
Research article

Sentiment-driven Markov mixture modeling of asset returns

Mostafa Abdolahi Moghadam Salkoye and Roman N. Makarov*

Department of Mathematics, Wilfrid Laurier University, 75 University Av. West, Waterloo, Canada

* **Correspondence:** Email: rmakarov@wlu.ca.

Abstract: This paper proposes a sentiment-based framework to forecast financial asset returns that combines financial news sentiment analyses with a Markov finite mixture model. Daily sentiment states are constructed from financial news headlines using the FinBERT model and are linked to the distributional characteristics of asset returns. Treating sentiment as a Markov process, the study examines how transitions across sentiment regimes shape the return distribution and contribute to the return heterogeneity. To capture this heterogeneity, we develop a finite Markov mixture distribution, conduct goodness-of-fit assessments, and then integrate the resulting sentiment-driven specification into a FIGARCH-based forecasting framework. Empirical results across 37 assets from different markets indicate that incorporating a four-state, headline-derived Markov sentiment structure outperforms traditional FIGARCH benchmarks across all metrics for Close-to-Close return forecasts. Meanwhile, the traditional models and classification-based approaches remain competitive for Open-to-Close and Open-to-Open horizons in specific metrics. These findings highlight the nuanced, time-dependent nature of news sentiment measures and how they are incorporated into asset price forecasting models.

Keywords: sentiment analysis; FinBERT*; Markov mixture distribution; FIGARCH[†] model; asset returns modeling

JEL Codes: C52, C53, C38, G17, G14.

1. Introduction

In financial modeling, the accurate characterization of asset returns distribution is crucial for risk assessment, portfolio optimization, derivative pricing, and predictive modeling. Deviations from assumed distributions can lead to substantial errors and misinformed decisions, making the accurate modeling of asset returns a cornerstone of robust and reliable financial systems. Predictive models for

*Financial Bidirectional Encoder Representations from Transformers (BERT)

[†]Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

asset returns, such as regression models or machine learning algorithms, often assume an underlying distribution for their error terms or target variables. A well-characterized return distribution improves the robustness and interpretability of these models.

Asset returns are influenced by various factors, such as market dynamics, macroeconomic indicators, and investor sentiment. Sentiment derived from financial news has emerged as a significant driver of market behavior, thereby reflecting the collective perceptions and actions of market participants. Thus, incorporating a sentiment analysis to enhance the financial return modeling has garnered considerable attention in both academic and practical settings.

Financial data often exhibit complex characteristics, such as non-normality, heteroskedasticity, and volatility clustering, which challenge traditional statistical models. In this context, the adoption of finite mixture distributions provides a flexible and robust framework to model heterogeneity in return data. Finite mixture models allow for combining multiple probability distributions, each representing distinct market regimes or states, which is particularly relevant when market sentiment is incorporated as a driving factor. The primary property of mixture distributions is that the shape of the density is extremely flexible, allowing it to be multimodal and exhibit considerable skewness or additional humps. This model provides a flexible approach to describe an unrefined data structure by summarizing the characteristics of the data in terms of the number, location, and spread of the mixture components (Frühwirth-Schnatter, 2006). By linking sentiment states to return distributions, financial models can more accurately capture the nuanced behavior of asset prices and improve the prediction accuracy.

Many studies have been conducted on financial news sentiment analyses, focusing on predicting sentiment of specific textual data or implementing overall sentiment scores or labels in prediction and classification models. However, no study has examined the impact of financial news sentiment evaluated by a financial-specific sentiment classification model on the asset return distributions within a finite mixture framework. This paper focuses on developing an innovative approach to modeling financial return data by integrating a sentiment analysis of financial news headlines as a Markov process and finite mixture distributions to enhance return prediction models. The proposed methodology begins by labeling daily return data based on sentiment states derived from a financial news analysis. Sentiment states are modeled as two- or four-state Markov processes. By associating returns with sentiment-driven classifications treated as a Markov process, the study investigates whether samples, which correspond to different sentiment states, follow distinct probability distributions. This step is pivotal to understand the interplay between sentiment and return dynamics.

To address the complexities of return distributions, a finite Markov mixture model with appropriate components and known allocations informed by sentiment states is designed. The mixture model aims to capture the heterogeneity in the returns while accounting for variations driven by market sentiment. Unlike traditional regime-switching models (e.g., MS-GARCH or HMMs[‡]) that endogenously infer states from price data, which often result in statistical black boxes with heavy computational burdens, our framework uses an observable Markov mixture approach. In this study, we utilize a finance-specific NLP[§] model (FinBERT) to directly label samples as prior information to allocate observations to specific components of a mixture model. By externally assigning regimes prior to parameter estimation, we anchor the mixture components to a tangible informational driver; this ensures a transparent economic interpretation while simplifying the parameter estimation. The goodness-of-fit of the proposed mixture

[‡]Hidden Markov Models

[§]Natural Language Processing

distribution is thoroughly tested to ensure its suitability for financial return data. Finally, the study integrates the Markov mixture distribution into a FIGARCH model for returns and compares its performance with that of traditional FIGARCH models. By evaluating these models under different assumptions, the research seeks to highlight the advantages of incorporating sentiment-driven finite Markov mixture distributions in predictive modeling.

The key contributions of this paper are as follows:

- A novel approach to labeling daily return data using the FinBERT model, thereby utilizing financial-specific NLP to directly assign observable states as prior information and enabling the investigation of relationships between sentiment and return distribution.
- Design and implementation of an observable finite Markov mixture distribution to model the heterogeneity in return data, which explicitly avoids the black-box nature and computational burdens of standard latent regime switching models.
- Rigorous testing of the goodness-of-fit of the mixture model and its integration into a FIGARCH framework for return prediction.
- A comparison of the performance of the proposed model against traditional FIGARCH models, thus demonstrating the impact of incorporating sentiment analyses and Markov mixture distributions.

This study contributes to the intersection of financial econometrics, machine learning, and sentiment analyses by proposing a sentiment-based framework to model daily returns. The results have significant implications for investors, financial analysts and researchers seeking to better understand the influence of market sentiment on asset price dynamics and to improve the accuracy of financial prediction models.

The remainder of this paper is organized as follows: Section 2 discusses related work on the integration of sentiment analyses into financial modeling; Section 3 describes the data used in our study; the proposed sentiment-based Markov switching models for asset returns are constructed in Section 4; the sentiment-adjusted Markov-switching FIGARCH model is developed in Section 5; Section 6 discusses computational results using empirical data for 37 assets drawn from various markets; and Section 7 concludes the paper.

2. Related work

The integration of sentiment analyses into financial modeling has gained substantial traction in recent years, driven by the growing availability of textual data and advances in natural language processing (NLP). Previous studies have demonstrated that sentiment derived from financial news, social media, and corporate reports can provide valuable insights into market dynamics and improve the predictability of asset returns. Wan et al. (2021) investigated the correlation of sentiment in financial news networks and its relationship with market movements. They constructed a network-based framework that analyzed sentiment interconnections between financial news articles and thereby demonstrated how these correlations could predict asset price fluctuations and market trends. Case & Clements (2021) fine-tuned a BERT-based model to analyze sentiment in financial, economic, and national news articles, demonstrating that daily financial news sentiment predicts short-term market returns, while economic news sentiment influences monthly returns.

Sentiment analyses have been widely applied to understand the impact of textual data on financial markets. Early research focused on sentiment extraction using lexicon-based approaches, such as

the Loughran–McDonald dictionary, which provides domain-specific sentiment indicators for finance. Recent advances in machine learning and NLP have enabled the use of pre-trained language models, such as BERT and its financial domain adaptation, FinBERT, to derive more nuanced sentiment scores from textual data (Huang et al., 2023). These sentiment measures have been shown to strongly correlate with market movements, thus providing a useful tool to predict asset returns and volatility. Tetlock (2007) demonstrated that the tone of news articles can predict market volatility, while Kelly & Ahmad (2018) developed a domain-specific sentiment analysis approach using multiple dictionaries tailored to financial terminology. They showed that sentiment extracted from the equity and oil market news could predict next-day returns and improve trading strategies, thus outperforming a simple buy-and-hold approach. Xiang et al. (2025) proposed an adaptive multimodal prompt-tuning model for few-shot sentiment analyses, thereby integrating text and visual modalities. By dynamically generating multimodal prompts and integrating them with pre-trained language models, their approach improves the semantic comprehension of multimodal inputs. Dridi et al. (2019) proposed FineNews, which is a fine-grained supervised sentiment analysis approach that predicts bullish and bearish sentiments in financial microblogs and news using lexical and semantic features. Their study highlighted the effectiveness of semantic features, such as BabelNet synsets and semantic frames, which outperformed unsupervised methods in sentiment prediction. Xu et al. (2019) investigated the impact of investor sentiment on realized range-based volatility using a regime-switching framework. Their findings demonstrated that incorporating investor sentiment significantly improves the accuracy of volatility forecasting, with robust results across various forecast windows. Despite these advances, there is a need to systematically link sentiment to the statistical properties of return distributions, such as their probability density functions, to better capture heterogeneity in market behavior.

Finite mixture models have emerged as a powerful tool to model the complexity of financial data, particularly when returns exhibit non-normal behavior, such as non-zero skewness and heavy tails. These models assume that the observed data are generated from a mixture of underlying distributions, each representing a distinct market regime or state. By capturing this heterogeneity, finite mixture models provide a flexible framework to model returns that traditional single-distribution approaches cannot.

Empirical studies have demonstrated the efficacy of finite mixture models in various financial applications. Hamilton (1989) introduced regime-switching models to account for structural changes in macroeconomic and financial data. Subsequent research has extended these concepts by incorporating finite mixture distributions to model volatility clustering and extreme events in asset returns. Shi & Ho (2021) employed an MRS-FIEGARCH model and RavenPack data to show that macroeconomic and firm-specific news sentiment significantly affects intraday stock return volatility states. Shi & Ho (2015) proposed a 3S-FIGARCH model with Markov regime-switching to address structural changes, thus improving the estimation of long-memory volatility in high-frequency data. Shi & Feng (2016) showed that public news arrivals reduce the IVOL-return relation by 50%, using CRSP and RavenPack data (2000–2011). Feng et al. (2022) utilized an MRS-LMGARCH model with RavenPack data to show that news sentiment significantly influences volatility states in Japanese TOPIX Core 30 stocks. However, no studies have investigated the integration of sentiment states as prior information to allocate observations to specific components of a mixture model, a key focus of this paper.

Regression models are a cornerstone of financial predictions, often assuming normally distributed errors for simplicity. However, this assumption is frequently violated in financial data due to heavy tails and asymmetric distributions. To address these limitations, researchers have combined finite mixture models

with regression frameworks, allowing for more robust parameter estimations under non-normal conditions.

Young & Hunter (2010) employed mixture regression models to account for heterogeneity in financial time series, thereby improving the return prediction performance. Yao et al. (2014) proposed a robust mixture regression model using the t-distribution to handle outliers and heavy-tailed errors, with adaptive trimming of high-leverage points and profile likelihood to optimize the degrees of freedom. Despite these advances, the integration of sentiment analyses into mixture regression models remains underexplored. This paper aims to bridge this gap by using sentiment-driven allocations within a finite-mixture regression framework.

Although significant progress has been made in sentiment analyses, mixture modeling, and regression-based financial predictions, several gaps remain. First, most existing studies treat sentiment as a standalone variable without directly linking it to the underlying distributional characteristics of asset returns. Second, finite mixture models have been widely used in financial modeling but are rarely informed by external factors such as sentiment states. Third, the comparison between sentiment-driven mixture regression models and traditional regression approaches has not been thoroughly investigated.

This research contributes to the literature by addressing these gaps through the following innovations: utilizing financial news sentiment analyses to label return data and investigate its influence on return distributions. Designing a finite mixture distribution informed by sentiment states enables a more accurate characterization of heterogeneous return dynamics and integrating the proposed mixture distribution into a regression framework and comparing its performance with traditional normal regression models. By combining sentiment analyses, mixture modeling, and regression frameworks, this study aims to provide a comprehensive approach to modeling financial returns, thus advancing the current state of research at the intersection of financial econometrics and machine learning.

3. Data and sentiment analysis

In this study, we aim to model daily asset log-returns using a finite Markov mixture framework. The components of the Markov mixture model correspond to sentiment states of the market interpreted from reported financial news headlines.

By defining the sentiment state for each day and labeling daily log returns, we can convert incomplete return data into complete data, thereby simplifying calculations in the mixture framework. To this end, we collected a dataset of financial news headlines for the period from March 2022 to February 2025, comprising 347,471 distinct headlines covering various assets across different markets. The dataset comprises all companies in the Dow Jones Industrial Average (DJIA) index, the 10 largest companies in the NASDAQ index, the two largest cryptocurrencies, and Brent oil in the commodity market. In the cryptocurrency market, Bitcoin and Ethereum are selected as the two most significant assets. News headlines covering the entire cryptocurrency market are collected to analyze sentiment around these two assets. From the commodity market, Brent oil is selected, and all news headlines related to macroeconomics, monetary policies, and energy transportation topics are collected during the time period to evaluate news sentiment around Brent oil.

The headline collection procedure was asset-specific and performed as follows. For each of the 34 equity assets, news headlines were retrieved via the AlphaVantage News & Sentiments API using a ticker-specific query parameter (e.g., `tickers=AAPL` for Apple Inc.). This API endpoint returns articles that AlphaVantage has pre-tagged with the queried ticker symbol through its internal entity recognition

and contextual-tagging pipeline, thus ensuring that only headlines directly pertaining to the target company are included in the asset's corpus. For the two cryptocurrency assets, a market-wide collection strategy was adopted. Headlines were gathered using cryptocurrency-sector keywords alongside specific tickers, as sentiment in this market is predominantly systemic and typically reported at the aggregate level rather than strictly per currency. For Brent Oil, headlines were retrieved using topical keywords related to macroeconomics, monetary policies, and energy transportation, thus reflecting the main informational drivers of Brent price movements. Following collection, a preprocessing step removed API-generated prefixes, suffixes, and any non-headline text fragments to ensure that only genuine editorial content was forwarded to the FinBERT model for sentiment scoring.

To ensure the data-driven estimation of Markov transitions and mixture allocations remains robust, our sample requires a baseline density of news signals. However, we selected a broad spectrum of market attention assets ranging from highly covered assets to those with lower news flow. Robustness of our model over this diverse set of assets demonstrates its applicability across different informational environments. The full list of included assets and the precise number of collected headlines for each, sorted by volume within their respective markets, is provided in Table 1. On some days, there were no headlines for many companies, so we employed techniques to handle missing data, which are further discussed in the next sections. This dataset has been merged with financial information gathered from Yahoo Finance.

Table 1. Overview of individual assets and the corresponding number of news headlines collected from AlphaVantage, sorted by the number of news headlines within each market.

Market	Ticker	Headlines	Ticker	Headlines
Stock	MSFT	21,770	AXP	4,659
	GS	21,145	CVX	4,474
	TSLA	20,763	UNH	3,849
	GOOG	19,448	CSCO	3,834
	META	19,419	PG	3,690
	NVDA	18,704	DIS	3,675
	NFLX	18,342	V	3,354
	WMT	16,112	CAT	2,428
	AMZN	14,756	AMGN	2,248
	BA	11,806	HON	2,120
	NKE	7,488	IBM	2,075
	KO	7,357	MCD	2,069
	VZ	6,846	AAPL	1,241
	JNJ	6,697	MMM	1,017
	COST	6,340	SHW	715
	AVGO	5,126	TRV	694
	HD	5,048	CRM	251
Cryptocurrency	CRYPTO:BTC	7,187		
	CRYPTO:ETH	6,948		
Commodity	Brent Oil	63,776		
Total		347,471		

In this research, we utilized the FinBERT API integrated by the Huggingface community (Araci, 2019) to analyze the sentiment polarity of each news headline and to evaluate the corresponding sentiment scores and probabilities of each headline belonging to each sentiment polarity class. Before sentiment extraction, noise in the collected news headlines should be removed. There are different prefixes and suffixes in the headlines reported by AlphaVantage that should be removed, as they don't contribute any analytical value. Additionally, there may be some texts that are incorrectly detected as headlines due to extraneous noise collected during the process, which should be removed from the data.

The FinBERT model outputs three probability scores (positive, negative, and neutral) for each headline, thereby assigning it to the class with the highest probability. Based on these classifications, the daily sentiment state is determined through a two-step procedure. First, the dominant class between positively and negatively classified headlines defines the daily label. Neutral headlines carry no directional signal and do not alter this initial count. However, if there is no dominant class (a tie), then the average combined probability score of all headlines for that day (including neutral ones) is calculated. The day is labeled Negative if this overall average is less than zero, and Positive if it is greater than or equal to zero. Additionally, days with no available news are assigned to the Positive state, thus reflecting a calm market environment devoid of negative shocks. Therefore, the Positive state functions as a comprehensive non-negative category capturing positive, neutral, and no-news days.

A new binary variable S_t , defined as

$$S_t = \begin{cases} 1 \equiv \text{Negative sentiment,} \\ 2 \equiv \text{Positive sentiment (} \equiv \text{Non-Negative),} \end{cases}$$

represents the sentiment state at time t for a particular asset and determines the allocation of observations to each component of the mixture model. As indicated above, the positive sentiment includes both positive and neutral news, as well as all cases with no reported news.

4. Finite sentiment-based Markov mixture log return distributions

Numerous studies have modeled financial markets as Markov processes that switch between states (Haas et al., 2004; Ang & Timmermann, 2012; Guidolin & Timmermann, 2006; Hamilton, 1989; Dua & Tuteja, 2021). As we are aware of the market's different behaviors across states, once we identify and predict the market state, we can more accurately estimate future movements. The set of states can be defined on the basis of various factors that drive the market fluctuations. The prevalent criterion is market volatility, which has been investigated in the literature many times. However, the sentiment of news headlines is not considered a source of market shifts across different states, although it has a significant impact on market fluctuations. Unlike standard Hidden Markov Models (HMMs) in which state memberships and parameters are jointly inferred as latent variables, our framework treats the sentiment state S_t as an observable variable externally assigned via FinBERT. In our modeling framework, we first define the state sequence through an NLP analysis and subsequently estimate the state-conditional return distributions. This two-step identification strategy avoids the convergence issues typical of fully latent models and explicitly ties market regimes to observable news sentiment.

4.1. Two-state Markov chain model

In this section, we assume that the news sentiment follows a homogeneous Markov chain $\{S_t\}_{t \geq 0}$ with two sentiment states labeled as $1 \equiv \text{Negative}$ and $2 \equiv \text{Positive}$, so that S_t represents the market sentiment state at time t . This Markov chain is characterized by the 2-by-2 transition probability matrix $\mathbf{P} = \begin{bmatrix} 1 - p_1 & p_1 \\ p_2 & 1 - p_2 \end{bmatrix}$ with entries defined by the following probabilities:

$$p_1 = \mathbb{P}(S_t = 2 \mid S_{t-1} = 1) \text{ and } p_2 = \mathbb{P}(S_t = 1 \mid S_{t-1} = 2).$$

The state-transition diagram for this two-state Markov chain is shown in Figure 1.

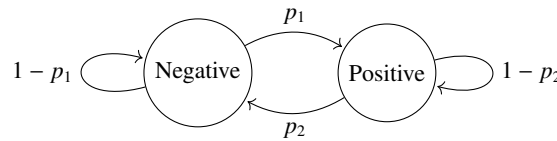


Figure 1. The state-transition diagram for the two-state Markov chain model.

To estimate the transition probabilities, we use the fact that the sojourn time T_i , which is defined as the time the Markov chain spends in state i before switching to another, follows a geometric probability distribution: $T_i \sim \text{Geom}(p_i)$, $i = 1, 2$. Thus, $\mathbb{E}[T_i] = 1/p_i$ holds. Using historical data, we compute the average sojourn time for each state: $\bar{T}_1$ for the negative state and $\bar{T}_2$ for the positive state. After that, we set the following:

$$\begin{aligned} \hat{p}_1 &= (\bar{T}_1)^{-1} \\ \hat{p}_2 &= (\bar{T}_2)^{-1} \end{aligned} \implies \hat{\mathbf{P}} = \begin{bmatrix} 1 - \hat{p}_1 & \hat{p}_1 \\ \hat{p}_2 & 1 - \hat{p}_2 \end{bmatrix}.$$

The stationary probability distribution, $\boldsymbol{\pi} = [\pi_1, \pi_2]$, is computed by solving $\boldsymbol{\pi} \hat{\mathbf{P}} = \boldsymbol{\pi}$. For example, the results for the Apple stock are as follows:

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.3728 & 0.6271 \\ 0.1247 & 0.8752 \end{bmatrix}, \quad \hat{\boldsymbol{\pi}} = [0.1659 \quad 0.8340].$$

4.2. Four-state Markov chain model

Here, we assume the state S_t at time t depends on the previous two states, S_{t-1} and S_{t-2} . This structural choice is motivated by the observation that financial markets are often not perfectly efficient, and the informational content of news sentiment may take time to be fully absorbed into asset prices. By modeling the transition dynamics over a two-day window, we effectively incorporate lagged sentiment into the framework without requiring traditional autoregressive terms. Our behavioral premise is that public news sentiment is generally reflected in asset prices within a 48-hour period. This model can be represented as a Markov chain with the following four states:

$$\begin{aligned} 1 &\equiv \text{Negative} + \text{Negative}, & 2 &\equiv \text{Negative} + \text{Positive}, \\ 3 &\equiv \text{Positive} + \text{Negative}, & 4 &\equiv \text{Positive} + \text{Positive}, \end{aligned}$$

where state $Y + T$ is composed of yesterday's sentiment state Y and today's sentiment state T .

Since the process cannot transition from state 1 to states 3 and 4, from state 2 to states 1 and 2, from state 3 to state 3 and 4, as well as from state 4 to states 1 and 2, the 4-by-4 transition probability matrix takes the following form:

$$\mathbf{P} = \begin{bmatrix} 1-p_1 & p_1 & 0 & 0 \\ 0 & 0 & p_2 & 1-p_2 \\ 1-p_3 & p_3 & 0 & 0 \\ 0 & 0 & p_4 & 1-p_4 \end{bmatrix},$$

with parameters $p_1, p_2, p_3, p_4 \in (0, 1)$. The state-transition diagram for the four-state Markov chain is shown in Figure 2.

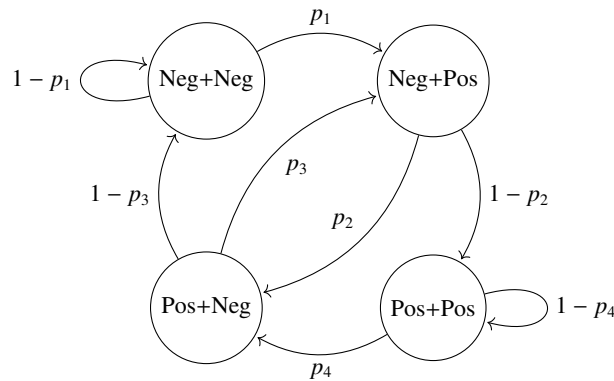


Figure 2. The state-transition diagram for the four-state Markov chain model.

The probabilities p_1 and p_4 can be estimated using the average sojourn times, $\bar{T}_1$ and $\bar{T}_4$, respectively, because when the process is in states 1 or 4, it either remains there or switches to another state. However, we cannot apply the same method to estimate the probabilities p_2 and p_3 , since when the process is in state 2 or 3, it definitely leaves that state at the next transition, thereby making the sojourn time in those states equal to 1. Therefore, to estimate the probabilities of transition from states 2 $\equiv$ Negative + Positive and 3 $\equiv$ Positive + Negative, we use the proportion of transitions originating from each state. The estimated probabilities are as follows:

$$\hat{p}_1 = \frac{1}{\bar{T}_1}, \quad \hat{p}_2 = \frac{N_{23}}{N_2}, \quad \hat{p}_3 = \frac{N_{32}}{N_3}, \quad \hat{p}_4 = \frac{1}{\bar{T}_4},$$

where N_{ij} is the total number of transitions from state i to state j , and N_i is the total number of transitions from state i to another state.

The stationary probability distribution for this model, $\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \pi_4]$, is computed by solving $\boldsymbol{\pi}\hat{\mathbf{P}} = \boldsymbol{\pi}$. The results for the Apple stock are as follows:

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.2954 & 0.7045 & 0 & 0 \\ 0 & 0 & 0.2027 & 0.7973 \\ 0.4189 & 0.5811 & 0 & 0 \\ 0 & 0 & 0.1120 & 0.8880 \end{bmatrix}, \quad \hat{\boldsymbol{\pi}} = [0.0612 \quad 0.1029 \quad 0.1029 \quad 0.7330].$$

4.3. Markov Mixture Log-Return distribution

We propose a Markov mixture distribution framework, where the market sentiment state, S_t , follows a Markov process with K states and transition probability matrix $\mathbf{P} = (p_{ij})_{i,j=\overline{1,K}}$. Therefore, the probability

distribution of asset log return, r_t , conditional on the state S_t , is given by one of the K distributions $\mathcal{D}(\theta_1), \mathcal{D}(\theta_2), \dots, \mathcal{D}(\theta_K)$:

$$r_t | \{S_t = k\} \sim \mathcal{D}(\theta_k).$$

The unconditional probability density function (PDF) of r_t can be written as a mixture of conditional PDFs as follows:

$$f(r_t; \theta, \pi) = \sum_{k=1}^K f(r_t; \theta | S_t = k) \mathbb{P}(S_t = k; \pi) = \sum_{k=1}^K f(r_t; \theta_k) \pi_k, \quad (1)$$

where $f(r; \theta_k)$ is the PDF of the distribution $\mathcal{D}(\theta_k)$, the vector of distribution parameters is $\theta = [\theta_1, \theta_2, \dots, \theta_K]$, and the Markov chain follows the stationary distribution with probabilities $\pi = [\pi_1, \pi_2, \dots, \pi_K]$ solving $\pi = \pi \mathbf{P}$.

An important advantage of this mixture formulation is its inherent ability to handle imbalanced distributions of sentiment states. In financial news, it is common for positive or calm market days to significantly outnumber days with severe negative shocks. In a standard modeling framework, such an imbalance might lead to classification bias or forecasting instability. However, in our Markov mixture framework, the empirical frequency of each state is intrinsically captured by the stationary probability distribution, π . Because the overall estimated distribution is a mixture weighted by these π_k values, the model dynamically adjusts to the frequency of each state.

We aim to determine whether the return dynamics follow different distributions across states, identify the best-fitting parametric distribution for each state, and estimate the distribution parameters for each state. Three different probability distributions, namely, the Gaussian (normal), Laplace, and Student's t distributions, are investigated in this study. They are known to be the best fit for stock returns in the literature (Choi & Yoon, 2020). The parameters are estimated by maximizing the likelihood functions, using closed-form solutions for the Gaussian and Laplace distributions, and a numerical approach for the t -distribution. Additionally, the unconditional (mixture) distribution is computed as a linear combination of different conditional distributions.

Moreover, we estimated the parameters of each underlying distribution, thereby assuming that log returns are not influenced by the sentiment Markov process S_t . Here, the resulting distribution is called the *overall distribution*. The calculation results for the Apple stock are summarized in Table 2.

Table 2. Estimated daily parameters for conditional distributions for each sentiment state, as well as for the overall distributions for the entire sample without considering sentiment states for the Apple stock. Here, $\hat{\mu}$ is the estimated location, $\hat{\sigma}$ is the estimated scale, and $\hat{\nu}$ is the estimated number of degrees of freedom.

Distribution	Parameter	Negative State	Positive State	Overall
Gaussian	$\hat{\mu}$	−0.001127	0.005023	0.000684
	$\hat{\sigma}$	0.018763	0.017031	0.015582
Laplace	$\hat{\mu}$	−0.000702	0.003683	0.0000003
	$\hat{\sigma}$	0.012984	0.023932	0.009243
	$\hat{\nu}$	3.228969	4.470832	3.279459
t-Distribution	$\hat{\mu}$	−0.000518	0.004251	0.001214
	$\hat{\sigma}$	0.012278	0.012035	0.012063

To select the appropriate distribution, the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1939) was applied to test different possible null hypotheses. It is a non-parametric test used to compare a sample distribution with a theoretical probability distribution (one-sample KS test) or to compare two sample distributions (two-sample KS test) and to check whether two samples came from the same (but unknown) probability distributions (Remillard, 2016). The results of the KS test are summarized in Table 3.

Table 3. Results of the Kolmogorov–Smirnov test for various null hypotheses concerning the dependence of the log-return distribution on sentiment state for the apple stock.

KS null hypothesis (H_0)	KS statistic	Result	Interpretation
Negative state returns follow the Gaussian distribution	0.4720	Reject H_0	Negative state returns don't follow the Gaussian distribution
Positive state returns follow the Gaussian distribution	0.4814	Reject H_0	Positive state returns don't follow the Gaussian distribution
The entire sample follows the Gaussian distribution	0.0761	Reject H_0	The entire sample does not follow the Gaussian distribution
Negative state returns follow the Laplace distribution	0.4673	Reject H_0	Negative state does not follow the Laplace distribution
Positive state returns follow the Laplace distribution	0.4779	Reject H_0	Positive state does not follow the Laplace distribution
The entire sample follows the Laplace distribution	0.0628	Reject H_0	The entire sample does not follow the Laplace distribution
Negative state returns follow the t-distribution	0.0214	Fail to reject H_0	Negative state returns follow the t-distribution
Positive state returns follow the t-distribution	0.0360	Fail to reject H_0	Positive state returns follow the t-distribution
The entire sample follows the t-distribution.	0.0184	Fail to reject H_0	The entire sample follows the t-distribution
Negative and Positive states returns are drawn from the same distribution	0.1571	Reject H_0	Negative and Positive states returns are drawn from different distributions

The results in Table 2 and Table 3 indicate that the distribution's parameters depend on the sentiment state. Moreover, the Gaussian and Laplace distributions do not provide a good fit to the conditional distributions (for both negative and positive states) or to the overall distribution. However, the Student's t-distribution works very well for each of the negative and positive states, as well as for the entire sample. Clearly, the log returns for the negative and positive states are drawn from different distributions.

The Kulback–Leibler (KL) divergence (Kullback & Leibler, 1951) is computed for pairs of fitted and empirical distributions to assess how they diverge from each other. Table 4 presents the KL score for each pair of distributions. The unconditional (mixture) distributions in Table 4 are computed by (1). An analysis of KL divergence scores indicates that the t-distribution is the most suitable model for the Apple stock returns data set, outperforming both the Gaussian and Laplace distributions. This finding highlights the importance of selecting an appropriate probability model that accurately captures the data's characteristics in each sentiment state.

Table 4. KL divergence scores for pairs of fitted and empirical sentiment-state-dependent log return distributions calculated for the apple stock.

Pair of distributions	D_{KL}
Unconditional Gaussian distribution vs. empirical distribution	0.006087
Unconditional Laplace distribution vs. empirical distribution	0.006062
Unconditional t -distribution vs. empirical distribution	0.000303

Figure 3 illustrates how different fitted distributions capture the variations in the sample histogram.

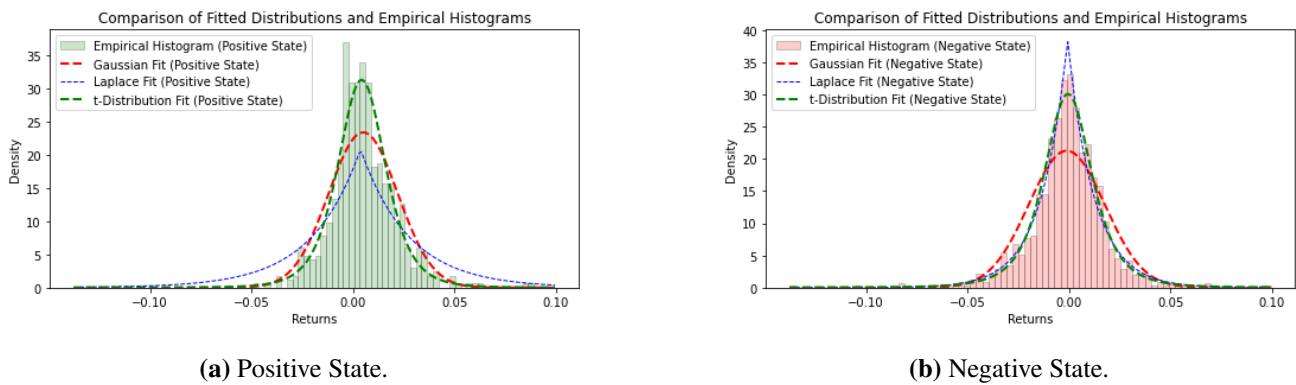


Figure 3. Comparison of fitted distributions for log returns in different states calculated for the Apple stock.

5. Finite Sentiment-Adjusted Markov-Switching FIGARCH (SA-FIGARCH) model

In this section, we have enhanced the volatility prediction model FIGARCH(1,1) by incorporating sentiment states, allowing the innovation term to follow a Markov switching t -distribution with parameters that depend on the sentiment state. Based on the Markov chain model constructed in Section 4, we propose a Markov switching return prediction time series model that assumes a distinct Student's t -distribution for residuals in each sentiment state. The model is described as follows:

$$\begin{aligned}
 r_t | S_t &= \mu_{S_t} + \epsilon_t, \\
 \epsilon_t &= \sigma_t \cdot z_t, \quad z_t \sim t(0, 1, \nu_{S_t}), \\
 \sigma_t^2 &= \omega_{S_t} + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2 + \left(\sum_{i=1}^q \gamma_i \epsilon_{t-i}^2 \right) (1 - L)^d,
 \end{aligned} \tag{2}$$

where r_t is log-return at time t , S_t is the latent sentiment state variable (Markov Sentiment State), μ_{S_t} is the expected return in state S_t , ϵ_t is the shock term at time t , σ_t^2 is the conditional variance at time t , $z_t \sim t(0, 1, \nu_{S_t})$ is a standardized Student's t -distributed innovation term with mean 0, variance 1, and ν_{S_t} degrees of freedom, ω_{S_t} is the constant term in the variance equation for state S_t , α is the impact of past squared shocks (ARCH term), β is the persistence of past volatility (GARCH term), L is the lag operator, and d is the fractional differencing parameter.

To make the model more realistic and practical, we assume that only sentiment information or news headlines from previous days are available, so our return prediction at time t should only rely on

information up to time $t - 1$. Therefore, the conditional PDF of r_t given the state S_{t-1} can be written as follows:

$$f_{r_t|S_{t-1}}(r|k) = \sum_j f_{r_t|S_t}(r|j) \mathbb{P}(S_t = j | S_{t-1} = k) = \sum_j f_{r_t|S_t}(r|j) p_{kj}.$$

By averaging with respect to the sentiment state at time t , we can write the predicted value of r_t given the state S_{t-1} as an average of the predicted values of r_t conditional on S_t :

$$r_t | \{S_{t-1} = k\} = \sum_{j=1}^K p_{kj} r_t | \{S_t = j\}. \quad (3)$$

Return prediction time series models typically assume that error terms follow a normal distribution with stationary deviations, thereby allowing the use of the ordinary least squares (OLS) method to estimate parameters (Baissa & Rainey, 2020). This can be an unrealistic assumption, as observed in Section 4.3, since fat-tailed distributions may provide a better fit to the data. Additionally, Markov-switching return-prediction models in the literature have not considered market sentiment, as inferred from financial news, as a dynamic factor in the Markov process.

6. Empirical results

6.1. News sentiment Markov process parameters

In this study, we tested the proposed framework on 37 different assets from various markets. We used 90% of the samples for training and the remaining 10% to test the models. The sentiment analysis and labeling of each day is conducted using the FinBERT library as discussed in Section 3. The transition probability matrix and stationary probability distributions for Markov chains with two and four states are calculated as described in Sections 4.1 and 4.2, respectively. The results for the two-state approach are summarized in Table 5, and those for the four-state framework are summarized in Table 6.

Based on the results in Table 5, on average, most assets tend to switch from a negative state to a positive one and remain in that state; additionally, there is a higher long-term probability that they will remain in a positive state. However, this is a little different for TSLA, GS, and BA. Those tend to persist in a negative state if the current state is negative. Additionally, in the long term, the probability of being in a negative state for these three companies is higher than average.

The results in Table 6, which only include non-zero probabilities, indicate that, on average, when assets experience two consecutive positive days based on sentiment extracted from news, they tend to have one negative day before returning to positive. Conversely, when assets have two consecutive negative days, they are more likely to continue exhibiting negative sentiment. Furthermore, over the long term, negative days tend to be more persistent and are more likely to be followed by additional negative days.

Table 5. Two-state Markov model parameters and stationary probabilities.

Ticker	p_{11}	p_{12}	p_{21}	p_{22}	π_1	π_2
AAPL	0.3729	0.6271	0.1248	0.8752	0.1660	0.8340
TRV	0.3760	0.6240	0.1317	0.8683	0.1742	0.8258
IBM	0.3248	0.6752	0.1293	0.8707	0.1607	0.8393
MSFT	0.3226	0.6774	0.0632	0.9368	0.0853	0.9147
MCD	0.5118	0.4882	0.2000	0.8000	0.2906	0.7094
AVGO	0.2739	0.7261	0.1990	0.8010	0.2151	0.7849
CRM	0.3571	0.6429	0.2813	0.7188	0.3043	0.6957
JNJ	0.2636	0.7364	0.1600	0.8400	0.1785	0.8215
HD	0.5065	0.4935	0.3675	0.6325	0.4268	0.5732
MMM	0.4468	0.5532	0.1310	0.8690	0.1914	0.8086
GOOG	0.4048	0.5952	0.0371	0.9629	0.0587	0.9413
NVDA	0.4094	0.5906	0.1544	0.8456	0.2072	0.7928
CSCO	0.3374	0.6626	0.1912	0.8088	0.2239	0.7761
KO	0.2484	0.7516	0.2127	0.7873	0.2205	0.7795
META	0.2258	0.7742	0.0389	0.9611	0.0478	0.9522
CAT	0.3096	0.6904	0.2556	0.7444	0.2702	0.7298
CVX	0.3424	0.6576	0.3558	0.6442	0.3511	0.6489
GS	0.5618	0.4382	0.2784	0.7216	0.3885	0.6115
VZ	0.2153	0.7847	0.1945	0.8055	0.1987	0.8013
PG	0.3254	0.6746	0.2712	0.7288	0.2867	0.7133
WMT	0.4403	0.5597	0.2802	0.7198	0.3336	0.6664
AMGN	0.3767	0.6233	0.2612	0.7388	0.2953	0.7047
V	0.2932	0.7068	0.2495	0.7505	0.2609	0.7391
TSLA	0.7037	0.2963	0.1725	0.8275	0.3679	0.6321
NFLX	0.4398	0.5602	0.2729	0.7271	0.3276	0.6724
NKE	0.4647	0.5353	0.4617	0.5383	0.4631	0.5369
COST	0.4103	0.5897	0.2771	0.7229	0.3197	0.6803
BA	0.5233	0.4767	0.3318	0.6682	0.4104	0.5896
SHW	0.3525	0.6475	0.1308	0.8692	0.1680	0.8320
AMZN	0.4500	0.5500	0.2068	0.7932	0.2732	0.7268
AXP	0.3147	0.6853	0.2519	0.7481	0.2688	0.7312
HON	0.2642	0.7358	0.1270	0.8730	0.1472	0.8528
UNH	0.5000	0.5000	0.3103	0.6897	0.3829	0.6171
DIS	0.4846	0.5154	0.3447	0.6553	0.4008	0.5992
CRYPTO:BTC	0.5000	0.5000	0.1069	0.8931	0.1761	0.8239
CRYPTO:ETH	0.3452	0.6548	0.2333	0.7667	0.2627	0.7373
BRENT	0.4197	0.5803	0.2063	0.7937	0.2622	0.7378

Table 6. Four-state Markov model parameters and stationary probabilities.

Ticker	p_{11}	p_{12}	p_{23}	p_{24}	p_{31}	p_{32}	p_{43}	p_{44}	π_1	π_2	π_3	π_4
AAPL	0.2955	0.7045	0.2027	0.7973	0.4189	0.5811	0.1120	0.8880	0.0612	0.1029	0.1029	0.7330
TRV	0.2340	0.7660	0.1923	0.8077	0.4615	0.5385	0.1226	0.8774	0.0656	0.1088	0.1088	0.7168
IBM	0.2368	0.7632	0.1392	0.8608	0.3671	0.6329	0.1278	0.8722	0.0522	0.1085	0.1085	0.7308
MSFT	0.3333	0.6667	0.1905	0.8095	0.3171	0.6829	0.0546	0.9454	0.0275	0.0578	0.0578	0.8570
MCD	0.4587	0.5413	0.2427	0.7573	0.5686	0.4314	0.1893	0.8107	0.1490	0.1418	0.1418	0.5673
AVGO	0.2273	0.7727	0.1930	0.8070	0.2920	0.7080	0.1983	0.8017	0.0586	0.1551	0.1551	0.6312
CRM	0.0000	1.0000	0.5000	0.5000	0.5556	0.4444	0.2083	0.7917	0.1121	0.2018	0.2018	0.4843
JNJ	0.3529	0.6471	0.1263	0.8737	0.2316	0.7684	0.1663	0.8337	0.0470	0.1314	0.1314	0.6902
HD	0.5541	0.4459	0.4118	0.5882	0.4575	0.5425	0.3383	0.6617	0.2154	0.2099	0.2099	0.3649
MMM	0.3125	0.6875	0.1818	0.8182	0.5714	0.4286	0.1233	0.8767	0.0878	0.1056	0.1056	0.7010
GOOG	0.2778	0.7222	0.1600	0.8400	0.5000	0.5000	0.0324	0.9676	0.0242	0.0349	0.0349	0.9059
NVDA	0.4516	0.5484	0.3636	0.6364	0.3793	0.6207	0.1162	0.8838	0.0847	0.1224	0.1224	0.6705
CSCO	0.4000	0.6000	0.2315	0.7685	0.3056	0.6944	0.1794	0.8206	0.0750	0.1472	0.1472	0.6306
KO	0.3000	0.7000	0.2645	0.7355	0.2314	0.7686	0.1964	0.8036	0.0544	0.1646	0.1646	0.6164
META	0.0000	1.0000	0.2500	0.7500	0.2917	0.7083	0.0304	0.9696	0.0108	0.0370	0.0370	0.9151
CAT	0.2787	0.7213	0.2647	0.7353	0.3235	0.6765	0.2500	0.7500	0.0832	0.1855	0.1855	0.5457
CVX	0.3068	0.6932	0.3491	0.6509	0.3669	0.6331	0.3595	0.6405	0.1219	0.2304	0.2304	0.4172
GS	0.6038	0.3962	0.3952	0.6048	0.5081	0.4919	0.2308	0.7692	0.2172	0.1694	0.1694	0.4440
VZ	0.1613	0.8387	0.2281	0.7719	0.2301	0.7699	0.1864	0.8136	0.0428	0.1559	0.1559	0.6454
PG	0.4118	0.5882	0.2766	0.7234	0.2837	0.7163	0.2691	0.7309	0.0933	0.1934	0.1934	0.5199
WMT	0.4579	0.5421	0.3309	0.6691	0.4265	0.5735	0.2606	0.7394	0.1469	0.1868	0.1868	0.4795
AMGN	0.3704	0.6296	0.3134	0.6866	0.3806	0.6194	0.2427	0.7573	0.1113	0.1841	0.1841	0.5206
V	0.3036	0.6964	0.2889	0.7111	0.2889	0.7111	0.2365	0.7635	0.0765	0.1844	0.1844	0.5546
TSLA	0.7105	0.2895	0.4177	0.5823	0.7000	0.3000	0.1214	0.8786	0.2624	0.1085	0.1085	0.5206
NFLX	0.5234	0.4766	0.3358	0.6642	0.3806	0.6194	0.2493	0.7507	0.1462	0.1831	0.1831	0.4877
NKE	0.4843	0.5157	0.4917	0.5083	0.4475	0.5525	0.4360	0.5640	0.2151	0.2479	0.2479	0.2890
COST	0.3711	0.6289	0.3333	0.6667	0.4380	0.5620	0.2556	0.7444	0.1313	0.1885	0.1885	0.4917
BA	0.6369	0.3631	0.3636	0.6364	0.4056	0.5944	0.3160	0.6840	0.2177	0.1949	0.1949	0.3925
SHW	0.1395	0.8605	0.1899	0.8101	0.4684	0.5316	0.1219	0.8781	0.0592	0.1088	0.1088	0.7231
AMZN	0.5495	0.4505	0.3545	0.6455	0.3670	0.6330	0.1659	0.8341	0.1215	0.1491	0.1491	0.5803
AXP	0.3651	0.6349	0.2761	0.7239	0.2985	0.7015	0.2437	0.7563	0.0864	0.1838	0.1838	0.5460
HON	0.2500	0.7500	0.1538	0.8462	0.2692	0.7308	0.1232	0.8768	0.0389	0.1084	0.1084	0.7444
UNH	0.5286	0.4714	0.3381	0.6619	0.4786	0.5214	0.2977	0.7023	0.1938	0.1909	0.1909	0.4244
DIS	0.5455	0.4545	0.3974	0.6026	0.4267	0.5733	0.3171	0.6829	0.1940	0.2066	0.2066	0.3928
CRYPTO:BTC	0.4815	0.5185	0.2963	0.7037	0.5185	0.4815	0.0851	0.9149	0.0887	0.0887	0.0887	0.7338
CRYPTO:ETH	0.2759	0.7241	0.2182	0.7818	0.3818	0.6182	0.2378	0.7622	0.0907	0.1720	0.1720	0.5653
BRENT	0.5122	0.4878	0.3125	0.6875	0.3514	0.6486	0.1787	0.8213	0.1097	0.1522	0.1522	0.5859

6.2. Return forecasting models

Full information model

To evaluate the performance of the sentiment labeling process and the practicality of the proposed prediction model, we tested the model under simplifying assumptions. Specifically, we assumed that the sentiment state at each time t (the period for which we aim to predict the log return) is deterministically determined from news headlines using the sentiment analysis and labeling procedure described in Sections 3 and 4. Once the sentiment state process S_t is derived from the daily news headlines, we predict the log return of that day using the SA-FIGARCH model (2).

The model's performance is assessed using three metrics: root mean squared error (RMSE), mean absolute error (MAE), and directional accuracy (DA). These metrics evaluate how well the model accurately predicts test samples. Based on the results in Table 7, the SA-FIGARCH model demonstrates strong predictive performances for nearly all assets in our study, with consistently low RMSE and MAE values. Moreover, it achieves high DA for most assets, exceeding 80% for four of them. These findings suggest that the model is both effective and feasible, making it suitable for comparison with other models or with alternative assumptions of the same model in future analyses.

Table 7. Prediction results of the full information model that assumes the sentiment state of each day is identified beforehand through collective news headlines sentiment before predicting the Open-to-Close log return of that day.

Ticker	RMSE	MAE	DA	Ticker	RMSE	MAE	DA
MSFT	0.012	0.010	0.507	UNH	0.019	0.015	0.452
GS	0.016	0.011	0.521	CSCO	0.012	0.009	0.521
TSLA	0.034	0.027	0.562	PG	0.011	0.009	0.507
GOOG	0.017	0.014	0.486	DIS	0.017	0.013	0.425
META	0.021	0.017	0.462	V	0.011	0.008	0.507
NVDA	0.030	0.022	0.458	CAT	0.016	0.012	0.493
NFLX	0.020	0.015	0.479	AMGN	0.018	0.013	0.479
WMT	0.012	0.010	0.521	HON	0.014	0.011	0.589
AMZN	0.017	0.014	0.534	IBM	0.014	0.010	0.397
BA	0.022	0.018	0.493	MCD	0.011	0.009	0.507
NKE	0.018	0.014	0.466	AAPL	0.013	0.010	0.486
KO	0.011	0.009	0.493	MMM	0.019	0.014	0.507
VZ	0.015	0.011	0.548	SHW	0.016	0.012	0.411
JNJ	0.012	0.010	0.534	TRV	0.014	0.011	0.528
COST	0.012	0.009	0.616	CRM	0.030	0.024	0.600
AVGO	0.027	0.020	0.562	CRYPTO:BTC	0.029	0.023	0.375
HD	0.014	0.011	0.507	CRYPTO:ETH	0.038	0.029	0.531
AXP	0.015	0.012	0.397	BRENT	0.021	0.018	0.459
CVX	0.014	0.011	0.493				

Average SA-FIGARCH model for log returns

This model is based on the estimated Markov processes for news sentiment state and the SA-FIGARCH model introduced in Section 5. To make practical and realistic predictions for log returns r_t at time t , we assume that information, including news headlines, is available up to time $t - 1$. Given S_{t-1} sentiment state at time $t - 1$ and Markov Process probabilities, the predicted value of r_t is computed as follows:

$$r_t | \{S_{t-1} = k\} = \sum_{j=1}^K p_{kj} \cdot (r_t | \{S_t = j\}). \quad (4)$$

If there is no related news available at time $t - 1$ for a particular asset, stationary probabilities π are used to compute the predicted log return as follows:

$$r_t = \sum_{j=1}^K \pi_j \cdot (r_t | \{S_t = j\}). \quad (5)$$

Logistic regression predicted state model

Additionally, we trained a logistic regression (LR) model that implements news headlines frequency variables up to time $t - 1$ to classify the sentiment state at time t . This model is fitted for comparison with our proposed Markov-switching SA-FIGARCH model. Included variables in this model are as *number of news headlines in each positive, negative, and neutral class*, plus *average sentiment score of positive, negative, and neutral headlines*, and also *sentiment score of the combination of all headlines*. The FinBERT library provides all the variables mentioned in this section. Given $\hat{S}_t$, the sentiment state estimated by the LR model, r_t , is calculated as follows:

$$\hat{r}_t | \{\hat{S}_t = k\} = \mathbb{E}[r_t | S_t = k], \quad k = 1, 2, \dots, K, \quad (6)$$

where $\mathbb{E}[r_t | S_t = k]$ is the expected return given that the sentiment state is k using a two- or four-state model discussed in Sections 4.1 and 4.2.

Additionally, we fitted the FIGARCH model introduced by Baillie et al. (1996) as a benchmark model. All prediction models are examined with three types of log returns as follows:

$$r_{OC,t} = \ln\left(\frac{C_t}{O_t}\right), \quad r_{CC,t} = \ln\left(\frac{C_t}{C_{t-1}}\right), \quad r_{OO,t} = \ln\left(\frac{O_{t+1}}{O_t}\right),$$

which are named Open-to-Close, Close-to-Close, and Open-to-Open log returns, respectively. The results of all prediction models for Open-to-Close log returns are summarized in Table 8, for Close-to-Close log returns are summarized in Table 9, and for Open-to-Open log returns are summarized in Table 10.

Based on the results in Tables 8, 9, and 10 included in Appendix ??, for the Open-to-Close log returns, the FIGARCH benchmark model performs slightly better than the other models in terms of RMSE and MAE, while the four-state SA-FIGARCH model performs the best according to the DA criterion. For the Close-to-Close log returns, on average, the four-state SA-FIGARCH model outperforms the other models across all criteria. For the Open-to-Open log returns, on average, the logistic regression state-predicted model performs better than the other models in terms of MAE and DA, while the four-state SA-FIGARCH model performs best in terms of RMSE.

Table 8. Prediction results of two-state SA-FIGARCH model (2State), four-state SA-FIGARCH model (4State), logistic regression state predicted model (LR), and FIGARCH model for **Open-to-Close** log returns. The best value for each asset is highlighted.

Ticker	RMSE				MAE				DA			
	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH
MSFT	0.012	0.013	0.013	0.014	0.010	0.011	0.010	0.011	0.616	0.562	0.616	0.425
GS	0.016	0.015	0.017	0.017	0.012	0.012	0.012	0.013	0.521	0.507	0.507	0.466
TSLA	0.035	0.032	0.032	0.036	0.028	0.025	0.026	0.028	0.466	0.562	0.507	0.466
GOOG	0.015	0.016	0.017	0.016	0.012	0.012	0.013	0.012	0.528	0.500	0.479	0.611
META	0.019	0.018	0.020	0.019	0.015	0.015	0.015	0.014	0.585	0.431	0.462	0.400
NVDA	0.027	0.029	0.024	0.028	0.021	0.023	0.019	0.022	0.542	0.403	0.583	0.403
NFLX	0.020	0.020	0.018	0.019	0.015	0.014	0.013	0.015	0.370	0.548	0.507	0.466
WMT	0.011	0.010	0.012	0.011	0.009	0.008	0.010	0.009	0.452	0.411	0.384	0.507
AMZN	0.017	0.018	0.016	0.017	0.013	0.014	0.012	0.014	0.521	0.466	0.589	0.452
BA	0.022	0.021	0.021	0.021	0.018	0.017	0.017	0.017	0.479	0.479	0.534	0.438
NKE	0.015	0.013	0.015	0.014	0.012	0.011	0.012	0.012	0.534	0.562	0.521	0.575
KO	0.010	0.009	0.009	0.009	0.009	0.007	0.007	0.008	0.479	0.589	0.521	0.438
VZ	0.013	0.011	0.012	0.012	0.010	0.009	0.010	0.009	0.384	0.562	0.425	0.507
JNJ	0.012	0.011	0.012	0.011	0.009	0.009	0.009	0.009	0.507	0.452	0.534	0.493
COST	0.012	0.011	0.011	0.012	0.009	0.009	0.008	0.009	0.521	0.493	0.562	0.479
AVGO	0.024	0.022	0.026	0.025	0.018	0.017	0.019	0.019	0.507	0.562	0.521	0.370
HD	0.013	0.013	0.014	0.014	0.010	0.011	0.011	0.011	0.466	0.521	0.397	0.411
AXP	0.014	0.014	0.013	0.012	0.011	0.011	0.010	0.010	0.562	0.507	0.575	0.479
CVX	0.012	0.011	0.013	0.011	0.009	0.009	0.010	0.009	0.507	0.548	0.466	0.589
UNH	0.018	0.018	0.017	0.017	0.013	0.013	0.013	0.013	0.562	0.507	0.438	0.562
CSCO	0.011	0.011	0.011	0.010	0.009	0.008	0.009	0.008	0.479	0.521	0.548	0.534
PG	0.010	0.011	0.010	0.009	0.008	0.008	0.008	0.007	0.521	0.548	0.575	0.644
DIS	0.012	0.014	0.013	0.014	0.010	0.011	0.010	0.011	0.466	0.452	0.479	0.438
V	0.010	0.011	0.010	0.010	0.008	0.009	0.008	0.008	0.589	0.479	0.466	0.575
CAT	0.015	0.014	0.014	0.016	0.012	0.011	0.011	0.013	0.452	0.493	0.507	0.452
AMGN	0.017	0.017	0.017	0.017	0.012	0.011	0.012	0.012	0.493	0.452	0.603	0.534
HON	0.015	0.013	0.015	0.012	0.011	0.010	0.012	0.010	0.548	0.534	0.411	0.534
IBM	0.013	0.012	0.013	0.012	0.010	0.008	0.010	0.008	0.493	0.658	0.479	0.562
MCD	0.010	0.011	0.011	0.010	0.008	0.009	0.008	0.008	0.575	0.384	0.534	0.534
AAPL	0.012	0.012	0.015	0.012	0.010	0.009	0.011	0.010	0.597	0.542	0.500	0.556
MMM	0.015	0.015	0.016	0.016	0.012	0.011	0.012	0.012	0.479	0.562	0.603	0.397
SHW	0.015	0.014	0.015	0.014	0.011	0.010	0.012	0.012	0.507	0.603	0.411	0.466
TRV	0.016	0.016	0.016	0.015	0.012	0.012	0.012	0.012	0.514	0.528	0.403	0.417
CRM	0.021	0.038	0.036	0.025	0.018	0.028	0.030	0.023	0.600	0.800	0.600	0.600
CRYPTO:BTC	0.027	0.025	0.032	0.026	0.021	0.021	0.026	0.021	0.438	0.688	0.484	0.375
CRYPTO:ETH	0.037	0.036	0.041	0.034	0.032	0.028	0.033	0.026	0.438	0.656	0.344	0.500
BRENT	0.020	0.020	0.018	0.020	0.016	0.015	0.014	0.017	0.432	0.541	0.575	0.473
Average	0.017	0.017	0.017	0.017	0.013	0.013	0.013	0.013	0.506	0.530	0.504	0.490
# of wins	6	14	6	11	8	10	10	10	7	12	12	5

Table 9. Prediction results of two-state SA-FIGARCH model (2State), four-state SA-FIGARCH model (4State), logistic regression state predicted model (LR), and FIGARCH model for **Close-to-Close** log returns. The best value for each asset is highlighted.

Ticker	RMSE				MAE				DA			
	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH
MSFT	0.018	0.016	0.018	0.016	0.013	0.012	0.013	0.012	0.493	0.575	0.438	0.479
GS	0.022	0.022	0.023	0.021	0.014	0.014	0.015	0.013	0.438	0.562	0.425	0.589
TSLA	0.049	0.055	0.053	0.050	0.034	0.041	0.041	0.037	0.548	0.479	0.466	0.562
GOOG	0.019	0.020	0.020	0.021	0.015	0.015	0.016	0.016	0.514	0.486	0.535	0.542
META	0.021	0.022	0.023	0.019	0.016	0.017	0.017	0.015	0.492	0.462	0.508	0.600
NVDA	0.029	0.031	0.032	0.033	0.022	0.025	0.026	0.026	0.653	0.514	0.472	0.389
NFLX	0.025	0.023	0.025	0.026	0.018	0.016	0.017	0.017	0.507	0.521	0.452	0.548
WMT	0.013	0.011	0.014	0.013	0.010	0.009	0.011	0.011	0.521	0.575	0.479	0.507
AMZN	0.022	0.019	0.022	0.022	0.018	0.016	0.017	0.017	0.479	0.575	0.507	0.548
BA	0.022	0.022	0.024	0.023	0.018	0.018	0.020	0.018	0.493	0.507	0.466	0.562
NKE	0.016	0.017	0.017	0.016	0.013	0.014	0.014	0.014	0.507	0.479	0.466	0.534
KO	0.014	0.012	0.012	0.012	0.010	0.009	0.009	0.009	0.452	0.507	0.521	0.493
VZ	0.016	0.015	0.016	0.014	0.012	0.011	0.012	0.010	0.411	0.452	0.438	0.548
JNJ	0.013	0.013	0.012	0.012	0.010	0.011	0.010	0.009	0.479	0.438	0.616	0.466
COST	0.013	0.014	0.012	0.012	0.011	0.011	0.010	0.010	0.479	0.507	0.479	0.521
AVGO	0.045	0.047	0.047	0.047	0.028	0.029	0.027	0.029	0.493	0.479	0.603	0.507
HD	0.015	0.016	0.015	0.014	0.012	0.013	0.012	0.011	0.548	0.411	0.507	0.589
AXP	0.018	0.019	0.018	0.018	0.013	0.014	0.014	0.014	0.521	0.534	0.562	0.521
CVX	0.014	0.016	0.016	0.015	0.011	0.012	0.012	0.011	0.548	0.521	0.466	0.507
UNH	0.024	0.025	0.024	0.026	0.017	0.017	0.017	0.019	0.548	0.534	0.493	0.479
CSCO	0.014	0.014	0.013	0.013	0.010	0.010	0.010	0.009	0.548	0.466	0.479	0.589
PG	0.012	0.013	0.011	0.012	0.009	0.010	0.008	0.009	0.562	0.466	0.589	0.562
DIS	0.016	0.015	0.016	0.017	0.010	0.011	0.012	0.012	0.493	0.438	0.493	0.521
V	0.013	0.012	0.012	0.013	0.010	0.009	0.009	0.009	0.452	0.575	0.438	0.493
CAT	0.020	0.019	0.021	0.019	0.015	0.014	0.015	0.014	0.438	0.562	0.521	0.562
AMGN	0.021	0.018	0.019	0.016	0.014	0.013	0.014	0.012	0.479	0.356	0.575	0.534
HON	0.015	0.015	0.013	0.015	0.011	0.011	0.010	0.012	0.575	0.507	0.534	0.534
IBM	0.022	0.021	0.022	0.021	0.013	0.013	0.014	0.013	0.575	0.548	0.466	0.425
MCD	0.013	0.013	0.014	0.013	0.010	0.010	0.010	0.009	0.507	0.425	0.548	0.521
AAPL	0.016	0.017	0.016	0.015	0.013	0.013	0.013	0.012	0.403	0.528	0.472	0.500
MMM	0.019	0.016	0.019	0.018	0.014	0.012	0.014	0.014	0.384	0.521	0.466	0.493
SHW	0.018	0.016	0.019	0.016	0.014	0.012	0.015	0.012	0.438	0.534	0.466	0.534
TRV	0.019	0.019	0.019	0.018	0.013	0.014	0.013	0.012	0.514	0.444	0.500	0.556
CRM	0.028	0.035	0.026	0.036	0.025	0.028	0.024	0.028	0.200	0.600	0.400	0.400
CRYPTO:BTC	0.031	0.027	0.033	0.029	0.024	0.022	0.024	0.023	0.469	0.625	0.613	0.531
CRYPTO:ETH	0.032	0.042	0.043	0.037	0.026	0.035	0.036	0.027	0.656	0.469	0.406	0.531
BRENT	0.018	0.021	0.021	0.020	0.014	0.016	0.017	0.016	0.554	0.459	0.452	0.405
Average	0.020	0.021	0.021	0.020	0.015	0.016	0.016	0.015	0.498	0.507	0.489	0.519
# of wins	11	8	5	13	9	12	4	12	7	9	6	15

Table 10. Prediction results of two-state SA-FIGARCH model (2State), four-state SA-FIGARCH model (4State), logistic regression state predicted model (LR), and FIGARCH model for **Open-to-Open** log returns. The best value for each asset is highlighted.

Ticker	RMSE				MAE				DA			
	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH	2State	4State	LR	FIGARCH
MSFT	0.018	0.017	0.018	0.015	0.013	0.013	0.014	0.011	0.548	0.562	0.534	0.466
GS	0.019	0.019	0.020	0.020	0.013	0.013	0.015	0.014	0.507	0.562	0.575	0.562
TSLA	0.056	0.054	0.054	0.052	0.043	0.043	0.040	0.039	0.425	0.397	0.507	0.493
GOOG	0.022	0.022	0.024	0.022	0.018	0.016	0.018	0.017	0.431	0.444	0.465	0.444
META	0.023	0.020	0.023	0.020	0.018	0.015	0.018	0.017	0.477	0.477	0.462	0.462
NVDA	0.038	0.037	0.034	0.031	0.032	0.029	0.027	0.026	0.500	0.458	0.458	0.486
NFLX	0.024	0.028	0.028	0.027	0.017	0.018	0.020	0.019	0.534	0.534	0.479	0.466
WMT	0.013	0.013	0.013	0.012	0.010	0.010	0.010	0.009	0.466	0.452	0.589	0.521
AMZN	0.022	0.018	0.023	0.020	0.017	0.014	0.018	0.016	0.466	0.603	0.575	0.479
BA	0.022	0.022	0.025	0.022	0.017	0.018	0.020	0.018	0.575	0.521	0.466	0.466
NKE	0.018	0.018	0.017	0.017	0.015	0.014	0.014	0.015	0.452	0.425	0.616	0.425
KO	0.010	0.011	0.011	0.010	0.007	0.009	0.009	0.008	0.534	0.452	0.479	0.548
VZ	0.015	0.015	0.015	0.013	0.011	0.011	0.012	0.010	0.479	0.452	0.411	0.562
JNJ	0.012	0.013	0.011	0.011	0.009	0.011	0.009	0.009	0.603	0.438	0.616	0.575
COST	0.012	0.012	0.014	0.014	0.009	0.010	0.010	0.010	0.589	0.671	0.507	0.479
AVGO	0.038	0.039	0.036	0.037	0.027	0.027	0.026	0.024	0.452	0.534	0.521	0.616
HD	0.015	0.015	0.015	0.016	0.011	0.011	0.011	0.012	0.507	0.438	0.562	0.589
AXP	0.017	0.018	0.017	0.016	0.012	0.014	0.012	0.012	0.507	0.425	0.452	0.521
CVX	0.015	0.013	0.014	0.015	0.011	0.010	0.011	0.012	0.548	0.562	0.521	0.507
UNH	0.023	0.024	0.023	0.025	0.016	0.016	0.016	0.017	0.507	0.493	0.466	0.452
CSCO	0.012	0.013	0.014	0.013	0.009	0.010	0.010	0.009	0.603	0.521	0.452	0.548
PG	0.013	0.012	0.013	0.012	0.010	0.009	0.010	0.009	0.452	0.425	0.548	0.507
DIS	0.017	0.018	0.017	0.018	0.011	0.012	0.012	0.013	0.521	0.466	0.562	0.411
V	0.015	0.014	0.014	0.013	0.011	0.010	0.011	0.011	0.452	0.479	0.493	0.411
CAT	0.021	0.019	0.021	0.022	0.016	0.015	0.016	0.017	0.411	0.603	0.466	0.466
AMGN	0.022	0.023	0.021	0.020	0.016	0.014	0.014	0.014	0.493	0.493	0.562	0.534
HON	0.020	0.019	0.016	0.018	0.014	0.014	0.012	0.013	0.397	0.452	0.616	0.507
IBM	0.019	0.020	0.021	0.019	0.013	0.014	0.015	0.013	0.452	0.521	0.521	0.466
MCD	0.014	0.014	0.014	0.015	0.010	0.009	0.009	0.010	0.466	0.603	0.562	0.575
AAPL	0.015	0.016	0.015	0.016	0.012	0.012	0.011	0.012	0.514	0.653	0.569	0.458
MMM	0.018	0.019	0.019	0.016	0.013	0.014	0.013	0.012	0.507	0.507	0.603	0.644
SHW	0.019	0.021	0.019	0.019	0.015	0.016	0.014	0.013	0.479	0.411	0.479	0.507
TRV	0.018	0.019	0.019	0.019	0.013	0.014	0.014	0.014	0.500	0.486	0.569	0.556
CRM	0.045	0.031	0.027	0.049	0.034	0.024	0.022	0.042	0.400	0.800	0.400	0.200
CRYPTO:BTC	0.027	0.023	0.033	0.029	0.021	0.019	0.028	0.024	0.531	0.594	0.323	0.344
CRYPTO:ETH	0.039	0.041	0.038	0.038	0.030	0.034	0.029	0.030	0.594	0.469	0.625	0.406
BRENT	0.020	0.020	0.019	0.020	0.014	0.016	0.016	0.017	0.595	0.541	0.425	0.473
Average	0.021	0.021	0.021	0.021	0.016	0.016	0.016	0.016	0.499	0.511	0.514	0.490
# of wins	9	6	7	15	8	10	7	12	6	12	13	8

Moreover, the only observable difference in our models' performance between high- and low-attention assets occurs within the Open-to-Close time horizon. Specifically, incorporating sentiment through our proposed frameworks yields better predictive improvements for top high-attention assets during this period compared to other time horizons.

7. Conclusions

The study successfully demonstrated the usefulness of integrating sentiment analyses from financial news into predictive models of asset returns. Using sentiment-driven Markov mixture models, the research provides a solid framework to capture the complex and diverse nature of financial data. Adding sentiment states as a Markov process improves the understanding of asset price movements and significantly boosts the accuracy of return predictions for Close-to-Close trading horizons compared to traditional models. However, our empirical findings also reveal that the predictive advantage of the SA-FIGARCH framework depends on the return time horizon. For Open-to-Close returns, the standard FIGARCH model maintains a slight edge in error minimization (RMSE and MAE), while the proposed four-state model is superior in DA. For Open-to-Open returns, logistic regression state-predicted models also prove highly competitive. The findings offer valuable insights for investors, financial analysts, and researchers, thereby highlighting that the timing of trades relative to the assimilation of news sentiment is a crucial factor in financial forecasting. Future research could expand this framework to include more complex sentiment measures and refine its integration with intraday financial models.

Author contributions

Mostafa Abdolahi Moghadam Salkoye was responsible for data collection, software implementation, computational analysis, and preparation of the initial manuscript draft. Roman N. Makarov contributed to the methodology, critically reviewed and edited the manuscript, and provided overall supervision of the research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

This research was funded by the NSERC Discovery grant RGPIN-2020-04782.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Ang A, Timmermann A (2012) Regime changes and financial markets. *Annu Rev Financ Econ* 4: 313–337. <https://doi.org/10.1146/annurev-financial-110311-101808>

- Araci D (2019) FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*. <https://doi.org/10.48550/arXiv.1908.10063>
- Baillie RT, Bollerslev T, Mikkelsen HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J Econometrics* 74: 3–30. [https://doi.org/10.1016/S0304-4076\(96\)01824-9](https://doi.org/10.1016/S0304-4076(96)01824-9)
- Baissa DK, Rainey C (2020) When BLUE is not best: Non-normal errors and the linear model. *Polit Sci Res Meth* 8: 136–148. <https://doi.org/10.1017/psrm.2018.34>
- Case J, Clements A (2021) The impact of sentiment in the news media on daily and monthly stock market returns. *Data Min*, 1504: 180–195. Springer Singapore. https://doi.org/10.1007/978-981-16-8531-6_13
- Choi SY, Yoon JH (2020) Modeling and risk analysis using parametric distributions with an application in equity-linked securities. *Math Probl Eng* 2020: 9763065. <https://doi.org/10.1155/2020/9763065>
- Dridi A, Atzeni M, Reforgiato Recupero D (2019) FineNews: Fine-grained semantic sentiment analysis on financial microblogs and news. *Int J Mach Learn Cybern* 10: 2199–2207. <https://doi.org/10.1007/s13042-018-0805-x>
- Dua P, Tuteja D (2021) Regime shifts in the behaviour of international currency and equity markets: A Markov-switching analysis. *J Quant Econ* 19: 309–336. <https://doi.org/10.1007/s40953-021-00273-9>
- Feng L, Fu T, Shi Y (2022) How does news sentiment affect the states of Japanese stock return volatility? *Int Rev Financ Anal* 84: 102267. <https://doi.org/10.1016/j.irfa.2022.102267>
- Frühwirth-Schnatter S (2006) *Finite Mixture and Markov Switching Models*. Springer. <https://doi.org/10.1007/978-0-387-35768-3>
- Guidolin M, Timmermann A (2006) An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *J Appl Econometrics* 21: 1–22. <https://doi.org/10.1002/jae.824>
- Haas M, Mittnik S, Paoletta MS (2004) A new approach to Markov-switching GARCH models. *J Financ Econometrics* 2: 493–530. <https://doi.org/10.1093/jjfinec/nbh020>
- Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384. <https://doi.org/10.1093/jjfinec/nbh020>
- Huang AH, Wang H, Yang Y (2023) FinBERT : A large language model for extracting information from financial text. *Contemp Account Res*, 40: 806–841. <https://doi.org/10.1111/1911-3846.12832>
- Kelly S, Ahmad K (2018) Estimating the impact of domain-specific news sentiment on financial assets. *Knowl-Based Syst* 150: 116–126. <https://doi.org/10.1016/j.knosys.2018.03.004>
- Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. *Giorn Dell'inst Ital Degli Att*, 4: 89–91
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Methods Stat* 22: 79–86. <https://doi.org/10.1214/aoms/1177729694>

- Remillard B (2016) *Statistical Methods for Financial Engineering*. Chapman and Hall/CRC. <https://doi.org/10.1201/b14285>
- Shi Y, Feng L (2016) A discussion on the innovation distribution of the Markov regime-switching GARCH model. *Econ Model*, 53: 278–288. <https://doi.org/10.1016/j.econmod.2015.11.018>
- Shi Y, Ho KY (2015) Modeling high-frequency volatility with three-state FIGARCH models. *Econ Model* 51: 473–483. <https://doi.org/10.1016/j.econmod.2015.09.008>
- Shi Y, Ho KY (2021) News sentiment and states of stock return volatility: Evidence from long memory and discrete choice models. *Financ Res Lett* 38: 101446. <https://doi.org/10.1016/j.frl.2020.101446>
- Smirnov NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull Math Univ Moscou* 2: 3–14.
- Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. *J Financ* 62: 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- Wan X, Yang J, Marinov S, et al. (2021) Sentiment correlation in financial news networks and associated market movements. *Sci Rep* 11: 3062. <https://doi.org/10.1038/s41598-021-82338-6>
- Xiang Y, Zhang A, Guo J, et al. (2025) Adaptive multimodal prompt-tuning model for few-shot multimodal sentiment analysis. *Int J Mach Learn Cybernet*. <https://doi.org/10.1007/s13042-025-02549-5>
- Xu W, Wang J, Ma F, et al. (2019) Forecast the realized range-based volatility: The role of investor sentiment and regime switching. *Physica A* 527: 121422. <https://doi.org/10.1016/j.physa.2019.121422>
- Yao W, Wei Y, Yu C (2014) Robust mixture regression using the t-distribution. *Comput Stat Data Anal* 71: 116–127. <https://doi.org/10.1016/j.csda.2013.07.019>
- Young DS, Hunter DR (2010) Mixtures of regressions with predictor-dependent mixing proportions. *Comput Stat Data Anal* 54: 2253–2266. <https://doi.org/10.1016/j.csda.2010.04.002>



AIMS Press

©2026 M.Abdolahi Moghadam, R.N.Makarov, licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)