Sentiment Analysis of Financial News vs. Data

Do headlines reflect reality?

James van Doorn **|** INST762 **|** May 20, 2025

Problem Statement

- People look to headlines for clues about the stock market.
- However, headlines can be misleading and reflect emotion more than reality.
  - In financial news, I notice a persistent “doom-and-gloom”.
- My goal is to compare headlines vs indicators.

https://static.vecteezy.com/system/resources/previews/018/918/170/original/stock-market-icon-vector.jpg

Data Sources

- FinSen Dataset: financial news
  - Aggregation of (~15,000) financial news articles:
    - Titles
    - Tags (144 unique)
    - Timestamps (2007-07-18 to 2023-07-16)
    - Content
- MarketWatch S&P 500 Daily Performance

Initial Project Scope

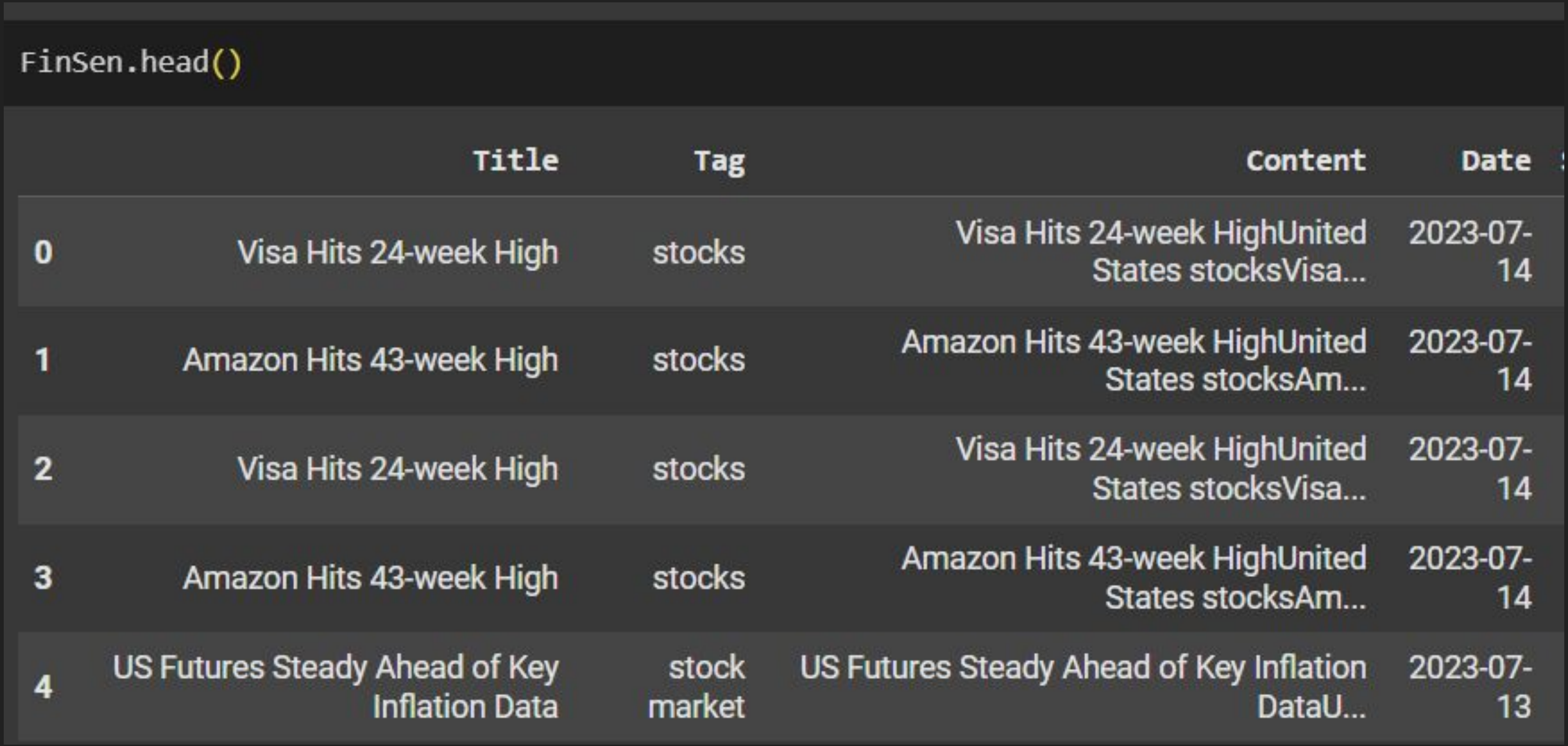
- Initially much larger: all of FinSen dataset article titles
- Attempted clustering to consolidate tags with similar meanings
    - Wanted to identify ideal indicator data to compare against

Tag Clustering Attempts

- Tried using these packages:
  - NLTK
  - Wordnet
  - Spacy
  - Sentence-Transformers
- Successfully found cosine similarity and distance, but struggled to apply them to the consolidation of tags

Revised Project Scope

- Setbacks with tag clustering forced me to narrow my scope to the stock market
- Stock-related articles still make up a plurality of FinSen articles (~4,500)
  - 8 unique tags containing the word “stock”
  - Ensures range and depth of data in analysis

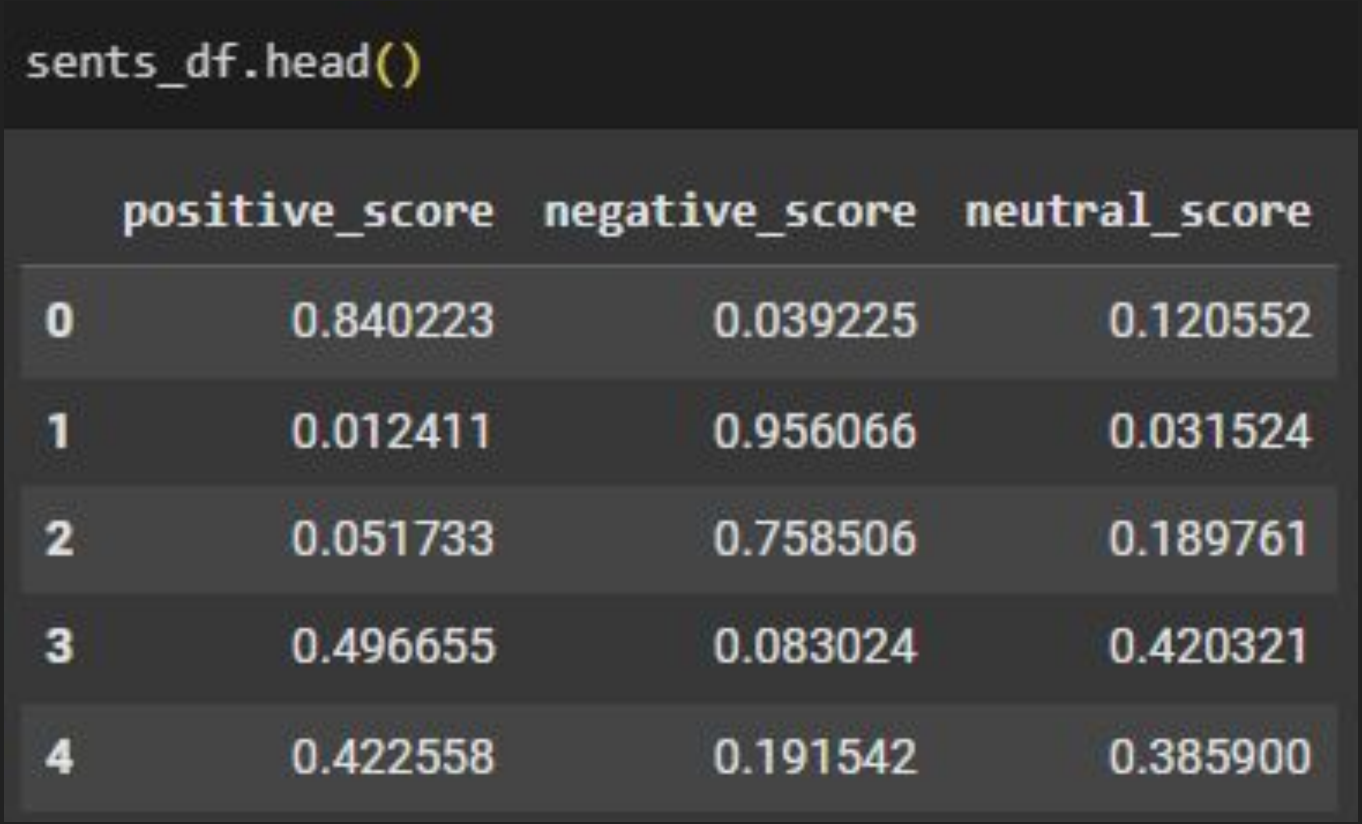


```
FinSen.head()
```

	Title	Tag	Content	Date
0	Visa Hits 24-week High	stocks	Visa Hits 24-week HighUnited States stocksVisa...	2023-07-14
1	Amazon Hits 43-week High	stocks	Amazon Hits 43-week HighUnited States stocksAm...	2023-07-14
2	Visa Hits 24-week High	stocks	Visa Hits 24-week HighUnited States stocksVisa...	2023-07-14
3	Amazon Hits 43-week High	stocks	Amazon Hits 43-week HighUnited States stocksAm...	2023-07-14
4	US Futures Steady Ahead of Key Inflation Data	stock market	US Futures Steady Ahead of Key Inflation DataU...	2023-07-13

Sentiment Analysis

- As people typically don’t read past the headline of an article, I decided to only perform sentiment analysis on the article titles.
  - Also saves time and computer resources
- Found FinBERT package trained on corpus of financial text
  - Simple loading in and generation



```
sents_df.head()
```

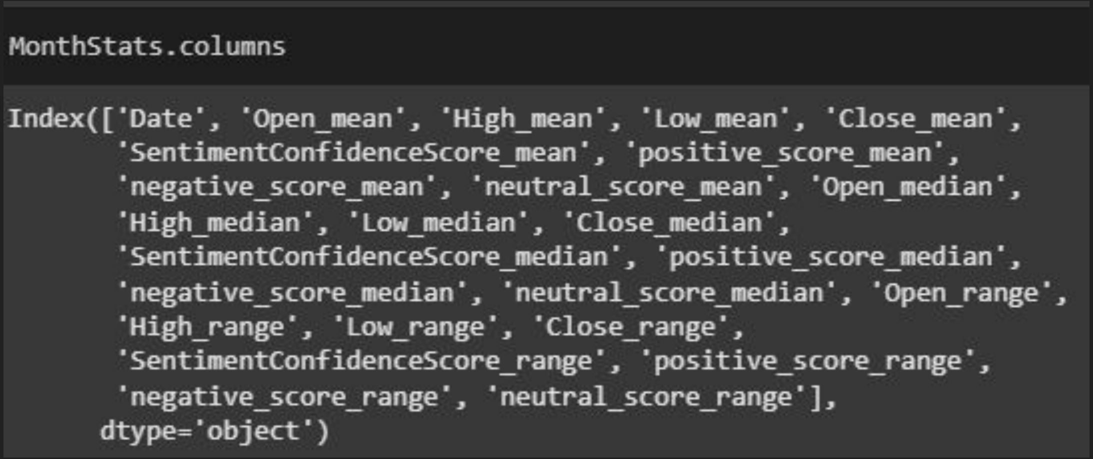
	positive_score	negative_score	neutral_score
0	0.840223	0.039225	0.120552
1	0.012411	0.956066	0.031524
2	0.051733	0.758506	0.189761
3	0.496655	0.083024	0.420321
4	0.422558	0.191542	0.385900

Comparison Pt. 1

- How can I compare sentiment scores and labels of stock news article titles to stock market data?
    - Stocks don’t trade on weekends, but people write articles on all days of the week
    - However, some weeks have no stock-related articles at all
        - Solution: Aggregate sentiment scores and stock data by month

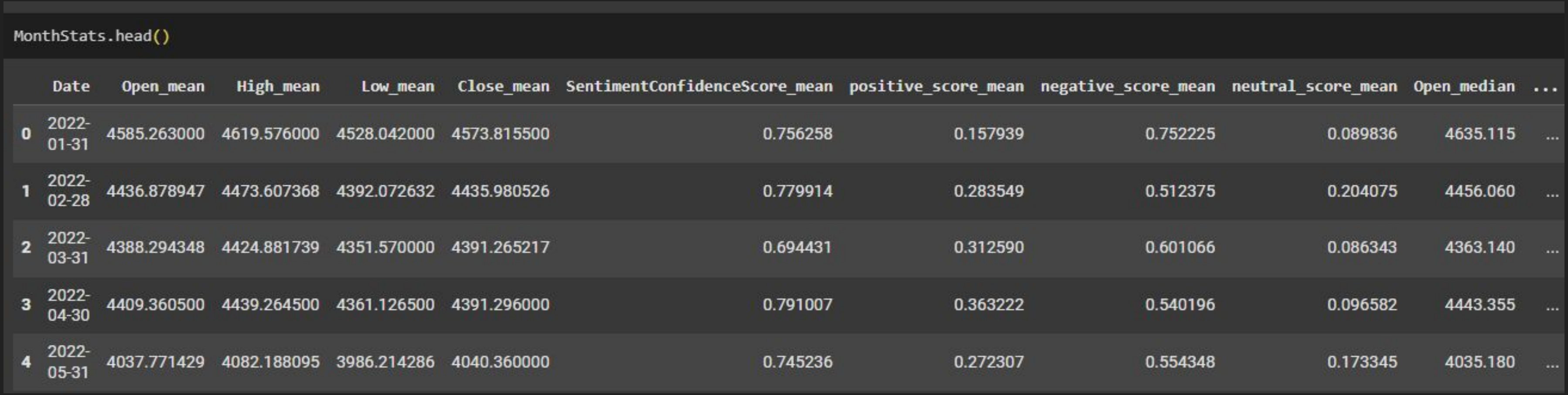
Comparison Pt. 2

- Aggregating leaves me with 19 rows of data to use, with each one an aggregate of a month’s (30 days starting at the listed date) data
  - Realize that open, high, and low stock data columns not useful for my purpose
    - Only relevant for intra-day stock trading



```
MonthStats.columns
```

```
Index(['Date', 'Open_mean', 'High_mean', 'Low_mean', 'Close_mean',
       'SentimentConfidenceScore_mean', 'positive_score_mean',
       'negative_score_mean', 'neutral_score_mean', 'Open_median',
       'High_median', 'Low_median', 'Close_median',
       'SentimentConfidenceScore_median', 'positive_score_median',
       'negative_score_median', 'neutral_score_median', 'Open_range',
       'High_range', 'Low_range', 'Close_range',
       'SentimentConfidenceScore_range', 'positive_score_range',
       'negative_score_range', 'neutral_score_range'],
      dtype='object')
```

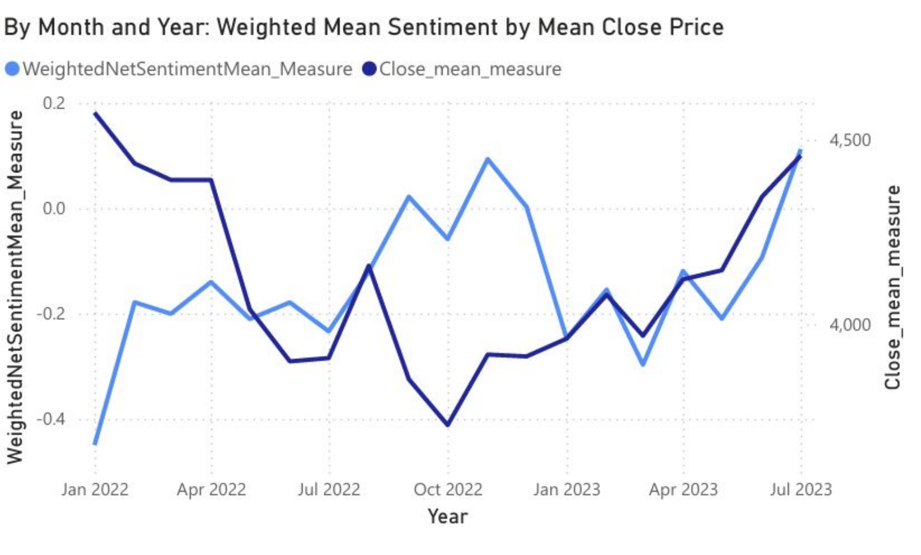


```
MonthStats.head()
```

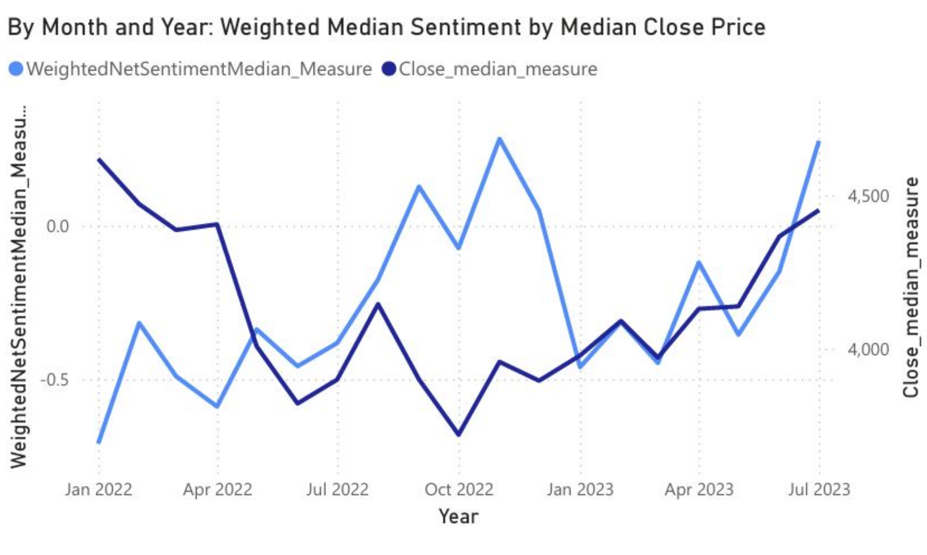
	Date	Open_mean	High_mean	Low_mean	Close_mean	SentimentConfidenceScore_mean	positive_score_mean	negative_score_mean	neutral_score_mean	Open_median	...
0	2022-01-31	4585.263000	4619.576000	4528.042000	4573.815500	0.756258	0.157939	0.752225	0.089836	4635.115	...
1	2022-02-28	4436.878947	4473.607368	4392.072632	4435.980526	0.779914	0.283549	0.512375	0.204075	4456.060	...
2	2022-03-31	4388.294348	4424.881739	4351.570000	4391.265217	0.694431	0.312590	0.601066	0.086343	4363.140	...
3	2022-04-30	4409.360500	4439.264500	4361.126500	4391.296000	0.791007	0.363222	0.540196	0.096582	4443.355	...
4	2022-05-31	4037.771429	4082.188095	3986.214286	4040.360000	0.745236	0.272307	0.554348	0.173345	4035.180	...

Comparison Pt.3:

- Turn on “Do Not Summarize” for every column in my dataset since each is already an aggregate
- Calculate weighted values based on sentiment confidence scores
    - Created initially as columns, not measures
        - Created additional challenges
        - Used “SELECTEDVALUE” to convert them all to measures

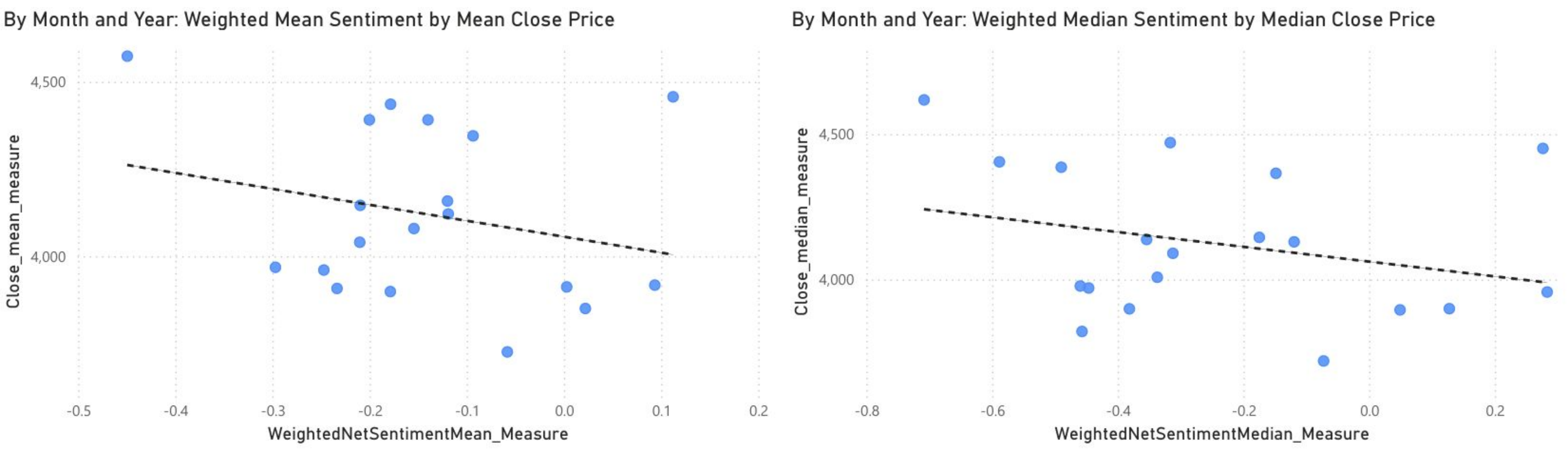


By Month and Year: Weighted Mean Sentiment by Mean Close Price



By Month and Year: Weighted Median Sentiment by Median Close Price

- Few months pulling average down
- Lag observed: sentiment increase precedes close price rise



By Month and Year: Weighted Mean Sentiment by Mean Close Price

By Month and Year: Weighted Median Sentiment by Median Close Price

- Slightly negative correlation between net sentiment of titles and close prices

Conclusion

- The sentiment of financial news articles has a very weak correlation with stock data
    - Most you could say is an increase in sentiment often precedes in an increase in the stock market
- Initial hypothesis basically unprovable
    - Comparing prices with sentiments is like comparing apples to oranges

Sources

Bert¶. BERT - transformers 3.0.2 documentation. (n.d.). https://huggingface.co/transformers/v3.0.2/model_doc/bert.html

EagleAdelaide. (n.d.). EagleAdelaide/finsen_dataset. GitHub. https://github.com/EagleAdelaide/FinSen_Dataset

Finbert - documentation quantconnect.com. FinBERT - QuantConnect.com. (n.d.). https://www.quantconnect.com/docs/v2/writing-algorithms/machine-learning/hugging-face/popular-models/finbert

Market activity U.S. market activity. MarketWatch. (n.d.). Download SPX Data | S&P 500 Index Price Data | MarketWatch

Prosusai/finbert · hugging face. ProsusAI/finbert · Hugging Face. (n.d.). https://huggingface.co/ProsusAI/finbert