By Group "Engravers"

1. Introduction: The Failure of Traditional Sentiment Analysis & Our Strategic Pivot

In the rapidly evolving landscape of Financial Technology, Natural Language Processing has long been hailed as the "Holy Grail" of alpha generation. For decades, quantitative researchers have sought to map textual data to asset prices, but the standard textbook approach—dictionary-based methods (e.g., Loughran-McDonald) or simple classifiers (e.g., BERT) tagging news as "Positive" or "Negative"—fails catastrophically when applied to complex macro assets like **Gold (XAU/USD)**.

Gold is not a stock. It has no quarterly earnings, no CEO scandals, and no product launches. It is a **macro-narrative asset**, its price driven by the intricate interplay of interest rates, geopolitical tension, inflation expectations, and central bank policies. A headline like "*US Employment Data Beats Expectations*" is "Positive" for the U.S. economy but often "Negative" for Gold—yet traditional sentiment models blindly predict "UP," leading to devastating trading losses. This is the **"Semantic Gap"** we set out to solve.

Generic news diluted signals, and company-specific events (e.g., product launches) lacked the transparent narrative link to price moves that macro assets exhibit. Recognizing Gold's unique dependence on universal macro drivers, we pivoted—hypothesizing that **Meta-Llama-3**'s emergent reasoning capabilities could decode the second-order financial logic chains traditional models miss. Our goal was not to build a mere classifier, but an **AI Macro Strategist** tailored to Gold's distinct dynamics.

2. The Data Engineering Challenge: Constructing a Decade of Truth

The foundation of any robust ML model is trustworthy data. We rejected static Kaggle datasets—plagued by survivorship bias, temporal gaps, and irrelevant content—and embarked on an ambitious data engineering initiative to reconstruct Gold's narrative history from **2020 to 2025** using custom-coded pipelines.

2.1 Targeted Data Collection: Solving Core Pain Points

To capture Gold's core drivers, we built a pipeline using `GNews` for textual data and `yfinance` for price data, focusing on keywords like "*Federal Reserve rate,*" "*Inflation CPI,*" and "*Geopolitical tension*". Two critical challenges emerged, and we solved them with concise, targeted code:

```
import pandas as pd
import yfinance as yf
from gnews import GNews
from tqdm import tqdm
import time
import datetime
from dateutil.relativedelta import relativedelta
import random
import os

OUTPUT_FILE = "gold_news_10years.csv"
GLOBAL_START = datetime.date(2020, 1, 1)
GLOBAL_END = datetime.date(2025, 12, 31)
```

```
KEYWORDS = [
    "Gold price", "Federal Reserve rate", "Inflation CPI",
    "Geopolitical tension", "US Dollar index", "Recession fears"
]
```

Challenge 1: Messy Date Formats

GNews returns inconsistent date strings (e.g., "Fri, 27 Dec 2024 GMT" vs. "2024/12/27"). We standardized them with a try-except block:

```
try:
    # Standard date format
    dt = pd.to_datetime(pub).strftime("%Y-%m-%d")
except:
    dt = current_date.strftime("%Y-%m-%d")
```

Challenge 2: Anti-Scraping Blocks

Frequent requests triggered Google's rate limits. We added strategic delays to stay compliant:

```
for kw in KEYWORDS:
    try:
        resp = google_news.get_news(kw)

        time.sleep(0.3)
    except: pass
```

After deduplication, we collected a large number of news articles. Each tagged with date, topic, and headline—capturing both high-impact events (e.g., 2020 Fed emergency rate cuts) and subtle narrative shifts (e.g., 2025 inflation concerns).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date	Topic	Headline												
2	2020-01-01	US Dollar index	Global Markets: Stocks end 2019 near record highs, dollar slides - Reuters												
3	2020-01-01	Gold price	'How good's the gold price?': Amid an economic slowdown can the value of gold keep shining? - Australian Broadcasting Corporation												
4	2020-01-01	Gold price	Gold in 2019: Lessons for the Year Ahead - Yahoo Finance												
5	2020-01-01	Recession fears	Hereâ€™s How 2019 Turned Out To Be A Historic Year For The Stock Market - Forbes												
6	2020-01-01	US Dollar index	Rupee kick starts New Year on positive note, settles 14 paise higher at 71.22 against US dollar - financialexpress.com												
7	2020-01-02	US Dollar index	Gold inches higher on lacklustre dollar - Business Day												
8	2020-01-02	US Dollar index	US Dollar Index Price Analysis: DXY starting 2020 with a bounce near 96.80 level - Forex Crunch												
9	2020-01-02	US Dollar index	Gold In 2019: Lessons For The Year Ahead - Investing.com												
10	2020-01-02	Inflation CPI	Which way will Indiaâ€™s growth, inflation, gold and Sensex move in 2020? - livemint.com												
11	2020-01-02	US Dollar index	The Dollarâ€™s Losses May Just Be Getting Started - Bloomberg.com												
12	2020-01-02	US Dollar index	Dollar bounces after end-2019 selloff, yuan shrugs off policy easing - curacaochronicle.com												
13	2020-01-03	US Dollar index	Shorting cannabis stocks was a billion-dollar idea in 2019 - MarketWatch												
14	2020-01-03	Geopolitical tension	Stocks Tumble and Oil Rises But Markets Mostly Shrug Off Killing of Iranâ€™s Top Military Leader - breitbart.com												
15	2020-01-03	Inflation CPI	Eleven States Record Retail Price Inflation Rate of Over 6% in November - TheWire.in												
16	2020-01-03	Recession fears	No. 5 on the list of Florida Politicians of the Decade: Bill Nelson - Florida Politics												
17	2020-01-03	Recession fears	Central Banks Are the Biggest Risk to the Economy in 2020 - Bloomberg.com												
18	2020-01-03	Federal Reserve rate	Home equity rates expected to remain low in 2020 - HousingWire												
19	2020-01-03	Federal Reserve rate	7 Best Fixed-Income Funds As Fed Keeps Rates Steady - US News Money												
20	2020-01-03	US Dollar index	How Every Asset Class, Currency, and Sector Performed in 2019 - Visual Capitalist												
21	2020-01-03	Federal Reserve rate	Fed Minutes to Shed Light on Interest Rate Consensus - The Wall Street Journal												
22	2020-01-03	Gold price	December 30 - January 3 - Gold Price												
23	2020-01-04	Federal Reserve rate	Low Interest Rates Worry the Fed. Ben Bernanke Has Some Ideas. (Published 2020) - The New York Times												
24	2020-01-04	Federal Reserve rate	The new tools of monetary policy - Brookings												
25	2020-01-04	Recession fears	Canada Housing Market Crash: 3 TSX REIT Stocks to Buy - The Motley Fool Canada												
26	2020-01-06	Geopolitical tension	Stop escalation, urges UN chief, as geopolitical tensions reach â€¨highest level this centuryâ€™ - UN News												
27	2020-01-06	Geopolitical tension	Saudi Aramco faces tough test less than a month after IPO - Nation Thailand												
28	2020-01-06	Geopolitical tension	Mondayâ€™s Daily Brief: Geopolitical strife, Australia bushfires, Burkina Faso, Cambodia updates - UN News												
29	2020-01-06	Geopolitical tension	6G can be the new frontier if we can work out how to use it - Tech in Asia												
30	2020-01-06	Gold price	Gold soars to highest price since 2013 as Trump and Iran escalate threats - Business Insider Africa												
31	2020-01-06	Geopolitical tension	Tension in Middle-East spooks market; what should investors do? - Moneycontrol												
32	2020-01-06	Gold price	India gold prices hit record high on safe-haven rush, weak rupee - Moneycontrol												
33	2020-01-06	Federal Reserve rate	Fed Adds $76.9 Billion in Overnight Money to Markets - The Wall Street Journal												
34	2020-01-06	Geopolitical tension	The Top 10 Geopolitical Risks for the World in 2020 - Time Magazine												
35	2020-01-06	Gold price	24K Gold Rate in India for January 2020 â€“ Week 1 - ClearTax Chronicles												

2.2 Labeling Logic: Eliminating Look-Ahead Bias

Financial data is noisy, so we implemented a **Next-Day Return (T+1)** labeling strategy to ensure the model only uses information available at prediction time. The core logic lies in our granular scoring function and date alignment:

Core Scoring Function

We mapped Gold's returns to a -5 to +5 scale, tied to its historical volatility:

```
def get_score_from_return(ret):
    # The threshold here can be adjusted based on the historical volatility of
gold
    # Gold volatility is smaller than that of individual stocks; generally, a
daily fluctuation exceeding 1.5% is considered significant
    r = ret * 100 # Convert to percentage

    if r > 2: return 5
    elif r > 1.4: return 4
    elif r > 0.8: return 3
    elif r > 0.4: return 2
    elif r > 0.1: return 1
    elif r < -2: return -5
    elif r < -1.4: return -4
    elif r < -0.8: return -3
    elif r < -0.4: return -2
    elif r < -0.1: return -1
    else: return 0 # -0.1 to 0.1 is considered absolutely neutral
```

Date Alignment to Avoid Non-Trading Days

We skipped news from weekends/holidays to ensure valid return matching:

```
dt = pd.to_datetime(date_str)

# Find the daily rate of return
if dt not in returns.index:
    continue

ret = returns.loc[dt].item() # Get Value

score = get_score_from_return(ret)
```

2.3 Data Quality: Distilling Signal from Noise

We filtered headlines shorter than 10 characters, removed non-English content (~3% of raw data), and handled missing prices by dropping gaps longer than 2 trading days. The final dataset contained **17,857 valid headline-return pairs**—spanning COVID-19 volatility, 2022 Fed hikes, and 2024's Gold bull run—balancing breadth (multiple market regimes) and depth (Gold-specific drivers).

3. The Scoring Innovation: From Binary to Spectrum

A critical flaw in traditional NLP is binary classification (Up/Down), which fails to capture market nuance: a slight inflation tick is not a war, and a minor dollar rally is not a Fed rate hike. Our **11-point Granular Scoring System** (-5 to +5) forces the LLM to learn both direction and impact magnitude—critical for building a strategy that differentiates between "hold" and "max leverage buy" signals.

Score	Sentiment Label	Gold Daily Return	Real-World Example
+5	Extreme Bullish	>2.0%	"Middle East War Breaks Out"
+3	Moderate Bullish	0.8% – 1.4%	"US CPI Surges Beyond Forecasts"
+1	Slight Bullish	0.1% – 0.4%	"Central Bank Gold Purchases Rise"
0	Neutral/Noise	-0.1% – 0.1%	"Analyst Upgrades Gold Miner Stock"
-1	Slight Bearish	-0.4% – -0.1%	"US Jobless Claims Fall"
-3	Moderate Bearish	-1.4% – -0.8%	"Inflation Cools More Than Expected"
-5	Extreme Bearish	< -2.0%	"Geopolitical Conflict Ends Abruptly"

The system explicitly labels a portion of the data as "neutral (0)," which teaches the model that **most news is noise**—a crucial lesson in preventing overtrading.

For example, our processed dataset is as follows:

Date	Headline	Score
2020-01-02	Gold inches higher on lacklustre dollar - Business Day	0
2020-01-02	US Dollar Index Price Analysis: DXY starting 2020 with a bounce near 96.80 level - Forex Crunch	0
2020-01-02	Gold In 2019: Lessons For The Year Ahead - Investing.com	0

4. The Hardware Advantage: Unleashing the high-performance graphics cards

To turn this vision into reality, we leveraged the **high-performance graphics cards**—a game-changer for LLM fine-tuning and data processing. Its VRAM enabled 4-bit quantization and high batch sizes, cutting preprocessing time from 4 hours to 20 minutes.

LoRA Configuration for Efficient Fine-Tuning

We trained Llama-3's parameters via LoRA, fitting the entire process on our high-performance graphics cards.

Prompt Formatting for Llama-3

We aligned our data with Llama-3's instruction-tuning format.

The powerful performance of high-performance graphics cards enabled rapid iteration—we tested three scoring thresholds and two hint structures in just a few days (rather than weeks) to ensure optimization for real-world applications.

5. Looking Ahead: From Blueprint to Execution

This first phase—architecting the data foundation, scoring system, and hardware pipeline—sets the stage for our next steps. In Blog Post 2, we'll detail how we tamed the "Bull Market Bias" (a side effect of 2025's Gold rally) and implemented the "Zero-Filter" aggregation algorithm to turn daily scores into actionable trades.

We've proven that building an AI Macro Strategist requires more than a powerful model—it demands aligning data collection, labeling, and scoring with the asset's unique logic. For Gold, that means focusing on macro narratives, capturing nuance, and eliminating noise. With our blueprint refined and hardware optimized, we're ready to move from design to deployment.

Thanks for reading.