Leveraging NLP to Extract Casual Events & Derive Investment Signal(s) from News Articles

Team 2

George Lyu, Samantha Armijo, Jiayu Shi, Sruthi Pappu, and Franck Brych

VANDERBILT UNIVERSITY

Agenda

Data Labelling

Model 1

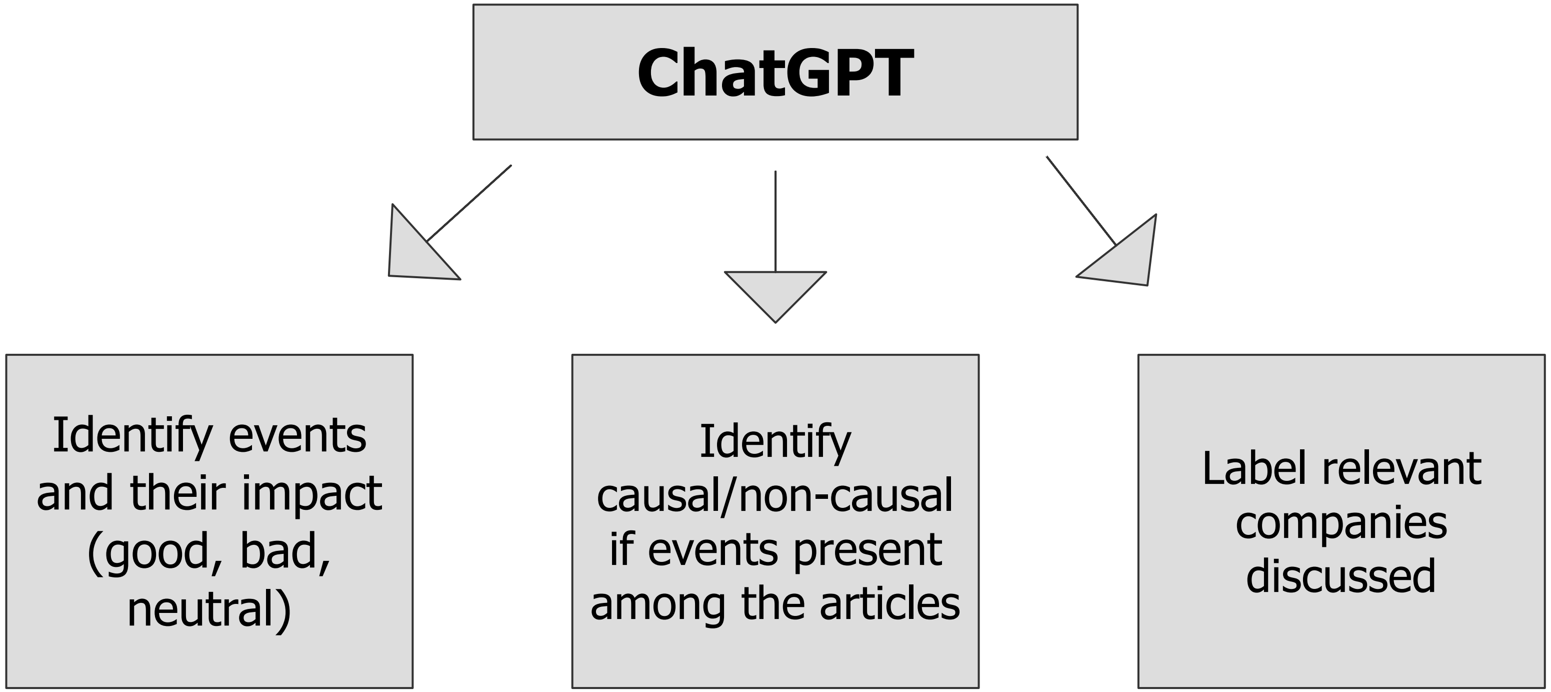
Model 2

AlphaLens / Portfolio

Next Steps: Graph-Based Representation & Analysis

Data Labelling

For each article:



Data Labelling

Gave ChatGPT entire dataset for all labelling.

Resulted in a lack of variability in the labelled data.

Provided ChatGPT with the entire dataset, tasking it to identify one event at a time before proceeding to the next.

Still fell short of variability.

Implemented few-shot learning. Gave entire dataset to ChatGPT to identify one event and impact at a time. Next, tasked to distinguish causal and non-causal relationships among the articles. Finally, label relevant companies for each article.

This approach significantly increased the variability of data in the identified events.

Data Labelling

Event	Count of Presence
Defaults	165
Mergers & Acquisitions	269
Revenue Discussions	1276
Margin/Profitability Discussions	379
Industry Competition	622

Agenda

Data Labelling

Model 1

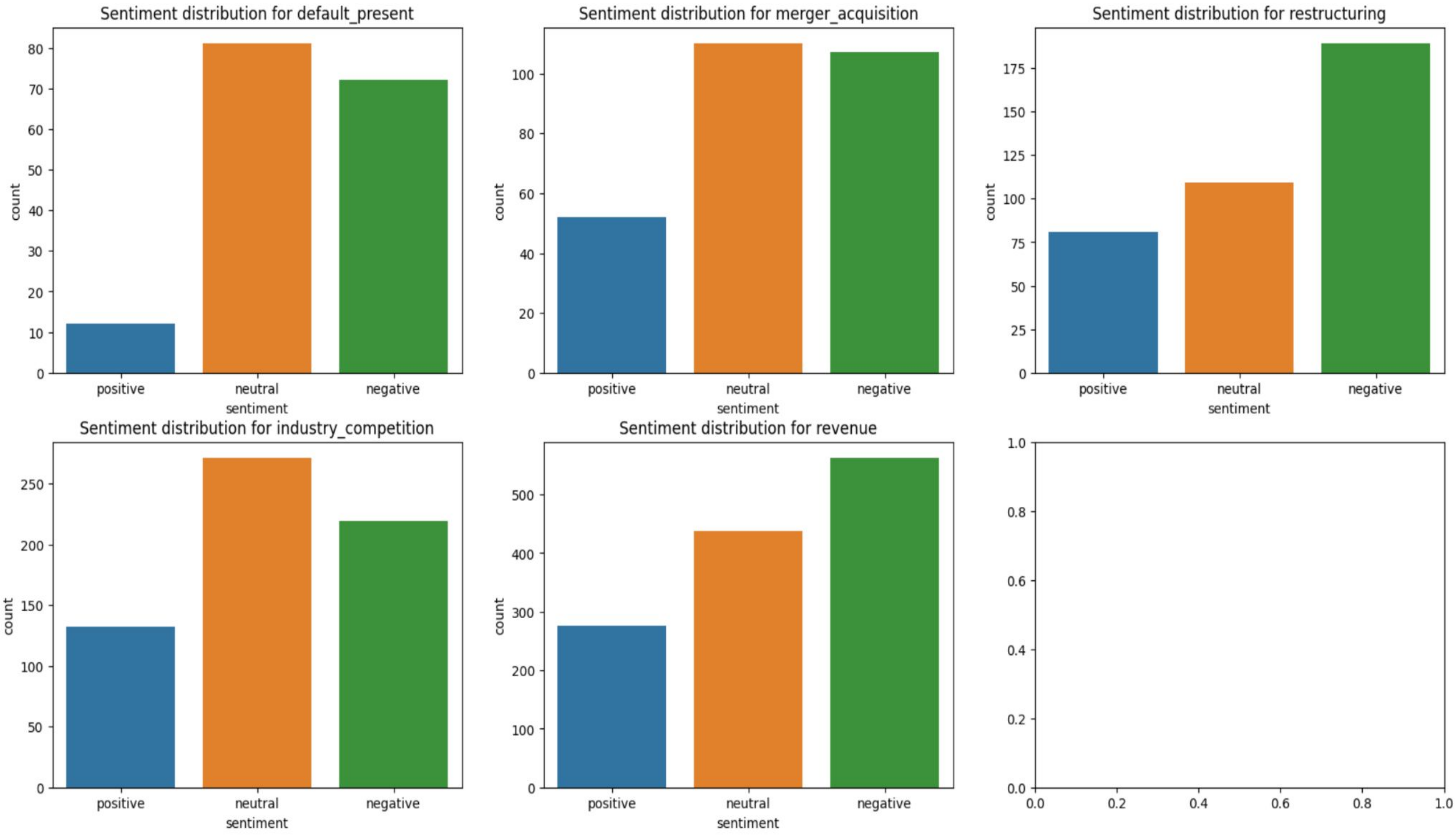
Model 2

AlphaLens / Portfolio

Next Steps: Graph-Based Representation & Analysis

DistilRoberta for Financial News

- Fine tuned version of Distil RoBERTa-base.
- On average Distil RoBERTa is twice as fast as Roberta-base.
- The training dataset consists of 4840 sentences from English language financial news categorised by sentiment.
- The model loss of 0.11 and an accuracy of over 98%.
- We split the data based on the article to get a better idea on the sentiment and the sentiment probability for each.



Sentiment distribution for default_present
Sentiment distribution for merger_acquisition
Sentiment distribution for restructuring
Sentiment distribution for industry_competition
Sentiment distribution for revenue
count
sentiment
positive
neutral
negative

Agenda

Data Labelling

Model 1

Model 2

AlphaLens / Portfolio

Next Steps: Graph-Based Representation & Analysis

FinBert

- Fine tuned version of bert-base-uncased on the dataset.
- The training dataset consists of 4840 sentences from English language financial news categorised by sentiment.
- When we ran it on the whole dataset, the model has an accuracy of about 87%.
- We split the data based on the impact type of the article to get the probability of the sentiment of each article and check the accuracy of each impact type.
- On the split data, the model had an average accuracy of 97% and an average loss of .45.

Agenda

Data Labelling

Model 1

Model 2

AlphaLens / Portfolio

Next Steps: Graph-Based Representation & Analysis

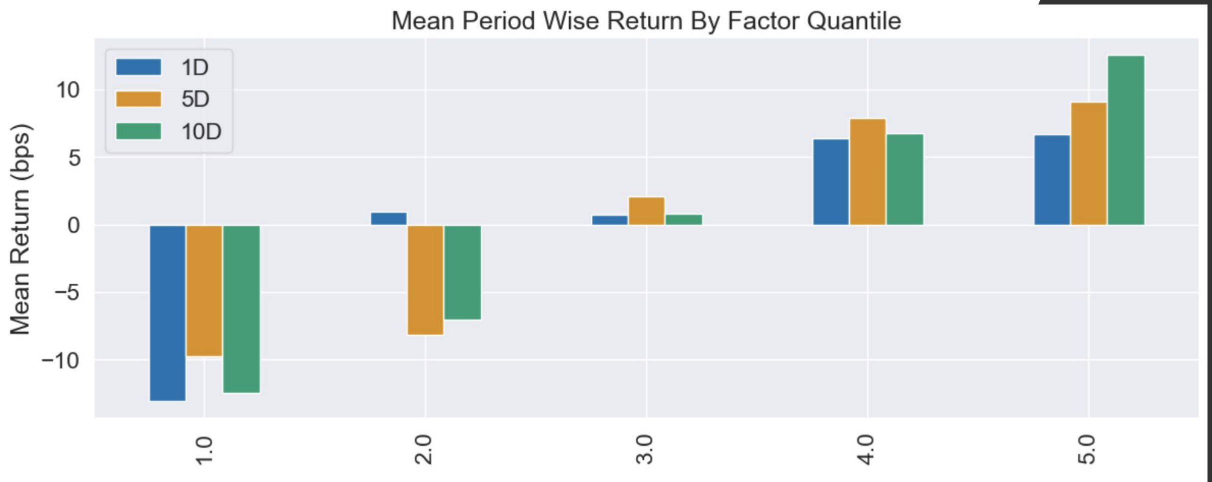
AlphaLens / Portfolio

Finbert Model **Default** Event:

1. **Positive Alpha** value for three quarters
2. Annual Return: **+6.14%** Return

Returns Analysis

	1D	5D	10D
Ann. alpha	1.612	1.366	1.438
beta	0.009	0.009	0.325
Mean Period Wise Return Top Quantile (bps)	6.652	9.073	12.559
Mean Period Wise Return Bottom Quantile (bps)	-13.059	-9.723	-12.435
Mean Period Wise Spread (bps)	19.711	18.636	24.573



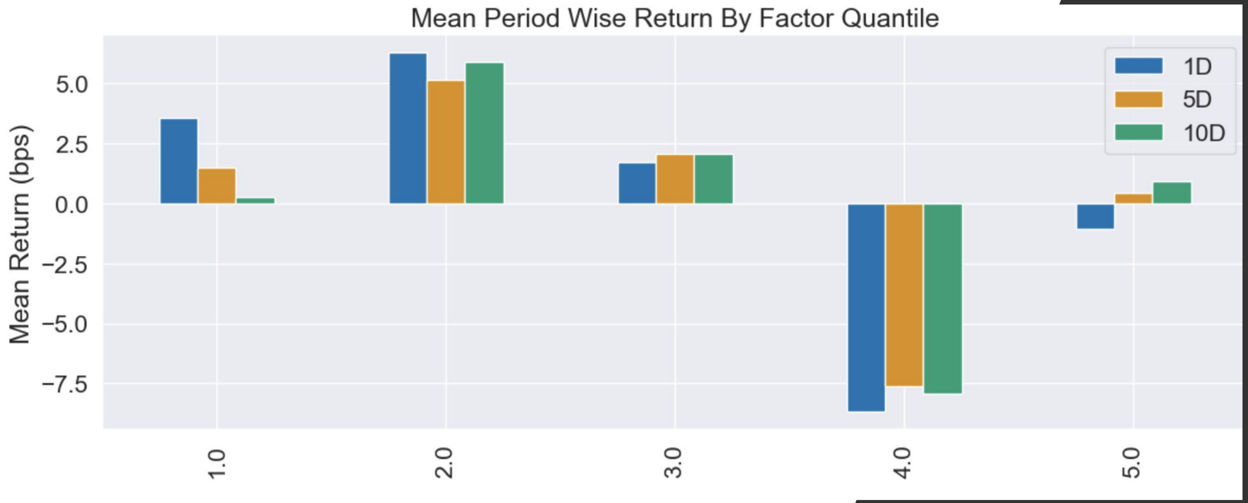
AlphaLens / Portfolio

Finbert Model **Mergers Acquisition** Event:

1. **Negative Return** on Quarter 4
2. Annual Return: **+0.125%** Return

Returns Analysis

	1D	5D	10D
Ann. alpha	-0.123	0.008	-0.027
beta	0.121	0.001	-0.065
Mean Period Wise Return Top Quantile (bps)	-1.055	0.445	0.930
Mean Period Wise Return Bottom Quantile (bps)	3.568	1.490	0.262
Mean Period Wise Spread (bps)	-4.623	-1.116	0.499



VANDERBILT UNIVERSITY

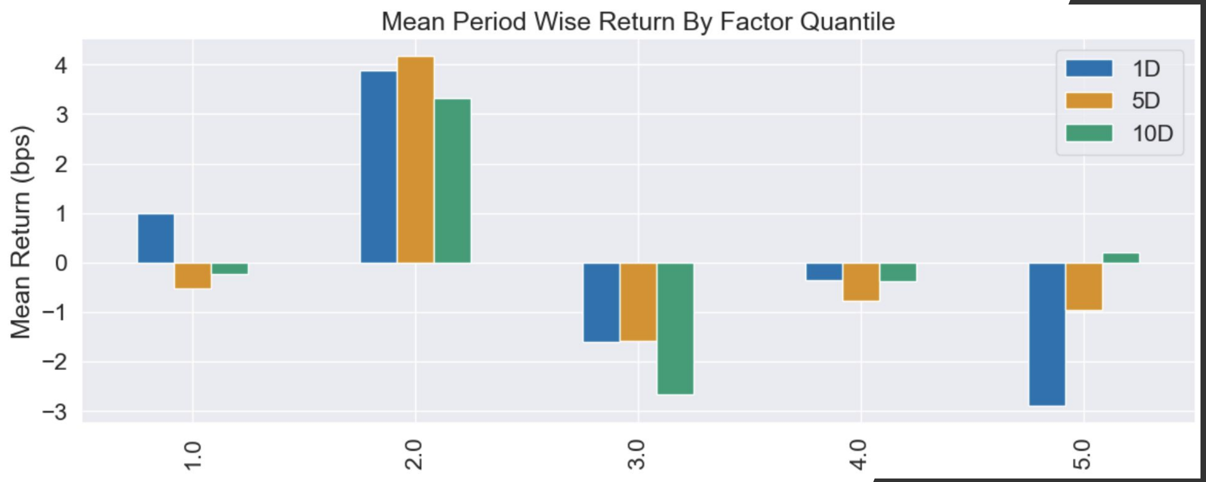
AlphaLens / Portfolio

Finbert Model **Industry Competition** Event:

1. **Negative Alpha** values
2. Annual Return: **+0.0098%** Return

Returns Analysis

	1D	5D	10D
Ann. alpha	-0.497	-0.250	-0.124
beta	0.536	0.398	0.259
Mean Period Wise Return Top Quantile (bps)	-2.892	-0.963	0.209
Mean Period Wise Return Bottom Quantile (bps)	1.001	-0.516	-0.235
Mean Period Wise Spread (bps)	-3.893	-0.467	0.392



VANDERBILT UNIVERSITY

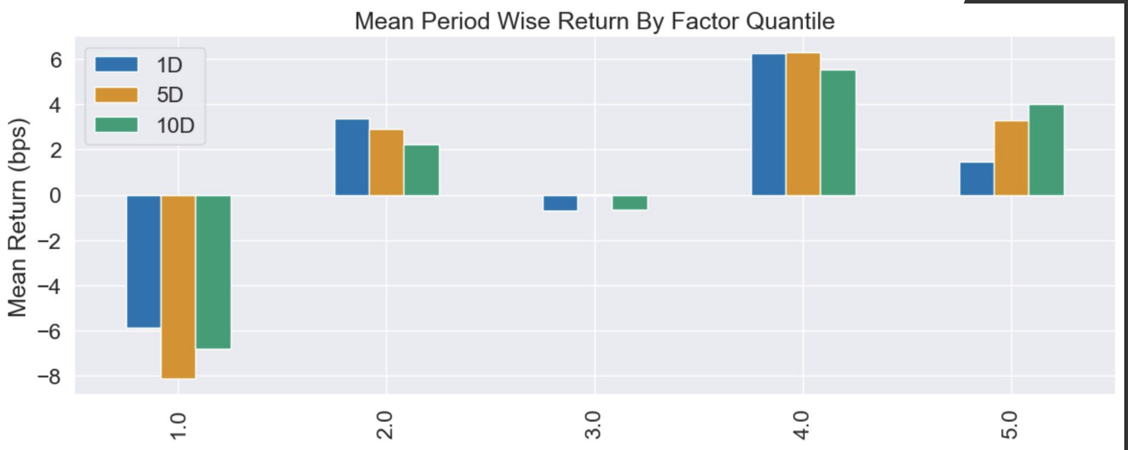
AlphaLens / Portfolio

Finbert Model **Margin Profitability** Event:

1. **Positive Alpha** value for 10D period
2. Annual Return: **+2.7%** Return

Returns Analysis

	1D	5D	10D
Ann. alpha	-0.089	-0.054	0.070
beta	-0.319	-0.142	-0.117
Mean Period Wise Return Top Quantile (bps)	1.464	3.297	4.000
Mean Period Wise Return Bottom Quantile (bps)	-5.877	-8.129	-6.797
Mean Period Wise Spread (bps)	7.341	11.494	10.831



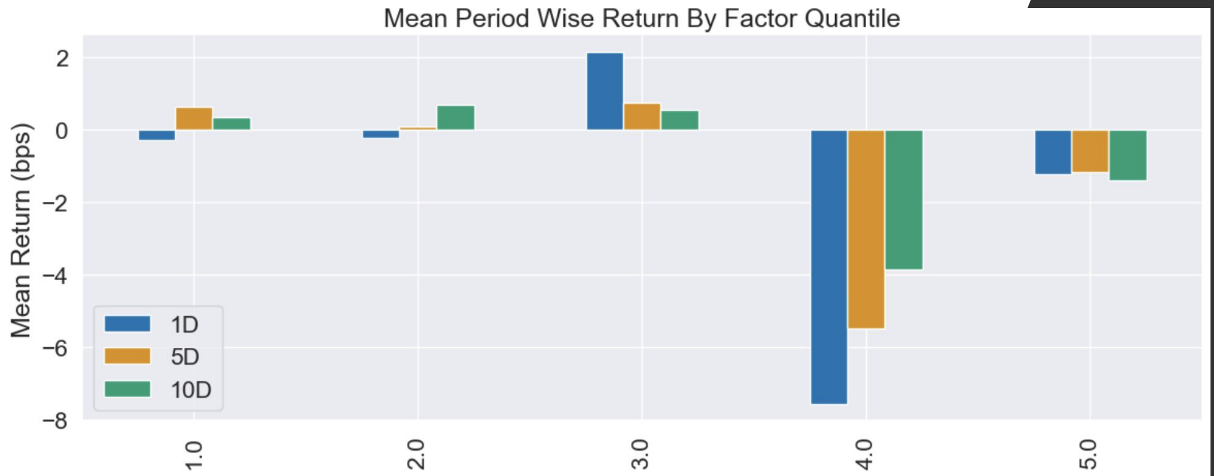
AlphaLens / Portfolio

Finbert Model **Revenue** Event:

1. **Negative Alpha** values
2. Annual Return: **-0.5%** Return

Returns Analysis

	1D	5D	10D
Ann. alpha	-0.035	-0.130	-0.060
beta	0.012	0.073	0.091
Mean Period Wise Return Top Quantile (bps)	-1.240	-1.176	-1.397
Mean Period Wise Return Bottom Quantile (bps)	-0.286	0.622	0.328
Mean Period Wise Spread (bps)	-0.954	-1.748	-1.638



Agenda

Data Labelling

Model 1

Model 2

AlphaLens / Portfolio

Next Steps: Graph-Based Representation & Analysis

Next Steps: Graph-Based Representation & Analysis

Construct a directed graph with event entities as nodes and causal links as directed edges, where the edge direction indicates the flow or order of causality.

Assign weights to edges based on the confidence of the causality, with stronger connections having higher weights.

Construct a directed graph where events and companies are represented as two separate nodes, and the impact of events on companies is depicted as directed edges.

Assign weights to the edges based on the confidence of the impact of the events. Stronger connections, indicating more significant impacts, will be assigned higher weights

Analyze graph properties such as centrality measures, clustering coefficients.

Employ link prediction techniques to forecast potential impacts on companies, identifying new and influential relationships

Thank you!

Questions.