metadata, applies deep relevance filtering, processes sentiment through LLM-based models, and transforms extracted signals into a normalized Directional Sentiment Pressure (DSP) score in the range [−3,+3].
This document outlines the architecture, mathematical formulation, data pipeline, weighting logic, relevance filtering, storage schema, and score computation methodology behind the DSP engine.

Table of Contents

1. Introduction

FinSight AI provides an end-to-end pipeline that transforms noisy, unstructured financial news into stable sentiment-based alpha signals. The DSP metric aggregates per-article directional impact into a normalized score that is invariant to article count, source distribution, and short-term sentiment imbalance.

The system was designed for:

- equity screening,
- intraday and swing-trading overlays,
- risk-aware decision support, and
- factor-model augmentation.

2. System Overview

FinSight AI consists of six stages:

1. **Data Ingestion**
2. **Relevance Filtering**
3. **LLM Sentiment Analysis**
4. **Dynamic Weighting**
5. **DSP Engine**
6. **Aggregation & Serving Layer**

Each stage is modular and independently replaceable.

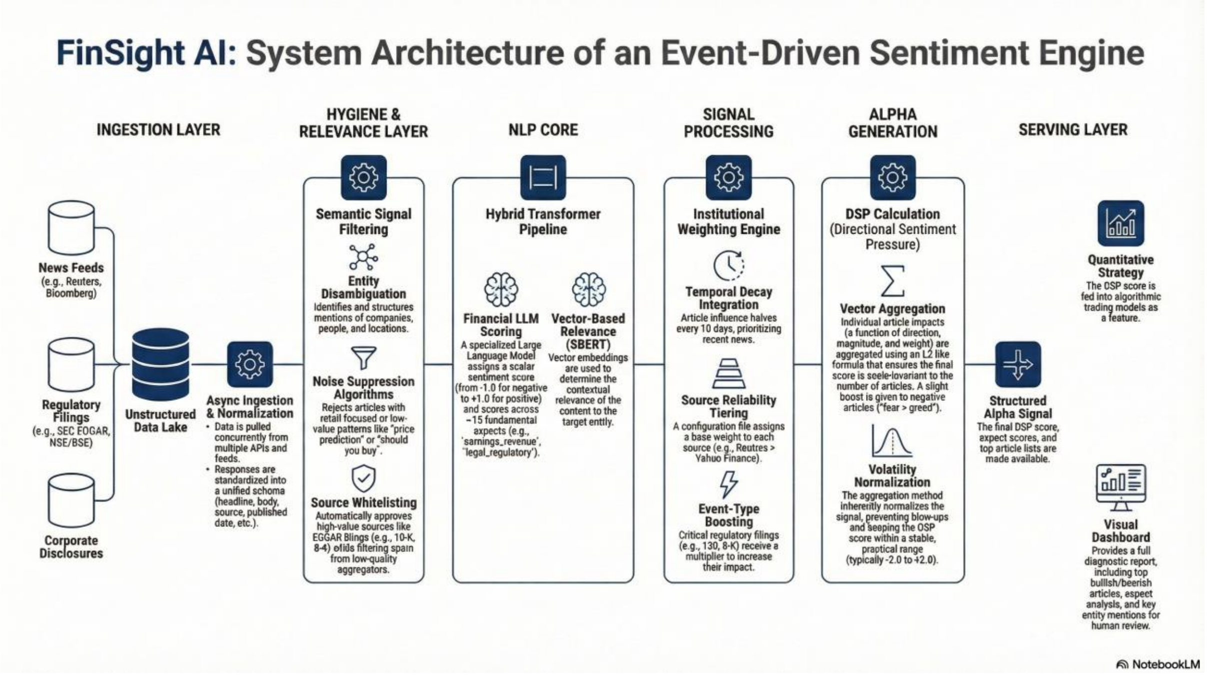

FinSight AI: System Architecture of an Event-Driven Sentiment Engine
INGESTION LAYER
HYGIENE & RELEVANCE LAYER
NLP CORE
SIGNAL PROCESSING
ALPHA GENERATION
SERVING LAYER
News Feeds (e.g., Reuters, Bioomberg)
Regulatory Filings (e.g., SEC EOGAR, NSE/BSE)
Corporate Disclosures
Unstructured Data Lake
Async Ingestion & Normalization
• Data is pulled concurrently from multiple APIs and feeds.
• Responses are standardized into a unified schoma (headline, body, source, published date, etc.).
Semantic Signal Filtering
Entity Disambiguation
Identifies and structures mentions of companies, people, and locations.
Noise Suppression Algorithms
Rejects articles with retail focused or low-value patterns like "price prediction" or "should you buy".
Source Whitelisting
Automatically approves high-value sources like EGGAR Blings (e.g., 10-K, 8-4) ofids filtering spain from low-quality aggregators.
Hybrid Transformer Pipeline
Financial LLM Scoring
A specialized Large Language Model assigns a scalar sentiment score (from -1.0 for negative to +1.0 for positive) and scores across ~15 fundamental axpects (e.g., 'sarnings_revenue', 'legal_regulatory').
Vector-Based Relevance (SBERT)
Vector embeddings are used to determine the contextual relevance of the content to the target entity.
Institutional Weighting Engine
Temporal Decay Integration
Article influence halves every 10 days, prioritizing recent news.
Source Reliability Tiering
A configuration file assigns a base weight to each source (e.g., Reutres > Yahuo Finance).
Event-Type Boosting
Critical regulatory filings (e.g., 130, 8-K) receive a multiplier to increase their impact.
DSP Calculation (Directional Sentiment Pressure)
Vector Aggregation
Individual article impacts (a function of direction, magnitude, and weight) are aggregated using an L2 like formula that ensures the final score is seele-lovariant to the number of articles. A slight boost is given to negative articles ("fear > greed").
Volatility Normalization
The aggregation method inhereritly normalizes the signal, preventing blow-ups and seeping the OSP score within a stable, praptical range (typically -2.0 to +2.0).
Structured Alpha Signal
The final DSP score, expect scores, and top article lists are made available.
Quantitative Strategy
The OSP score is fed into algorithmic trading models as a feature.
Visual Dashboard
Provides a full diagnostic report, including top bullish/beerish articles, espect analysis, and key entity mentions for human review.
NotebookLM

Figure 1: End-to-End System Overview

3. Data Ingestion Layer

3.1 News Sources

Tiered ingestion from:

- Tier-1 media (Reuters, Bloomberg, WSJ)
- Tier-2 portals (Yahoo Finance, MarketWatch)
- Indian media (Moneycontrol, Economic Times, LiveMint)
- Alternative sources (Biztoc, Benzinga, aggregator feeds)

3.2 Regulatory Filings

Supported filings:

- SEC: 8-K, 10-K, 10-Q, 13D, 13G, Form-4
- NSE/BSE filings for Indian equities

3.3 Normalization

Every document is normalized into:

{ headline, body, source, published_at, symbol, entities }

3.4 LLM Summaries

Long-form articles are compressed using internal LLM summarization for consistent downstream text length.

4. Relevance Engine

The relevance engine removes noise aggressively before any scoring occurs.

4.1 Normalization

Lowercasing, punctuation removal, entity cleanup.

4.2 Company Matching

- Symbol match (“AAPL”, “TSLA”)
- Name match (“Tesla”, “Apple Inc.”)
- Alias match (e.g., “Alphabet” → GOOG)

4.3 Noise Filtering

Articles are discarded if:

- they exhibit SEO-spam patterns (“boosts stake”, “should you buy”),
- they stem from low-quality sources (certain aggregator spam),
- they describe minor portfolio churn irrelevant to fundamentals.

4.4 Filing Override

Regulatory filings bypass relevance filters and are always included.

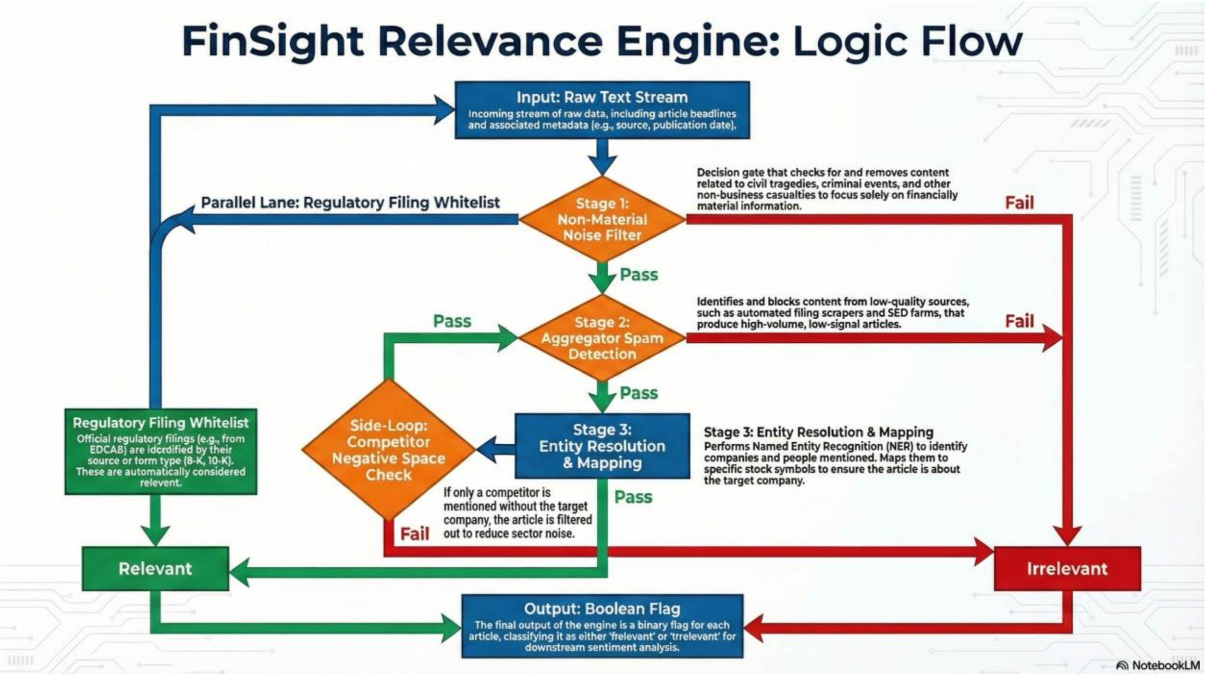

FinSight Relevance Engine: Logic Flow
Input: Raw Text Stream
Incoming stream of raw data, including article beadlines and associated metadata (e.g., source, publication date).
Parallel Lane: Regulatory Filing Whitelist
Stage 1: Non-Material Noise Filter
Decision gate that checks for and removes content related to civil tragedies, criminai events, and other non-business casualties to focus solely on financially material information.
Fail
Pass
Pass
Stage 2: Aggregator Spam Detection
Identifies and blocks content from low-quality sources, such as automated filing scrapers and SED farms, that produce high-volume, low-signal articles.
Fail
Pass
Regulatory Filing Whitelist
Official regulatory filings (e.g., from EDCAB) are idcrdified by their source or torm type (8-K, 10-K). These are automatically considered relevent.
Side-Loop: Competitor Negative Space Check
Stage 3: Entity Resolution & Mapping
Stage 3: Entity Resolution & Mapping
Performs Named Entity Recognition (NER) to identify companies and people mentioned. Maps them to specific stock symbols to ensure the article is about the target company.
If only a competitor is mentioned without the target company, the article is filtered out to reduce sector noise.
Pass
Fail
Relevant
Irrelevant
Output: Boolean Flag
The final output of the engine is a binary flag for each article, classifying it as either 'frelevant' or 'trrelevant' for downstream sentiment analysis.
NotebookLM

Figure 2: Relevance Engine

5. NLP & Sentiment Layer

5.1 LLM Sentiment

A fine-tuned sentiment model outputs:

sentiment_score ∈ [-1.0, +1.0]

5.2 Aspect Sentiment

Fifteen business dimensions are computed:

- earnings & revenue
- costs & margins
- liquidity
- dividends & buybacks
- leadership
- strategy
- M&A
- legal & regulatory
- competition
- product
- technology
- supply chain
- customer experience
- brand
- labor force

5.3 Entity Extraction

NER (Named Entity Recognition) models extract organizations, persons, and locations.

6. Dynamic Weighting Engine

Each article receives a numerical weight:

```
final_weight = ( base_source_weight
                × recency_factor
                × engagement_factor
                × filing_adjustment)
```

6.1 Base Source Weight

Loaded from internal trust model. (Redacted in public repo.)

6.2 Recency Decay

Non-linear temporal decay calibrated to the swing trading horizon

6.3 Engagement Factor

Engagement impact is normalized using a logarithmic scaling function with a proprietary volatility dampening coefficient to marginalize viral noise while preserving signal magnitude.

7. DSP Computation

A normalized, direction-weighted, magnitude-adjusted aggregation of validated news sentiment over a fixed time window.

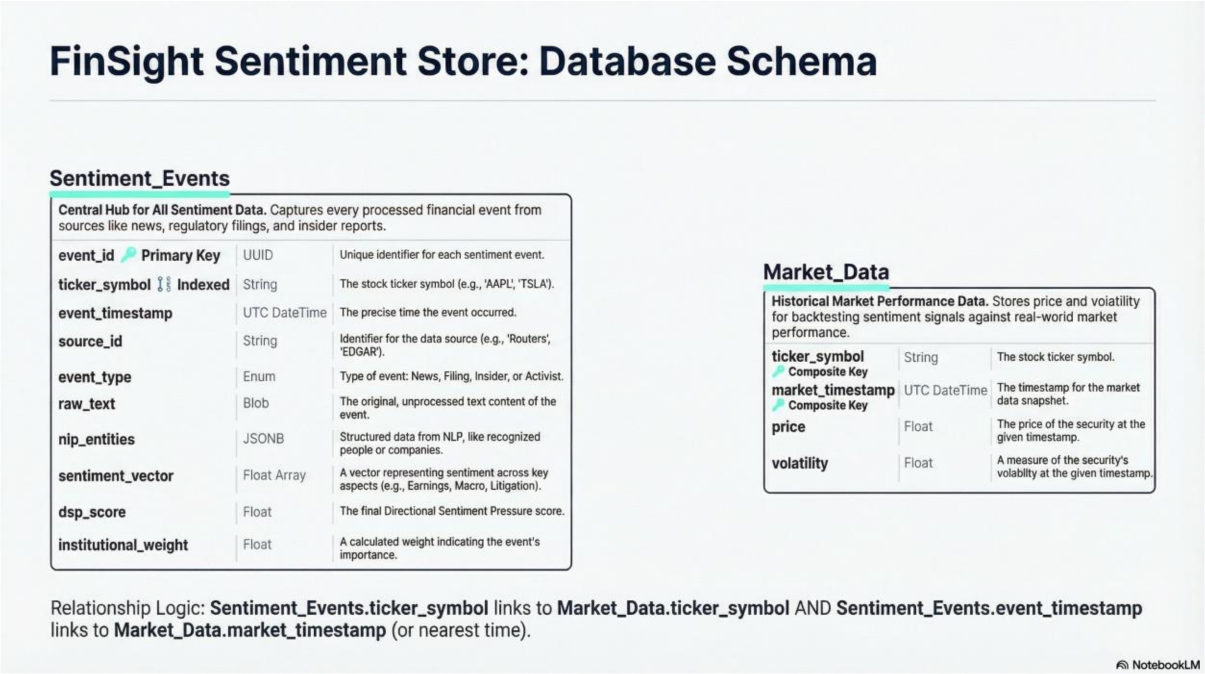
DSP is clipped to [−3.0, +3.0].

8. Data Storage Schema (SQLite)

The article table includes:

Field Name	Data Type
id	INTEGER PRIMARY KEY
symbol	TEXT
headline	TEXT
summary	TEXT
source	TEXT
published	TEXT (ISO-8601)

Field Name	Data Type
form_type	TEXT
entities	JSON
sentiment_score	REAL
sentiment_label	TEXT
aspect_sentiment	JSON
engagement	INTEGER
weight	REAL
weighted_score	REAL
fetch_date	TEXT



FinSight Sentiment Store: Database Schema

Sentiment_Events

Central Hub for All Sentiment Data. Captures every processed financial event from sources like news, regulatory filings, and insider reports.

event_id **Primary Key**	UUID	Unique identifier for each sentiment event.
ticker_symbol **Indexed**	String	The stock ticker symbol (e.g., 'AAPL', 'TSLA').
event_timestamp	UTC DateTime	The precise time the event occurred.
source_id	String	Identifier for the data source (e.g., 'Routers', 'EDGAR').
event_type	Enum	Type of event: News, Filing, Insider, or Activist.
raw_text	Blob	The original, unprocessed text content of the event.
nip_entities	JSONB	Structured data from NLP, like recognized people or companies.
sentiment_vector	Float Array	A vector representing sentiment across key aspects (e.g., Earnings, Macro, Litigation).
dsp_score	Float	The final Directional Sentiment Pressure score.
institutional_weight	Float	A calculated weight indicating the event's importance.

Market_Data

Historical Market Performance Data. Stores price and voiatility for backtesting sentiment signals against real-worid market performance.

ticker_symbol **Composite Key**	String	The stock ticker symbol.
market_timestamp **Composite Key**	UTC DateTime	The timestamp for the market data snapshet.
price	Float	The price of the security at the given timestamp.
volatility	Float	A measure of the security's volablity at the given timestamp.

Relationship Logic: **Sentiment_Events.ticker_symbol** links to **Market_Data.ticker_symbol** AND **Sentiment_Events.event_timestamp** links to **Market_Data.market_timestamp** (or nearest time).

NotebookLM

Figure 3: SQLite Schema

9. Implementation Notes

- System is fully asynchronous (asyncio)
- Caching avoids API throttling
- LLM calls are rate-limited
- Normalization ensures stability across volatile news cycles
- L2 normalization protects against volume imbalances

10. Limitations

- Non-English sources are minimally supported
- Model depends on LLM quality
- Corporate synonym dictionary must be maintained manually

11. Future Work

- Multi-asset joint sentiment modeling
- Cross-symbol entity-flow mapping
- Predictive regressions for return forecasting
- Regime-aware DSP scaling

12. References

- SEC EDGAR documentation
- Financial NLP research papers
- Entity extraction models