AS11 Report Naman Kedia

Zero is Not Hero - Paper Analysis

- ***Explain zero shot***

  Zero Shot Classification lets a model make predictions about new categories it has never seen before during training. It's like teaching someone general knowledge about language and then asking them to solve a specific problem they haven't practised - for example, a LLM trained on general language can understand and classify FOMC statements as hawkish or dovish even if it never specifically learned about monetary policy. This is helpful when we don't have lots of labelled examples to train with.
- ***Paper Summary***

  Recent AI models like ChatGPT can understand and analyse text without specific training (zero-shot) or minimal training (few-shot). The paper challenges the idea of whether zero-shot LLM inference is a strong enough technique to work with financial texts compared to RoBERTa, which was specifically trained on financial data.

  To break the data analysis and text interpretation there are some methods deployed like sentiment analysis, named entity recognition etc. which were deemed important from a stock market impacting point of view.

  The results concluded that ChatGPT performed well without training, but the paper showed that specially trained models still performed better, even though using these big AI models took more time.

Selecting examples in few-shot inference for optimal performance and tradeoffs

In this assignment, I explored two different scenarios for classifying statements based on their hawkishness. In the first scenario, I selected one example each for Hawkish, Dovish, and Neutral categories. In the second scenario, I sampled two examples each from the Hawkish and Dovish categories. These variations were designed to evaluate two key factors: (1) whether increasing the number of examples improves classification performance and (2) whether diversity in the examples (from different categories) is important for accurate classification.

I saw slightly better results in case 1 where there was homogeneity across label datasets.

To optimise the few-shot prompt, I could include a more balanced mix of examples across all classes (Hawkish, Dovish, and Neutral) to ensure better model generalisation. Adding more diverse examples and clarifying instructions would improve accuracy but may increase token usage. The tradeoff could be between clarity and prompt/token length leading to unnecessary overfitting.

Performance and Latency Analysis

Document insights about trade-offs between model size, performance, and latency.

Model	F1 Score	Precision	Recall	Accuracy	Latency	Valid Predictions
Zero Shot	**0.647**	**0.682**	**0.630**	**0.651**	**1.72**	**212/214**
Few Shot (1 each)	**0.682**	**0.676**	**0.69**	**0.678**	**1.73**	**214/214**
Few Shot (2 hawkish, 2 dovish example)	**0.66**	**0.71**	**0.65**	**0.66**	**.63**	**214/214**
Fomc - Roberta	**0.874**	**0.87**	**0.88**	**0.87**	**.63**	**214/214**

*(Latency includes network latency for LLM'S too)

Based on this table comparing zero-shot, few-shot, and RoBERTa models for FOMC statement classification, here are the key insights about trade-offs:

- Smaller, fine-tuned models like FOMC-RoBERTa outperform larger models like LLaMA 3 in both performance (F1 Score: 0.874 vs. 0.647-0.682) and latency (0.63s vs. 1.72-1.73s).
- Few shot models do perform slightly better than zero short model but also have longer token length and potentially more latency

- Despite the potential benefits of few-shot learning, adding more examples can increase latency without significantly improving performance.
- Additionally, fine-tuning for task-specific goals offers better prediction reliability and lower costs compared to general-purpose models.

This makes FOMC-RoBERTa the optimal choice for specialized tasks, especially one like this one.

Trading Strategy Based on Hawkish-Dovish Measure

In this strategy, the trading decision is driven by the hawkishness or dovishness of monetary policy, measured by a hawkish score for the previous year. The key steps in the strategy are outlined below:

1. **Hawkish Period**
    a. During hawkish periods, monetary tightening is expected, and the strategy aims to take advantage of more stable, low-risk sectors.
    b. The *Finance, Banking and Real Estate* sectors are chosen, as these industries tend to have more stability in volatile markets. The strategy selects 10 of the top stocks by marketCap from this sector, specifically targeting those with the most stable avg beta in previous year (i.e., a beta close to 0), indicating lower risk and less volatility relative to the market.

2. **Dovish Period**
    a. In dovish periods, monetary easing is expected, which typically benefits growth sectors.
    b. The *Services* and *Manufacturing* sectors are targeted, as they tend to benefit from easier monetary conditions and economic stimulus. The strategy selects 10 of the top stocks by marketCap from this sector, specifically targeting those with high previous returns and positive betas, indicating growth potential and potential strong performance in the upcoming period.

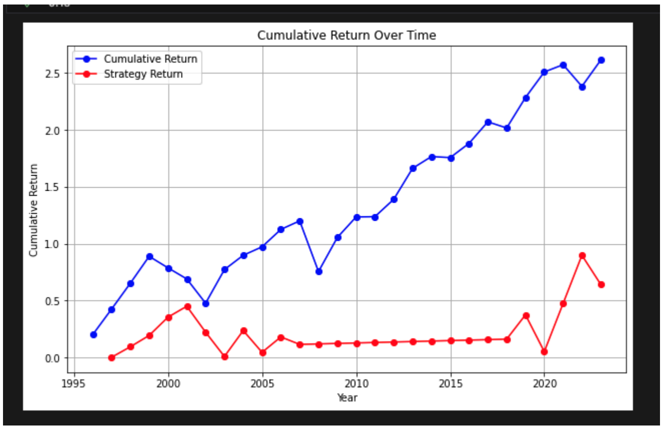
Insights from Comparing Cumulative Returns under Different Monetary Policy Stances

The trading strategy based on binary hawkish/dovish signals significantly underperformed the market, with cumulative returns of approximately 75% compared to the market's 250%. Although the strategy exhibited lower volatility, especially between 2005 and 2015, its simplistic reliance on hawkish and dovish signals may not have been adequate for capturing the complexity of market movements. This suggests that the strategy could benefit from more nuanced signal processing or the inclusion of additional factors to better account for market dynamics.

Reliability of Using LLM-based Classification for Financial Forecasting

Using LLM-based classification for financial forecasting may be reliable for broad trend identification, but for my strategy, it seems too simplistic to fully capture the intricacies of market behavior. The classification approach, while offering some insights, failed to improve the strategy's performance significantly, highlighting the potential need for more advanced forecasting methods or integration of other better nuanced strategies.

Market vs Strategy Cumulative Return

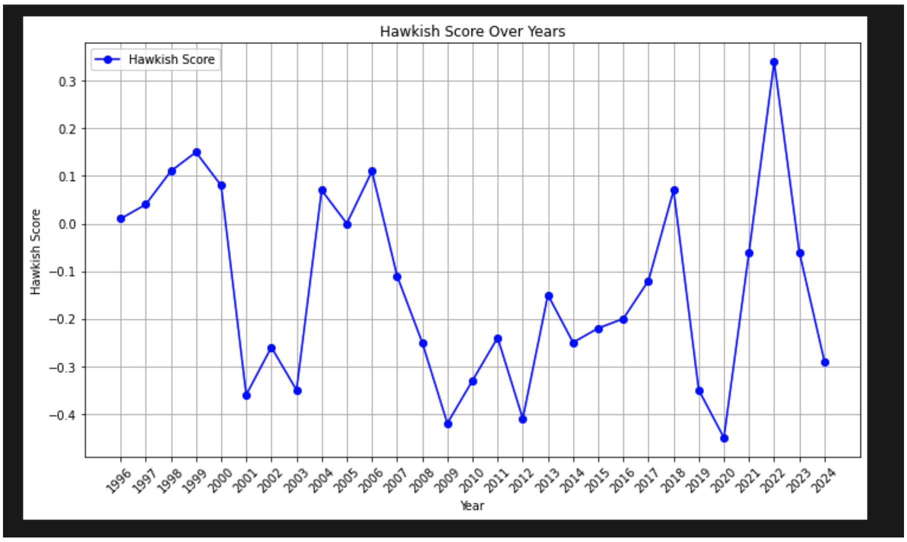


***************** End Of Main Report *********************

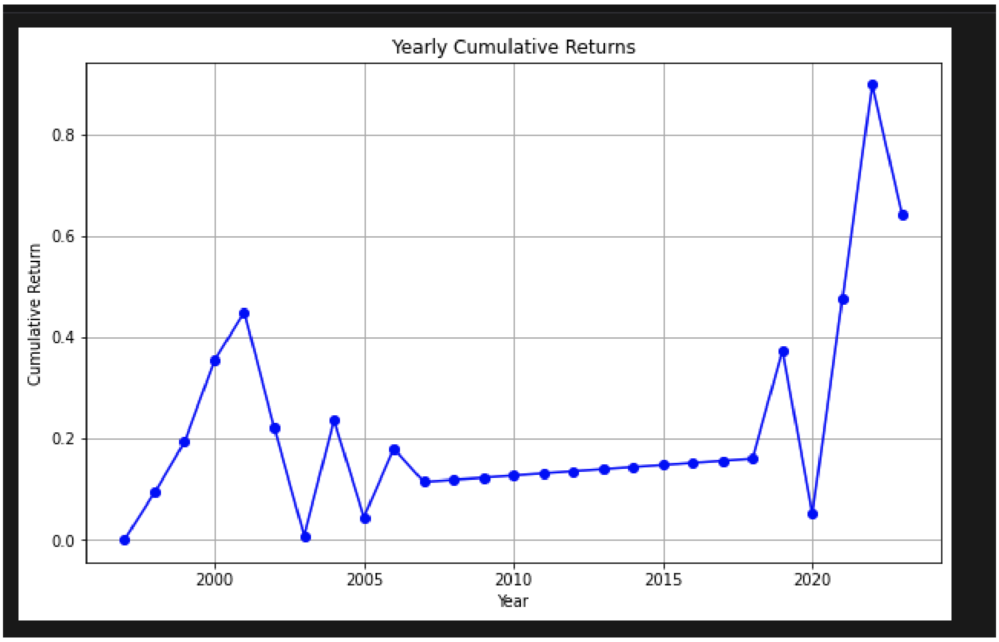
AS11 Report Naman Kedia

Plots and Other Metric Graphs

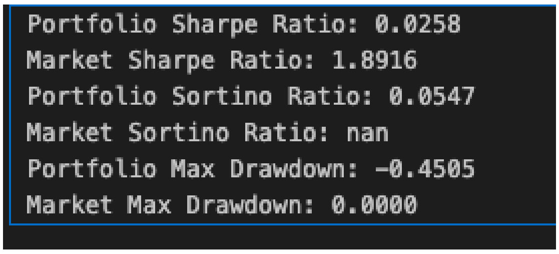
Hawkish Measure



Returns of Custom Trading Strategy



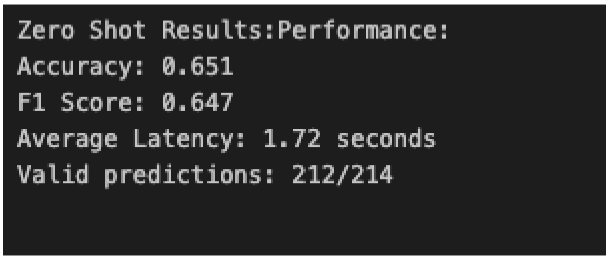
Other metrics of Custom Trading Strategy



```
Portfolio Sharpe Ratio: 0.0258
Market Sharpe Ratio: 1.8916
Portfolio Sortino Ratio: 0.0547
Market Sortino Ratio: nan
Portfolio Max Drawdown: -0.4505
Market Max Drawdown: 0.0000
```

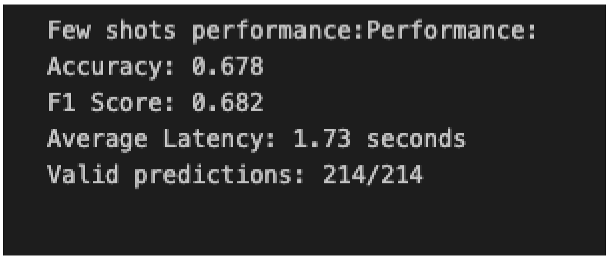
AS11 Report Naman Kedia

Zero Shot Metrics



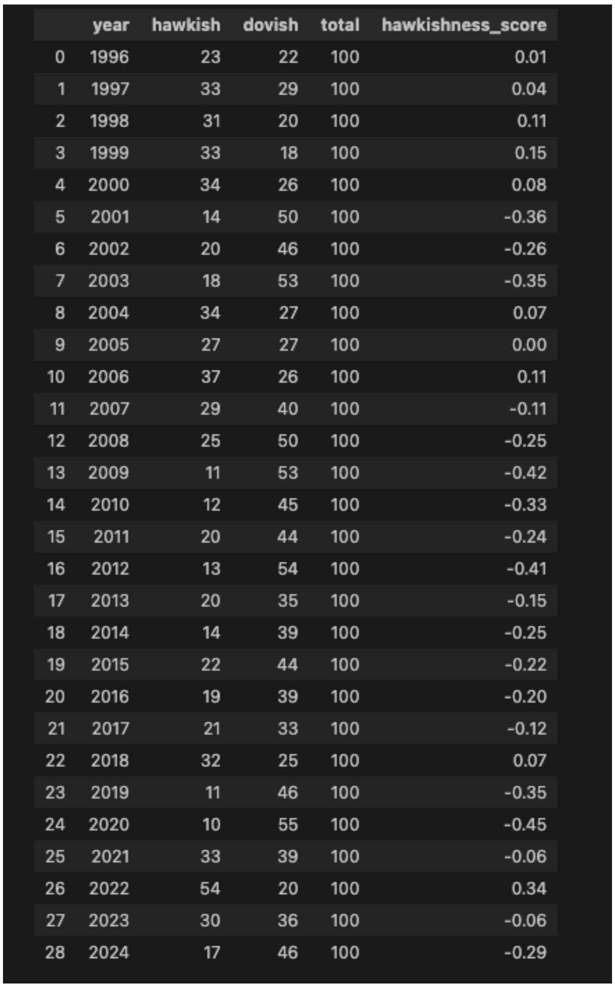
```
Zero Shot Results:Performance:
Accuracy: 0.651
F1 Score: 0.647
Average Latency: 1.72 seconds
Valid predictions: 212/214
```

Few Shots Metric



```
Few shots performance:Performance:
Accuracy: 0.678
F1 Score: 0.682
Average Latency: 1.73 seconds
Valid predictions: 214/214
```

Hawkish Score Result



	year	hawkish	dovish	total	hawkishness_score
0	1996	23	22	100	0.01
1	1997	33	29	100	0.04
2	1998	31	20	100	0.11
3	1999	33	18	100	0.15
4	2000	34	26	100	0.08
5	2001	14	50	100	-0.36
6	2002	20	46	100	-0.26
7	2003	18	53	100	-0.35
8	2004	34	27	100	0.07
9	2005	27	27	100	0.00
10	2006	37	26	100	0.11
11	2007	29	40	100	-0.11
12	2008	25	50	100	-0.25
13	2009	11	53	100	-0.42
14	2010	12	45	100	-0.33
15	2011	20	44	100	-0.24
16	2012	13	54	100	-0.41
17	2013	20	35	100	-0.15
18	2014	14	39	100	-0.25
19	2015	22	44	100	-0.22
20	2016	19	39	100	-0.20
21	2017	21	33	100	-0.12
22	2018	32	25	100	0.07
23	2019	11	46	100	-0.35
24	2020	10	55	100	-0.45
25	2021	33	39	100	-0.06
26	2022	54	20	100	0.34
27	2023	30	36	100	-0.06
28	2024	17	46	100	-0.29

References:

- **ChatGPT used for some coding components for eg. computing trading strategy ratios and metrics**