Cross-Sector Market Regime Forecasting with LLM-Augmented News Analysis

Timur Mudarisov
University of Luxembourg
Luxembourg
timur.mudarisov@uni.lu

Radu Valentin State*
University of Luxembourg
Luxembourg
radu.state@uni.lu

Zsofia Kraussl
City, University of London
Great Britain
zsofia.kraussl@uni.lu

Alexander Yakubov
University of Luxembourg
Luxembourg
alexander.yakubov@uni.lu

Tatiana Petrova
University of Luxembourg
Luxembourg
tatiana.petrova@uni.lu

Abstract

This paper investigates the utilization of news in predicting market regimes. The findings illustrate that employing an ensemble of multiple FinBERT models can outperform straightforward time-series prediction by 73% in accuracy and 110% in F1 score. The NLP models demonstrate strong performance across two different market-regime scenarios and show the ability to detect market shifts.

CCS Concepts

• **Applied computing** → **Forecasting**; • **Mathematics of computing** → **Markov processes**; • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Web mining**;

Keywords

Large Language Models, market-regimes, efficient market hypothesis

ACM Reference Format:
Timur Mudarisov, Radu Valentin State, Zsofia Kraussl, Alexander Yakubov, and Tatiana Petrova. 2024. Cross-Sector Market Regime Forecasting with LLM-Augmented News Analysis. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3677052.3698642

1 Introduction

The concept of market regimes is crucial in understanding market data. Market regimes reflect the impact of previous economic events and serve as important signals for traders. While some researchers believe that history repeats itself [3, 4, 8], it is also important to note that exogenous events can significantly influence the market [24]. As a recent example to underline this claim, the COVID-19 pandemic profoundly impacted many market sectors, leading to market inefficiency and to liquidity concerns of different asset classes [23, 26].

The ability to forecast upcoming market regimes, and to estimate their impact on market performance, is thus one of the crucial tasks for an investor. The problem of accurate forecasting is also a well-known and discussed research area in Finance. Scholarly articles have been emerging to address the phenomenon and to provide topical evidence to improve forecasting. The majority of these articles base their argumentation and assessment on the Efficient Market Hypothesis (EMH), as the universal, fundamentally agreed law of financial markets [22]. The most popular trend to model the underlying market dynamics instrumentalizes Markov chains [12, 15], implying that market-regime switching follows a Markov process.

Strict modeling assumptions, however, can imply inefficiency concerns, as many scholarly articles agree (see e.g. [13, 20, 21]). No wonder, therefore, that academic research in Finance has been embracing behavioral considerations for financial decision-making, and used behavioral explanations to improve modeling accuracy of market performance [14]. Investors' risk taking behavior influences how they make financial decisions, how they leverage on arbitrage opportunities, and consequently, how asset allocation among less and more risk-averse investors evolves. Furthermore, as a result of globalized financial markets and the digital transformation of interactions, it is no longer the individual decision-making that requires attention, but the effects of decision-making in relation to others. Perceptions, and the diffusion of signals in a broader societal context receives higher and higher attention in financial forecasting [9, 10, 25].

In line with the emergence of advanced analytical techniques that rely on artificial intelligence, the analytical design to address the problem of efficient forecasting started embracing Natural Language Processing (NLP) methods. Just as recommended by [14], NLP empowers researchers to address the societal implications of financial decision-making. It is capable of addressing perceptions that are created by externalities. In this paper, we treat news as such externalities. In fact, the market effects of news and of the perceptions generated by news have been studied before (see e.g. [19]). It is the development and diffusion of advanced analytical techniques that allow researchers to refine their modeling approach and design and to improve the efficiency of market forecasting.

*All authors contributed equally to this research.

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698642

Large Language Models (LLMs), such as OpenAI's GPT-4, have gained significant popularity in recent years due to their impressive ability to generate human-like text and understand context in a wide variety of applications. These models are being used in numerous fields, including customer service, content creation, and data analysis, highlighting their versatility and potential to transform industries. The usage of such models has already been widely studied [16]. The development of Machine Learning (ML) methods allows researchers to use models on big data and design models that distinguish different market regimes. Our paper adds to this line of academic research and examines the ensemble of several FinBERT models [5] for classifying natural deterministic market-regime scenarios. The above-described problem context in financial forecasting motivates our primary research idea: news reflects the asset price movements. As asset price movements reach a certain homogeneous dynamic, they create, as we define in this paper, market shifts, leading to the change of a market regime. In other words, we hypothesize that news carry signals that change pricing dynamics, forecasting, and therefore, cause market shifts. We also hypothesize that these changing pricing dynamics follow similar, cross-sectoral patterns across different financial markets. In other words, signals, and the perception of these signals carry consecutive effects across different asset classes, changing therefore the pricing dynamics in a homogeneous manner, leading to the change of a market regime. As a consequence, we present in this paper a model that forecasts market shifts. We utilize for our model-building news that are obtained from several financial news sources, such as Bloomberg [1] or CNN [2].

The rest of this paper is structured as follows. The next section describes the problem statement. Section 3 discusses the model design. Section 4 describes the experiment that we constructed to test our model. Section 5 discusses our results. Section 6 concludes.

2 Problem statement

The concept of the market regime was described in [17], where the author examined the model with hidden parameter S_t. In that paper, S_t is a stochastic process defined on probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$ with values in $[0, \dots, S-1]$. The author examined the case where $\mathbb{P}(S_t = i) = \pi_i$ are unknowns and defined a model to determine them. Generally speaking, and in line with our earlier described motivation, the market regime's process is extremely unpredictable, thus models that rely on constant probabilities can generate inefficient, potentially misleading results.

In this paper, we use the following description for the market regime process. Consider R_t and S_t as stochastic processes defined on the same probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_t, \mathbb{P})$, where R_t is observable monthly return and S_t is a hidden market-regime jump process. We assume that the following hypothesis holds:

Assumption 1. $R_{t'} \sim f_i(x)$ for $t' \in \{t : S_t = i\}$,

where $f_i(x)$ are the probability density functions. In other words, the return process follows its own distribution in each market regime.

[1] https://www.bloomberg.com/news
[2] https://edition.cnn.com/business/economy

Our primary aim is to develop the model, which estimates the value of S_{t+1}. Unfortunately, process S_t is extremely complex, so we use the deterministic estimation $\hat{S}_t$ of it, i.e., this is the partition of the R_t on timestamps, satisfying the Assumption 1. To develop a model that forecasts market regimes, we have decided to follow a two-step procedure:

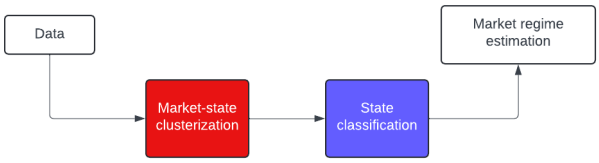

Figure 1: Model overview

First, we will analyze the time-series data and apply clusterization to label the series. Afterward, we will use a classification model to predict the labels generated from the clustering process. Our ultimate objective with this model is to minimize the cross-entropy loss (see Section 3.2).

As a consequence, to develop our model, we need to solve the following practical tasks:

(1) Establish the clustering method based on Assumption 1.
(2) Prepare the datasets that include information about economic events in the market.
(3) Establish the model that determines the trend and entertains the market-shift behavior.

3 Model

In the following subsections, we sequentially go through these practical tasks and describe the fundamental parts of our model, as depicted in Fig. 1.

3.1 Market regimes and clusterization

We consider clusterization to be the fundamental part of addressing different market regimes and thus exploring signs of market regime shifts. The application of many such methods for financial time series has been extensively examined [2]. Famous suitable machine learning algorithms, such as KMeans, Gaussian mixtures, and Hierarchical Clustering, are impossible to use for time-series clusterization. Usually, these techniques do not use the time-varying component of the data series, and remain very sensitive to the outliers. Statistical and econometrics techniques offer solutions to this problem. Markov Switching Models [12], Conditional Correlation (DCC-GARCH) models [7] and State Space models [11] are practical frameworks, suitable for the market regime detection. However, these models imply constant probability assumption, which assumption is not suitable for our modeling aim.

Based on the historical data, we propose a deterministic way for market-regime clusterization. Let us determine the problem statement first. As a fundamental requirement, we fix the time horizon of the model we are about to build. In this paper, we select both the monthly and quarterly time horizons.

Assume there is a market-regime process $S_t \in [0, \dots, S-1]$, where S is the total number of regimes. We want to develop a deterministic algorithm that estimates the given process using time-series data and satisfies Assumption 1. Let be $\{R_t\}_{t\in[1,\dots,T]}$ – the monthly return of the asset and $\{\sigma_t\}_{t\in[1,\dots,T]}$ – daily volatility based on the previous 21 trading days (we took the last day of the month), i.e.,

$$\sigma_t = \sqrt{\frac{\sum_{i=0}^{20}(r_{t-i} - \bar{r}_{[t-20,t]})^2}{20}}, \tag{1}$$

where r_t is the daily return, $\bar{r}_{[t-20,t]}$ is the mean return for previous 21 trading days. We suggest using the cumulative return: $CR_t = \sum_{i=1}^{t} R_i$, instead of R_t. Even though R_t satisfies the stationarity property, CR_t gives the trend properties of the asset.

Following the notation, we propose two types of clusterization:

(1) Return-based clusterization. The function which takes as input $\{ACR_t\}_{t\in[1,\dots,T-3]}$, where $ACR_t = \frac{1}{6}\sum_{i=0}^{5} CR_{t-i+3}$ (i.e. it's shifted cumulative return of the asset)
(2) Volatility-based clusterization. The function which takes as input $(\{ACR_t\}_{t\in[1,\dots,T]}\}, \{\sigma_t\}_{t\in[1,\dots,T]})$

We mentioned that S_t can take values from 0 to $S-1$. Hence, we need to determine the number of market regimes we distinguish. Empirical studies demonstrate that the optimal number of market regimes is 4 or 6 [11]. We shall therefore explore both cases to determine which number is better.

Using the algorithms described in Appendix A, we provide the clusterization example presented for the S&P 500 index for 1990–2024 (Fig. 2).

We use these clusterization algorithms to label the data and to create a target value for the classification model. Before proceeding to the next stage of model-building, we introduce an additional yet essential assumption:

Assumption 2. The market follows the same cross-sectoral market partition.

In other words, we assume that assets of the same fundamental nature inherit the same price formation dynamics. Therefore, we use one classification model for each market sector.

3.2 Classification model

After defining the fundamentals, we now describe our classification model. We have $\{R_t\}_{t\in[1,\dots,T]}$, $\{CR_t\}_{t\in[1,\dots,T]}$ and $\{\sigma_t\}_{t\in[1,\dots,T]}$ – time-series data, also we have $\{Z_t\}_{t\in[1,\dots,T]}$ – news data. We consider Z_t as the sequence of embeddings of the news (or simply the news), i.e. $\{Z_{t,1}, \dots, Z_{t,k_t}\}$, with $Z_{t,i} \in \mathbb{R}^{784}$ and k_t being the number of news for timestamp t. Next, we have $\{\hat{S}_t\}_{t\in[1,\dots,T]}$ – market-regime target value, i.e. $\hat{S}_t \in [0, \dots, S-1]$ – the value obtained from clusterization algorithm described in Appendix A. Denote the time-series inputs of the model as X_t. For each timestamp t, it is the vector containing different time-series parameters. We opt to develop an algorithm $CLF : X_t \times Z_t \to [0, \dots, S-1]$, which estimates the probabilities of the market regimes for the next month.

The general scheme is presented in Figure. 3. We discuss this scheme below in details.

3.2.1 Time-series classification. The bottom part of the block scheme corresponds to estimating the market-regime probabilities using the time-series data. We opt to design a model that predicts these probabilities for the following month using the features obtained from the time series. Our problem statement is therefore: We have $\{X_t\}_{t\in[1,\dots,T]}$ – time-series features obtained from the returns, cumulative returns, and volatility. Our aim is to find parameters $\hat{\theta}_{\min}$:

$$\hat{\theta}_{\min} = \underset{\theta}{\operatorname{argmin}} \left[\mathcal{L}(TS_\theta(X_t), \hat{S}_t)\right], \tag{2}$$

where $\mathcal{L}(\hat{y}_t, y_t)$ is cross-entropy loss, $TS_\theta(X_t)$ is a complex enough neural network, parametrized with θ. We combined the LSTM layer and 4-layer linear network with GeLU activation.

3.2.2 News data classification. The upper block is dedicated to the task of classifying market regimes by analyzing the news dataset. This is the most complex block of the model, and it contains five parts:

- Preprocessing is used to retrieve tokens from the news bodies. We use the standard tokenizer presented in [18].
- Selector model eliminates news with low financial sentiment. This part is needed to avoid using non-financial news. In our analysis, we are utilizing the FinBERT model to conduct sentiment analysis on the data [5].
- Scoring model assigns a score-value $q \in (0, 1)$ to each piece of news. This part is crucial for understanding which news will affect the dynamic of the asset. The higher the score, the more «informative» news for our analysis. In the same way, we use the LLM here. However, for the given problem, we used all-MiniLM-L6-v2 due to its lower complexity.
- Probability model is another FinBERT model, which maps the news to the vector of probabilities of market regimes: $FB_\chi(Z) : \mathbb{R}^{784} \to \mathbb{R}^S$. Here χ are parameters of FinBERT model. This model is the final part of the NLP block and predicts market regimes based on news datasets.
- The final prediction:

$$NC_\omega(Z_t) = \sum_{i\in\mathbb{I}_t} q_{t,i} FB_\chi(Z_{t,i}),$$

where $q_{t,i}$ is the score of the news $Z_{t,i}$, $\mathbb{I}_t$ is the subset of all news, which have «informative» sentiment, ω are parameters of final layer and $FB_\chi(Z_{t,i})$ is the next period probability vector of the market-regimes.

In the followings we describe all the modules of the block scheme as shown in Figure 3.

Selector model To exclude non-informative news, we need to understand if a given text contains financial information. Different approaches have been emerging to solve the problem. Before the LLMs, authors examined the word frequency features [1] or worked with word embeddings [6]. Unfortunately, the latter two methods are not complex for sentiment analysis problem. Development of the LLM allowed researchers to apply it to financial text analysis [5]. The authors, using LLM, classified the text into three classes: *positive*, *negative*, or *neutral*. The achieved results were impressive (up to 15% of improvement), due to the proper model architecture.

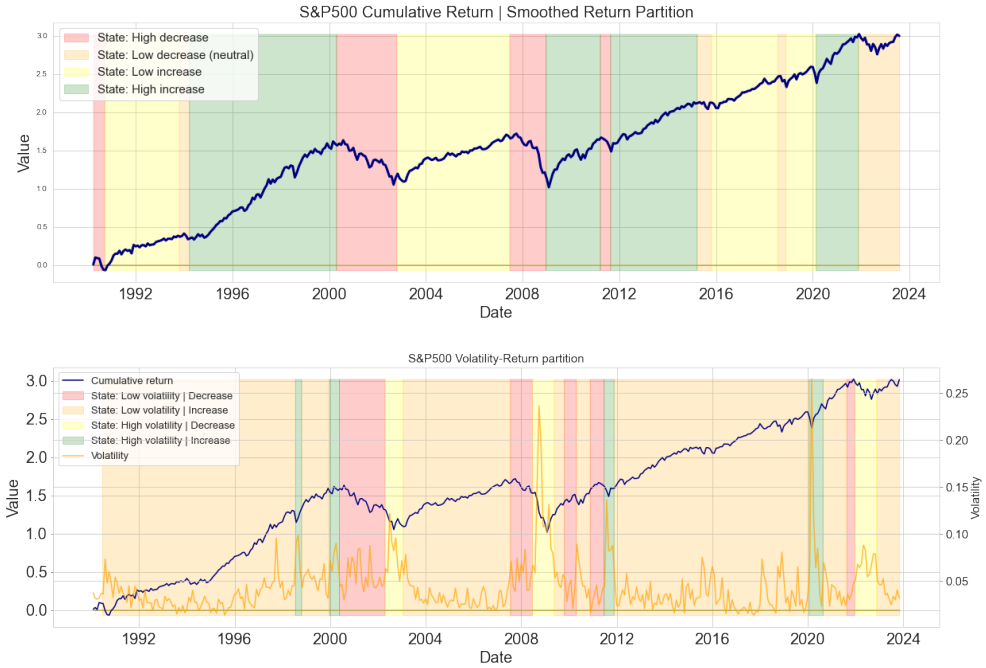

Figure 2: Market-regimes clusterizations for S&P500 financial index with 4 possible regimes: return-based method and volatility-based method

.

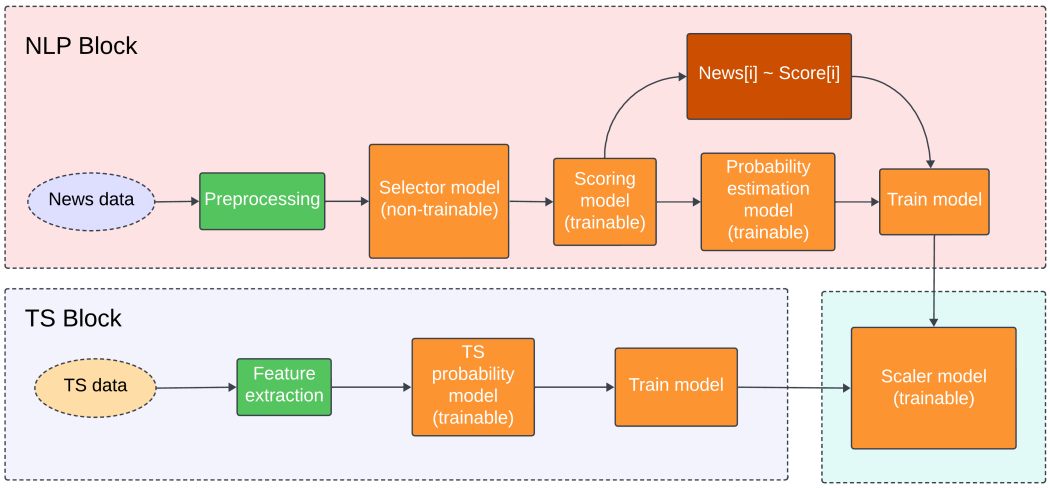

Figure 3: Model scheme

Therefore, we use this model to select news with positive or negative financial sentiment. We follow the next rule of thumb: The news is considered «informative» if and only if its sentiment score corresponding to the *positive* or *negative* class is higher than p. The latter parameter implies the strictness of the model, and can be optimized using cross-validation techniques.

Scoring model This is the essential, systemic part to reach optimal performance. Using the «informative» news $(Z_{t,1}, \ldots, Z_{t,k_t})$ obtained from the previous step, we develop a model that prioritizes the news based on its body. We use the concept of market movers, defined as follows:

Assumption 3. The extensive price movements during one day are correlated with news that occur more than once.

Let's focus on the inner idea of the given assumption. For the given asset, we can aggregate the moments of significant price movements and check the general news trends for the given moments. This allows us to correspond the asset dynamics with news bodies.

We formulate the following problem: based on «informative» news, we opt to develop a model that detects whether the news affects the market in the longer term, i.e., whether the (piece of) news is a market-mover. To solve the problem, we train the classification model with the labeled dataset and a FinBERT model $FB_\psi(Z)$ with ψ as the parameter of the model:

$$\hat{\psi}_{\min} = \underset{\psi}{\operatorname{argmin}} \left[\mathcal{L}_{\text{bin}}(FB_\eta(Z_{t,i}), \tilde{S}_{t,i} \right], \tag{3}$$

where $\tilde{S}_t$ is the market-mover label of the piece of news and the $\mathcal{L}_{\text{bin}}(\hat{y}_t, y_t)$ is binary cross-entropy loss. This model allows us to use scores, i.e. the probability $\mathbb{P}(FB_{\hat{\psi}_{\min}}(Z_{t,i}) = 1)$. Hence, we can set the scores (or weights) to the news to satisfy its ability to predict market behavior.

Probability model The problem statement is the following: Given $(Z_{t,1}, \ldots, Z_{t,k_t})$ as the vectors of «informative» news, the vector $(q_{t,1}, \ldots, q_{t,k_t})$ as corresponding scores, we opt to develop a FinBERT model FB_χ that fulfill:

$$\hat{\chi} = \underset{\chi}{\operatorname{argmin}} \left[\mathcal{L}(P_t, \hat{S}_t) \right], \tag{4}$$

where $P_t = \sum_{i \in \mathbb{I}_t} q_{t,i} FB_\chi(Z_{t,i})$, $q_{t,i}$ is the score of the news $Z_{t,i}$, $\mathbb{I}_t$ is the subset of all news, which have «informative» sentiment and $FB_\chi(Z_{t,i})$ is the next period probability vector of the market-regimes.

Scaler layer The scaler function is the final module of the block scheme in Fig. 3, which combines two previous parts. Using predictions $P_{1,t} = TS_{\hat{\theta}_{\min}}(X_t)$ and $P_{2,t} = NC_{\hat{\omega}_{\min}}(Z_t)$ we obtain two numbers w_1 and w_2, representing the model confidence in each prediction. The final probability vector is obtained using the following formula: $P_t = w_{1,t}P_{1,t} + w_{2,t}P_{2,t} \in \mathbb{R}^S$. Hence, we have $SC(P_1, P_2) : \mathbb{R}^S \times \mathbb{R}^S \to \mathbb{R}^2 = (w_1, w_2)$ and our aim is to find $\hat{h}_{\min}$:

$$\hat{h}_{\min} = \underset{h}{\operatorname{argmin}} \left[\mathcal{L}(w_{1,h,t}P_{1,t} + w_{2,h,t}P_{2,t}, \hat{S}_t) \right] \tag{5}$$

It is suggested that a simple 4-layered neural network with GeLU activation shall be used for the scaler layer.

4 Experiment

After describing our model and its structure in detail, we now turn to our experiment that we use to address the performance of our proposed model design. To start with, we describe below the data we instrumentalize.

4.1 Datasets

We employed different datasets for our study. In the following section, we highlight the data structure for each dataset.

(1) Time-series dataset D_1. This dataset contains the data for S&P500 index. The data was collected from January 1990 to July 2024. The fields are `(date, return, close)`.
(2) Daily time-series dataset D_2. This dataset contains the data for the S&P500 index at daily intervals. The data was collected from January 2010 to July 2024. The overall fields are `(date, return, close)`.
(3) Scoring model news dataset S_1. This dataset contains news titles for Bloomberg, CNBC, Investing, SeekingAlpha, and BBC sources for "significant" days. The methodology is following (see Appendix B.1 for more details):
   (a) Using, $r_t \in D_2$ we find the $t : |r_t - \bar{r}_{t-21,t-1}| \geq 3\sigma_t$, where $\bar{r}_{t-21,t-1} = \frac{1}{21}\sum_{i=1}^{21} r_{t-1}$ and σ_t is the corresponding standard deviation. Let's denote these moments as $\mathbb{T}$.
   (b) For each, $t \in \mathbb{T}$ we collect `(date, title, body)` for all news from the mentioned sources.
   (c) Next, we label the news as «informative» if its title is similar to different news from different sources. For similarity, we use **all-MiniLM-L6-v2** model.
   (d) We obtain a dataset of `(date, title, body, label)` for significant dates.
(4) General model news dataset S_2. This dataset contains the titles and bodies of news from Bloomberg, The Economist, CNBC, and BBC sources from January 2010 to July 2024.

4.2 Experiment design

Experiments were conducted for the dates from January 2019 to January 2024. Although we have a larger dataset, we used these dates based on scoring model results. The latter model is trained on the previous three years. We believe the following hypothesis holds:

Assumption 4. Informative news follows different patterns during different large segments of time.

This means that informative news from different dates are covered by different topics.

Model complexity parameters are presented in Table 1.

Table 1: Model complexity

Model name	Num. of parameters (mln.)	Comments
Selector model	110	Non-trainable
Scoring model	22	Trainable
Probability model	110	Trainable
Time series model	0.8	Trainable
Scaler model	0.6	Trainable

For all environmental setups, we used PyTorch library. The training process follows the parameters given in Table 2. The datasets were divided into training and testing samples randomly to avoid time-consuming model overturning. Each learning process took 8–10 minutes on the NVIDIA Tesla T4 GPU.

5 Results and discussion

In the given section, we describe our results.

Table 2: Parameters of model

Parameter	Value	Comments
lr	5e-3	Learning Rates
optimizer	Adagrad	Model optimizer
p	0.7	Selector parameter
λ	1e-2	Weight decay
rs	42	Random Seed

5.1 Baseline

The model does not have any closely related baseline methods. As discussed, most well-known approaches use probability models to estimate market regimes. However, we opted for pre-determined market clustering, which can differ from probability models. Comparing the model with [5] is not feasible due to the different problem statements. Although it's possible to train the FinBERT model to classify the given classes on the entire dataset, the result would be non-existent due to the amount of data and spam news.

We will use two natural methods for our baseline prediction: last-state prediction and time series (ts) prediction. Due to the persistence of market regime partition, the last-state prediction method has high accuracy and a high average weighted F1 score. However, it cannot be used in practice due to the non-deterministic nature of the current market regime. It can be considered a state-of-the-art solution, assuming EMH holds.

The ts-prediction method is essential for facilitating a comparison between our model and time-series forecasting, emphasizing the importance of the NLP component. In the subsequent results section, we perform a thorough analysis to compare the predictions produced by our model with those from ts-prediction, ultimately assessing the benefits obtained from using the combined model. Nonetheless, it is important to note the best possible result for the specified problem.

As mentioned earlier, if we assume that EMH holds, the best course of action is to utilize the last-moment regime. However, because we are unable to ascertain the true regime value at the current moment, it cannot be used as a valid model. Nevertheless, it can be valuable in establishing an upper limit for model performance. The last-regime prediction achieves 91% accuracy and a $wF1$ score between 0.92 and 1, which will be defined in the next section.

5.2 Our results

In the following section, we provide the results of our model for different market clusterization algorithms and two cases of 4 and 6 regimes. We added the states' meanings to make them more reasonable.

We examined the classical classification metrics: accuracy and weighted version of the F1 score.

$$wF1 = \sum_{i=0}^{S-1} w_i F1_i,$$

where

$$w_i = \frac{\text{number of samples with i-th class}}{\text{total number of samples}}$$

and $F1_i$ is standard F1 score for i-th class. The results are presented in Table 3. We marked the results with boldface that were better than those of the time-series prediction. In the given table, we have the following columns:

(1) *Model*: type of the model.
(2) *Regime type*: type of the clusterization method. RB is return-based, and VB is volatility-based.
(3) *Regimes*: number of regimes in clusterization.
(4) *Accuracy*: accuracy score out-of-sample.
(5) *wF1*: weighted F1 score.

Table 3: Results

Model	Regime type	Regimes	Accuracy	wF1
TS	RB	4	55%	0.39
TS	RB	6	55%	0.39
TS	VB	4	64%	0.49
TS	VB	6	37%	0.36
TS + NLP	RB	4	**81**%	**0.82**
TS + NLP	RB	6	**72**%	**0.68**
TS + NLP	VB	4	**82**%	**0.78**
TS + NLP	VB	6	**64**%	**0.66**

Upon reviewing the result table, it's evident that our model has outperformed the prediction based solely on the time-series dataset. In the case of a larger amount of regimes, the model outperforms ts-prediction by $\frac{64-37}{37} \approx 73\%$ and by $\frac{72-55}{55} \approx 31\%$. In the context of the weighted F1 score metric, the model gives $\frac{82-39}{39} \approx 110\%$ of upgrade in case of 4 market regimes and return-based clusterization.

Given these findings, the model surpasses the average prediction obtained from the TS Model and can effectively identify market shifts. However, the model still exhibits some limitations, including false market shifts and frequent changes in trends (as elaborated in the next subsection).

6 Conclusion

In this paper, we explored the use of financial news for market-regime detection and presented a practical approach for categorizing market regimes based on asset return and volatility time series. Using a combination of FinBERT models, we demonstrated that our approach, following a step-wise, structural process provides valuable results. Our approach, and the underlying model design, outperformed the traditional time-series model for different market regimes, using different partition strategies, namely showing a 73% improvement in accuracy for the 6-regimes scenario using volatility-based, and 110% improvement for the weighted F1 score for the 4-regimes scenario using return-based partitions.

Nevertheless, our model has limitations. Even though it distinguishes between different market regimes, it still produces forecasting errors and does not fully align with the persistence property of market regimes. In practice, this leads to frequent changes in an asset trend. We believe that retraining the scoring model may help reduce forecasting errors, since the scoring model corresponds to the model confidence in the given piece of news. Furthermore, the

time-dependency of the model is evident due to the scoring module in the process depicted in Fig. 3.

Acknowledgments

We would like to express our sincere gratitude to Evgeny Polyachenko, Alexander Kalinichenko, and the Finmars team for engaging in a highly insightful discussion and providing valuable advice.

References

[1] Basant Agarwal and Namita Mittal. 2016. *Machine Learning Approach for Sentiment Analysis*. Springer International Publishing, Cham, 21–45. https://doi.org/10.1007/978-3-319-25343-5_3
[2] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. 2015. Time-series clustering – A decade review. *Information Systems* 53 (2015), 16–38. https://doi.org/10.1016/j.is.2015.04.007
[3] Andrew Ang and Geert Bekaert. 2004. How regimes affect asset allocation. *Financial Analysts Journal* 60, 2 (2004), 86–99.
[4] Andrew Ang and Allan Timmermann. 2012. Regime changes and financial markets. *Annu. Rev. Financ. Econ.* 4, 1 (2012), 313–337.
[5] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs.CL] https://arxiv.org/abs/1908.10063
[6] Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications* 77 (2017), 236–246. https://doi.org/10.1016/j.eswa.2017.02.002
[7] Tim Bollerslev. 1990. Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized Arch Model. *The Review of Economics and Statistics* 72, 3 (1990), 498–505. http://www.jstor.org/stable/2109358
[8] Arthur F. Burns and Wesley C. Mitchell. 1946. *Measuring Business Cycles*. NBER, USA.
[9] Tim S Campbell and William A Kracaw. 1980. Information production, market signalling, and the theory of financial intermediation. *the Journal of Finance* 35, 4 (1980), 863–882.
[10] Maria Jose Roa Garcia. 2013. Financial education and behavioral finance: new insights into the role of information in financial decisions. *Journal of economic surveys* 27, 2 (2013), 297–315.
[11] Massimo Guidolin and Allan Timmermann. 2007. Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control* 31, 11 (2007), 3503–3544. https://doi.org/10.1016/j.jedc.2006.12.004
[12] James D Hamilton and Raul Susmel. 1994. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics* 64, 1 (1994), 307–333. https://doi.org/10.1016/0304-4076(94)90067-1
[13] Bruce E. Hansen. 1992. The Likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of GNP. *Journal of Applied Econometrics* 7, S1 (1992), S61–S82.
[14] David Hirshleifer. 2015. Behavioral finance. *Annual Review of Financial Economics* 7, 1 (2015), 133–159.
[15] Chang-Jin Kim. 1994. Dynamic Linear Models with Markov-Switching. *Journal of Econometrics* 60, 1-2 (1994), 1–22.
[16] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. arXiv:2311.10723 [q-fin.GN] https://arxiv.org/abs/2311.10723
[17] Georg Lindgren. 1978. Markov Regime Models for Mixed Distributions and Switching Regressions. *Scandinavian Journal of Statistics* 5, 2 (1978), 81–91. http://www.jstor.org/stable/4615692
[18] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the American Society for Information Science and Technology* 1 (04 2014). https://doi.org/10.1002/asi.23062
[19] Michael T Maloney and J Harold Mulherin. 2003. The complexity of price discovery in an efficient market: the stock market reaction to the Challenger crash. *Journal of corporate finance* 9, 4 (2003), 453–479.
[20] Zacharias Psaradakis, Martin Sola, and Fabio Spagnolo. 2004. On Markov Error-Correction Models, with an Application to Stock Prices and Dividends. *Journal of Applied Econometrics* 19, 1 (2004), 69–88.
[21] Barbara Rossi. 2005. Testing Long-Horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle. *International Economic Review* 46, 1 (2005), 61–92.
[22] Paul Samuelson. 1965. Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review, 6:2* 1 (1965), 41–49.
[23] Nuhu A Sansa. 2020. The Impact of the COVID-19 on the Financial Markets: Evidence from China and USA. *Electronic Research Journal of Social Sciences and Humanities* 2 (2020), 11.
[24] Pavel Savor and Mungo Wilson. 2014. Asset pricing: A tale of two days. *Journal of Financial Economics* 113, 2 (2014), 171–201.
[25] Adam Hale Shapiro, Moritz Sudhof, and Daniel J Wilson. 2022. Measuring news sentiment. *Journal of econometrics* 228, 2 (2022), 221–243.
[26] Jingjing Wang and Xiaoyang Wang. 2021. COVID-19 and financial market efficiency: Evidence from an entropy-based analysis. *Finance Research Letters* 42 (2021), 101888.

A Market regimes clusterization

In the given appendix, we describe the algorithms for market-regime partition. As we discussed, the definition of market regimes is pretty vague. The further algorithms were organized to satisfy Assumption 1 and the econophysics ideas provided in [11].

A.1 Return based partition

The algorithm presented below corresponds to the return-based partition. Its general meaning is to divide the data by time segment with approximately the same rate of increase.

Algorithm 1 Return based partition

Require: $S > 0$, ACR_t, $T = 6$ (default)
  Determine $t : ACR_t \geq ACR_{t+k}$ and $ACR_t \geq ACR_{t-k}$ with $k \in [1, 3]$ {Call them up}
  Determine $t : ACR_t \geq ACR_{t+k}$ and $ACR_t \geq ACR_{t-k}$ with $k \in [1, 3]$ {Call them $down$}
  Get $points = \{t_1 < \ldots < t_k\}$, with $t_i \in up$ or $t_i \in down$
  for $i \leq k - 1$ **do**
    Train linear regression $y = \hat{A}X + \hat{B}$, where $X = [0, \ldots, t_{i+1} - t_i]$ and $y = [0, \ldots, ACR_{t_{i+1}} - ACR_{t_i}]$
    Set $\alpha_i = \hat{A}$
  end for
  $diff = \dfrac{\max(\alpha_i) - \min(\alpha_i)}{S}$
  for $i \leq k - 1$ **do**
    Set $\hat{S}_t = j : \alpha_j \geq \min(\alpha_i) + diff \cdot j$ and $\alpha_j \leq \min(\alpha_i) + diff \cdot (j + 1)$ for $t_i \leq t \leq t_{i+1}$
  end for

Consider 2 first figure as an example of the given algorithm.

A.2 Volatility based partition

The algorithm presented below corresponds to the volatility-based partition. Its general meaning is the following: divide the data by time segment at approximately the same rate of increase.

Algorithm 2 Volatility based partition

Require: $S > 0$, ACR_t, σ_t, $T = 6$ (default)
Determine $z_{80\%}(\sigma_t)$, 80% percentile of the volatility time series
Determine $t : \sigma_t \geq z_{80\%}(\sigma_t)$
Determine $t : \sigma_t \leq z_{80\%}(\sigma_t)$
Determine first points of high and low values: $\{t_1, \ldots, t_k\}$
for $i \leq k - 1$ **do**
  For segment $[t_i, t_{i+1}]$ train a KMeans for R_t data with 2,3 clusters
  Get the optimal number of classes based on the Silhouette score. Call it k_{opt}.
  Train Linear Regression for each class and for the whole segment. Call R_1 and $R_{opt} = \frac{1}{N}\sum_{i \leq N} R_i^2$ their r-squared scores.
  if $R_1 \geq R_{opt}$ **then**
    Set α_i - slope coefficient of Linear Regression
  else
    Make a partition of $[t_i, t_{i+1}]$ on the segments of the same class value. Call it $[t_i, t_{i_1}] \cup \ldots \cup [t_{i_m}, t_{i+1}]$.
    For each $[t_{i_k}, t_{i_{k+1}}]$, set α_{i_k} - slope of corresponding Linear Regression
  end if
end for
Call $[t_0, \ldots, t_N]$ and $[\alpha_1, \alpha_N]$ – corresponding time segments partition and slope values
$diff = \frac{\max(\alpha_i) - \min(\alpha_i)}{S}$
for $i \leq N - 1$ **do**
  Set $\hat{S}_t = j : \alpha_j \geq \min(\alpha_i) + diff \cdot j$ and $\alpha_j \leq \min(\alpha_i) + diff \cdot (j + 1)$ for $t_i \leq t \leq t_{i+1}$
end for

B Implementation details

Here, we describe several model implementation details.

B.1 Scoring model

In the given subsection, we describe the implementation details of our proposed scoring model. As mentioned earlier, the general aim of this part is to weigh the news based on its significance, satisfying Assumption 3. The algorithm we proposed in the Datasets section is simple outlier detection. We provide several informative points corresponding to the given model.

(1) Find appropriate outliers (dates of outlier events) based on the asset's daily return mean value and deviation for the last 21 trading days. We suggest to use factor 3 to avoid non-significant timestamps.
(2) For the given timestamps, we analyze the news titles from different sources. We choose titles due to its similarity. Bodies may contain different topics inside, and LLM may not trace the news similarity due to complexity of the model.
(3) We decided to use the **all-MiniLM-L6-v2** model due to the low complexity of the model and its ability to correctly identify similar news by titles.
(4) Similarity function we used is cosine similarity determined by:
$$\cos(x, y) = \frac{\langle x, y \rangle}{\|x\|\|y\|}, \tag{6}$$
where x, y vector embeddings, and $\langle x, y \rangle$ is simple Euclidean dot-product.
(5) The criteria were the following: two news are similar if the cosine similarity of their embedding vectors is higher than 0.8. The threshold is empirical and a question of possible research.