An Extension of FinEntity: Entity-level Sentiment Classification for Financial Texts

Niranjan Vijaya Krishnan
Princeton University
nv2608@princeton.edu

Lily Weaver
Princeton University
lw3612@princeton.edu

Chaeyoung Lee
Princeton University
cl5587@princeton.edu

Abstract

An effective way to gain an understanding of the sentiment directed towards a financial asset is through recent news articles. However, with the amount of information released daily, keeping track of sentiment changes is challenging. To streamline this process, large language models can be used for named entity recognition (NER) and sentiment classification. In this project, we build upon the FinEntity paper (Tang et al., 2023) by benchmarking more recent language models (DeBERTa, RoBERTa, GPT-4o, Qwen, LLaMA, etc.)—we test with and without a Conditional Random Fields (CRF) layer (Lafferty et al., 2001) for open source models and various prompting strategies for closed source models. Our results show that DeBERTa-CRF outperforms all other models in both precision and F1 scores. Furthermore, we replicate the cryptocurrency case study from FinEntity and extend it to commodities such as oil, gold, copper, and silver, analyzing how entity-level sentiment classification aligns with market trends. The code required for this paper is available at https://github.com/niruvk/FinEntity-Extension.

1 Introduction

Named Entity Recognition (NER) serves as a financial tool for extracting insights from unstructured data such as earnings reports and financial news. It is a foundational tool in NLP to derive insights on specific entities such as companies, financial instruments, and economic indicators. Models such as FinBERT (Araci, 2019) are typically trained to derive sentiment at the sequence-level, however NER enables fine-tuning and allows models to derive sentiment at the entity-level, allowing for more accurate sentiment results. Example 1 demonstrates a scenario where entity-level sentiment extraction is beneficial.

> **Example 1.** *"TSLA stock rose despite the overall market showing signs of a recession with SPY and QQQ down 20%."* A sequence-level sentiment model would classify the overall sentiment as negative. However, an entity-level model would assign:
>
> - TSLA: positive
> - SPY: negative
> - QQQ: negative

Thus, our chosen paper by Tang et al. (2023) introduces an entity-level sentiment classification dataset, called FinEntity, that annotates financial entity spans and their sentiment (positive, neutral, and negative) in financial news. The authors demonstrated that fine-tuning general-purpose transformers (e.g., BERT, FinBERT) could achieve strong performance, and that adding structured prediction layers like Conditional Random Fields (CRF) further improved precision in boundary-sensitive cases. Building on this work, our project explores the effectiveness of more recent models, including RoBERTa, DeBERTa, Qwen, LLaMA, and GPT-4o, on the FinEntity task. We evaluate each model with and without a CRF layer and test large language models (LLMs) like GPT-4o in few-shot and one-shot settings. Our goal is to assess how modern pretrained models, fine-tuned models, and prompting strategies compare in ability to accurately perform financial named entity recognition and sentiment classification. In addition to benchmarking, we conduct a case study applying financial NER to commodities, expanding on the cryptocurrency- focused analysis from the FinEntity paper. Sentiment analysis in commodity markets is particularly relevant given their sensitivity to geopolitical events, supply chain shocks, and speculative behavior.

2 Related Work

Our work most closely builds upon the entity-level sentiment classification work done by Tang et al.

(2023). Their main contributions are creating a financial entity-level sentiment classification dataset called FinEntity, testing pre-trained language models (PLMs) and ChatGPT on entity-level sentiment classification, and conducting a correlation analysis and prediction experiment with Bitcoin prices and daily sentiment scores.

2.1 Financial Sentiment Classification

Natural language processing techniques have been used to identify the sentiment expressed toward an entity in financial news data, albeit at the sequence level (eg. sentences, paragraphs, news articles). Yadav et al. (2020) used an unsupervised approach for classifying financial news based upon sentiment indicators. Its unit of classification was an entire financial news article. They extracted parts of speech (POS) phrases from the article, calculated semantic orientation (SO) using point wise mutual information-information retrieval (PMI-IR), and assigned a classification label to the financial news article based on the average SO of the phrases. Frankel et al. (2022) compared machine learning and dictionary-based methods for analyzing disclosure sentiment of both 10-K reports and earnings conference call transcripts. The unit of classification was also an entire report or transcript. However, financial news articles tend to discuss multiple entities with differing sentiments expressed toward each, which makes long sequence-level sentiment classification an oversimplification of the individual sentiments expressed toward entities. Thus, a finer method that consists of entity recognition and sentiment classification is needed.

2.2 FinEntity Dataset

To this end, Tang et al. (2023) constructed an entity-level sentiment classification dataset, called FinEntity, which consists of sentences and the corresponding financial entity span labels and associated sentiment.

Example 2.
Content: *"On the positive side, Siemens is rallying 6% after a boom in quarterly orders and packaging maker Huhtamaki is also up by 6% after profit beat expectations."*
Annotations:

- {'end': 107, 'tag': 'Positive', 'value': 'Huhtamaki', 'start': 98, 'label': 'Positive'}
- {'end': 29, 'tag': 'Positive', 'value': 'Siemens', 'start': 22, 'label': 'Positive'}

The content of the dataset was collected from a financial news dataset from Refinitiv Reuters Database. It was made sure that there was a balanced distribution of positive/neutral/negative samples and that 80% of the sequences contained more than one entity. The BILOU annotation scheme was used to identity entity spans and each annotated BILU entity was given a sentiment label—resulting in thirteen possible labels (B/I/L/U-positive/neutral/negative and the O label). Each token in the input sequence was assigned one of the thirteen. Twelve annotators were recruited for the task such that each example was annotated by three annotators. Cross-checks and consistency-checks were done to ensure data quality. The final dataset contains of 979 total samples, which includes 2,131 total entities (503 positive entities, 498 negative entities, 1,130 neutral entities).

2.3 Selecting Models

Tang et al. (2023) fine tuned and tested the performance of open source models (BERT, FinBERT) with and without a conditional random field (CRF) layer and tested the performance of ChatGPT 3.5 using zero-shot and few-shot prompting. BERT, which stands for Bidirectional Encoder Representations from Transformers, was developed by Devlin et al. (2019) at Google and presented a novel approach of understanding the context of a word based on both its left and right surroundings. FinBert, developed by Araci (2019), is a specialized version of BERT that was trained on a large corpus of financial text. FinBERT is tailored for financial language processing as the model is familiar with specialized finance terms and language from its exposure to financial texts. A conditional random field (CRF) layer commonly takes as input the output of the BERT model and outputs the most likely sequence of labels (Lafferty et al., 2001). The CRF layer improves prediction by considering the relationship between output labels and learning the transition scores between each output label. The transition from B-positive to I-positive should be very likely, whereas the transition from B-positive to I-negative should be very unlikely. Considering the fact that we are using BILOU + pos/neg/neutral sequence tagging, using a CRF layer to enforce valid tag transitions is crucial for

good performance. OpenAI's gpt-3.5-turbo model was released in 2022 and supports code generation, basic reasoning, and text tasks. However, they are known to be less accurate in multi-step reasoning and math. We expand upon this list by testing more recent models such as DeBERTa, RoBERTa, Llama, and Qwen, as well as the gpt-4o model. DeBERTa, which stands for Decoding-enhanced BERT with Disentangled Attention, was introduced by Microsoft in 2021 (He et al., 2021). DeBERTa separates the positional and content embeddings, which helps the model understand relative position and semantic meaning independently. RoBERTa, which stands for robustly optimized BERT approach, was developed by Facebook AI in 2019 (Liu et al., 2019). It uses the same BERT architecture, but trains it on a larger training corpus by removing the next sentence prediction objective, applying dynamic masking, and increasing batch sizes and training time. LLaMA, which stands for Large Language Model Meta AI, was released by Meta in 2024, is a autoregressive transformer model trained on public data and designed to be lightweight and efficient for research use (Touvron et al., 2023). Qwen, developed by Alibaba in 2024, is another open-source model trained on Chinese, English, and multimodal corpora (Bai et al., 2023). OpenAI's GPT-4o model was released in 2024 with an entirely new architecture from prior GPT-4 models (OpenAI et al., 2024). GPT-4o is compatible with multi-modal inputs and outputs.

3 Methodology

3.1 Model Benchmarking on FinEntity

We begin by fine-tuning and evaluating a diverse set of language models on the FinEntity dataset. In addition to the baselines used in the original FinEntity paper (BERT, FinBERT, GPT-3.5), we incorporate more recent models, including enhanced BERT variants (DeBERTa and RoBERTa), decoder-only models (Qwen and LLaMA), and GPT-4o. For each open source model, we test variants with and without a Conditional Random Field (CRF) decoding layer. The Conditional Random Field (CRF) decoding layer improves token-level sequence labeling by modeling dependencies between adjacent labels, which is particularly beneficial in named entity recognition (NER) tasks. For GPT-4o, we evaluate both one-shot and few-shot prompting settings as well as fine-tuning. Model performance is evaluated using F1 score, with results reported per sentiment class (positive, negative, neutral) as well as aggregated using micro, macro, and weighted averages.

3.2 Commodity Correlation Analysis

Then, to further evaluate real-world applications of financial NER, we conduct a commodity-focused case study as an extension of Tang et al. (2023) cryptocurrency analysis. We focus on oil, gold, copper, and silver and perform sentiment analysis using the Bloomberg Financial News dataset, a dataset of 446,762 financial news articles scraped from Bloomberg between 2006 and 2013 (Benayoun, 2024; Philippe Remy, 2015). FinBERT is employed for both sequence-level and entity-level sentiment analysis to maintain consistency with the original FinEntity framework and to eliminate confounding factors arising from model differences. Our process consists of the following steps:

3.2.1 Sequence-Level Analysis

We extract sentences with target keywords for each commodity (e.g., "oil", "WTI", "OPEC" for oil) from articles on the same date and feed them to FinBERT to obtain sequence-level sentiment scores (positive/negative/neutral).

3.2.2 Entity-Level Analysis

We run articles containing commodity-related keywords through our fine-tuned FinBERT-CRF model to obtain a set of entities along with their corresponding sentiment score for each day.

3.2.3 Correlation Analysis

For each day and each commodity, we obtain a net sentiment score by summing all sentiment values (positive = +1, negative = -1, neutral = 0) across relevant sentences or entities. The resulting time series of daily sentiment scores is normalized using min-max scaling to allow for comparison across commodities and methods.

To quantify the relationship between sentiment and commodity prices, we compute the maximal information coefficient (MIC). MIC is a non-parametric measure of association that captures a wide range of linear and non-linear relationships between paired variables. Unlike Pearson or Spearman correlations, MIC can detect both monotonic and non-monotonic dependencies, making it well-suited for identifying complex patterns in financial time series (Reshef et al.,

2011). Commodity price data is obtained via the yfinance Python library and aligned with the sentiment time series over the 2012–2013 period.

4 Results

4.1 Model Benchmarking on FinEntity

We evaluated numerous models and their CRF counterparts on their performance on the FinEntity Dataset. As shown in Figure 1, models with an added CRF layer typically outperform their base versions. Comparing the F1 scores across all models, DeBERTa-CRF (He et al., 2021) attained the highest scores for most metrics, including Negative F1 score of 0.88, Positive F1 score of 0.94, and Micro, Macro, and Weighted Average F1 scores of 0.89. Tang et al. (2023) found FinBERT-CRF to be the best model for entity-level sentiment. This is assumed to be because FinBERT is pre-trained on a large corpus of financial text, which is applicable for entity-level sentiment classification in financial texts. However, DeBERTa has several improvements over BERT: it is trained on a much larger corpus, including sources such as wikipedia, it has an enhanced mask decoder, and utilizes disentangled attention (He et al., 2021; Devlin et al., 2019). This gives DeBERTa an advantage in precise context understanding, which is essential for entity-level sentiment classification. The CRF layer additionally augments the model by modeling dependencies in output labels (Lafferty et al., 2001). This explains the result why DeBERTa-CRF performs the best amongst all other models.

Open-weight decoder-only LLMs such as LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023), and GPT-4o consistently underperform compared to other models. F1 scores for these models range from 0.54 - 0.87, indicating that they struggle in fine-tuning compared to the BERT models. This is because decoder-only models specialize in autoregressive tasks such as predicting the next token. As a result, these models emphasize future tokens, because they are unidirectional and context is lost for sentiment classification. This is why encoder-based models such as BERT and DeBERTa perform much better. They are bidirectional, making them much better and sentiment classification and understanding context.

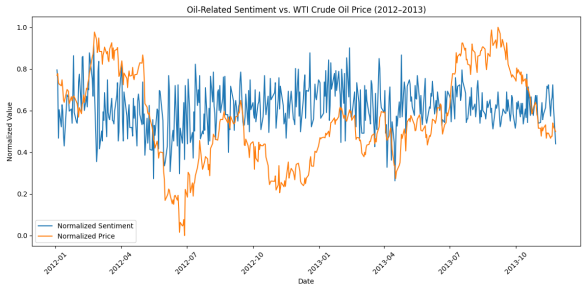

Figure 1: Normalized Sequence-level Sentiment (Orange) and Price (Blue) for Oil (2012-2013)

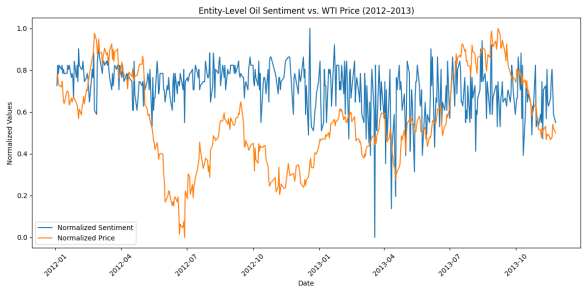

Figure 2: Normalized Entity-level Sentiment (Orange) and Price (Blue) for Oil (2012-2013)

4.2 Commodity Correlation Analysis

To evaluate the correlation between sentiment analysis and commodity price, we calculated the Maximum Information Coefficient (MIC) from using both sequence-level FinBERT and entity-level FinBERT-CRF.

Comparing sequence-level and entity-level sentiment correlation, we find that entity-level sentiment shows greater correlation for copper and silver and the same correlation for oil. However, entity-level sentiment for gold has weaker correlation than sequence-level correlation. We hypothesize that this is because gold is heavily tied to the market. As a result, when it appears in articles, the sentences that contain gold related keywords are able to gain information from the overall market in sequence-level sentiment classification. This allows for better correlation for sequence-level sentiment than entity-level sentiment. Correlation with oil prices being the same for sequence-level and entity-level sentiment can be attributed to the correlation between oil news and oil price being low in general. Oil correlation performed the worst of the three commodities, indicating that sentiment-classification is not the best suited for explaining oil prices.

Model	Negative	Positive	Neutral	Micro Avg	Macro Avg	Weighted Avg
BERT	0.75	0.81	0.82	0.80	0.80	0.80
BERT-CRF	0.82	0.81	0.81	0.81	0.81	0.81
FinBERT	0.83	0.81	0.84	0.83	0.83	0.83
FinBERT-CRF	**0.88**	0.84	0.82	0.84	0.85	0.84
GPT-3.5 (zero)	0.58	0.39	0.71	0.59	0.56	0.59
GPT-3.5 (few)	0.62	0.73	0.61	0.67	0.65	0.68
DeBERTa	0.67	0.72	0.81	0.75	0.73	0.75
DeBERTa-CRF	**0.88**	**0.94**	0.87	**0.89**	**0.89**	**0.89**
RoBERTa	0.87	0.92	0.88	**0.89**	**0.89**	**0.89**
RoBERTa-CRF	0.82	0.86	**0.90**	0.87	0.86	0.87
LLaMA	0.68	0.79	0.59	0.72	0.69	0.72
LLaMA-CRF	0.70	0.79	0.60	0.73	0.71	0.72
Qwen	0.66	0.76	0.57	0.71	0.68	0.70
Qwen-CRF	0.67	0.79	0.58	0.72	0.70	0.70
GPT-4o (zero)	0.54	0.61	0.60	0.60	0.58	0.60
GPT-4o (few)	0.58	0.75	0.71	0.71	0.68	0.71
GPT-4o (fine-tuned)	0.79	0.81	0.87	0.85	0.83	0.85

Table 1: Entity-level Sentiment Classification Results

The correlation analysis reveals that the performance of sequence-level versus entity-level sentiment classification can vary depending on the asset. Entity-level sentiment classification typically outperforms sequence-level, however there are exceptions such as gold where sequence-level sentiment performs better. Furthermore, the oil correlation analysis shows that when there is low general correlation between news sentiment and price, neither sentiment-level nor sequence-level performs better over the other.

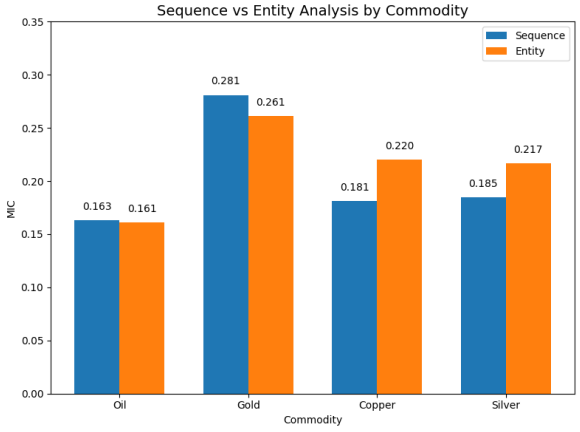

Figure 3: Correlation Between Sentiment Score and Commodity Price for Sequence-level and Entity-level sentiment

5 Conclusion

Our study shows that adding a Conditional Random Field (CRF) layer to a base model significantly enhances its performance on entity-level sentiment classification tasks in financial texts. When benchmarking all models on the FinEntity dataset, DeBERTa-CRF achieved the highest F1 scores across most sentiment classes, outperforming FinBERT-CRF and all other models. This shows how important architecture, training corpus, and sequence modeling is in NER and entity-level sentiment classification tasks.

We additionally found that decoder-only models such as LLaMA, Qwen, and GPT perform worse compared to encoder-based models such as FinBERT and RoBERTa. This is because decoder-based models focus on autoregressive tasks and emphasizes future context rather the past. Meanwhile, BERT based models look at both directions for context and focuses on understanding the text, making it the clear choice for sentiment classification.

The commodity correlation analysis further reveals that while entity-level sentiment is typically better than sequence-level sentiment at explaining asset prices, the performance is actually dependent on the asset type. For example, gold shows higher correlation with sequence-level sentiment, due to its representation of the overall market. However, when news sentiment and price are weakly correlated, as seen in oil, neither sequence-level nor entity-level sentiment classification performs better over the other.

Overall, conducting financial analysis with entity-level sentiment classification is actually dependent on the asset being analyzed. Factors

that impact the performance of entity-level news sentiment include how the asset is presented in media, how correlated with the market the asset is, and to what degree asset price is influenced by media.

6 Limitations

In our benchmarking on FinEntity, the biggest limitation is that we did not perform extensive hyperparameter tuning for the additional models introduced (e.g., DeBERTa, RoBERTa, Qwen, LLaMA) and instead reused the same settings applied to baseline models in the original FinEntity paper. This means our results may not reflect the full potential of newer models, especially those with different optimization characteristics or pretraining objectives.

With regards to our correlation analysis, we observe correlations between sentiment and commodity prices, but our analysis is descriptive and does not imply causality. Changes in sentiment may reflect price movements instead of driving them, especially in commodities influenced by macroeconomic or geopolitical events. Our analysis also assumes same-day alignment between sentiment and price, which may not account for lagged market reactions or anticipatory sentiment. More sophisticated time-series modeling (e.g., Granger causality, VAR) might be able to offer deeper insights.

7 Future Work

We plan to benchmark additional large language models such as Claude and Gemini in zero-shot, few-shot, and fine-tuned settings on FinEntity. Additionally, we aim to expand DeBERTa and RoBERTa models by pre-training them on large amounts of financial text. This will create versions of FinDeBERTa and FinRoBERTa that can then be used to train on FinEntity.

Instead of conducting case studies on cryptocurrency and stocks, we can evaluate entity-level sentiment classification's performance on other assets such as stocks, bonds, real-estate, and derivatives.

Additionally, the case study can be further expanded into trading applications. Using entity-level news sentiment, we could explore price prediction, P/L in portfolios, risk management, and more. This allows us to analyze entity-level sentiment classification's performance in real-world applications.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Jinze Bai et al. 2023. Qwen technical report.

Dan Benayoun. 2024. Processed dataset of financial news articles from bloomberg (2006-2013).

Jacob Devlin et al. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Richard Frankel et al. 2022. Disclosure sentiment: Machine learning vs. dictionary methods. *Manage. Sci.*, 68(7):5514–5532.

Pengcheng He et al. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

John Lafferty et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann Publishers Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. 2024. Gpt-4 technical report.

Xiao Ding Philippe Remy. 2015. Financial news dataset from bloomberg and reuters. `https://github.com/philipperemy/financial-news-dataset`.

David N Reshef et al. 2011. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Yixuan Tang et al. 2023. Finentity: Entity-level sentiment classification for financial texts.

Hugo Touvron et al. 2023. Llama: Open and efficient foundation language models.

Anita Yadav et al. 2020. Sentiment analysis of financial news using unsupervised approach. *Procedia Comput. Sci.*, 167:589–598.

A Additional Commodity Time-Series Graphs

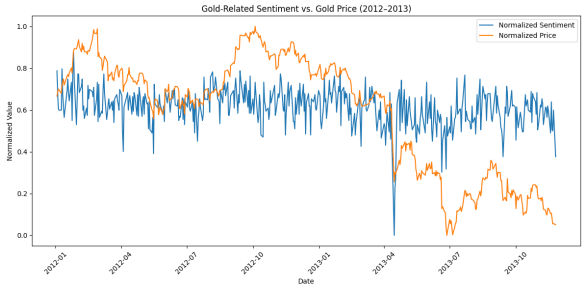

Figure 4: Normalized Sequence-level Sentiment (Orange) and Price (Blue) for Gold (2012-2013)

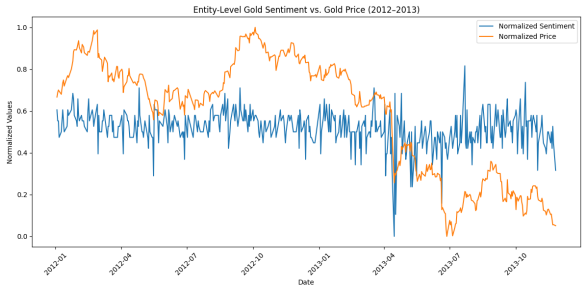

Figure 5: Normalized Entity-level Sentiment (Orange) and Price (Blue) for Gold (2012-2013)

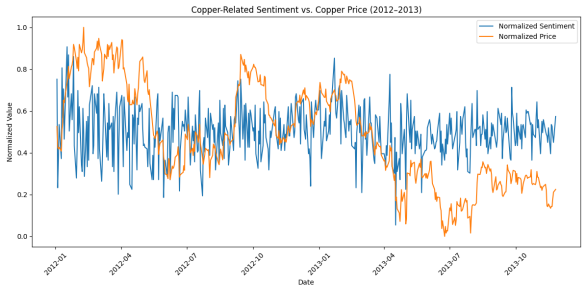

Figure 6: Normalized Sequence-level Sentiment (Orange) and Price (Blue) for Copper (2012-2013)

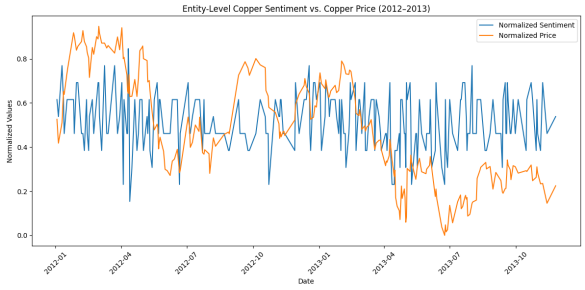

Figure 7: Normalized Entity-level Sentiment (Orange) and Price (Blue) for Copper (2012-2013)

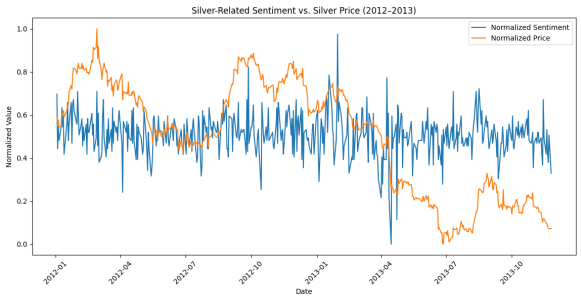

Figure 8: Normalized Sequence-level Sentiment (Orange) and Price (Blue) for Silver (2012-2013)

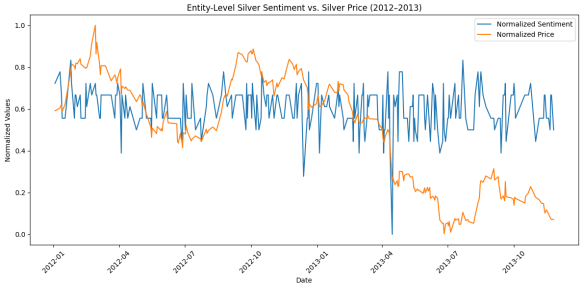

Figure 9: Normalized Entity-level Sentiment (Orange) and Price (Blue) for Silver (2012-2013)