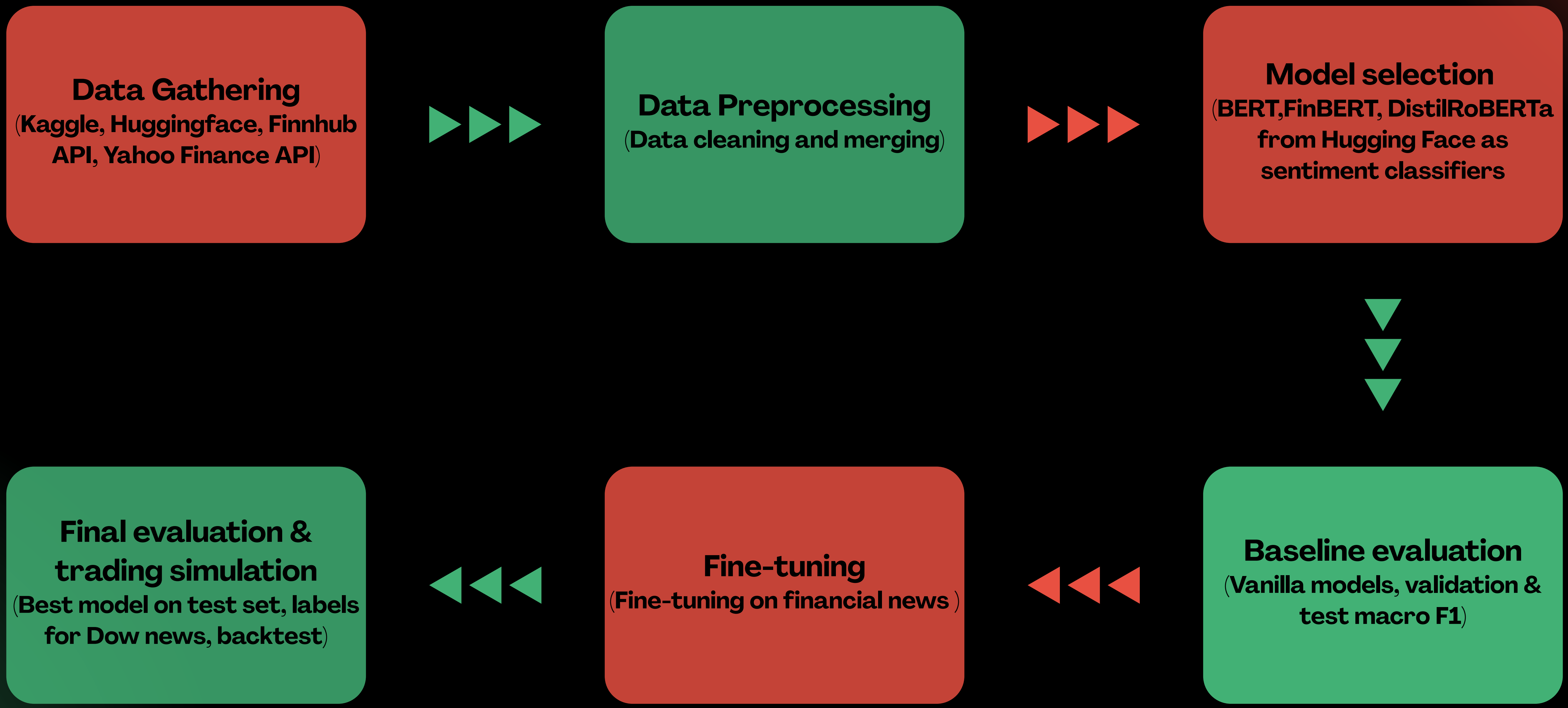
QUANTITATIVE INVESTING

LLMs for financial news sentiment analysis

▸▸▸ Why people use **LLMs** in trading?

- Financial news trigger price reactions
- Traders face thousands of headlines and articles per day
- LLMs can scan, filter, summarize and make decisions much faster than humans
- Sentiment models turn unstructured text into numerical scores, that can be used in algorithmic trading strategies
- Automated sentiment extraction allows reaction within seconds of news releases
- Turning real-time sentiment into signals can generate alpha (excess return) before markets fully incorporate the information

Project Overview



Data

kaggle

Finnhub

Hugging Face

yahoo! finance

Wikipedia
The Free Encyclopedia

- Fine-tuning set consisting of 3 smaller datasets with news text and manually verified sentiment labels
  - Strongly imbalanced (skewed) label distribution
  - Undersampled “positive” and “neutral” labels
  - 70/15/15 split into train, validation and test sets
- Live news accessed via Finnhub API (timestamp, stock ticker and short summary) used for the backtesting with a helper function
- Historical stock prices for our stock universe accessed via yFinance API with a helper function
- Ticker universe (list of stocks) scraped from Wikipedia

Model selection
why these transformers?

- sentence-level sentiment classification for financial news
  - Need models that are:
    - Strong baselines for text classification
    - Available as open-source weights on Hugging Face
    - Reasonably small and fast enough for backtesting / (future) live use
  - For each we compare:
    - Vanilla (pretrained only) vs. finetuned on our labeled news dataset
    - Validation performance (macro F1) and final Test performance
  - Therefore we use three related transformer models:
    - BERT (general-purpose baseline)
    - FinBERT (finance-specific BERT)
    - DistilRoBERTa (lighter, faster model)

Why BERT/FinBERT/DistilRoBERTa?

- BERT
  - General-purpose transformer baseline for text classification
  - Pretrained on large generic corpora (Wikipedia + BookCorpus)
  - Well-studied, many examples and tools available
  - Gives us a fair reference: "How far do we get without finance-specific pretraining?"

- FinBERT
  - BERT architecture further pretrained on financial text (earnings calls, SEC filings, financial news)
  - Learns finance-specific vocabulary and phrasing (guidance, downgrade, beat/miss, etc.)
  - We expect better sentiment understanding in edge cases: mixed guidance, macro risks, etc.
  - Serves as our "domain-expert" model for this task

- DistilRoBERTa
  - Compressed version of RoBERTa: fewer parameters, faster inference
  - Already finetuned on financial news sentiment out of the box
  - Good candidate for low-latency trading use (cheaper to run, still strong accuracy)
  - Lets us study the trade-off: slightly smaller model vs accuracy vs speed

Fine-tuning

from vanilla models to financial sentiment

Concept

- Use our labeled financial news dataset (75 train / 15 val / 15 test split)
- Fine-tune only on the training split
- Select best checkpoint by macro F1 on the validation set
    - treats all three classes equally (negative / neutral / positive), so performance on minority classes matters, not just the dominant label.

Mechanism

```
batch_size = 16
num_epochs = 3

training_args = TrainingArguments(
    output_dir="./bert_base_uncased_fin_sentiment",
    eval_strategy="epoch",      # evaluate on val at end of each epoch
    save_strategy="epoch",
    load_best_model_at_end=True,
    metric_for_best_model="macro_f1",
    greater_is_better=True,

    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    num_train_epochs=num_epochs,

    logging_dir="./logs",
    logging_steps=50,

    report_to="none",
)

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_ds_tok,
    eval_dataset=val_ds_tok,
    compute_metrics=compute_metrics,
)
```

results before and after finetuning

BERT

	precision	recall	f1-score	support
negative	0.8661	0.8856	0.8757	577
neutral	0.8309	0.7920	0.8110	577
positive	0.8424	0.8628	0.8525	576
accuracy			0.8468	1730
macro avg	0.8465	0.8468	0.8464	1730
weighted avg	0.8465	0.8468	0.8464	1730

FinBERT

anilla FinBERT – classification report (val):

	precision	recall	f1-score	support
negative	0.0552	0.0485	0.0517	577
neutral	0.1070	0.1005	0.1037	577
positive	0.1806	0.2135	0.1957	576
accuracy			0.1208	1730
macro avg	0.1143	0.1209	0.1170	1730
weighted avg	0.1142	0.1208	0.1170	1730

Fine-tuned FinBERT – classification report (val):

	precision	recall	f1-score	support
negative	0.8862	0.8908	0.8885	577
neutral	0.8211	0.8111	0.8160	577
positive	0.8655	0.8715	0.8685	576
accuracy			0.8578	1730
macro avg	0.8576	0.8578	0.8577	1730
weighted avg	0.8576	0.8578	0.8577	1730

DistilRoBERTa

Vanilla DistilRoBERTa – classification report (val):

	precision	recall	f1-score	support
negative	0.8801	0.7504	0.8101	577
neutral	0.6705	0.8076	0.7327	577
positive	0.7882	0.7431	0.7650	576
accuracy			0.7671	1730
macro avg	0.7796	0.7670	0.7693	1730
weighted avg	0.7796	0.7671	0.7693	1730

Fine-tuned DistilRoBERTa – classification report (val):

	precision	recall	f1-score	support
negative	0.8872	0.8856	0.8864	577
neutral	0.8135	0.8163	0.8149	577
positive	0.8696	0.8681	0.8688	576
accuracy			0.8566	1730
macro avg	0.8567	0.8567	0.8567	1730
weighted avg	0.8567	0.8566	0.8567	1730

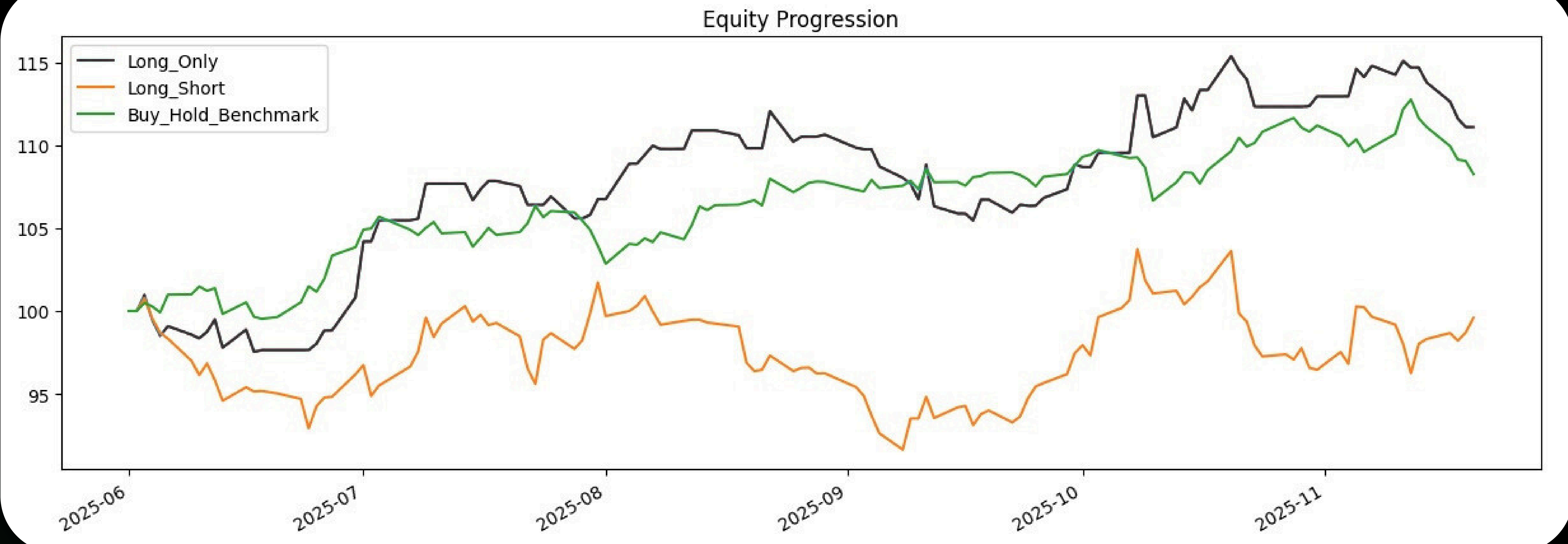
- Fine-tuning improves macro F1
- FinBERT gains the most (finance-specific pretraining + task-specific labels).
- We keep FinBERT, BERT and DistilRoBERTa for final comparison on the test set.

Results on **Test set**

- Macro F1 (test set):
  - BERT = 84.6% | FinBERT = 86.5% | DistilRoBERTa = 85.1%
  - FinBERT is the best overall model → highest macro F1 and accuracy (0.865 vs. 0.846 / 0.851).
  - Finance-specific pretraining helps → FinBERT especially improves negative/positive sentiment compared to vanilla BERT.
  - DistilRoBERTa is slightly weaker than FinBERT but close → good speed/accuracy trade-off.

```
===== BERT (finetuned) - TEST =====

Accuracy: 0.8461538461538461
              precision    recall  f1-score   support

    negative     0.8881    0.8958    0.8920       576
     neutral     0.8010    0.8038    0.8024       576
    positive     0.8491    0.8388    0.8439       577

    accuracy                         0.8462      1729
   macro avg     0.8461    0.8462    0.8461      1729
weighted avg     0.8461    0.8462    0.8461      1729


===== FinBERT (finetuned) - TEST =====

Accuracy: 0.86524002313476
              precision    recall  f1-score   support

    negative     0.8957    0.9097    0.9027       576
     neutral     0.8360    0.8056    0.8205       576
    positive     0.8625    0.8804    0.8714       577

    accuracy                         0.8652      1729
   macro avg     0.8647    0.8652    0.8648      1729
weighted avg     0.8647    0.8652    0.8648      1729


===== DistilRoBERTa (finetuned) - TEST =====

Accuracy: 0.8507807981492193
              precision    recall  f1-score   support

    negative     0.8885    0.8993    0.8939       576
     neutral     0.8105    0.8021    0.8063       576
    positive     0.8524    0.8510    0.8517       577

    accuracy                         0.8508      1729
   macro avg     0.8505    0.8508    0.8506      1729
weighted avg     0.8505    0.8508    0.8506      1729
```

Trading Results



Stat	Long_Only	Long_Short	Buy_Hold_Benchmark
Start	2025-06-01	2025-06-01	2025-06-01
End	2025-11-20	2025-11-20	2025-11-20
Risk-free rate	0.00%	0.00%	0.00%
Total Return	11.11%	-0.41%	8.27%
Daily Sharpe	1.59	0.04	1.73
Daily Sortino	2.99	0.07	3.11
CAGR	25.07%	-0.86%	18.38%
Max Drawdown	-5.89%	-9.92%	-3.99%
Calmar Ratio	4.25	-0.09	4.61

THANK YOU!

Resources

- Guo, T., & Hauptmann, E. (2024). Fine-Tuning Large Language Models for Stock Return Prediction Using Newsflow. RAM Active Investments. Retrieved from https://www.ram-ai.com/sites/default/files/2024-08/202408_fine-tuning-large-language-models-for-stock-return-prediction.pdf

- Hauptmann, E., Betrix, V., Jamet, N., Guo, T., & Piquet, L.-A. (2024). Financial Sentiment Analysis with Large Language Models: An Introductory & Comparative Study on News Flow. RAM Active Investments. Retrieved from https://www.ram-ai.com/sites/default/files/2024-04/202404_ramai_financial-sentiment-analysis-with-llm.pdf

- Thakar, C. (2023, July 24). Sentiment Analysis for Trading – Part I. IBKR Quant. Interactive Brokers. Retrieved from https://www.interactivebrokers.com/campus/ibkr-quant-news/sentiment-analysis-for-trading-part-i/

- Moody's. (2024, November 8). The Power of News Sentiment in Modern Financial Analysis. Retrieved from https://www.moodys.com/web/en/us/insights/digital-transformation/the-power-of-news-sentiment-in-modern-financial-analysis.html