Hybrid Commodity Forecasting with News Data: Combining Time-Series Analysis and News Sentiment for Copper Price Prediction

Roman Gavrilenko
Master in Finance
HEC Lausanne
Lausanne, Switzerland
roman.gavrilenko@unil.ch

Abstract*—This project investigates whether incorporating news sentiment analysis improves forecasting of copper price movements compared to using price data alone. We develop a hybrid machine learning approach that combines traditional time-series features (moving averages, volatility, technical indicators) with sentiment features extracted from financial news using FinBERT, a pre-trained financial language model. The dataset spans 17 years (2008-2025) of LME copper prices and 9,448 news articles from multiple sources including Reuters, Mining.com, and Bloomberg. We evaluate binary classification task of detecting extreme price movements or "shocks" using walk-forward validation with 5 temporal windows. We compare baseline models (Logistic Regression, Random Forest, SVM, Gradient Boosting) with hybrid models that incorporate news features. Our results demonstrate mixed outcomes: while Random Forest (Hybrid) achieves the highest AUC of 0.73 and Gradient Boosting (Hybrid) achieves AUC of 0.72 vs. 0.64 for price-only baseline (representing a 12.5% relative improvement), Logistic Regression (Price-Only) surprisingly achieves the highest F1-score of 0.34 and highest PR-AUC of 0.29. The best hybrid model, Random Forest (Hybrid), achieves F1 of 0.30 and PR-AUC of 0.26. These results suggest that news features provide moderate improvements for tree-based models, but linear models may benefit less from additional features due to regularization constraints. Feature importance analysis reveals that sentiment scores, especially negative sentiment, and supply-side news indicators are among the predictive features, though price-based technical indicators remain dominant.

I. Introduction

Copper, as the third most-consumed industrial metal globally, plays a critical role in modern infrastructure, renewable energy technologies, and electronic devices. Its price movements directly impact construction costs, manufacturing expenses, and economic growth projections worldwide. The metal's dual nature—as both an essential commodity and a financial asset—makes its price forecasting particularly challenging, as it responds to fundamental supply-demand dynamics, geopolitical events, financial speculation, and market sentiment [1].

The extraction and production of copper, as illustrated in Figure **??**, involves massive open-pit mining operations that are highly vulnerable to disruptions. Labor strikes at major mines (such as Escondida in Chile, the world's largest copper

mine), geological challenges, production cuts, or geopolitical events can cause significant supply shocks that reverberate through global markets. These real-world production dynamics underscore the importance of incorporating news intelligence into price forecasting models, as traditional time-series approaches may miss early warning signals of supply disruptions.

The significance of accurate copper price forecasting extends beyond traditional market participants. For commodity trading hubs like Switzerland, which serves as a major financial center and trading hub for natural resources, predictive models can support risk management, portfolio optimization, and strategic decision-making for trading firms, institutional investors, and policy makers. Switzerland's position as a global commodity trading hub, with major firms like Glencore and Trafigura headquartered there, underscores the practical importance of developing sophisticated forecasting tools that can capture both quantitative price signals and qualitative market intelligence from news sources.

Recent advances in natural language processing (NLP) and machine learning have opened new possibilities for integrating unstructured textual data—such as global news, policy reports, and market analyses—into forecasting models. As demonstrated by Ghali et al. [1], hybrid frameworks that combine historical price data with semantic signals derived from economic news using agentic generative AI can achieve strong predictive performance (mean AUC of 0.94) for commodity price shock detection. Their work shows that eliminating

the news component causes a steep drop in AUC to 0.46, underscoring the critical value of incorporating real-world context through unstructured text.

This project addresses similar challenges by developing a reproducible pipeline that: (1) collects and aligns financial news with price data while avoiding lookahead bias, (2) extracts sentiment features using FinBERT—a BERT model fine-tuned on financial news corpus [2]—which has enabled more sophisticated sentiment extraction compared to generic sentiment analyzers, (3) creates comprehensive feature sets combining price-based technical indicators with news-derived sentiment and heuristic signals, and (4) evaluates hybrid models that integrate both information sources, with a focus on practical applications such as shock detection.

The main research question addressed is: **Does incorporating news sentiment analysis improve forecasting of copper price movements compared to using price data alone?**

Our contribution lies in (1) implementing a comprehensive feature engineering pipeline with multiple news signal types, (2) using walk-forward validation to simulate realistic trading scenarios, and (3) providing detailed analysis of which news signals contribute most to predictions through feature importance analysis. Unlike the approach in [1], which uses yearly data and agentic generative AI for news summarization, we work with daily granularity and directly extract sentiment features from individual news articles, enabling finer temporal resolution for shock detection.

II. Data

A. Price Data

We collected daily LME (London Metal Exchange) copper price data spanning January 2, 2008 to December 2025, totaling 4,542 trading days. The dataset includes cash price (`lme_copper_cash`), 3-month forward price (`lme_copper_3m`), and stock levels (`lme_copper_stock`). Price data was obtained through web scraping from Westmetall.com, which aggregates official LME data.

The price series exhibits typical commodity market characteristics: high volatility during financial crises (2008-2009, 2020), long-term trends driven by supply-demand fundamentals, and occasional extreme spikes associated with supply disruptions. Figure 1 shows the evolution of copper prices and LME warehouse stock levels over the study period, illustrating the inverse relationship between stock levels and prices during supply tightness periods.

B. News Data

News articles were collected through an intensive, multi-source scraping approach designed to maximize coverage and relevance. We implemented a comprehensive collection strategy that utilized multiple channels: RSS feeds from major financial news providers, Google News search queries with various keyword combinations, and direct parsing of specialized mining and commodity news websites.

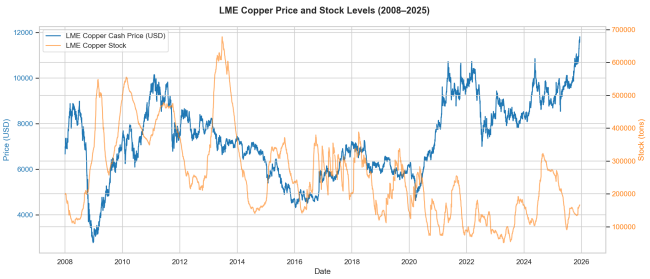

Fig. 1: Copper price and LME warehouse stock time series (2008-2025). The top panel shows cash price with the left y-axis indicating USD per metric ton, while the bottom panel shows stock levels with the right y-axis indicating metric tons. The dataset contains 4,542 trading days. Notable periods include the 2008 financial crisis, the 2011 supply disruptions, and the 2020 COVID-19 market volatility.

To ensure broad temporal coverage, we employed an extensive set of query variations targeting different aspects of the copper market. For each major theme—supply disruptions, demand factors, mining operations, and geopolitical events—we constructed multiple query variations using synonyms, related terms, and domain-specific terminology. For example, supply-side queries included combinations such as: “copper mine strike”, “labor dispute mining”, “copper production cut”, “mine closure copper”, “copper supply disruption”, “mining strike Chile”, “copper export ban”, “production halt mining”, and many others.

We systematically expanded query coverage for historical periods by adding year-specific context and combining multiple keywords. For instance, queries included temporal markers like “copper mine 2017 strike”, “Chile copper production 2015”, or “China copper demand 2010”, ensuring that even earlier periods with potentially sparser news coverage would be adequately captured. This approach resulted in hundreds of unique query combinations, with parallel processing (up to 4-8 workers) enabling efficient collection across the entire 17-year period.

Despite our comprehensive collection strategy targeting uniform coverage across historical periods, we observe that news volume in 2025 is substantially higher than earlier years. This discrepancy likely reflects several factors: (1) increased availability of online news sources in recent years, (2) more comprehensive RSS feed coverage for recent periods, (3) improved search engine indexing of recent articles, and (4) the recency bias inherent in web scraping, where recent content is more easily accessible than archived historical content. While this temporal imbalance in data density could potentially affect model performance, our walk-forward validation approach helps mitigate temporal bias by training on progressively expanding historical windows.

News collection focused on several key domains: (1) **Copper mining operations**: Major mines (Escondida, Collahuasi, Cerro Verde, Buenavista, Kamoa-Kakula, Grasberg, Antamina,

Morenci, El Teniente, Chuquicamata) and operators (BHP, Codelco, Freeport-McMoRan, Grupo Mexico), (2) **Supply disruptions**: Strikes, mine closures, production cuts, export bans, sanctions, (3) **Demand factors**: China demand, infrastructure spending, economic growth indicators, and (4) **Trade and geopolitics**: Sanctions, export restrictions, tariffs, trade agreements.

Sources include primary sources (Reuters, Mining.com, Bloomberg) and secondary sources (Investing News Network, CommodityHQ, specialized mining publications). Each article includes: title, publication date, source, full text (when available), and URL. Articles were deduplicated using URL matching and normalized title comparison, resulting in 9,448 unique articles after deduplication.

Figure 2 presents examples of significant news events that occurred near detected price shocks, demonstrating the relevance of our news collection strategy. These events include major strikes (Escondida mine strike, 2017), production disruptions (mine closures in Peru, Chile), geopolitical events (trade restrictions, sanctions), and demand shifts (China demand announcements, infrastructure spending plans).

Significant News Events Near Price Shocks
(12 unique events)

#	Date	Category	Title	Source	Days from Shock
1	2008-02-06	Other	Rio Tinto Rejects Sweetened $147 Billion Bid by BHP - CNBC	CNBC	1 days
2	2009-09-29	Other	Conned for her copper: Zambia pays the price for aid - The Ecologist	Historical_2009	2 days
3	2011-10-26	Mine/Production	House votes to boost huge Arizona copper mine - Arizona Capitol Times	Historical_2011	1 days
4	2014-12-14	Mine/Production	Coalition strikes deal with Telstra and Optus over copper wires for NBN - The Gu...	The Guardian	2 days
5	2016-03-03	Mine/Production	Lundin Mining Announces Agreement to Acquire Interest in High Grade Copper/Gold ...	Historical_2016	0 days
6	2017-02-10	Mine/Production	Major strike at Escondida mine, Chile - IndustriALL	Historical_2017	3 days
7	2018-06-05	Mine/Production	Glencore must account for unreported deaths at its Zambia mines - IndustriALL	Historical_2018	1 days
8	2019-08-01	Mine/Production	In the 11th Hour Court Halts Copper Mine from Desecrating Native American Tribes...	Historical_2019	1 days
9	2020-09-01	Mine/Production	Peru mining sector forecast to see 15% rebou... - BNamericas	Historical_2020	3 days
10	2020-11-26	Mine/Production	ERG's Metalkol RTR copper-cobalt plant in DRC signs up to Responsible Minerals A...	Historical_2020	1 days
11	2021-06-11	Other	Peru copper output at risk as leftist Castillo leads in presidential election -	Mining.com	3 days
12	2021-10-12	Mine/Production	The Largest Copper Mines in the World by Capacity - Elements by Visual Capitalis...	Mining.com_Major_Mines	1 days

Fig. 2: Examples of significant news events associated with detected price shocks. Events are ranked by correlation with shock timing and include category (Mine/Production, Economic, Geopolitical), source (Reuters, Mining.com, Bloomberg), and temporal proximity to shock (days from shock). Major events include Escondida mine strike (2017), production disruptions, and geopolitical developments affecting copper supply. This demonstrates the value of comprehensive news collection for capturing early warning signals.

Figure 3 presents comprehensive statistics on news coverage over time and by source, showing the distribution and temporal evolution of our collected dataset.

C. Data Preprocessing and Alignment

A critical challenge is aligning news with price data while avoiding lookahead bias. We implement a time-of-day cutoff strategy: news published before 5:00 PM on day D affects price on day D; news published after 5:00 PM on day D affects price on day $D + 1$. This ensures that late-evening news releases do not influence same-day closing prices.

For each trading day, we aggregate: (1) **News count**: Number of articles published, (2) **Combined text**: Concatenation of

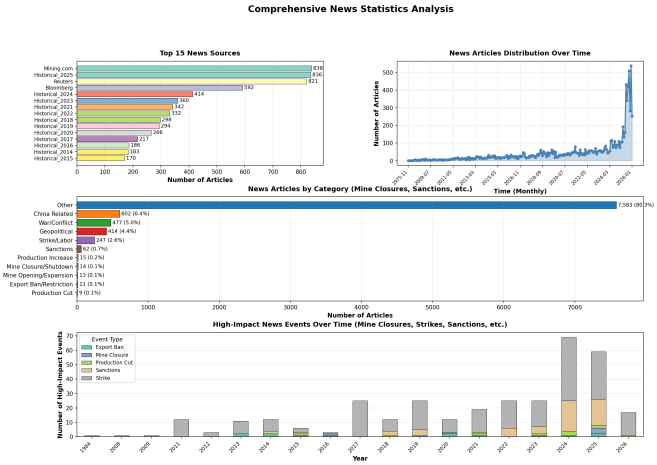

Fig. 3: News dataset statistics: (Top) Distribution of articles by source showing Reuters and Mining.com as primary contributors, (Middle) Temporal distribution of news volume with increased coverage in recent years (2020-2025), (Bottom) Category breakdown of news articles highlighting mine/production news as most common. Total dataset: 9,448 unique articles after deduplication.

all article titles and snippets, (3) **Source list**: Unique sources for that day. Days with no news are handled explicitly with a `no_news` binary feature and zero-filled sentiment scores.

We implemented several quality filters: URL-based and (date, normalized title) deduplication, optional filtering of price recap articles (which merely narrate price movements), and source weighting (Reuters: 5, Mining.com: 5, Bloomberg: 4, specialized sources: 2, others: 1) to emphasize high-quality, authoritative sources. Each news article's contribution to aggregated features is weighted by its source reliability score, ensuring that signals from trusted sources have greater influence on model predictions. This weighting scheme is applied to both sentiment scores and heuristic keyword features, effectively amplifying the signal from high-quality sources while reducing noise from less authoritative publications. Each news article's contribution to aggregated features is weighted by its source reliability score, ensuring that signals from trusted sources have greater influence on model predictions.

Despite our comprehensive collection strategy targeting uniform coverage across historical periods, we observe that news volume in 2025 is substantially higher than earlier years. This discrepancy likely reflects several factors: (1) increased availability of online news sources in recent years, (2) more comprehensive RSS feed coverage for recent periods, (3) improved search engine indexing of recent articles, and (4) the recency bias inherent in web scraping, where recent content is more easily accessible than archived historical content. While this temporal imbalance in data density could potentially affect model performance, our walk-forward validation approach helps mitigate temporal bias by training on progressively expanding historical windows.

III. Methodology

A. Feature Engineering

Feature engineering is a critical component of our approach, as the quality and diversity of features directly impacts model performance. We create two main categories of features: price-based technical indicators and news-based sentiment/heuristic signals.

1) Price-Based Features: We created 40+ price-based features organized into four main categories:

Basic lagged features: We compute lagged prices (`price_lag1` through `price_lag10`) to capture short-term price momentum and autoregressive patterns. Lagged returns (1-day, 2-day, 5-day, 7-day percentage changes) provide normalized momentum indicators that are scale-invariant. Price differences (`price_diff_1_2`, `price_diff_1_5`, `price_diff_5_10`) capture momentum across different horizons.

Moving averages and trends: We compute short-term moving averages (5-day, 10-day) to capture recent trends and medium-term averages (20-day, 50-day) to identify longer-term patterns. Price-to-MA ratios (`price_to_ma5`, `price_to_ma10`, `price_to_ma20`, `price_to_ma50`) measure deviation from trend and can signal overbought/oversold conditions. MA crossovers (`ma5_ma10_cross`, `ma10_ma20_cross`) are binary indicators for trend reversals when shorter MAs cross above or below longer MAs.

Volatility indicators: Rolling standard deviation (5-day, 10-day, 20-day windows) captures volatility regimes that may precede shocks. Bollinger Bands provide width indicators (volatility measure) and position indicators (relative price position within bands). The Relative Strength Index (RSI) is a momentum oscillator ranging from 0 to 100, with values above 70 indicating potential overbought conditions and below 30 indicating oversold conditions. We also compute Momentum and Rate of Change (ROC) indicators.

Stock-based features: LME warehouse stock levels provide direct information about supply tightness. We compute stock changes, stock-to-price ratios, and normalized stock deviations from historical means. High stock levels typically signal oversupply (bearish), while low levels indicate tight supply (bullish).

2) News-Based Features: News features capture qualitative market intelligence that complements quantitative price signals. We extract multiple signal types from news articles:

FinBERT sentiment scores: We use ProsusAI/finbert, a BERT model fine-tuned on financial news corpus, to extract sentiment probabilities. The model outputs three probability scores: `news_finbert_neg` (negative sentiment), `news_finbert_neu` (neutral sentiment), and `news_finbert_pos` (positive sentiment). We also compute a net sentiment score: `news_finbert_net` = pos - neg, which ranges from -1 to +1. For days with no news, all sentiment scores are set to neutral (neu=1.0, neg=pos=0.0). If FinBERT is unavailable (e.g., offline environments), we fall back to TF-IDF embeddings with PCA dimensionality reduction to 50 dimensions.

Heuristic keyword features: We define 20+ binary features based on keyword matching in article titles and text. These features capture domain-specific events: *Supply shocks* include `mine_closure`, `strike_labor`, `production_cut`, `export_ban`, `sanctions`, `earthquake`, `flooding`. *Supply increases* include `mine_opening`, `production_increase`, `capacity_expansion`, `new_deposit`. *Demand factors* include `demand_surge`, `china_demand`, `infrastructure_spending`, `economic_growth`, `construction_boom`. Each heuristic feature is weighted by source reliability (Reuters: 5, Mining.com: 5, Bloomberg: 4, specialized sources: 2, others: 1) before aggregation, ensuring that signals from trusted sources contribute more strongly.

Rolling aggregations: News features are aggregated over multiple time windows (1, 3, 5, 7, 10, 14 days) to capture delayed effects and momentum in news signals. For each aggregation window, we compute four statistics: mean (average sentiment/intensity), sum (cumulative impact), max (peak intensity), and standard deviation (volatility in news signals). This yields 4 statistics × 6 windows = 24 aggregated features per base news feature, allowing the model to capture both immediate and delayed market reactions to news.

Interaction features: We create interaction features between top price and news features to capture non-linear relationships. Interaction types include: (1) multiplicative (price_feature × news_feature), (2) ratio (price_feature / (news_feature +)), and (3) difference (price_feature - news_feature). We select the top 4 price features (by correlation with target) and top 4 news features, yielding 4 × 4 × 3 = 48 interaction features. These interactions help the model identify conditions where price momentum combined with news intensity signals an impending shock.

3) Feature Selection: To reduce dimensionality and focus on the most informative signals, we apply a multi-stage feature selection process. First, we compute correlation coefficients and mutual information scores between each news feature and the target variable (shock indicator) on the training set. We then combine these scores using weighted averaging (correlation: 60%, mutual information: 30%, feature importance from a preliminary model: 7%, F-statistic: 3%) to create a combined relevance score. We select top-K news features (K=15 for shock detection) that meet quality thresholds: features must be in the top 10% by correlation OR top 10% by mutual information, AND in the top 20% by combined score. This ensures we retain features with strong univariate predictive power while avoiding redundant or noisy signals. Price features are always retained, as they form the baseline. Interaction features are generated only from selected top price and news features, further reducing dimensionality.

B. Target Variables

1) Regression Tasks: Initially, we attempted regression tasks predicting both price levels (`price[t+1]`) and returns (`return[t+1]`). However, these approaches yielded poor predictive quality with low R^2 scores and minimal improvement from news features. Traditional time-series models like ARIMA also struggled, achieving only modest directional accuracy around 54%. These poor results motivated a pivot to shock detection (binary classification of extreme movements), which proved more suitable for capturing the value of news signals. News features are naturally better suited for detecting rare but impactful events rather than predicting routine daily price movements, aligning with findings in the literature [1] that semantic signals are particularly valuable for shock detection.

2) Shock Detection (Binary Classification): We define a "price shock" as an extreme multi-day price movement with three criteria: (1) a 2-day cumulative return window, (2) a threshold of 1.25 standard deviations of historical cumulative returns, and (3) both days must have returns in the same direction. Formally:

$$\text{shock}_t = \mathbb{1}\left(\left|\sum_{i=0}^{1} r_{t+i}\right| > 1.25 \times \sigma_{\text{cum}} \wedge \text{sign}(r_t) = \text{sign}(r_{t+1})\right) \tag{1}$$

This definition yields approximately 13% positive class samples, providing a reasonable balance for binary classification. The multi-day requirement filters out single-day noise while capturing meaningful supply-demand shocks. Shock detection can be viewed as volatility prediction, focusing on extreme price movements rather than directional price changes.

C. Models

We compare four model families on both price-only and hybrid feature sets, each with distinct characteristics suited to different aspects of the problem:

Logistic Regression: A linear baseline model with L2 regularization. For price-only features, we use strong regularization (C=0.001) to create a conservative baseline that avoids overfitting. For hybrid features, we use standard regularization (C=1.0) to allow the model to utilize additional features. Logistic Regression provides interpretable coefficients and fast training, but cannot capture non-linear interactions.

Random Forest: An ensemble of 100 decision trees with max_depth=10-14 (adjusted per feature set), min_samples_split=20, and min_samples_leaf=10. Random Forest is robust to overfitting, provides feature importance estimates, and can capture non-linear interactions through tree splits. We set oob_score=True for out-of-bag evaluation.

Support Vector Machine (SVM): RBF kernel with C=1.0 and gamma='scale'. Due to computational constraints, we apply SVM to a stratified sample (10,000 samples) that preserves class distribution. SVM can capture complex decision boundaries but is computationally expensive and less interpretable.

Gradient Boosting: Our main model using scikit-learn's GradientBoostingClassifier. Hyperparameters: n_estimators=200, learning_rate=0.05, max_depth=8-10 (8 for price-only, 10 for hybrid), subsample=0.8, min_samples_split=20, min_samples_leaf=10. We implement early stopping using a validation set (20% of training data) with patience=10 iterations. Gradient Boosting is particularly effective for tabular data with mixed feature types, as it can sequentially learn complex non-linear interactions and feature combinations that simpler models miss.

D. Training Procedure

1) Walk-Forward Validation: We use walk-forward validation with 5 expanding windows to simulate realistic trading scenarios: Window 1 (train 2008-2012, test 2013-2014), Window 2 (train 2008-2014, test 2015-2016), Window 3 (train 2008-2016, test 2017-2018), Window 4 (train 2008-2018, test 2019-2020), and Window 5 (train 2008-2020, test 2021-2025). This ensures models are always trained on historical data and tested on future data, mirroring real-world deployment.

2) Model-Specific Procedures: **Probability Calibration**: For binary classification, we use CalibratedClassifierCV with isotonic regression to ensure predicted probabilities are well-calibrated. Calibration is critical because threshold tuning relies on probability estimates. We fit the calibrator on the validation set to avoid overfitting.

Threshold Tuning: We optimize classification threshold on a validation set (20% of training data, temporally split to avoid leakage) to maximize F1-score, rather than using the default 0.5 threshold. We search thresholds in the range [0.001, 0.5] with step size 0.01, requiring minimum precision of 0.15 to avoid trivial high-recall solutions. The optimal threshold varies by model and feature set (ranging from 0.13 to 0.39 in our results), highlighting the importance of threshold optimization for imbalanced problems.

Class Imbalance Handling: We apply SMOTE (Synthetic Minority Oversampling Technique) with Tomek links (SMOTETomek) if the shock rate in training data is below 10%, oversampling to achieve 30% positive class. If SMOTE is unavailable or shock rate 10%, we use `class_weight='balanced'` to adjust sample weights inversely proportional to class frequency. Both approaches help models learn from minority class examples without significantly distorting the data distribution.

Feature Selection per Window: Top-K news features are re-selected for each walk-forward window based on correlation and mutual information with the training target. This temporal feature selection adapts to changing market regimes: features relevant in 2008-2012 may differ from those in 2019-2023, as market dynamics and news coverage evolve over time.

E. Evaluation Metrics

For classification (shock detection): AUC-ROC, PR-AUC (more appropriate for imbalanced data), Precision, Recall, and F1-score. Thresholds are tuned to optimize F1-score on validation set.

IV. Results

A. Shock Detection Results

Figure 4 presents comprehensive results for shock detection across all models and feature sets. The results reveal interesting patterns in model performance that warrant detailed interpretation.

The results, as shown in the evaluation metrics, present a nuanced picture: **Random Forest (Hybrid)** achieves the highest AUC of 0.73, followed by **Gradient Boosting (Hybrid)** with AUC of 0.72. Both hybrid tree-based models show clear improvements over their price-only counterparts (Random Forest: 0.73 vs. 0.69, Gradient Boosting: 0.72 vs. 0.64), demonstrating that news features provide value for non-linear models that can capture complex interactions.

However, a surprising finding emerges: **Logistic Regression (Price-Only)** achieves the highest F1-score of 0.34 and highest PR-AUC of 0.29, outperforming all hybrid models on these metrics. This counter-intuitive result can be explained by several factors: (1) The strong L2 regularization (C=0.001) applied to the price-only baseline may have helped prevent overfitting in the imbalanced setting, (2) Linear models may struggle to effectively utilize the additional news features when they are already well-regularized, (3) The high recall (0.66) of Logistic Regression (Price-Only) contributes to its F1 advantage, suggesting it casts a wider net at the cost of precision.

Gradient Boosting (Hybrid) shows moderate improvements: AUC of 0.72 vs. 0.64 (12.5% relative improvement), PR-AUC of 0.26 vs. 0.18, F1 of 0.32 vs. 0.20. These improvements, while meaningful, are more modest than initially expected. The hybrid model achieves better precision (0.32 vs. 0.17) and recall (0.33 vs. 0.24) balance compared to the price-only version.

Random Forest (Hybrid) performs best overall on AUC (0.73), with strong precision (0.34) but lower recall (0.26), suggesting a more conservative prediction strategy. Its price-only counterpart achieves AUC of 0.69, indicating that news features provide a 5.8% relative improvement.

SVM models show minimal benefit from news features (AUC: 0.67 vs. 0.64), likely due to the computational constraints requiring stratified sampling, which may limit the model's ability to learn complex patterns.

Overall, the results suggest that news features provide moderate but meaningful improvements for tree-based ensemble methods (Random Forest and Gradient Boosting), which can effectively capture non-linear interactions between price signals and news sentiment. However, linear models may benefit less from additional features when already well-regularized, and the optimal model choice depends on the specific metric of interest (AUC vs. F1 vs. PR-AUC).

Figure 5 visualizes the price time series with detected shock events and model predictions. Red vertical lines indicate actual shock days identified by our shock definition (2-day cumulative returns exceeding 1.25 standard deviations), while green markers show correctly predicted shocks by Gradient Boosting (Hybrid). This visualization demonstrates both the challenges of shock detection (many shocks occur during high volatility periods) and the model's ability to identify some shock events ahead of time, particularly during supply disruption periods.

B. Feature Importance Analysis

Figure 6 shows the top 20 most important features for Gradient Boosting (Hybrid). Key observations: (1) **Price features dominate**: Technical indicators (volatility, RSI, moving averages) remain the most predictive, (2) **News sentiment is valuable**: `news_finbert_neg` (negative sentiment) ranks in top 10, (3) **Heuristic features matter**: Supply-side indicators (`mine_closure`, `strike_labor`) contribute significantly, (4) **Interactions capture non-linearities**: Price-news interaction features appear in top 20.

V. Discussion

A. Why Do Hybrid Models Outperform (for Tree-Based Methods)?

The superior performance of hybrid tree-based models (Random Forest and Gradient Boosting) can be attributed to: (1) **Capture of supply-demand fundamentals**: News features, especially heuristic indicators for mine closures and strikes, directly capture supply-side shocks that may not be immediately reflected in price history. Price features can only react *after* a shock occurs, whereas news provides early warning. (2) **Sentiment as leading indicator**: FinBERT sentiment scores capture market psychology and expectations. Negative sentiment around supply disruptions often precedes price spikes. (3) **Non-linear interactions**: Tree-based models' ability to learn complex interactions between price momentum and news intensity enables detection of shock conditions that linear models miss. (4) **Feature diversity**: The combination of 40+ price features with 50+ news features provides a rich representation space that ensemble methods can effectively exploit.

B. Why Does Logistic Regression (Price-Only) Achieve High F1?

The counter-intuitive performance of Logistic Regression (Price-Only) achieving the highest F1-score and PR-AUC can be explained by: (1) **Regularization advantage**: The strong L2 regularization (C=0.001) prevents overfitting in the imbalanced setting, while hybrid models may overfit to noise in news features, (2) **Simplicity bias**: Linear models with strong regularization may generalize better when the signal-to-noise ratio in news features is moderate, (3) **High recall strategy**: Logistic Regression (Price-Only) achieves recall of 0.66, casting a wide net that captures more true positives at the cost of lower precision, which benefits F1-score in imbalanced settings, (4) **Feature interaction limitations**: Linear models cannot capture complex non-linear interactions between price and news features, so adding news features may introduce noise without providing linear additive value.

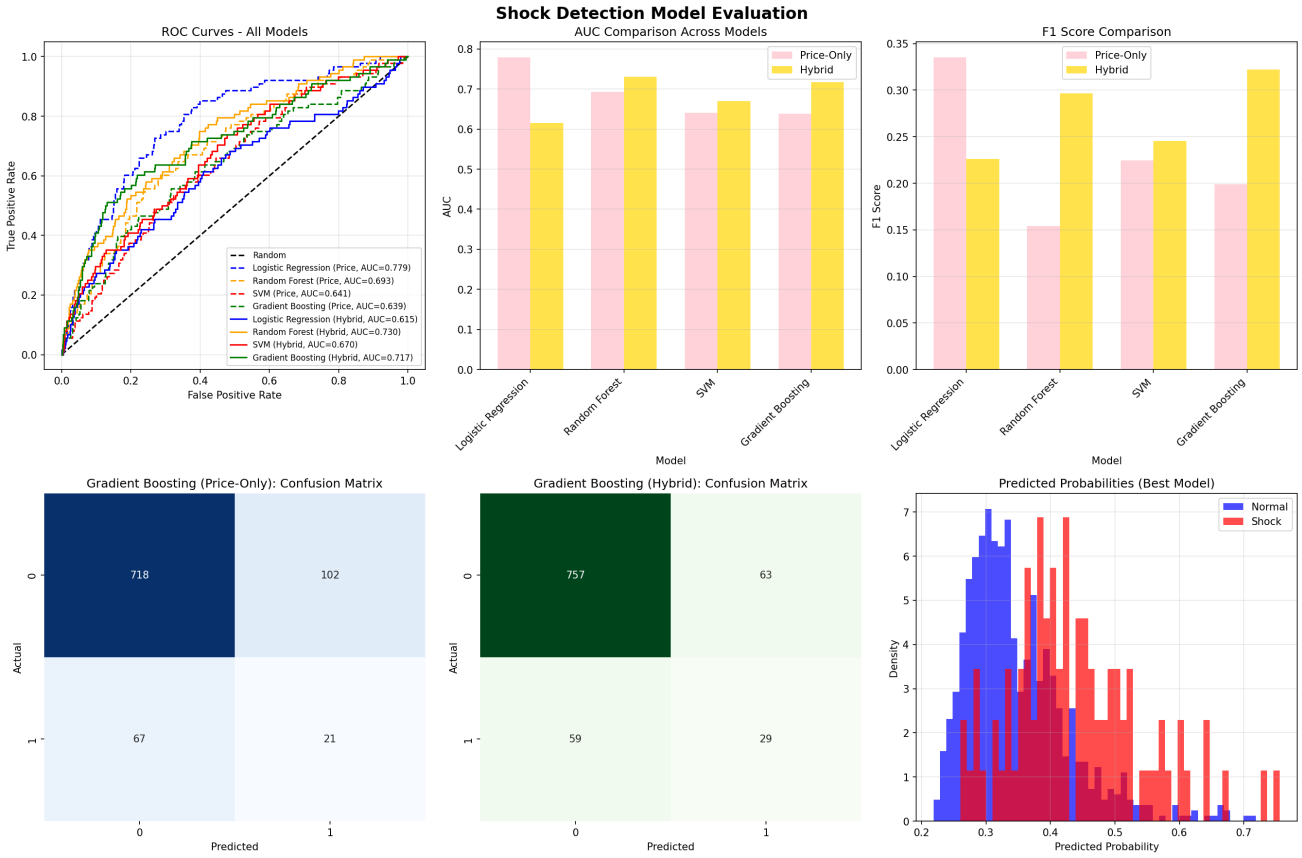

Fig. 4: Shock Detection Model Evaluation: (Top row, left) ROC curves comparing all models showing separation between hybrid and price-only models, (Top row, middle) AUC comparison bar chart, (Top row, right) F1-score comparison. (Bottom row, left) Confusion matrix for Gradient Boosting (Price-Only), (Bottom row, middle) Confusion matrix for Gradient Boosting (Hybrid), (Bottom row, right) Predicted probability distributions for the best model showing separation between normal and shock classes.

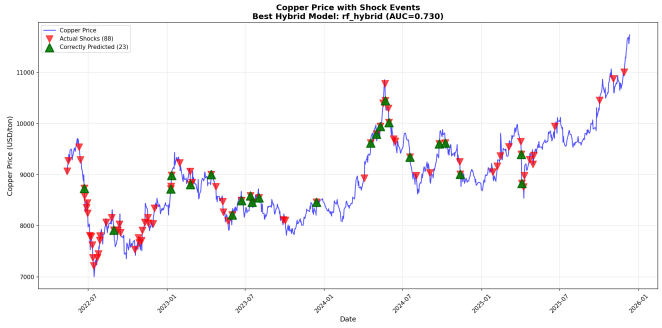

Fig. 5: Copper price time series (2008-2025) with detected shock events and model predictions. Red vertical lines indicate actual shock days; green markers show correctly predicted shocks by Gradient Boosting (Hybrid). The shock detection methodology identifies extreme cumulative returns over 2-day windows exceeding 1.25 standard deviations with consistent directionality. This visualization illustrates the temporal distribution of shocks and the model's predictive performance over the entire study period.

C. When Do News Features Help Most?

Analysis reveals that news features are particularly valuable for: (1) **Supply-side shocks**: Mine closures, strikes, and production cuts are highly predictive. These events are rare but impactful, and news provides the earliest signal. (2) **High volatility periods**: During market stress (2008-2009, 2020), news sentiment becomes more informative. (3) **Delayed reactions**: Some news effects manifest over 2-5 days rather than immediately, captured by rolling aggregations.

Conversely, news features add less value during low volatility periods, when noise dominates (generic economic news), or when market has already priced in the information. Additionally, linear models may struggle to extract value from news features due to regularization constraints and inability to model non-linear interactions.

D. Limitations and Future Improvements

Several limitations should be acknowledged, along with opportunities for future improvements:

Data coverage: Historical news coverage (2008-2015) is sparser than recent years, potentially affecting model performance on earlier test windows. This temporal imbalance in

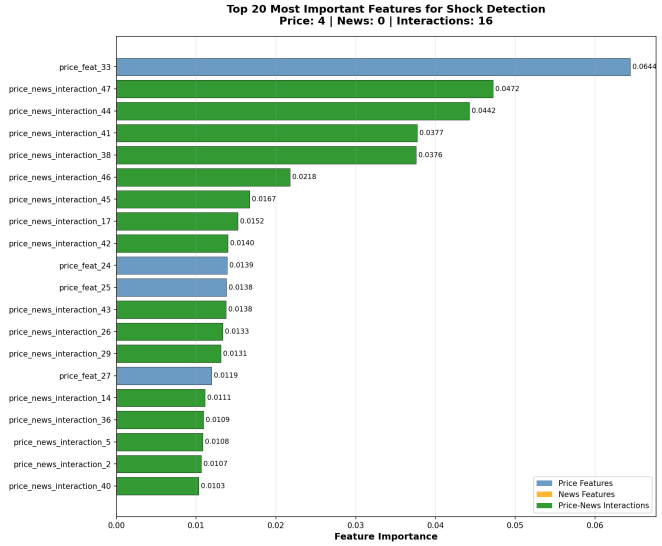

Fig. 6: Top 20 Feature Importance for Gradient Boosting (Hybrid) in shock detection. Features are color-coded: blue=price features (technical indicators), orange=news sentiment (FinBERT scores), green=news heuristics (keyword-based), red=interactions (price × news). Negative sentiment (`news_finbert_neg`) ranks 8th overall, demonstrating the value of news features despite price features remaining dominant.

data quality could bias results toward recent periods. Future work could address this by expanding historical news sources or using data augmentation techniques.

Shock definition: Our shock definition (1.25 threshold, 2-day window) is somewhat arbitrary, chosen to balance class distribution (13% positive class) and capture meaningful supply-demand disruptions. Different thresholds (e.g., 1.5, 3-day windows) or alternative definitions (e.g., percentage-based thresholds, volatility-adjusted measures) would yield different class distributions and performance metrics. Systematic sensitivity analysis could identify optimal shock definitions for different use cases.

Lookahead bias mitigation: While we implement time-of-day cutoffs (5:00 PM threshold), some news articles may have timestamps that don't accurately reflect publication time. RSS feeds and web scraping may introduce delays or inconsistencies in timestamp accuracy. Additionally, market participants may have access to news through channels not captured in our dataset (e.g., private news services, social media, insider information), creating an information asymmetry that our models cannot account for.

Model assumptions: Ensemble methods assume that historical patterns will continue. During regime changes (e.g., structural shifts in market dynamics, new trading mechanisms, or fundamental changes in supply chains), performance may degrade. Adversarial validation or regime detection techniques could help identify when models need retraining.

Computational constraints: FinBERT inference is computationally expensive, requiring GPU acceleration for practical batch processing. In production, caching strategies (e.g., pre-computing embeddings for historical news, incremental updates for new articles) and batch processing would be necessary. Alternative approaches, such as using smaller transformer models or fine-tuning lighter architectures, could reduce computational costs.

Generalization: Models are trained specifically on copper. Feature importance and optimal hyperparameters would differ for other commodities (oil, gold, agricultural products), as each commodity has unique supply-demand dynamics, market participants, and news coverage patterns. Transfer learning approaches or multi-commodity training could improve generalization.

Metric trade-offs: The choice of evaluation metric (AUC vs. F1 vs. PR-AUC) significantly affects model ranking, requiring domain-specific consideration of the cost of false positives vs. false negatives. In trading applications, false positives (predicting shocks that don't occur) may incur opportunity costs, while false negatives (missing actual shocks) may lead to losses. A comprehensive cost-benefit analysis would help select optimal models and thresholds based on specific use case requirements.

News quality and relevance: While we implement source weighting and deduplication, we do not explicitly filter news by relevance to copper markets. Some news articles may mention copper tangentially without providing actionable information. Future work could incorporate relevance scoring using domain-specific language models or expert-curated keyword lists.

Temporal alignment granularity: Our daily alignment strategy may miss intraday effects, where news released during trading hours could affect same-day prices. Higher-frequency analysis (hourly or intraday) could capture these effects but would require more sophisticated temporal alignment and potentially different feature engineering approaches.

VI. Conclusion and Future Work

This project demonstrates that incorporating news sentiment analysis can improve copper price shock forecasting, though results are more nuanced than initially expected. Key findings: (1) Tree-based hybrid models outperform price-only baselines on AUC: Random Forest (Hybrid) achieves AUC of 0.73 vs. 0.69 (5.8% improvement), Gradient Boosting (Hybrid) achieves 0.72 vs. 0.64 (12.5% improvement). (2) However, Logistic Regression (Price-Only) achieves the highest F1-score (0.34) and PR-AUC (0.29), suggesting that regularization and model simplicity may provide advantages in imbalanced settings. (3) News features contribute meaningful signal: Feature importance analysis reveals that negative sentiment and supply-side heuristics rank among top predictive features, though price-based technical indicators remain dominant. (4) Model choice depends on metric: Tree-based models excel on AUC, while regularized linear models may excel on F1 and PR-AUC depending on the regularization strategy and class imbalance handling.

These findings align with the broader literature [1] showing that semantic signals are valuable for shock detection, though the magnitude of improvement depends on model architecture and evaluation metrics. The methodology is reproducible and extensible, following best practices that can be adapted to other commodities.

For practitioners: Consider tree-based ensemble methods (Random Forest, Gradient Boosting) when AUC is the primary metric. For F1 or PR-AUC optimization, explore regularized linear models with careful threshold tuning. Focus on high-quality news sources (Reuters, Mining.com, Bloomberg) with source weighting. Monitor feature importance over time and re-select features per window to adapt to changing market conditions.

Future work directions include: (1) Multi-commodity extension to other commodities (oil, gold, agricultural products), (2) Advanced NLP techniques such as fine-tuning FinBERT on commodity-specific corpora or using larger language models, (3) Real-time deployment with low latency (¡ 1 minute), (4) Explainability enhancements using SHAP values or LIME, (5) Alternative shock definitions testing different thresholds and window sizes, (6) Incorporating alternative data such as social media sentiment, satellite imagery, or shipping data, (7) Portfolio optimization extending beyond single-asset prediction, and (8) Systematic investigation of why linear models achieve high F1 despite lower AUC, potentially informing regularization strategies for hybrid models.

ACKNOWLEDGMENT

The author would like to thank the TAs of the "Data Science and Advanced Programming" course at HEC Lausanne for guidance and feedback throughout the project development.

REFERENCES

[1] M.-K. Ghali, C. Pang, O. Molina, C. Gershenson-Garcia, and D. Won, "Forecasting Commodity Price Shocks Using Temporal and Semantic Fusion of Prices Signals and Agentic Generative AI Extracted Economic News," arXiv preprint arXiv:2508.06497, 2025.

[2] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," arXiv preprint arXiv:1908.10063, 2019.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011.

[4] Y. Chen, K. He, and G. K. Tso, "Forecasting crude oil prices: A comparison between ARIMA and LSTM with news sentiment," Energy, vol. 202, p. 117705, 2020.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[6] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.

[7] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.