Macro Sentiment Index: Can Financial News Predict Future Market Volatility?

Xiang Feng, Mingyang Gao, Yihe Liu
CMPT 983

What Is VIX?

- VIX = Volatility Index, derived from S&P 500 options
- Measures expected 30-day market volatility
- Often called the “Fear Index”
- High VIX → high uncertainty & risk
- A reactive indicator — responds after volatility is priced in

Motivation: Why Might Sentiment Predict VIX?

- Macro news often arrives earlier than market reactions
- Central-bank briefings, government statements, and market analyses frequently precede volatility
- Negative or uncertain news can increase future market stress
- Investor sentiment shapes risk perception and option pricing
- News sentiment may provide an early-warning signal
- Text-based indicators can complement VIX's reactive nature

Hypothesis

News sentiment captures shifts in market mood

Aggregate sentiment can form a Macro Sentiment Index (MSI)

Hypothesis:

- MSI is correlated with VIX
- Which will fluctuate first, MSI or VIX

Overview of Our Data Pipeline

Collect daily financial & macro news from multiple sources

1. RSS (for 7 days), “yahoo_sp500, yahoo_treasury, … ”
2. API (for 30 days) https://newsapi.org News API
3. Kaggle (for 5 years) Financial News Dataset

   https://www.kaggle.com/datasets/yogeshchary/financial-news-dataset?resource=download

   https://www.kaggle.com/datasets/notlucasp/financial-news-headlines?resource=download

4. VIX (yfinance)

Sentiment Model (FinBERT)

Use FinBERT (ProsusAI/finbert) trained on financial text

Classifies each headline into:

- Positive / Negative / Neutral

Extract soft probabilities

Data:

- Normalize and hash URLS to prevent processing duplicates.
- Handle missing (fill, forward filling)
- Store and standardize the timestamps, such as aligning to NY time for more intuitive VIX comparison. (RSS, API)

Source Debiasing

- Compute sentiment score:

$$Reward_{i,norm} = \frac{Reward_i - \mu_{source}}{\sigma_{source}},$$

  Reward i: The raw sentiment value of a news item (the i-th item).

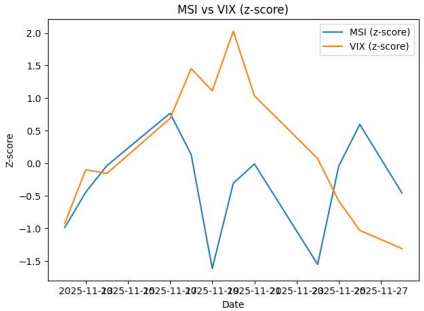
  Example: FinBERT output pos - neg.

- Different news outlets have different tone biases
  - Some outlets are consistently optimistic, others pessimistic
  - Raw sentiment scores reflect outlet style, not market conditions
- We correct this by subtracting each outlet's average sentiment
  - Debiased score = sentiment – mean(sentiment of that source)

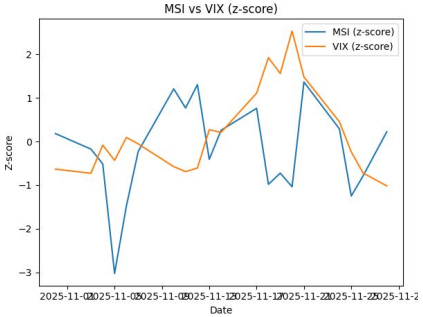
Daily MSI Construction

- Convert timestamps to daily dates (only for Kaggle data)
- Aggregate debiased sentiment per day
- Daily features include:
  - Mean sentiment
  - Median sentiment
  - Number of news articles
- Standardize MSI via z-score $z = \frac{x - \text{mean}}{\text{std}}$
- Additional rolling features:
  - 3-day rolling mean — captures short-term trend
  - 3-day rolling standard deviation — captures sentiment instability

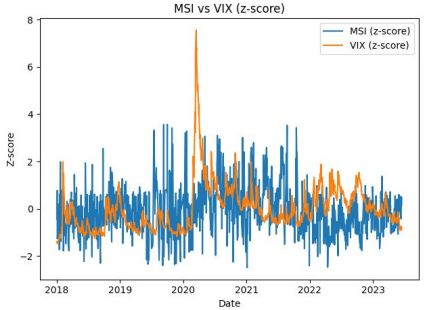
Visual Comparison: MSI vs VIX



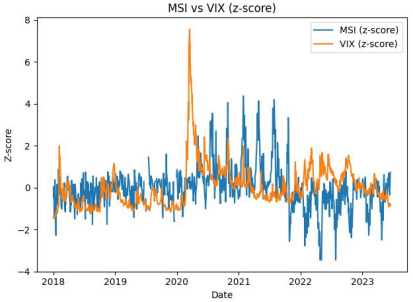
7 Days (RSS)



30 Days (API)



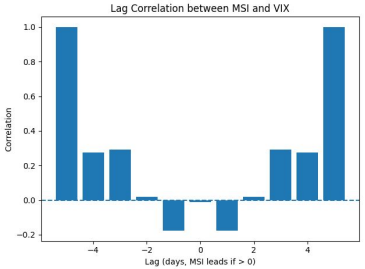
5 Years (Kaggle)



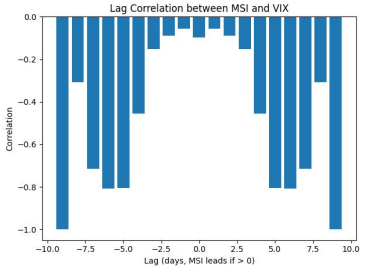
5 Years (Kaggle, noise reduction)

- Some local segments appear to move together, but no consistent or sustained relationship is observed
- Visual inspection suggests no clear leading pattern

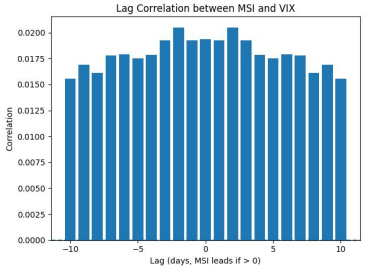
Lag Correlation



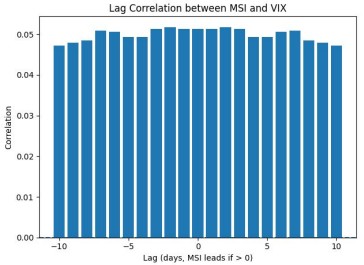
7 Days (RSS)



30 Days (API)



5 Years (Kaggle)

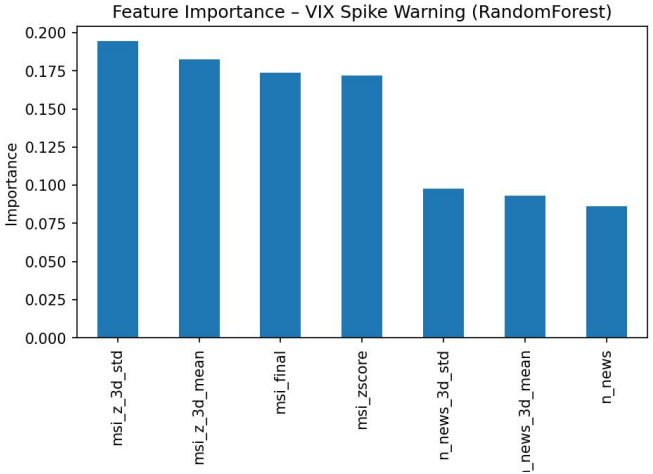


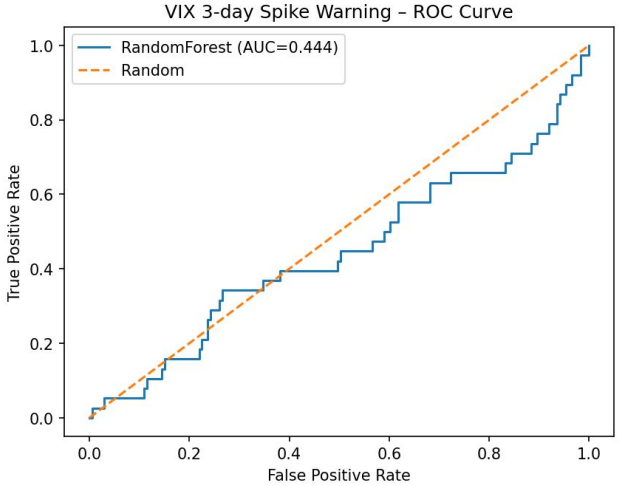
5 Years (Kaggle, noise reduction)

- Shift MSI forward/backward by k days
- Compute correlation between MSI(t + k) and VIX(t)
- Correlation remains near zero across all lags

Spike Warning Model

- Goal: Predict whether VIX will spike in the next 3 days
- Model: Random Forest classifier
  - Robust to noise
  - Works well on small datasets
  - Provides feature importance
- Result:
  - Model performs close to random (AUC ≈ 0.44)





Summary of Experiments

- MSI does not lead VIX
  - Lag correlation shows no forward predictive relationship
- MSI fails to predict VIX spikes
  - RandomForest AUC ≈ 0.44 → below random guessing
- Sentiment volatility ≠ market volatility
  - Short-term sentiment swings correlate weakly but not predictively
- Overall result
  - MSI is not a leading one (maybe a lagging indicator?)

Experiments on a Different Dataset

- A different dataset: https://www.kaggle.com/datasets/notlucasp/financial-news-headlines?resource=download
- Headlines and Descriptions Extracted from

  CNBC, Guardian, Reuters.

- **Spearman** and **Kendall Correlation** Test:

  Spearman: -0.02696

  Kendall: -0.01743

cnbc_headlines.csv (682.32 kB)

Detail | Compact | Column

Headlines	Time	Description
Jim Cramer: A better way to invest in the Covid-19 vaccine gold rush	7:51 PM ET Fri, 17 July 2020	"Mad Money" host Jim Cramer recommended buying four companies that are supporting vaccine developers...
Cramer's lightning round: I would own Teradyne	7:33 PM ET Fri, 17 July 2020	"Mad Money" host Jim Cramer rings the lightning round bell, which means he's giving his answers to c...

MSI-VIX Causality

- Potential **lagged or dynamic influence** CANNOT be captured by **correlation**
  - Correlation -> How they move together?
- Motivates the **causality analysis** between MSI-VIX
  - Causality Analysis -> Whether one helps predict the other?
  - **Granger Causality**, designed to test predictive, time-lag dependent causality.

Granger Causality Test

- **Definition:**

  If historical values of A improve the prediction of B, **A Granger-causes B**

$$B_t = c + \sum_{i=1}^{p} \emptyset_i B_{t-i} + \varepsilon_i$$

$$B_t = c + \sum_{i=1}^{p} \emptyset_i B_{t-i} + \sum_{j=1}^{p} \theta_j A_{t-j} + u_t$$

  , **theta_j** -> how much contribution of **A_t-j** for predicting **B_t**

  , more accurate **B_t** relies on **A_t-1, A_t-2 ..., A_t-j**

VIX Granger-causes MSI

H0: **MSI** NOT Granger-causes **VIX**, if $p < 0.05$, significant, reject H0

H0': **VIX** NOT Granger-causes **MSI**, if $p < 0.05$, significant, reject H0

Lags	F-test p value	Results (**H0**)
1	0.1796	Not significant
2	0.1900	Not significant
3	0.2726	Not significant
4	0.3736	Not significant

Lags	F-test p value	Results (**H0'**)
1	0.0188	Significant
2	0.0539	Marginal Significant
3	0.0390	Significant
4	0.0215	Significant

Economic Thoughts

- **the VIX of 1 day earlier** helps predicting **MSI more accurately**

- **Market digests information faster than narrative formation**
- **Emotion of news follows the market rather than leading it**
- **Understanding volatility is more useful for anticipating sentiment**

Possible Future Work

- Rethinking whether sentiment index is not representable to VIX
- OR the selected dataset is not representable?

- Data Enhancement by exploring and introducing the datasets of
  - investor advisory, risk alert ...
  - market commentary from institutions, analyst, and etc.

Thank You !