Sentiment-Enhanced Forecasting:

A Framework for Integrating Market Sentiment into Financial Prediction Models

From class project → reusable benchmark framework and TSLA case study using FNSPID + FinBERT

Welcome

Presenters: Iman Hamdan
Advisors: Dr. Andrew Van Benschoten

Agenda

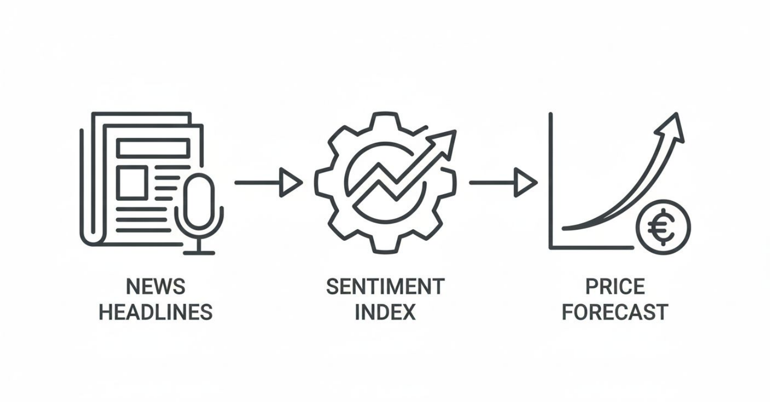
Motivation & research gap
Framework & methodology
TSLA case study (results)
Discussion & next steps

Reminders
Ask us questions in the chat!

Motivation: Why Sentiment + Forecasting?

Why Should Forecasting Care About Sentiment?

- Financial decisions increasingly depend on **unstructured news and text**
- Traditional price-only models ignore **market tone and narrative**
- Prior work shows sentiment can matter, but:
    - Uses **proprietary or ad-hoc datasets**
    - Often lacks **transparent methodology** and **economic evaluation**
- We want a **reproducible framework** that tests *how much* sentiment really adds



Still don’t know *how much* sentiment really helps once we control for price information on a standardized benchmark.

Benchmark & Dataset: FNSPID

Benchmark Testbed: FNSPID
Financial News Dataset

Use the **FNSPID Financial News Stock Price Integrated Dataset** as benchmark

Contains:

- Time-aligned **news + stock prices** for **multiple U.S. equities**
- Predefined splits and a published baseline study

Phase1: focus on **TSLA** as a detailed case study
Phase 2: extend to other tech equities (AAPL, NVDA, AMZN…).

FNSPID Dataset Description

Ticker	Period	# Articles
AAPL	2020-01 to 2023-12	15,200
AAPL	2020-01 to 2023-12	11,800
GOOG	2020-01 to 2023-12	9,500
TSLA	2020-01 to 2023-12	8,900
AMZN	2020-01 to 2023-12	8,900
META	2020-01 to 2023-12	7,300
NFLX	2020-01 to 2023-12	6,300
NVDA	2020-01 to 2023-12	5,800

What did FNSPID actually find?

1. FNSPID tests **LSTM, CNN, GRU, RNN, TimesNet (Transformer)** on **5, 25, 50 US stocks** using a **GPT 1–5 sentiment score**.
2. For **LSTM / CNN / GRU / RNN**, models **with and without sentiment look almost identical** (MAE, MSE, R^2 barely change)
3. Only the **Transformer (TimesNet)** shows a **clear boost** when sentiment is included (higher R^2, lower MAE/MSE).
4. FNSPID concludes that **sentiment "doesn't help much"** – my work asks whether that result is driven by a **noisy GPT index** rather than sentiment itself.

Dataset	A-Sen.	A-Sen.	A-Sen.	A-Non.	A-Non.	A-Non.	B-Sen.	B-Sen.	B-Sen.	B-Non.	B-Non.	B-Non.
Name	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
LSTM	.02599	.00157	.87115	.02530	.00148	**.88016**	.02677	.00160	**.86811**	.02523	.00142	**.88181**
CNN	.06180	.00712	.48205	.04913	.00475	.61811	.04236	.00354	.71668	.04522	.00398	.66687
GRU	.02474	.00143	.88588	.02494	.00141	.88302	.02631	.00154	.86756	.02470	.00139	.87746
RNN	.04152	.00355	.72957	.03353	.00251	.81128	.04315	.00339	.54265	.03898	.00291	.65470
Transformer	**.01801**	**.00058**	**.87260**	**.01883**	**.00060**	.86659	**.01700**	**.00060**	.84659	**.01007**	**.00021**	.94629
TimesNet	.02847	.00148	.63407	.02225	.00089	.81824	.03441	.00194	.51742	.02697	.00129	.69189

FNSPID Table 3 – test metrics with and without GPT-based sentiment features.

Research Gap & Hypothesis

- Existing FNSPID paper:
  - Uses **prompted GPT sentiment ratings** on a narrow 1–5 scale
  - Limited diagnostic analysis of the **sentiment index quality**

- Our focus:
  - Replace ad-hoc scores with **FinBERT-based, transformer sentiment indices**
  - Provide **transparent, modular** code that others can reuse
  - Benchmark our composite index against both the original FNSPID GPT scores and alternative sentiment models.

Hypothesis H1:

H1: *Transformer-based, finance-tuned sentiment indices (FinBERT and composite variants) improve stock-return forecast performance and outperform **existing GPT-based sentiment setups** when evaluated on the FNSPID benchmark.*

Contributions (What's New?)

From:

- **Generic GPT 3.5 1–5 news ratings as the sentiment proxy**
- **Single, monolithic pipeline that mixes scraping, scoring, and forecasting**
- **Evaluation focused mainly on error metrics (MSE/MAE) with limited diagnostics**
- **Results reported for a benchmark with weak transparency and reproducibility**

To:

- **Transformer-based, domain-specific sentiment indices**
    - FinBERT + extended transformer variants, calibrated and validated
- **Modular, reusable framework**
    - Clear stages: data → sentiment index → forecasting models → stats → trading
- **Systematic comparison of price-only vs sentiment-augmented deep models**
    - LSTM, GRU, Transformer on FNSPID (TSLA case, extendable to multiple equities)
- **Statistical and economic evaluation**
    - Diebold–Mariano tests, Sharpe ratios, equity curves for trading-strategy performance

Insights from prior work

Challenges with existing sentiment-forecasting studies

- Heavy reliance on **generic GPT or lexicon scores** not tuned to finance
- **Black-box pipelines** that mix data collection, sentiment, and modeling
- Limited use of **standardized benchmarks** (few papers use FNSPID)
- Focus on a **single asset or short horizon** with weak out-of-sample testing
- Little discussion of **economic value** (trading performance, risk-adjusted returns)

“The code isn’t reusable – every paper has its own custom pipeline.”

“It’s hard to tell whether sentiment is helping, or if it’s just noise.”

Current sentiment research is scattered, hard to reproduce, and rarely tied to trading

“Great ML models, but no connection to trading or portfolio decisions.”

“I don’t see how the sentiment index was actually built or validated.”

"Any final work will need to use a standardized benchmark and show that transformer-based sentiment indices can outperform current state-of-the-art sentiment models."

Andrew Van Benschoten
Research Advisor

Our Research **North Star (Goals)**

Build a transparent, reusable framework to test whether transformer-based sentiment indices improve stock-return forecasts.

Simplify

Clean, modular pipeline for FNSPID news & prices (easy to reuse on new equities)

Sentiment index as a first-class object

Design and validate a composite FinBERT-based sentiment index

Forecasting & trading impact

Compare price-only vs sentiment-enhanced models using both statistics and a TSLA trading strategy

What stays fixed across experiments? Not Changing?

Benchmark dataset

- FNSPID integrated news–price dataset as the common testbed.

Prediction target

- Daily TSLA return / price over the same time window

Train–test protocol

- Rolling-window train / test with identical splits across models

Evaluation metrics

- MSE, MAE, R^2, hit-rate, Sharpe ratio & drawdown for strategies

Controlled Experimental Design – what stays fixed across models

Challenges in building a robust sentiment index

Why GPT 1–5 Scores Are Not Enough

This is why we move to FinBERT + composite transformer-based indices with real diagnostics

Noisy sentiment labels

Generic LLM scores on a 1–5 scale can be unstable and hard to interpret.

Domain mismatch

General-purpose language models
miss finance-specific tone, jargon,and event structure.

Weak diagnostics

Many studies report only forecast errors,without ablations, significance tests,or economic interpretation.

Data & Modeling Overview

Our framework separates the problem into modular stages:

1. Collecting and cleaning benchmark data,
2. Building a transformer-based sentiment index,
3. Training deep learning forecasters
4. Evaluating both statistical and economic performance.
   This makes the pipeline easier to reproduce and extend to new equities.

FNSPID news–price dataset

TSLA focus (phase 1)

plan to extend to additional equities

Sentiment Modeling

FinBERT + transformer variants

Daily composite index (per ticker) with diagnostics

Forecasting & Evaluation

LSTM / GRU / Transformer

Price-only vs sentiment-augmented

DM tests + Sharpe-based trading

Economic Evaluation: TSLA Trading Strategy

Long–flat trading rule

- Each day, use the model's **next-day TSLA return forecast**.
- If the forecast return > 0 → **go long TSLA** for that day.
- Otherwise, **stay in cash (flat)**.

Price-only vs sentiment-enhanced

- Run the same rule **twice** with identical data/splits:
- **Price-only model** (baseline).
- **Price + FinBERT sentiment index** model.
- This isolates the **incremental value of sentiment**.

Economic metrics

- **Sharpe ratio** (risk-adjusted return).
- **Cumulative return** & **max drawdown**.
- **Hit rate** = % of correctly signed forecasts.

TSLA

case study

Phase 1

Extend the benchmark

Phase 2

New transformer-based sentiment index

Rebuild & Add

- Rebuild the **TSLA FNSPID pipeline** in a clean repo and **reproduce baseline results**.
- Add a **FinBERT daily sentiment index** and compare price-only vs sentiment-augmented models for TSLA.

Compare & Forecast

- Generalize to **all FNSPID stocks** and compare **price-only, GPT index, FinBERT index**.
- Evaluate both **forecast accuracy** and a simple **TSLA trading strategy** (Sharpe, drawdown, cumulative return).

Complete Footprint

- Design a **composite transformer-based sentiment index** and plug it into the same framework.
- Test across multiple equities and **target a publishable paper + open-source toolkit**.

Q&A