NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Unsupervised Clustering and Ensemble Decision Strategies in Cryptocurrency Trading

A SOM-Based Hybrid Model for Signal Generation

Catarina Alexandra Gouveia Andrade de Oliveira

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Unsupervised Clustering and Ensemble Decision Strategies in Cryptocurrency Trading

A SOM-Based Hybrid Model for Signal Generation

by

Catarina Alexandra Gouveia Andrade de Oliveira

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Professor Fernando Bação, PhD

Universidade Nova de Lisboa, Information Management School

Lisboa, Portugal

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 14th, 2025

Catarina Alexandra Gouveia Andrade de Oliveira

DEDICATION

To my family, whose unwavering support, encouragement, and love have made this journey possible. To my friends, for their companionship and belief in me. And above all, to my parents, whose investment in my education laid the foundation has given me the opportunity for every step I have taken, this achievement is as much theirs as it is mine.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who supported me throughout the development of this thesis. I am thankful to my supervisor, Professor Fernando Bação, and to Professor Farina Pontejos, for their guidance, constructive feedback, and encouragement during every phase of this work. You made every idea feel possible and gave me the confidence to pursue them. A heartfelt thank you to my family for your patience, motivation, and presence during every step of this demanding journey. Mom, thank you for reading every single page, your dedication has always meant the world to me. To my boyfriend, in a world of shifting trends, you've been my most reliable constant, your support has been more stable than any market I've studied. To my closest friends, who reminded me to take breaks or simply listen when I needed it, thank you for being there. Lastly, I am grateful to the open-source developers whose tools laid the foundation for this project.

ABSTRACT

This study presents a hybrid financial modeling framework that combines technical indicators and sentiment analysis to generate trading signals for the BTC-USD market using Self-Organizing Maps (SOMs). The proposed pipeline integrates two data sources: numerical price data and financial news, from which sentiment scores are extracted using FinBERT. After feature engineering and normalization, three SOMs are trained independently, one using only technical features, one using only sentiment features, and one combining both. A comprehensive grid search is performed to optimize the feature selection and SOM hyperparameters. The resulting cluster assignments are translated into trading signals (buy, hold, sell) and evaluated using four ensemble strategies: majority, unanimous, weighted, and aggressive voting, and, among these, the weighted ensemble is further optimized by testing various weight combinations for each signal source. The framework is tested over a defined out-of-sample period, and its performance is assessed using metrics such as cumulative return, sharpe ratio, and maximum drawdown. The results demonstrate that combining SOM-based clustering with ensemble decision strategies can yield interpretable and competitive trading signals in volatile markets like cryptocurrency. This work highlights the relevance of unsupervised learning techniques and multi-source data integration in financial forecasting and trading automation.

KEYWORDS

Bitcoin; Ensemble Learning; Sentiment Analysis; Self-Organizing Maps; Technical Indicators;

Sustainable Development Goals (SDG):

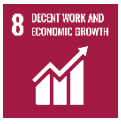

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
BMU	Best-Matching Unit
BTC-USD	Bitcoin ($)
FinBERT	Financial Bidirectional Encoder Representations from Transformers
HTML	HyperText Markup Language
LLM	Large Language Model
MACD	Moving Average Convergence Divergence
NLP	Natural Language Processing
PF	Profit Factor
QE	Quantization Error
RFE	Recursive Feature Elimination
RSI	Relative Strength Index
SOM	Self-Organizing Map
SR	Sharpe Ratio

1. INTRODUCTION

Financial markets are complex systems, shaped by a dynamic interplay of macroeconomic factors, investor sentiment, and unpredictable external shocks. In this volatile environment, traditional analytical models often struggle to capture the non-linear dependencies and evolving patterns embedded in financial time series. As a result, the financial industry has increasingly turned to machine learning techniques capable of processing large volumes of data and uncovering hidden structures without predefined rules.

Among unsupervised learning models, Self-Organizing Maps (SOMs) are recognized for their capacity to reduce dimensionality while preserving the topological characteristics of the data, which makes them particularly suitable for applications in finance, where interpretability and pattern recognition are essential. Initially developed for general-purpose clustering and visualization, SOMs have found renewed relevance in tasks such as market regime identification, anomaly detection, and strategy optimization in financial contexts.

The present study builds upon these foundations by applying SOMs to develop and evaluate trading signals derived from distinct market states. By clustering historical data into regimes, typically associated with bullish, bearish, or sideways behavior, SOMs can provide a structured view of market conditions. The resulting clusters are then analyzed for their predictive characteristics and translated into actionable signals, such as, in the financial context, buy, sell, or hold recommendations. To enhance the reliability of these signals, the system introduced by this thesis incorporates financial indicators, sentiment analysis derived from financial news sources, and ensemble logic to refine decision-making.

By integrating sentiment-based insights using financial natural language processing tools and examining the role of momentum indicators, this approach aims to account for both quantitative and qualitative dimensions of market behavior, recognizing that prices are often influenced not just by fundamentals, but also by perception and psychological factors. In this regard, news sentiment analysis complements the technical structure of SOMs, offering a broader informational base for signal generation.

Furthermore, this research emphasizes robustness by segmenting the evaluation according to different market regimes. Rather than relying on a single aggregated test set, the study separately assesses strategy performance in rising, falling, and stagnant markets. This approach provides a more nuanced view of each model's adaptability and highlights the importance of regime-awareness in trading system design.

The contribution of this thesis is twofold. On one hand, it introduces a comprehensive pipeline that covers data preprocessing, feature selection, SOM training, signal extraction, and strategy evaluation under realistic trading constraints. On the other hand, it presents an empirical comparison of multiple approaches, technical SOMs, sentiment-driven methods, hybrid

combinations, and ensemble systems, benchmarking them against traditional trading heuristics and a passive buy-and-hold strategy.

The central research objectives are to examine whether SOMs can effectively segment market data into meaningful regimes, whether these regimes can be mapped to profitable trading actions, and how their performance compares across methodologies and market conditions. In achieving these goals, the study aims to contribute with a rigorous, data-driven approach to the development of interpretable trading models grounded in unsupervised learning.

Ultimately, this thesis aligns with broader trends in financial data science, where the integration of machine learning and domain expertise is paving the way for more adaptive, transparent, and effective investment strategies. By demonstrating the practical viability of SOM-based systems, this study offers both theoretical insights and applied contributions to the field of quantitative finance.

2. LITERATURE REVIEW

The application of advanced machine learning techniques in financial trading has attracted substantial research interest over recent decades. Among these techniques, SOMs, a type of unsupervised neural network, have demonstrated their capability in clustering and visualizing high-dimensional financial data.

Eugene Fama's Efficient Market Hypothesis (EMH), introduced in 1970, proposes that all available information is already reflected in asset prices, making it impossible to reliably forecast future price movements. According to this hypothesis, prices follow a random walk, and no predictive model can consistently outperform the market (Fama, 1970). This view was supported by studies like Qian & Rasheed (2007), which emphasized the inherent unpredictability of price changes resulting from the randomness of information inflows, such as news events.

However, the assumption of market efficiency has been challenged. Research demonstrates that financial markets often exhibit non-linear and dynamic behaviors, contrary to the strong form of EMH (Huang et al., 2005). Machine learning has emerged as a powerful tool for capturing these complexities, leveraging non-linear algorithms to detect patterns and improve price prediction accuracy.

Deep learning architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been explored in financial time series forecasting, but their reliance on labeled datasets poses challenges (Fischer & Krauss, 2018). This is where unsupervised learning techniques, such as SOMs, become valuable, as they can identify structures in data without explicit labels, offering a flexible approach to market analysis.

SOMs, introduced by Kohonen (1982), are a type of artificial neural network designed to perform unsupervised learning, particularly effective at clustering and visualizing high-dimensional datasets in a low-dimensional space. Due to their capacity to uncover hidden structures and offer insights into complex datasets, SOMs have been widely adopted across multiple fields. In healthcare and bioinformatics, they have supported the analysis of genomic data, facilitated patient clustering for personalized medicine, and contributed to disease classification (Larrañaga et al., 2006). In the areas of image and speech processing, SOMs have enhanced facial recognition systems, improved handwriting analysis, and enabled more accurate phoneme classification, bolstering performance in pattern recognition tasks (Kohonen, 2001). The field of cybersecurity has also benefited from SOMs through their use in anomaly detection systems, where they help cluster normal and abnormal behaviors to identify potential cyber threats (Li et al., 2019). In financial and economic forecasting, SOMs have been applied to detect fraud, support customer segmentation in banking services, and refine risk assessment models (Deboeck & Kohonen, 1998). Moreover, they have proven valuable in market and consumer behavior analysis by explaining purchasing patterns and

informing the development of more effective marketing strategies (Vesanto & Alhoniemi, 2000).

SOMs are particularly valuable in the financial domain due to their capacity to uncover hidden structures in complex market data, thereby facilitating the identification of trading patterns and market anomalies. One notable application lies in clustering market data, where SOMs are employed to group similar stocks or market states, enabling traders to detect correlations and uncover potential arbitrage opportunities (Deboeck & Kohonen, 1998). This approach has proven especially effective in sectoral analysis, as it allows for the classification of stocks into meaningful categories based on performance indicators. Additionally, SOMs are used to visualize financial data by reducing its dimensionality, producing intuitive representations of market dynamics that enhance decision-making (Leinweber & Madhavan, 2001). Such visualizations can reveal structural shifts in the market, providing early warnings of possible trend reversals. Furthermore, SOMs contribute to the generation of trading signals by processing technical indicators such as moving averages, Bollinger Bands, and MACD. These models, as demonstrated by Enke & Mehdiyev (2013), have shown particular utility in high-frequency trading environments, where rapid pattern recognition is essential for effective trade execution.

Applications such as those described by Matos, Marques and Cardoso (2014) and Johnsson (2012) underscore SOMs' flexibility in analyzing stock market series, particularly in detecting non-linear patterns. Their research illustrated how SOMs could differentiate between bullish and bearish market regimes, allowing traders to adjust their strategies accordingly. Additionally, Mohamed (2019) demonstrated the use of SOMs for stock clustering combined with return prediction models, bridging unsupervised and supervised techniques. This hybrid approach enabled more accurate forecasting by first grouping stocks with similar behavior and then applying regression techniques within each cluster.

Investor sentiment plays a crucial role in shaping asset prices and trading decisions. Sentiment analysis, the extraction of emotions and opinions from textual data, has increasingly been leveraged to anticipate short-term market movements. Among the key contributions in this area, Harder & Fazlija (2024) demonstrated the predictive power of financial news sentiment on commodity prices. Similarly, Bollen et al. (2011) established a significant correlation between Twitter sentiment and stock market returns, laying the foundation for incorporating real-time textual data into trading algorithms.

Several studies have explored hybrid approaches that integrate sentiment with traditional financial indicators. For example, Belamfedel Alaoui et al. (2025) proposed a deep learning framework that combines sentiment signals with technical indicators, achieving improved predictive accuracy. More recent research by Griemsmann (2022) focused on sentiment analysis for commodities, showing how online media sentiment can generate actionable trading signals. In parallel, Abe & Nakagawa, (2020) developed a SOM-based approach that incorporated news sentiment with technical indicators to enhance predictive accuracy and

profitability, while Cuello (2024) highlighted the usefulness of dimensionality reduction techniques, such as SOMs, in processing sentiment features and extracting meaningful insights.

Pei et al. (2023) highlighted the utility of SOMs in visualizing clusters based on financial and sentiment data emphasizing their adaptability in market forecasting tasks. By mapping sentiment-driven investor behavior, this approach helped identify shifts in market dynamics before they manifested in price movements. Abe & Nakayama (2018), for instance, showed the potential of combining heterogeneous data sources to enhance model performance. However, the specific application of ensemble SOMs enriched with sentiment-based features remains largely underexplored, suggesting a promising direction for future research.

While SOMs are particularly effective for clustering and visualizing complex datasets, ensemble methods are widely recognized in the financial domain for enhancing model robustness and predictive accuracy. These techniques operate by aggregating the outputs of multiple models, thereby mitigating individual biases and improving overall performance. A common strategy involves voting mechanisms, in which predictions from several models are combined through majority or weighted voting schemes, as discussed by (López de Prado, 2020). Other popular methods include bagging and boosting, which involve training multiple model instances on varying data subsets to reduce variance or bias, according to (Dietterich, 2000). Hybrid approaches have also gained traction, where unsupervised models such as SOMs are first used to extract latent features, which are then input into supervised models like Support Vector Regression (SVR) for tasks such as stock market forecasting, as demonstrated by (Hsu et al., 2009).

Despite the widespread use of ensemble methods in machine learning, their application in conjunction with SOMs, particularly in trading, remains relatively limited. However, the concept of ensemble SOMs, wherein multiple SOMs are trained on diverse data sources such as technical indicators, sentiment scores, or market regimes, presents a promising avenue for further exploration. Recent advancements, such as the deep fuzzy SOM proposed by Pei, Luo, and Liu (2023), have begun to integrate ensemble-like methodologies within SOM-based frameworks, highlighting their potential to significantly improve decision-making in financial trading environments.

2.1. Research Gaps and Opportunities

Despite the notable progress in the application of SOMs and sentiment analysis within financial trading, several key research gaps persist. One of the most evident limitations lies in the limited exploration of ensemble approaches involving multiple SOMs trained on heterogeneous datasets. The integration of such ensembles remains a promising yet underdeveloped area, particularly in the context of complex and noisy financial environments.

Furthermore, while sentiment analysis has gained traction in recent years, few studies have implemented cutting-edge natural language processing (NLP) techniques capable of capturing more nuanced and context-aware sentiment signals. The use of pretrained transformer models or financial domain-specific language models is still rare, representing a missed opportunity for enhancing predictive accuracy.

Another critical area requiring further investigation concerns the capacity of SOM-based systems to adapt to regime shifts in financial markets. Dynamic market conditions, including transitions between bullish, bearish, and sideways phases, challenge the robustness and generalizability of static models. The ability of SOM ensembles to detect and adapt to such structural changes remains largely unexplored.

Scalability and computational efficiency also emerge as pressing concerns. Training multiple SOMs on high-dimensional data can be resource-intensive, raising questions about the balance between model complexity and real-time applicability in live trading environments. Moreover, enhancing the explainability of ensemble SOM output continues to be a challenge, particularly for practitioners seeking interpretable signals that can support decision-making in real-world scenarios.

Previous works such as those by Cervelló-Royo et al. (2015) and Zhong and Enke (2017) underscore the potential of combining technical indicators with sentiment-driven strategies. However, these contributions do not fully address the design and implementation of ensemble SOM architectures capable of leveraging this hybrid approach. Consequently, there is a clear opportunity for this thesis to advance the state of the art by proposing a framework that integrates diverse data sources, employs sophisticated sentiment analysis techniques, and introduces ensemble methodologies to improve model robustness and interpretability.

Through this contribution, the research aspires to close existing gaps and offer novel insights into the development of SOM-based trading strategies that are not only effective but also adaptable and transparent.

3. METHODOLOGY

This research adopts a quantitative approach to investigate how SOMs can be used to develop data-driven trading strategies. By focusing on structured numerical and textual data from the cryptocurrency market, specifically Bitcoin, the study aims to explore the clustering potential of SOMs, and the effectiveness of ensemble models built on technical, sentiment-based, and hybrid feature sets.

3.1. RESEARCH FRAMEWORK

The study is structured around a multi-stage pipeline that reflects the end-to-end process of data-driven financial modeling, illustrated in Figure 1 and follows a multi-stage pipeline, which is detailed bellow the figure.

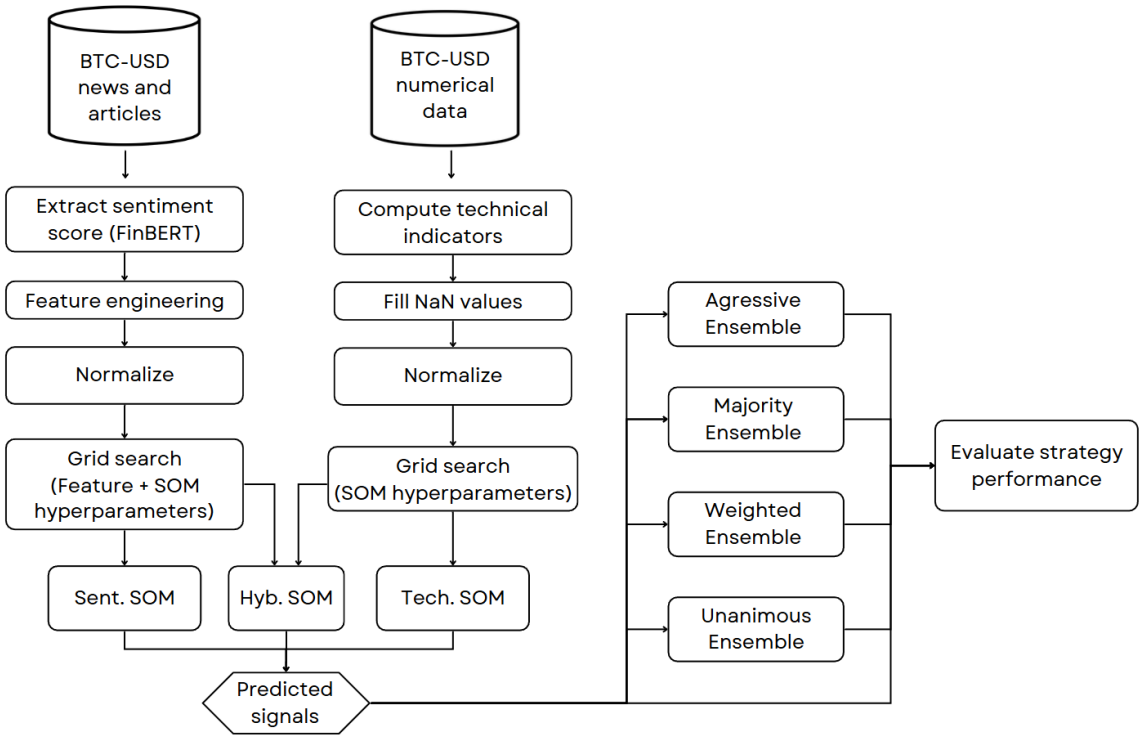

Figure 1 - Methodological framework used for predicting trading signals based on technical and sentiment data using SOMs and ensemble strategies

1. Data collection and preparation: gathering and cleaning historical price data and sentiment information relevant to Bitcoin, ensuring consistency and completeness.
2. SOM training: training three separate SOMs, one based purely on technical indicators, another on sentiment features, and a third using a hybrid of both.
3. Trading signal generation: translating cluster memberships into actionable trading signals such as buy, hold, or sell.

4. Ensemble construction: combining predicted signals from the different SOMs using a range of ensemble strategies to improve robustness and predictive quality.
5. Evaluation: measuring model performance using metrics such as cumulative return, Sharpe ratio, and win rate over a test period.

The decision to use SOMs is grounded in their ability to reduce dimensionality while preserving topological relationships, an essential property when trying to discover latent patterns in high-dimensional financial data.

3.2. Data Collection and Preparation

Bitcoin (BTC) was selected as the target asset for this study primarily due to its inherent volatility and pronounced sensitivity to shifts in public sentiment. Unlike traditional financial instruments, which are often influenced by macroeconomic fundamentals or institutional actors, BTC operates in a decentralized and sentiment-driven environment. Its price movements are frequently catalyzed by news events, social media trends, and changes in retail investor perception, making it an ideal candidate for testing sentiment-augmented trading strategies. These characteristics align particularly well with the capabilities of SOMs, which are experts in detecting complex and nonlinear patterns in market behavior, as such, BTC provides a rich context for evaluating the effectiveness of SOM-based modeling in the presence of dynamic and sentiment-reactive financial data.

Figure 2 illustrates the daily closing price of the exchange rate of Bitcoin against the US dollar (BTC-USD) pair across the full analysis period, providing the base for technical indicator extraction and trading signal generation.

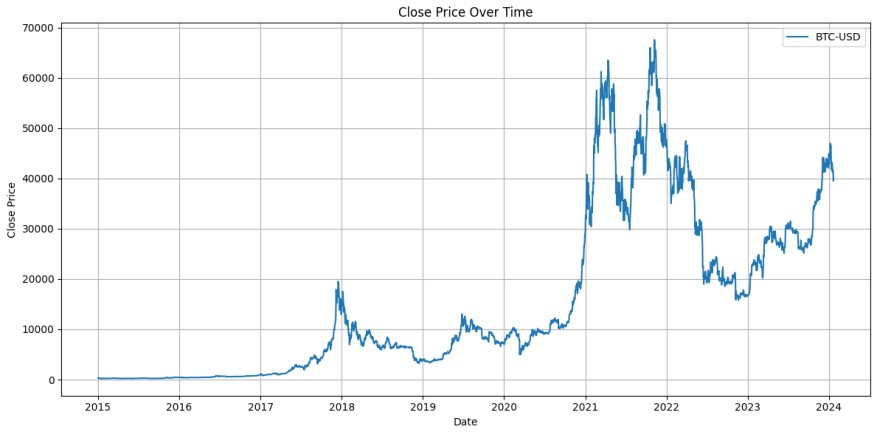

Figure 2 - Time series of the BTC-USD asset's closing price used as the base for technical analysis and trading signal generation

This study integrated two primary data sources. First, market data was obtained through the yfinance API, a Python library that facilitates access to historical market data from Yahoo Finance, which provided historical Bitcoin price and volume data from 2015 to 2024. This

dataset included daily values for open, high, low, close, and volume, forming the basis for technical analysis (Aroussi, 2025).

Second, sentiment data was initially intended to be collected in real time via APIs and web scraping tools targeting Yahoo Finance News, Google News, and Twitter (X). However, due to the limitations of free-tier access to these platforms, this study adopted a pre-collected and publicly available dataset of financial news headlines and tweets from Hugging Face (Schau, 2024). To extract sentiment signals from this textual data, FinBERT, a transformer-based model specifically fine-tuned for financial sentiment analysis (Araci, 2019) was employed. FinBERT is particularly effective in capturing the nuances of sentiment in financial texts, outperforming general-purpose sentiment classifiers in this domain, such as VADER (Hutto & Gilbert, 2014) and TextBlob (Loria, 2025). The model assigns each text snippet a positive, neutral, or negative label along with a confidence score, which were later aggregated and processed as daily sentiment indicators.

A wide array of technical indicators was computed using the pandas-ta Python library. These indicators were selected to cover diverse market characteristics, including momentum, trend strength, volatility, and volume dynamics. Formulas for each indicator are available in Annex A, and the complete list of indicators considered before feature selection is presented in Appendix A.

Momentum indicators such as the Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Commodity Channel Index (CCI), Williams %R, Momentum (MOM), Rate of Change (ROC), and the Ultimate Oscillator (UO) were employed to capture the speed and magnitude of price movements. Trend-following indicators like Simple Moving Averages (SMA) and Exponential Moving Averages (EMA), computed over 50 and 200 days, were used to identify medium- and long-term directional movement. Although the Parabolic SAR was initially considered, it was ultimately discarded due to its poor performance in non-trending or ranging markets like Bitcoin, where it frequently produced false signals.

Volatility indicators included the Average True Range (ATR) and Bollinger Bands, calculated over a 5-day moving average with a 2-standard deviation envelope. Volume-based indicators such as the On-Balance Volume (OBV) and Chaikin Money Flow (CMF) were incorporated to detect accumulation or distribution pressure. Once computed, all technical indicators were merged with price data into a unified dataset, aligned on a daily frequency.

On the other hand, the sentiment analysis pipeline began with textual data sourced from the above cited dataset of Bitcoin-related headlines and tweets, compiled from Yahoo Finance and over 360 distinct media outlets, including Bloomberg, CNBC, Reuters, CoinDesk, and The Guardian. Each text entry underwent cleaning via regular expressions, removing HTML tags, hyperlinks, excess whitespace, and bracketed content. This preprocessed text was then fed into FinBERT, a transformer-based model fine-tuned for financial sentiment classification,

which returned a sentiment positive, neutral, or negative label and an associated confidence score for each instance.

To synchronize textual data with market data, all sentiment outputs were aggregated daily. For each date, the average sentiment score and average sentiment label score, numerically encoded as -1, 0, and 1, were computed. In addition, counts and proportions of each sentiment class were calculated, resulting in features such as avg_sentiment_score, avg_sentiment_label_score, pct_positive, pct_negative, and total.

Missing values in the sentiment features were imputed with zeros. A new composite score, sentiment_score_combined, was created as a weighted average of the raw sentiment score, 70%, and the label score, 30%. A polarization metric was also introduced, defined as the difference between the proportion of positive and negative articles. Finally, the total daily sentiment volume was log-transformed to mitigate skewness due to the large range of number of articles in each day. All sentiment-derived features, including sentiment_score_combined, sentiment_polarization, and the log-transformed total volume, were standardized using z-score normalization.

Before training the model, the complete dataset underwent a series of preprocessing steps to ensure consistency, integrity, and suitability for the SOM ensemble learning framework. Given that the dataset combined numerical and textual features, preprocessing strategies were tailored to each type.

The numerical dataset was constructed using historical daily price data from Yahoo Finance, with values across the full analysis period. After computing technical features, all numeric columns were explicitly cast to float to ensure compatibility and avoid type inconsistencies. Missing values were handled using a combination of forward fill, backward fill, and linear interpolation, except in the early periods of long-lookback indicators (e.g., EMA-200), where their lack was intrinsic due to historical requirements.

No explicit outlier removal was performed, as visual inspection and statistical summaries revealed no anomalies inconsistent with typical financial time series behavior. All features, whether derived from prices or sentiment, were normalized using Min-Max scaling, which maps all values to the [0, 1] range. This approach was chosen to preserve the relative distribution of values while ensuring compatibility with the SOM's distance-based learning mechanism.

3.3. SOM Training Process

After preprocessing, the final dataset was structured into a unified DataFrame containing both technical indicators and sentiment scores. The dataset was chronologically split into a training set, from 2015-01-01 to 2021-12-31, comprising 2,557 daily observations and a test set, from

2022-01-01 to 2024-01-23, with 753 observations. This temporal split preserves the integrity of the financial time series, ensuring no data leakage.

Although unsupervised learning with SOMs does not require labeled targets or a traditional validation process, the training set was further divided into sub-training, from 2015-01-01 to 2020-01-01, and validation, from 2020-01-02 to 2021-12-31, with 1827 and 730 observations respectively, for the purpose of feature and parameter selection through grid search. This internal split the selection of optimal configurations while strictly keeping the test set unseen for unbiased final evaluation. The test set was then used to evaluate real-world generalization and the performance of the trading strategies derived from SOM clustering.

In this study, three distinct SOMs were trained, one per dataset type technical, sentiment, and hybrid, each configured with parameters optimized by performing grid search to enhance pattern separation and model stability and reduce multicollinearity or noise. The best-performing set of features and correspondent SOM hyper parameters (grid size, sigma, learning rate and number of iterations), were retained for further modeling, outlined in Table 1 and Table 2 respectively, with additional visualizations presented in Appendix B.

Som Type	**Selected features**
Technical	Open, CCI, EMA_50, ROC, OBV, CMF
Sentiment	z_sentiment_score_combined, z_sentiment_polarization, z_log_total_sentiment
Hybrid	Open, CCI, EMA_50, ROC, OBV, CMF, z_sentiment_score_combined, z_sentiment_polarization, z_log_total_sentiment

Table 1 - Selected features per strategy type

Model	Technical	Hybrid	Sentiment
SOM size	(20, 20)	(15, 15)	(20, 20)
Sigma	1.50	1.50	0.50
LR	0.50	0.10	0.50
It.	500	1000	1500
CR	3.86	2.62	2.17
SR	1.83	1.35	1.87
MD	0.29	0.29	0.08

Table 2 - SOM hyper parameters selected for training, as output of the grid search

Note: Features for each SOM type in Table 2 were listed in Table 1. SOM size indicates the grid dimensions of the SOM. Sigma and LR refer to the neighborhood radius and learning rate, respectively. It. denotes the number of training iterations. CR = Cumulative Return, SR = Sharpe Ratio, MD = Maximum Drawdown. Features list the input variables used for each SOM model, with the hybrid model incorporating additional sentiment-based indicators.

For the technical SOM, a grid search with 100 trials on different feature subsets was implemented. Then, each SOM's performance of each subset was quantified using the Cumulative Return from a simulated trading strategy on the validation dataset. This process not only improved prediction accuracy and generalization but also enhanced model interpretability and computational efficiency.

For the sentiment SOM, only the engineered features were manually selected, undergoing these same features into grid search to test multiple combinations of hyperparameters to select the ones that would maximize the Cumulative Return, following the same strategy as for the technical SOM.

Lastly, the features used to train the hybrid SOM were simply a combination of both features for the two types of SOM, to linearly identify if a blend of the best parameters of each SOM would result in a better performing hybrid one. Once again, this SOM went through a grid search to select the best-performative hyperparameters for these features.

Following the training phase, the features selected cluster interpretation was performed using U-Matrix visualizations (Figure 3), which reveal the relative distances between neurons. Additionally, density heatmaps were generated to inspect the distribution of data instances across the map, although these are included in the Appendix C due to space and interpretability considerations. Each square on the SOM grid represents a neuron that aggregates similar input vectors, so, proximity on the map reflects similarity in the underlying data. Clusters were then labeled as bullish, bearish, or neutral based on the average future return of the data points assigned to each node.

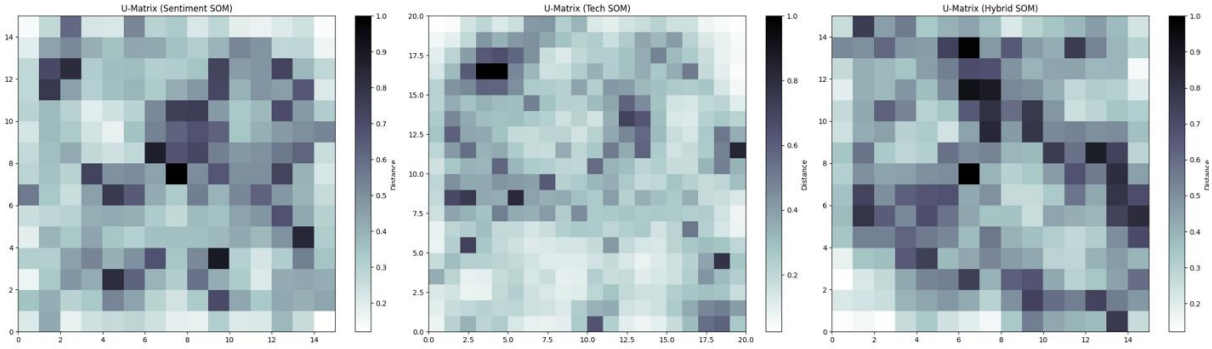

Figure 3 - U-Matrixes for the Sentiment, Technical and Hybrid SOM

As we can see on the U-Matrixes on Figure 3, all SOMs have successfully differentiated the input space into several distinct regions. Areas with darker colours represent greater distances between neighbouring nodes, indicating stronger changes in the underlying sentiment patterns. These high-distance zones typically correspond to cluster boundaries and suggest meaningful shifts in sentiment characteristics. Conversely, lighter zones show regions where the SOM identified similar sentiment inputs, leading to tighter cluster groupings. This

topological visualization confirms that the SOM can capture nuanced variations in financial sentiment, which is critical for identifying shifts in market behaviour based on news and social media input.

3.4. TRADING SIGNAL GENERATION

The transformation of unsupervised clusters into meaningful trading signals was executed through a structured, multi-phase methodology. Each Self-Organizing Map (SOM) model was trained on a subset of features, technical, sentiment, or hybrid, resulting in a topological mapping of the input data into a low-dimensional grid. This process grouped similar market states into spatially coherent clusters, which were then interpreted as latent regimes typically aligned with bullish, bearish, or neutral market dynamics.

To assign trading actions to these clusters, a forward-looking analysis was performed. Each cluster was evaluated based on its average future return over a predefined time horizon (e.g., 5 days) and its associated standard deviation. These two statistics were combined to produce a composite performance score for each cluster, calculated as a weighted average of its mean return and Sharpe ratio. Rather than relying on fixed thresholds, the decision boundaries for signal assignment were dynamically derived from the empirical distribution of these scores. Specifically, clusters with scores above the 75th percentile were interpreted as strong bullish signals and labeled as 'buy', those below the 25th percentile were considered bearish and labeled as 'sell', and the remainder were assigned a neutral 'hold' action.

The overall logic of the SOM-based signal assignment is summarized in Figure 4, which presents a schematic decision tree outlining the flow from cluster statistics to signal labeling.

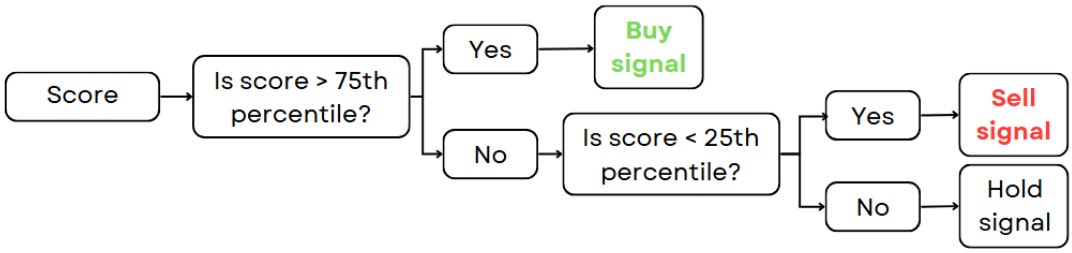

Figure 4 - Signal assignment tree for the SOM-based strategy based on future return and Sharpe score

This visual representation reinforces the data-driven, statistically grounded nature of the signal generation process and contrasts with more heuristic-based models like momentum or Fear & Greed. Each market observation is assigned to a cluster, which is then evaluated based on its average future return and Sharpe ratio. A composite score is computed and compared to dynamic percentile thresholds to generate a trading signal: buy, sell, or hold.

This labeling process resulted in a decision rule for each SOM node, which could then be applied to unseen data. During testing, the input features were scaled using the same normalization parameters fitted on the training set, ensuring consistency. Each new market observation was projected onto the trained SOM and matched its best-matching unit (BMU). The trading signal was then inferred by referencing the cluster-to-action dictionary previously derived from the training data.

Figure 5 illustrates the resulting trading signals over time, respectively for the SOMs trained on the combined hybrid feature set. See Appendix D for additional signal visualizations from the sentiment-only and technical-only SOMs. These figures demonstrate the dynamic adaptation of signals under changing market conditions and reveal the consistency and differentiation in signal behavior across model types.

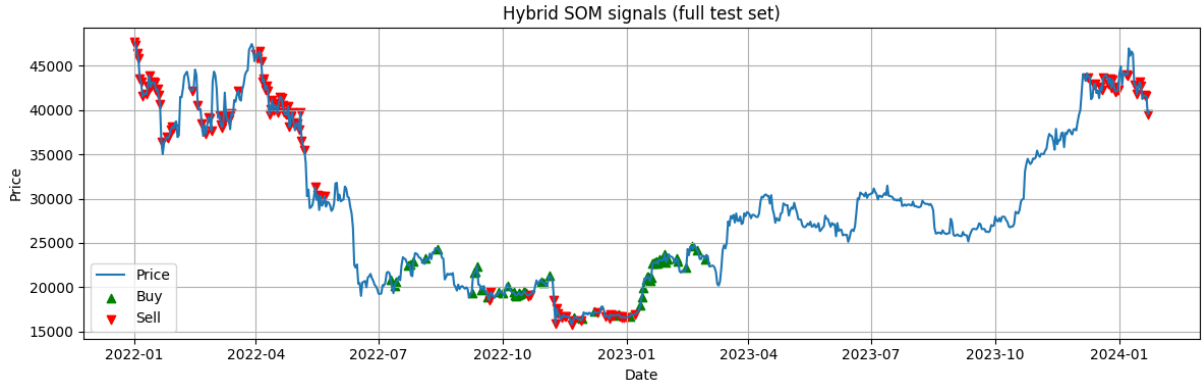

Figure 5 - Signals generated from the clusters outputted from the SOM trained with both technical and sentiment data

To increase robustness and reduce sensitivity to noise, additional filtering mechanisms were tested in parallel. For instance, sentiment-based strategies leveraged polarity scores extracted from FinBERT, with buy or sell actions activated based on confidence thresholds in sentiment direction. Similarly, momentum-based strategies were implemented using smoothed percentage changes in price over a rolling window, triggering buy or sell signals when these exceeded calibrated bounds. These alternative strategies were not only evaluated individually but also integrated into ensemble decision-making frameworks alongside SOM outputs.

Crucially, all strategies were simulated under realistic backtesting conditions, including transaction costs, slippage, and capital constraints. While exposure limits helped prevent overleveraging, it may also be valuable to explore the integration of additional risk management tools, such as stop-loss and take-profit mechanisms, to further enhance protection against volatile market behavior and improve the robustness of real-world applications.

Furthermore, the impact of transaction costs was considered during the design of the signal assignment logic. While signals were generated solely based on cluster characteristics and forward-looking return estimates, thresholds for triggering 'buy' and 'sell' actions were

calibrated conservatively to ensure that expected returns exceeded plausible transaction costs. This precaution reduced the likelihood of overtrading and helped ensure that the strategy remained viable under realistic market frictions.

Together, these components formed a flexible simulation pipeline in which each strategy, SOM-based, sentiment-based, momentum-driven, or heuristic, could be independently evaluated and compared to a passive buy-and-hold benchmark. Performance was assessed using cumulative return, Sharpe ratio, and maximum drawdown, enabling a rigorous comparison of effectiveness and stability across models and market regimes.

3.5. ENSEMBLE STRATEGY CONSTRUCTION

Rather than relying on a single SOM to produce trading signals, this study employed an ensemble learning approach. Signals from the three SOMs, trained respectively on technical indicators, sentiment features, and a hybrid of both, were aggregated using a series of decision-making strategies designed to enhance robustness and mitigate overfitting. This multi-model ensemble leverages the complementary perspectives of distinct data representations, increasing the likelihood of generalizable patterns.

Four ensemble strategies were implemented and tested. The first was majority voting, in which the final daily signal is the most frequently predicted class (buy, hold, or sell) across the three SOMs. Second, a weighted voting approach assigned different importance levels to each SOM, with weight distributions reflecting either their standalone backtest performance or domain relevance, as technical signals being more predictive during trending markets, for example.

Two additional strategies modeled distinct risk profiles. The unanimous strategy generated a signal only when all SOMs agreed, making it a more conservative approach. Conversely, the aggressive strategy issued a trading signal if any SOM suggested either 'buy' or 'sell'; otherwise, it defaulted to 'hold'. Finally, a probabilistic strategy was developed to capture the relative vote distribution of the SOMs, outputting probabilities for each class. While not directly used in backtesting, this strategy provides a foundation for future research using probabilistic or threshold-based models.

The logic behind each ensemble strategy is illustrated in Figure 6, which summarizes how signals from the three SOMs are combined under each rule-based decision process. Each block represents a different logic rule applied to the outputs of the technical, sentiment, and hybrid SOMs: aggressive, majority, weighted, and unanimous.

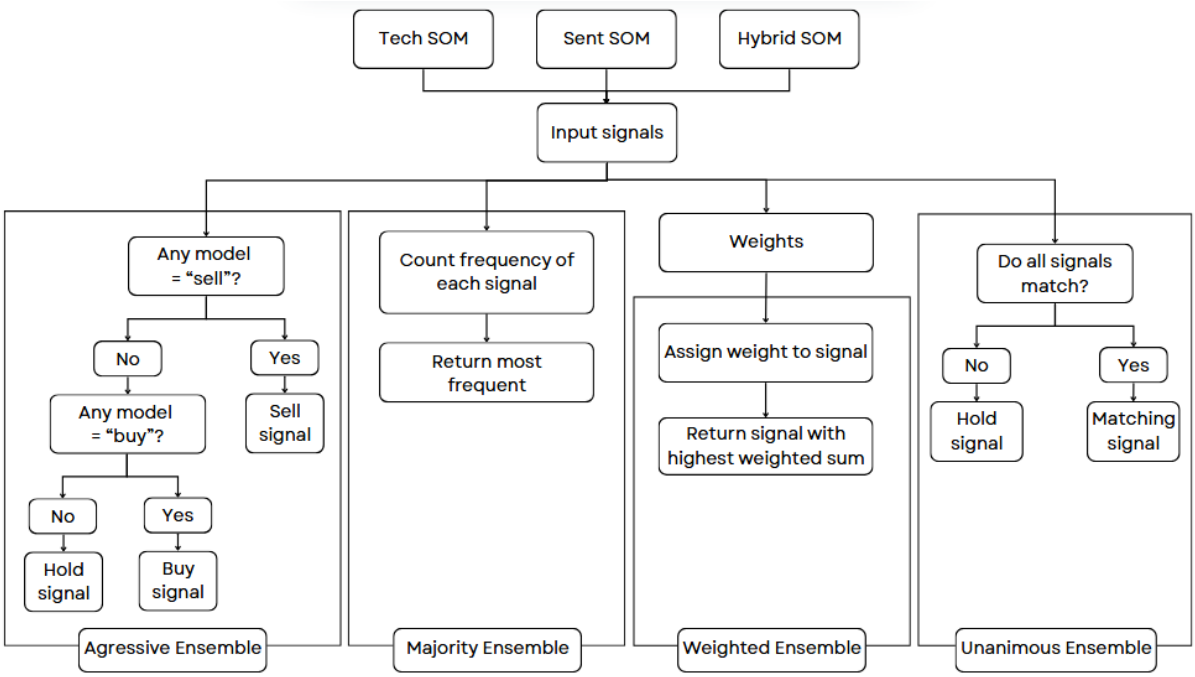

Figure 6 - Overview of ensemble voting strategies used to combine SOM signals

Each ensemble method was then evaluated on the out-of-sample test data from 2022 to 2024, using a standardized backtesting framework that simulated trades and calculated cumulative returns, Sharpe ratios, and drawdowns. A dedicated visualization compared the performance of all ensemble strategies against the baseline SOMs and external benchmarks, is presented and discussed in the Results chapter.

3.6. Evaluation Metrics

The effectiveness of the proposed trading system, based on SOMs, is assessed through a comprehensive evaluation framework that integrates both predictive performance and financial viability. This dual perspective ensures that the system is not only capable of uncovering meaningful structures in the input data but also capable of producing reliable and robust trading signals under realistic market conditions.

On the predictive side, three main aspects are considered. First, quantization error is used to evaluate how accurately the input vectors are represented by their best-matching units on the SOM grid. A lower quantization error indicates that the model effectively captures the structure of the input space. Second, topological preservation is assessed through U-Matrix visualizations, which illustrate the relative distances between neighboring neurons and allow for a qualitative assessment of whether the SOM maintains the spatial relationships inherent in the data. A well-structured map will show smooth transitions and coherent clusters, suggesting that similar inputs are consistently mapped to adjacent neurons. Third, the stability of the clustering process is examined by retraining each SOM multiple times with different

random seeds and comparing the resulting activation patterns. This procedure helps evaluate the consistency of the learned representations and whether the model is robust to initialization variability.

From a financial standpoint, the system is evaluated using standard performance metrics derived from backtesting in out-of-sample data. The cumulative return provides a straightforward measure of profitability over time, while the Sharpe ratio captures the quality of those returns in relation to the risk incurred, by comparing excess return to volatility. Additionally, maximum drawdown is considered to quantify the system's exposure to potential losses, representing the most significant peak-to-trough decline during the testing period and serving as an indicator of capital risk.

The trading simulator used in this evaluation incorporates real-world constraints, including transaction costs and partial capital allocation, thus offering a realistic view of how the strategy might perform in live market conditions. Although elements such as stop-loss rules and dynamic position sizing were not implemented in the current version, the system architecture was designed to accommodate these extensions in future iterations. Such enhancements would contribute to more advanced risk management and allow for further testing of the system's practical deployability.

To complement the overall evaluation, the system is also tested for robustness across different market regimes. Market phases are identified based on rolling return windows and categorized into bullish, bearish, and sideways segments. Each SOM model, technical, sentiment-based, and hybrid, is trained independently within each regime, followed by separate out-of-sample testing for that specific market condition. This process enables a more granular understanding of how each model adapts to different structural contexts and whether it can maintain stable and profitable behavior under changing dynamics.

3.7. BACKTESTING STRATEGY AND BENCHMARKING

To assess the practical viability of the trading signals generated by the Self-Organizing Map (SOM) models, a comprehensive backtesting framework was employed. This framework not only simulates real-world trading conditions, such as transaction costs, limited capital deployment, and regime-specific modelling, but also compares the SOM strategy against established benchmark approaches. These benchmarks provide a reference point to evaluate whether the unsupervised learning method delivers a tangible advantage in trading performance.

The first benchmark consists of a traditional Buy and Hold strategy. In this approach, an investor purchases Bitcoin at the beginning of the test period and holds the position

throughout. It represents a passive investment model and is often used as a baseline to determine whether active trading strategies can yield superior returns.

In contrast, the second benchmark is based on a sentiment-oriented model inspired by the Fear and Greed index commonly used in behavioural finance. A custom indicator was developed using three key features: recent price momentum (measured as the 3-day rolling average of returns), market stability (inferred from the inverse of short-term volatility), and relative trading volume (comparing short- and long-term moving averages). These components are individually ranked and averaged to produce a composite score between 0 and 100. Low values, indicating fear, trigger buy signals, whereas high values, reflecting excessive greed, prompt sell decisions. Intermediate values result in no action, allowing the model to remain neutral. The logic behind this benchmark is illustrated in Figure 7, which outlines the decision process used to determine trading actions based on the computed index.

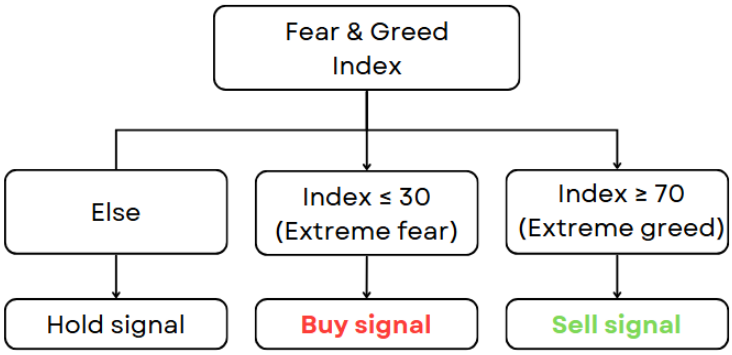

Figure 7 - Decision Tree for the Fear and Greed strategy

The third benchmark follows a momentum-based strategy, a well-known trend-following technique grounded in the assumption that positive price trends tend to persist over short horizons. The trading signal is derived from the percentage price change over a 10-day window. If the momentum exceeds a predefined threshold, typically around +2%, a buy signal is generated, while a decline below -2% triggers a sell signal. Values within this range lead to a hold decision. This rule-based logic is summarized visually in Figure 8, offering a clear overview of how the model translates momentum into trading actions.

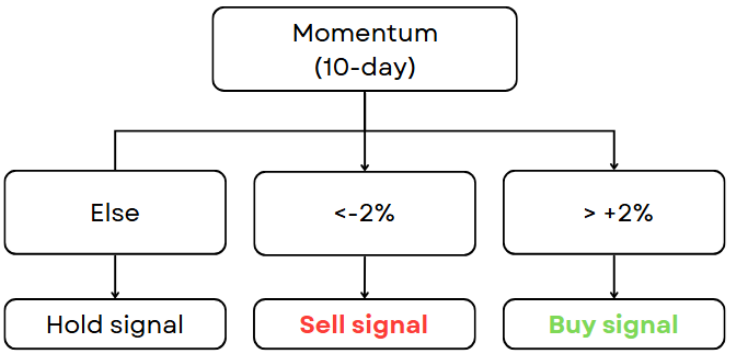

Figure 8 - Decision Tree for the Momentum-based trading strategy

These benchmark strategies serve as robust comparators against which the SOM-based system is evaluated. While Buy and Hold offers a simplistic passive alternative, the Fear & Greed and Momentum models embody more reactive, interpretable heuristics commonly used in practice. Together, they help contextualize the SOM's performance, especially when analysed across different market regimes. The subsequent chapter presents a detailed comparison of these approaches in terms of cumulative return, Sharpe ratio, and drawdown, providing insights into the relative effectiveness and resilience of each strategy under various market conditions.

3.8. SOFTWARE AND TOOLS

The implementation of the trading system and all supporting analysis was carried out using a fully Python-based technology stack, selected for its versatility, open-source availability, and extensive support within the data science and financial research communities. The core environment was developed in a Jupyter Notebook, which allowed for a seamless combination of code, visualizations, and explanatory text, facilitating both exploratory data analysis and iterative model refinement.

For numerical computations, data manipulation, and preprocessing, the project made extensive use of NumPy and Pandas, which enabled efficient handling of large financial time series and complex feature engineering routines. The scikit-learn library played a central role in scaling data and computing auxiliary metrics such as quantization error and clustering stability. Clustering and SOM-specific logic was implemented through a customized adaptation of the MiniSom library, which provided an intuitive interface for training SOMs and allowed for the extraction of neuron activations and mappings.

The sentiment analysis component relied on FinBERT, a transformer-based language model pre-trained for financial sentiment classification. Preprocessed textual data from financial news and headlines was passed through FinBERT using the Transformers library by Hugging Face, allowing for the generation of sentence-level sentiment scores aligned with market context. These scores were subsequently incorporated as input features for SOM training or used directly in sentiment-based trading strategies.

Visualization and diagnostics were handled using Matplotlib and Seaborn, providing insightful representations of market regimes, SOM grids, trading signals, and cumulative returns. To evaluate trading performance, a custom backtesting engine was developed in Python, integrating realistic assumptions such as transaction costs, partial allocations, and optional risk controls like stop-loss and take-profit logic. This backtesting framework also computed key financial metrics, Sharpe ratio, drawdown, profit factor, and win rate, which were critical in benchmarking strategy robustness.

The project architecture was modular, supporting reproducibility and scalability. All experiments were organized into distinct scripts for training, evaluation, and comparative analysis, with a single notebook containing the main workflow. Intermediate models and results persisted using standard serialization formats such as CSV, ensuring continuity across sessions and compatibility with future extensions. A detailed overview of the code repository structure and implementation details is provided in Appendix G.

Overall, the toolchain adopted in this work provided the necessary flexibility to explore a wide array of experimental settings, from regime-specific modeling to ensemble strategies, while maintaining reproducibility and analytical rigor. The exclusive use of open-source libraries further reinforces the accessibility of the approach, allowing for transparent validation and future adaptation in both academic and applied financial settings.

4. EMPIRICAL STUDY

The empirical study was conducted to evaluate the practical applicability, robustness, and financial performance of the proposed SOM-based trading system across multiple market conditions. This study involved a structured pipeline beginning with data preparation and feature engineering, followed by model training, signal generation, strategy evaluation, and comparative analysis against both traditional and alternative benchmarks.

4.1. DATASET AND PREPROCESSING

The dataset comprised historical data on price and volume, technical indicators, and sentiment scores derived from financial news headlines using FinBERT, into three differente sets: technical, sentiment and hybrid. Each of these sets served as the basis for training a dedicated SOM, enabling the unsupervised clustering of historical market conditions into latent states, without reliance on pre-labeled data.

4.2. TRAINING, REGIME SEGMENTATION AND SIGNAL GENERATION

The study was conducted using a train-validation-test split, with the training and validation sets used for model fitting and cluster-to-signal mapping, and the test set remaining unseen until final performance evaluation. Additionally, the test period was segmented into three market regimes bullish, bearish, and sideways, based on smoothed price trends and volatility filters. The bullish regime is characterized by sustained upward trends, the bearish by persistent downtrends and sideways by low-volatility intervals of time. This segmentation allowed for regime-specific training and performance comparisons, offering insight into how the system adapts to changing market dynamics.

Following training, each SOM was used to project the training and validation samples onto the 2D neuron grid. Each neuron, or node, therefore represents a group of similar market conditions based on the selected features. To make these clusters actionable, each neuron was assigned a trading signal label: buy (b), sell (s), or hold (h).

This labeling was based on the average forward return of all samples from the training and validation sets that were mapped to that neuron. Specifically, if the average future return of a node's associated samples exceeded a positive threshold (e.g., > +0.5%), it was labeled as a buy signal. If the average return was below a negative threshold (e.g., < -0.5%), it was assigned a sell signal. Otherwise, the signal was considered hold. The thresholds were empirically defined to balance risk and actionability.

This approach ensures that signal generation is entirely independent of the test data, avoiding information leakage and better reflecting real-world deployment conditions. It also benefits from a richer sample size by using both training and validation data, improving the statistical reliability of the average return per neuron.

To complement the topological insights from the U-Matrixes, the trained SOMs were also visualized with labeled trading signals, buy, hold, or sell, assigned to each neuron based on the average future returns of the data points mapped to it. This visualization translates the latent clustering into actionable signals and highlights how different regions of the map are associated with distinct trading behaviors. For instance, areas associated with strong positive returns were labeled as buy, those with strong negative returns as sell, and neutral zones as hold. This mapping reinforces the interpretability of the SOM, showing not only how market conditions are grouped but also how those groups translate into strategy decisions.

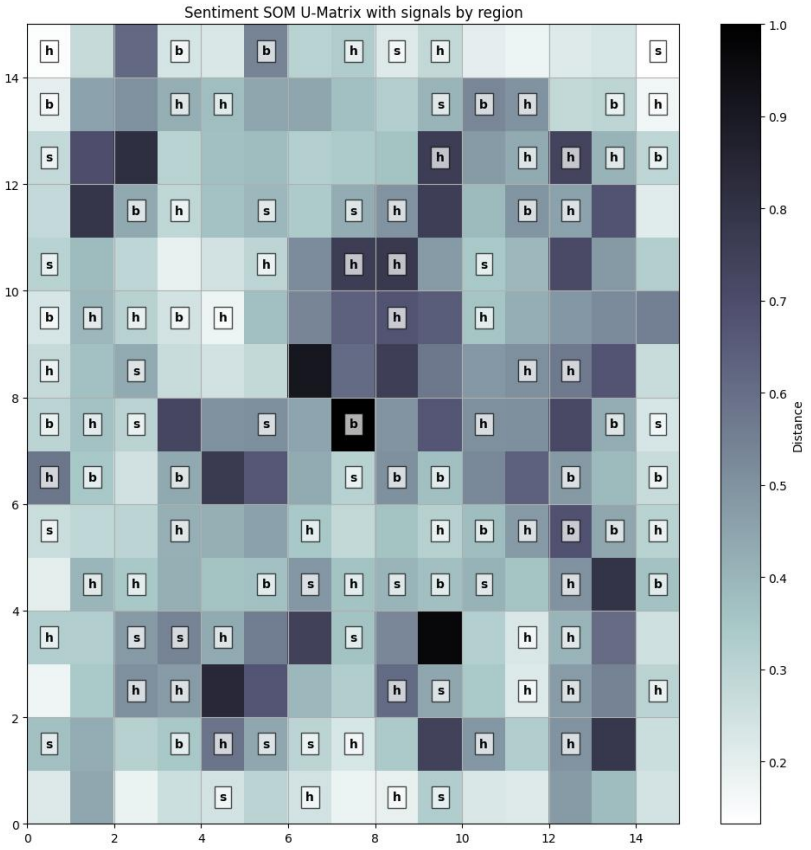

Figure 9 - Labeled SOMs with associated trading signals for the Sentiment SOM

Figure 9 presents the U-Matrix of the Sentiment SOM, annotated with the dominant trading signals in each neuron's region. The map shows that areas with high inter-neuron distance (darker shades) tend to coincide with signal transitions, indicating that those regions separate distinct market behaviors. For example, concentrated clusters of 'b' (buy) signals appear in regions with higher cohesion (lighter backgrounds), suggesting consistent positive future returns for sentiment patterns mapped to those neurons. Conversely, 's' (sell) signals are more frequent near high-distance areas, reinforcing their association with volatility or regime shifts. This reinforces the model's ability to organize sentiment-based information into meaningful trading strategies. Additional U-Matrix visualizations for the Technical and Hybrid SOMs, annotated with their respective signal mappings, are provided in Appendix E for reference.

Once each SOM was labeled, the test dataset was projected onto the same maps. Each new observation was matched to its Best Matching Unit (BMU), i.e., the neuron with the closest feature profile, and inherited the associated signal (b, s, or h). This allowed for real-time

simulation of how the model would behave on unseen data, using only the knowledge encoded during training.

The results on the test set, using the Hybrid SOM and summarized in the tables below, highlight the performance of the SOM-based strategy compared to traditional benchmarks under each regime. (Note: PF = Profit Factor; # = Number of Trades)

Strategy	Return	Sharpe	Drawdown	WinRate	PF	# Trades
Hybrid in bull regime SOM	1.59	3.15	5.90%	37.61%	2.05	3
Fear & Greed	0.96	-0.06	27.04%	33.76%	0.98	20
Momentum	1.12	0.71	18.19%	38.03%	1.18	29
Buy & Hold	1.05	—	—	—	—	—

Table 3 - Performance metrics during bull market regime

During bullish phases, Table 3, the SOM hybrid strategy outperformed both Fear & Greed and Momentum strategies in terms of Sharpe ratio and drawdown control, despite a low number of trades. This indicates the model selectively identifies high-conviction opportunities. The Buy & Hold strategy, while generally effective in uptrends, offered lower risk-adjusted returns.

Strategy	Return	Sharpe	Drawdown	WinRate	PF	# Trades
Hybrid in bear regime SOM	1.28	0.85	25.87%	36.97%	1.34	25
Fear & Greed	1.34	1.67	6.58%	24.17%	1.86	11
Momentum	1.10	0.57	13.42%	18.48%	1.25	41
Buy & Hold	0.87	—	—	—	—	—

Table 4 - Performance metrics during bear market regime

In bearish conditions, Table 4, the Hybrid SOM strategy maintained a competitive edge in profitability, though the Fear & Greed index led in Sharpe ratio and drawdown minimization. This suggests that sentiment-based heuristics are particularly effective in downturns. The Buy & Hold strategy significantly underperformed, reaffirming the need for active management.

Strategy	Return	Sharpe	Drawdown	WinRate	PF	# Trades
Hybrid in sideways regime SOM	1.00	0.17	41.10%	45.45%	1.05	26
Fear & Greed	0.83	-0.43	31.94%	35.71%	0.88	28
Momentum	1.23	0.86	21.47%	24.03%	1.34	55
Buy & Hold	0.87	—	—	—	—	—

Table 5 - Performance metrics during sideways market regime

Under sideways conditions, Table 5, Momentum emerged as the best performer in terms of return and Sharpe ratio, with the Hybrid SOM offering a balanced alternative. The increased drawdown seen in the SOM approach may indicate over-sensitivity to noise during range-bound markets. However, its higher win rate points to a stable success rate across trades.

These segmented insights highlight the strengths and limitations of each strategy under different market conditions, reinforcing the case for regime-aware portfolio construction.

4.3. STRATEGY EVALUATION AND BENCHMARKING

Trading signals were then used to simulate trades under a fixed set of backtesting rules, which included partial capital allocation, transaction costs, and basic risk controls.

To assess relative performance, financial metrics such as cumulative returns, maximum drawdown, Sharpe ratio, and win/loss ratio, were used. The SOMs' performances were contextualized by following benchmark strategies against multiple baselines: a Buy & Hold strategy, a Fear & Greed Index-based strategy, a Momentum-based approach using rolling returns, and sentiment-only strategies derived from FinBERT scores. Additionally, ensemble strategies were tested, combining outputs from different SOMs using majority voting, weighted confidence, or unanimous agreement mechanisms.

The results revealed meaningful distinctions between strategies across market regimes. In bullish periods, momentum and buy-and-hold approaches tended to dominate, while in

bearish and sideways markets, the SOM-based strategies, particularly those incorporating sentiment, exhibited more resilient performance and improved drawdown control. Ensemble strategies generally offered more stable outcomes, reinforcing the advantage of model diversification.

Furthermore, the SOM models were evaluated not only in terms of profitability but also through unsupervised learning metrics such as quantization error, topological preservation, and cluster stability. These provided insight into the internal consistency and representational quality of the trained models, adding an additional layer of validation beyond financial returns.

Overall, the empirical study confirmed the viability of using SOMs as the foundation for a systematic trading strategy. By integrating diverse sources of information and testing under varied market conditions, the system demonstrated adaptability and potential for real-world application, while also highlighting key areas, such as generalization and interpretability, that warrant further development.

4.4. DAILY SIGNAL INTERPRETATION AND PRATICAL IMPLICATIONS

To complement the aggregate performance metrics, we examine a representative decision day from the ensemble framework. The date 2022-11-08 was selected for in-depth analysis due to its significance in financial sentiment dynamics. On this day, Binance announced a potential acquisition of FTX amid a liquidity crisis, which was later retracted, triggering heightened uncertainty and a sharp drop in cryptocurrency markets, Bitcoin dropped to its lowest level in two years (Hajric & Shen, 2022). Figure 10 provides the broader context by marking this decision point within the BTC-USD price timeline.

Figure 10 - BTC-USD price chart with the selected decision day (2022-11-08) highlighted

This event revealed a divergence between model responses: while the Hybrid SOM promptly issued a "Sell" signal, the Sentiment and Technical SOMs signaled "Hold", as did most ensemble strategies. This divergence reflects a lag in sentiment-based signals relative to

market-driven features, highlighting how different SOM configurations capture and react to emerging risks.

To understand the rationale behind the hybrid model's decision, we examine its U-Matrix. As shown in Figure 11, the input vector for 2022-11-08 was mapped to the BMU at coordinates (14, 5), a region historically associated with negative return profiles during the training phase. The surrounding high-distance borders in the U-Matrix indicate a well-separated cluster, suggesting strong signal confidence. This supports the model's bearish stance on that date.

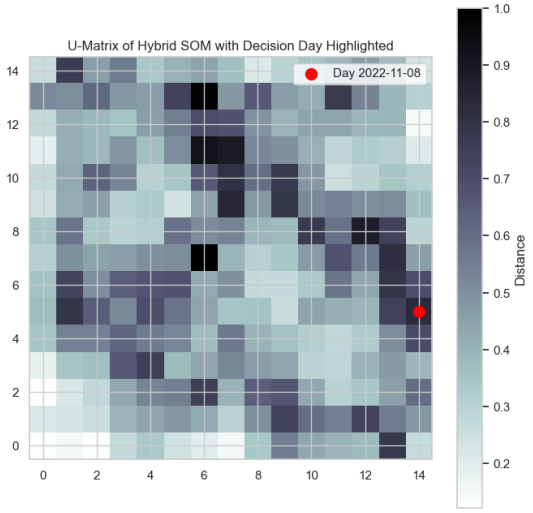

Figure 11 - U-Matrix of the Hybrid SOM showing bearish cluster activation on 2022-11-08

In contrast, the Technical SOM (Figure 12) mapped the input vector to a region associated with neutral return expectations, at coordinates (18, 0), aligning with the resulting Hold decision. Despite visual separation in the U-Matrix, this BMU did not correspond to historically bearish outcomes, thus justifying a more conservative signal.

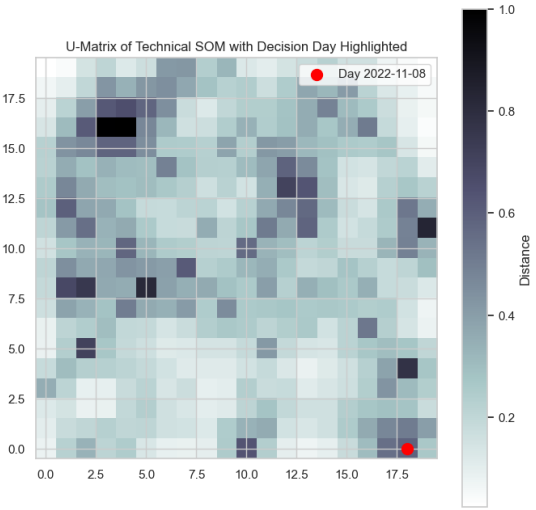

Figure 12 - U-Matrix of the Technical SOM showing neutral signal activation on 2022-11-08

The decision outputs from each individual SOM and the ensemble strategies are summarized in the following Table 6. This divergence across models illustrates the modular behavior and adaptability of the ensemble framework. The Aggressive Ensemble, which is more sensitive to minority signals, aligned with the Hybrid SOM's bearish prediction and issued a Sell. Conversely, the Majority, Weighted, and Unanimous strategies required broader agreement across models and therefore maintained a Hold stance.

SOM	Signal
Techical	Hold
Sentiment	Hold
Hybrid	Sell
Majority	Hold
Weighted	Hold
Unanimous	Hold
Agressive	Sell

Table 6 - Daily trading signals from individual SOMs and ensemble strategies

This case study exemplifies several key takeaways that underscore the practical value of the proposed approach. First, model interpretability is enhanced through U-Matrix visualizations, which offer a transparent explanation for why a specific SOM issued a particular signal by linking cluster boundaries to historical performance. Second, the clear separation observed in the Hybrid SOM's U-Matrix highlights that some signals carry a higher degree of confidence than others, an insight that can inform risk-adjusted decision-making and adaptation to shifting market conditions.

To determine which signal to follow in this case, we consider both model confidence and the broader context. Although most models issued a "Hold" signal, the Hybrid SOM demonstrated the strongest conviction, supported by a well-defined cluster historically linked to negative returns. This suggests high signal reliability, particularly under conditions of structural market stress such as the FTX crisis. A risk-aware investor aiming to avoid significant downturns may have benefited from following the Hybrid SOM's or Aggressive Ensemble's "Sell" recommendation. In scenarios marked by major news shocks and asymmetric information, models that integrate both technical and sentiment features, like the Hybrid SOM, may offer superior early-warning capability.

Finally, the flexibility of ensemble strategies becomes evident through their ability to modulate the level of consensus required for action. This allows the system to cater to different investor risk profiles: while conservative participants may favor Unanimous or Weighted strategies, more aggressive traders may prefer approaches that respond quickly to minority signals. The system thus delivers real-time, interpretable, and diversified trading signals that can serve as the foundation for both automated agents and interactive decision-support tools.

5. RESULTS AND DISCUSSION

The results of the empirical study offer a detailed perspective on the efficacy, adaptability, and robustness of the proposed SOM-based trading system across different feature domains, market conditions, and strategy combinations. This section synthesizes the key findings from the experimental setup, contextualizing them within the broader landscape of algorithmic trading and financial forecasting.

The initial stage of evaluation focused on comparing SOM models trained on different types of features, technical indicators, sentiment scores, and hybrid combinations.

The Hybrid SOM consistently demonstrated the strongest performance among individual SOM models, achieving the highest cumulative return (1.38) and Sharpe ratio (0.78). This indicates that combining technical and sentiment features can yield a more balanced and effective representation of market dynamics. While it had a modest number of trades (32), the Hybrid SOM maintained strong stability, as evidenced by a moderate drawdown of 19.14% and a solid profit factor of 1.21.

In contrast, the Sentiment SOM achieved a return of 1.18 and a Sharpe ratio of 0.51, suggesting that sentiment features alone provide valuable signals but are somewhat more volatile. Its high number of trades (256) and lower win rate (30.41%) reflect the noisier nature of sentiment-driven data.

The Technical SOM, although more conservative with only 3 trades, yielded a return of 1.14 and a Sharpe ratio of 0.30. It had the highest win rate (47.01%) but also the highest drawdown (40.16%), suggesting it struggled during more volatile or sideways market phases.

For benchmarking purposes, baseline strategies such as Fear & Greed, Momentum, and Buy & Hold were also evaluated. All three lagged behind the SOM-based models in terms of risk-adjusted returns. The Buy & Hold approach produced a return of 0.84, while both Fear & Greed and Momentum yielded returns of 0.98 with negligible Sharpe ratios (0.03), and drawdowns close to 30%. These results further emphasize the added value of SOM-based strategies, particularly when enhanced with sentiment-aware components.

The integration of multiple SOM models through ensemble methods, such as majority voting, weighted aggregation, and aggressive unanimous filtering, proved instrumental in enhancing strategy stability and minimizing susceptibility to noise-induced overreactions.

Among ensemble strategies, the Ensemble Weighted method yielded the highest cumulative return of 1.46 and Sharpe ratio of 0.92, while maintaining a relatively low drawdown of 15.30%. This configuration, which gives more weight to signals from the most consistent individual models, achieved a balanced combination of profitability and risk control.

The Ensemble Unanimous strategy also performed exceptionally well, with a return of 1.30, a Sharpe ratio of 0.88, and the lowest drawdown across all models (8.37%), though at the cost of only two trades. This indicates a very selective but highly precise decision-making behavior.

The Ensemble Aggressive model, which is more permissive in aggregating signals, generated a return of 1.25 and a Sharpe of 0.58, executing 296 trades. While more active, it remained effective with a drawdown of just 16.56%.

Importantly, the ensemble mechanisms smoothed out the inherent volatility of individual models and mitigated the effects of model-specific noise. These findings support the hypothesis that aggregating the outputs of weakly correlated SOMs leads to more resilient and generalizable trading frameworks, for which the increased robustness is visually evident in the smoother and more stable cumulative returns of ensemble strategies compared to individual SOMs (Figure 13).

Figure 13 - Cumulative returns of all tested strategies (SOMs, ensembles, benchmarks) on BTC-USD, out-of-sample from 2022 to 2024

The weak correlations between individual SOMs are empirically confirmed in the correlation matrix (Figure 23, Appendix F), which illustrates the complementary behavior of each model type. By distributing decision-making across multiple perspectives (technical, sentiment, hybrid), the ensemble system reduced overfitting to any single information source and delivered a more reliable performance profile under varying market conditions.

Strategy	Return	Sharpe	Drawdown	WinRate	PF	# Trades
Tech SOM	1.14	0.30	40.16%	47.01%	1.06	3
Sentiment SOM	1.18	0.51	14.44%	30.41%	1.14	256
Hybrid SOM	1.38	0.78	19.14%	30.68%	1.21	32
Fear & Greed	0.98	0.03	31.27%	24.97%	1.01	39
Momentum	0.98	0.03	29.98%	27.36%	1.01	136
Ensemble Majority	1.19	0.41	25.82%	37.05%	1.10	52
Ensemble Weighted	**1.46**	**0.92**	15.30%	32.27%	1.24	66
Ensemble Unanimous	1.30	0.88	**8.37%**	28.95%	**1.25**	2
Ensemble Aggressive	1.25	0.58	16.56%	33.86%	1.15	296
Buy & Hold	0.84	—	—	—	—	—

Table 7 - Performance metrics for each trading strategy

Table 7 (Note: PF = Profit Factor, # = Number of) summarizes the performance metrics for each tested strategy, including cumulative return (Return), Sharpe ratio (Sharpe), maximum drawdown (Drawdown), win rate (WinRate), profit factor (PF), and the total number of trades (# Trades). These indicators offer a multifaceted view of profitability, risk, and operational behavior. The Hybrid SOM and Ensemble Weighted strategies stand out for delivering the best risk-adjusted returns.

When analyzed across bullish, bearish, and sideways regimes, the SOM-based strategies revealed varying degrees of efficacy. In bull markets, momentum-based and buy-and-hold strategies typically achieved the highest cumulative returns due to persistent upward trends. However, SOM strategies, especially hybrid and ensemble variants, remained competitive while offering superior drawdown control.

In bear markets, the SOM strategies outperformed traditional methods by a significant margin. Their unsupervised nature allowed them to detect subtle structural shifts in the data, resulting in more timely exit signals and reduced exposure to losses. Notably, the Fear & Greed and sentiment-based strategies also performed well in these conditions, likely due to their ability to capture behavioral biases and panic-driven news flows.

In sideways markets, characterized by mean-reverting behavior and limited directional trends, the SOM strategies demonstrated moderate success. While not as profitable as in trending markets, they generally outperformed momentum-based strategies, which tend to underperform in choppy conditions. The use of ensemble logic further enhanced performance stability during such periods.

These findings confirm the context-dependency of strategy performance and highlight the value of adapting models or signal thresholds dynamically based on detected market regimes.

The strategies were benchmarked using a consistent set of financial and algorithmic metrics. The cumulative return provided a measure of absolute profitability, while the Sharpe ratio and Profit Factor indicated the quality and consistency of returns relative to risk. The Maximum Drawdown served as a proxy for capital preservation, and the Win Rate and Number of Trades offered operational insight into signal reliability and trading frequency.

Across these metrics, ensemble SOMs, especially the weighted and hybrid configurations, offered a compelling balance between profitability and risk. The inclusion of transaction costs and position sizing logic in the simulation framework ensured that results were realistic and aligned with practical trading constraints.

Beyond financial performance, the SOMs were evaluated for topological preservation, quantization error, and cluster stability. The topological coherence of the maps confirmed that the SOMs preserved meaningful spatial relationships within the input data, a desirable property for market structure learning. Repeated training runs with different seeds revealed relatively stable cluster formations, suggesting that the SOMs were not overly sensitive to initialization. However, some variability remained in smaller maps or when using highly volatile input features, indicating a need for additional robustness testing in future iterations.

While the results are encouraging, several limitations must be acknowledged. The system's reliance on historical data and backtesting introduces the risk of overfitting and regime bias, especially when tested on the same market the model was trained on. Additionally, the quality of the sentiment signals was partially constrained by API rate limits and the availability of labeled news data, which may have introduced gaps in the sentiment time series.

Moreover, the interpretability of the SOM clusters remains a challenge. While efforts were made to map clusters to latent market states using statistical descriptors, the black-box nature of unsupervised learning can obscure the underlying economic rationale. Future work may explore techniques such explainable clustering models to enhance transparency.

This study contributes to the growing field of machine learning in quantitative finance, illustrating how unsupervised techniques like SOMs can uncover structure in noisy financial data and generate practical trading signals. The modular design of the framework allows for straightforward integration of additional signals or market domains, supporting cross-asset generalization. While initial tests focused on equities, the methodology is extensible to commodities, cryptocurrencies, and FX markets, provided that sufficient feature engineering and retraining are applied.

In practice, the system may serve both as a standalone trading model and as a signal enhancement module within a broader portfolio optimization strategy. Its emphasis on interpretability, flexibility, and ensemble learning aligns well with the needs of institutional and retail investors seeking data-driven decision tools.

6. CONCLUSION

This study explored the potential of SOMs as a foundation for intelligent trading signal generation. Through the implementation of technical, sentiment-based, and hybrid SOM configurations, the system demonstrated promising capabilities in identifying latent patterns in financial time series and translating them into practical buy, sell, or hold actions. The integration of multiple signal types, combined with a robust back testing framework, allowed for a meaningful evaluation of profitability, risk, and adaptability under varying market regimes.

Empirical results revealed that SOM-based strategies can outperform traditional benchmarks under various market conditions. Notably, the hybrid and ensemble models produced more balanced results, suggesting that combining diverse perspectives, such as market sentiment and technical indicators, enhances decision robustness. These findings were further supported by quantitative evaluations of the SOMs themselves, using measures such as quantization error and topological preservation, which confirmed the structural reliability of the trained models.

Despite these encouraging results, the study also identified several limitations that shape the direction of future work.

6.1. LIMITATIONS

Several limitations must be acknowledged, both in terms of methodological design and practical implementation. These constraints suggest opportunities for future refinement and should be considered when interpreting the system's general applicability.

A primary limitation arises from the availability and quality of data, particularly for the sentiment analysis component. The model's ability to generate meaningful sentiment-based signals depends heavily on timely and accurate financial news. However, access to such data was partially restricted due to limited API quotas and rate-limited endpoints, which constrained the volume and diversity of articles processed. This bottleneck potentially impacted the representativeness of the sentiment features, especially during high-volatility periods when news flow intensifies. Additionally, financial sentiment extraction from text inherently involves ambiguity, and while FinBERT offers domain-tuned capabilities, misclassifications or context loss can occur, particularly with nuanced or ironic headlines.

From a computational standpoint, training multiple SOMs across different market regimes, feature sets, and random seeds posed scalability challenges. Although MiniSom is efficient for small to medium-sized maps, the combination of repeated training and ensemble formation required significant resource management, particularly during stability analyses.

Another important consideration is the generalizability of the findings. While the model showed reasonable adaptability across bullish, bearish, and sideways regimes within a single asset or asset class, there is no guarantee that these results would extend uniformly to other instruments, markets, or economic conditions. Attempts to test for cross-asset generalization,

though conceptually motivated, were limited by data availability and time constraints. Moreover, structural breaks in the data, such as those introduced by major geopolitical events or shifts in market microstructure, could disrupt the latent patterns the SOM attempts to capture.

A further limitation relates to the interpretability of the SOM ensemble outputs. While the unsupervised nature of SOMs allows for flexible discovery of latent market structures, it also introduces challenges in explaining model decisions to end users or portfolio managers. The current approach to signal generation, based on aggregated cluster-level return statistics, offers some transparency but still lacks the fine-grained justifications that would be expected in a production-level decision-support system. More advanced techniques such as attention mechanisms or post-hoc interpretability tools may help address this gap in the future.

Finally, data leakage and overfitting remain potential concerns, despite careful separation of training and testing sets. The reliance on percentile-based thresholds for signal labeling, while intuitive, may be sensitive to regime shifts or outliers. Similarly, the performance evaluation assumes static transaction cost levels and uniform execution quality, which may diverge from live trading environments where slippage and liquidity constraints can vary dynamically.

6.2. FUTURE WORK

Recognizing these limitations is crucial for contextualizing the results and guiding future work. Addressing them would involve not only expanding the scope of data and testing but also incorporating more advanced modeling techniques, robust validation protocols, and explainability frameworks to make the system both more reliable and more transparent in practical settings.

First, enhancing the sentiment pipeline with access to richer and more timely data, or incorporating more advanced language models like LLMs fine-tuned on financial corpora, could significantly improve signal reliability. Comparing SOM-driven sentiment signals with those generated by transformer-based architectures would offer a valuable benchmark of each method's strengths.

Moreover, while the methodology was evaluated on held-out data within the same asset class, it has yet to be tested across broader market environments. The absence of cross-asset and cross-regime generalization poses a challenge to assessing the true scalability and transferability of the approach. Although the system is designed to support such extensions, further validation across multiple asset classes, including commodities, foreign exchange, and cryptocurrencies, is essential to confirm its generalizability.

The current implementation lacks certain operational features typically found in professional trading systems. In particular, the absence of dynamic stop-loss and take-profit mechanisms means the strategy may be exposed to avoidable losses during periods of volatility. Introducing these features would not only align the system more closely with real-world trading practices but could also improve risk-adjusted returns and capital preservation.

Looking ahead, the use of real-time market regime detection, rather than static regime classification, could enable dynamic strategy switching based on evolving market conditions. This would move the framework closer to adaptive trading systems capable of adjusting behavior in response to macroeconomic shifts or structural market changes, improving the usability of SOM-based systems in institutional settings.

Finally, transitioning from an experimental to an operational system will require improvements in scalability, infrastructure, and automation. Real-time data ingestion, live execution logic, and performance dashboards would be necessary components in actual trading environments.

In conclusion, this thesis presents a robust and flexible architecture for trading signal generation based on unsupervised learning, highlighting the effectiveness of combining technical and sentiment information within a SOM-based structure and laying the groundwork for future enhancements. While challenges remain, especially in terms of scalability, interpretability, and generalization, the modular nature of the framework positions it as a strong foundation for future research and practical development in algorithmic trading systems.

BIBLIOGRAPHICAL REFERENCES

Abe, M., & Nakagawa, K. (2020). *Deep Learning for Multi-factor Models in Regional and Global Stock Markets.* Springer International Publishing, Vol. 12331, 87–102. https://doi.org/10.1007/978-3-030-58790-1_6

Abe, M., & Nakayama, H. (2018). *Deep Learning for Forecasting Stock Returns in the Cross-Section*. https://doi.org/10.48550/arXiv.1801.01777

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. https://doi.org/10.48550/arXiv.1908.10063

Aroussi, R. (2025). *Ranaroussi/yfinance* [Python]. https://github.com/ranaroussi/yfinance

Belamfedel Alaoui, S., Hafid, A., Sayyouri, M., & Rahouti, M. (2025). Leveraging Machine Learning and Deep Learning Models for Enhanced Stock Price Prediction: A State-of-the-Art Analysis. *Distributed Computing and Artificial Intelligence, 21st International Conference,* Springer Nature Switzerland, 53–64. https://doi.org/10.1007/978-3-031-82073-1_6

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications*, *42*(14), 5963–5975. https://doi.org/10.1016/j.eswa.2015.03.017

Cuello, F. (2024). *Self-Organizing Maps*. Trendspider. https://trendspider.com/learning-center/self-organizing-maps/

Deboeck, G., & Kohonen, T. (Eds.). (1998). *Visual Explorations in Finance*. Springer. https://doi.org/10.1007/978-1-4471-3913-3

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Enke, D., & Mehdiyev, N. (2013). Stock Market Prediction Using a Combination of Stepwise Regression Analysis, Differential Evolution-based Fuzzy Clustering, and a Fuzzy Inference Neural Network. *Intelligent Automation & Soft Computing*, *19*(4), 636–648. https://doi.org/10.1080/10798587.2013.839287

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Griemsmann, B. (2022). *Sentiment analysis of online media—Extracting a trading signal for commodities*. https://run.unl.pt/handle/10362/142654

Hajric, V., & Shen, M. (2022). Crypto Markets Extend Drop as Doubts About FTX-Binance Deal Grow. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2022-11-08/cryptocurrencies-plunge-as-doubts-about-ftx-binance-deal-grow

Harder, P., & Fazlija, B. (2024). Using Financial News Sentiment for Stock Price Direction Prediction. *ResearchGate*. https://doi.org/10.3390/math10132156

Hsu, S.-H., Hsieh, J. P.-A., Chih, T.-C., & Hsu, K.-C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, *36*(4), 7947–7951. https://doi.org/10.1016/j.eswa.2008.10.065

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). https://doi.org/10.1609/icwsm.v8i1.14550

Johnsson, M. (Ed.). (2012). *Applications of Self-Organizing Maps*. InTech. https://doi.org/10.5772/3464

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (2001). *Self-Organizing Maps* (Vol. 30). Springer. https://doi.org/10.1007/978-3-642-56927-2

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Leinweber, D. J., & Madhavan, A. N. (2001). Three Hundred Years of Stock Market Manipulations. *The Journal of Investing*, *10*(2), 7–16. https://doi.org/10.3905/joi.2001.319457

Li, L., He, W., Xu, L., Ash, I., Anwar, M., & Yuan, X. (2019). Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. *International Journal of Information Management*, *45*, 13–24. https://doi.org/10.1016/j.ijinfomgt.2018.10.017

López de Prado, M. M. (2020). *Machine Learning for Asset Managers*. Cambridge University Press. https://doi.org/10.1017/9781108883658

Loria, S. (2025). *TextBlob: Simplified Text Processing*. https://textblob.readthedocs.io/en/dev/

Matos, D., Marques, N. C., & Cardoso, M. G. M. S. (2014). Stock market series analysis using self-organizing maps. *Revista de Ciências Da Computação*, *9*, 79–90.

Mohamed, A. (2019). *Artificial intelligence in investing: Stock clustering with self-organizing map and return prediction with model comparison*. https://lutpub.lut.fi/handle/10024/159794

Pei, D., Luo, C., & Liu, X. (2023). Financial trading decisions based on deep fuzzy self-organizing map. *Applied Soft Computing*, *134*, 109972. https://doi.org/10.1016/j.asoc.2022.109972

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25–33. https://doi.org/10.1007/s10489-006-0001-7

Schau, E. (2024). *BTC_yahoo.csv·* https://huggingface.co/datasets/edaschau/bitcoin_news/blob/main/BTC_yahoo.csv

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600. IEEE Transactions on Neural Networks. https://doi.org/10.1109/72.846731

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, 126–139. https://doi.org/10.1016/j.eswa.2016.09.027

APPENDIX A - FEATURES

Feature	Description
Close	Daily closing price of the asset
High	Daily highest price
Low	Daily lowest price
Open	Daily opening price
Volume	Daily trading volume
RSI	Relative Strength Index (momentum indicator)
MACD	Moving Average Convergence Divergence (trend indicator)
CCI	Commodity Channel Index (trend indicator)
WILLR	Williams %R (momentum indicator)
SMA_50	50-day Simple Moving Average
SMA_200	200-day Simple Moving Average
EMA_50	50-day Exponential Moving Average
EMA_200	200-day Exponential Moving Average
Bollinger_Upper	Upper band of Bollinger Bands
Bollinger_Lower	Lower band of Bollinger Bands
ATR	Average True Range (volatility indicator)
OBV	On-Balance Volume (volume-based momentum)
CMF	Chaikin Money Flow (volume and price pressure)
Momentum	Momentum indicator
ROC	Rate of Change
UO	Ultimate Oscillator (momentum indicator)
avg_sentiment_score	Mean FinBERT confidence score
avg_sentiment_label_score	Mean numeric sentiment label
pos_count	Count of positive news items
neut_count	Count of neutral news items
neg_count	Count of negative news items
total	Total number of sentiment samples per day
pct_positive	Percentage of positive sentiment
pct_neutral	Percentage of neutral sentiment
pct_negative	Percentage of negative sentiment
sentiment_score_combined	Weighted combination of sentiment scores
sentiment_polarization	Difference between positive and negative sentiment
log_total_sentiment	Log-transformed total sentiment count
z_sentiment_score_combined	Z-score normalized combined sentiment score
z_sentiment_polarization	Z-score normalized sentiment polarization
z_log_total_sentiment	Z-score normalized log total sentiment count

Table 8 - Features resulting from preprocessing

APPENDIX B - GRID SEARCH FOR SOM HYPERPARAMETERS

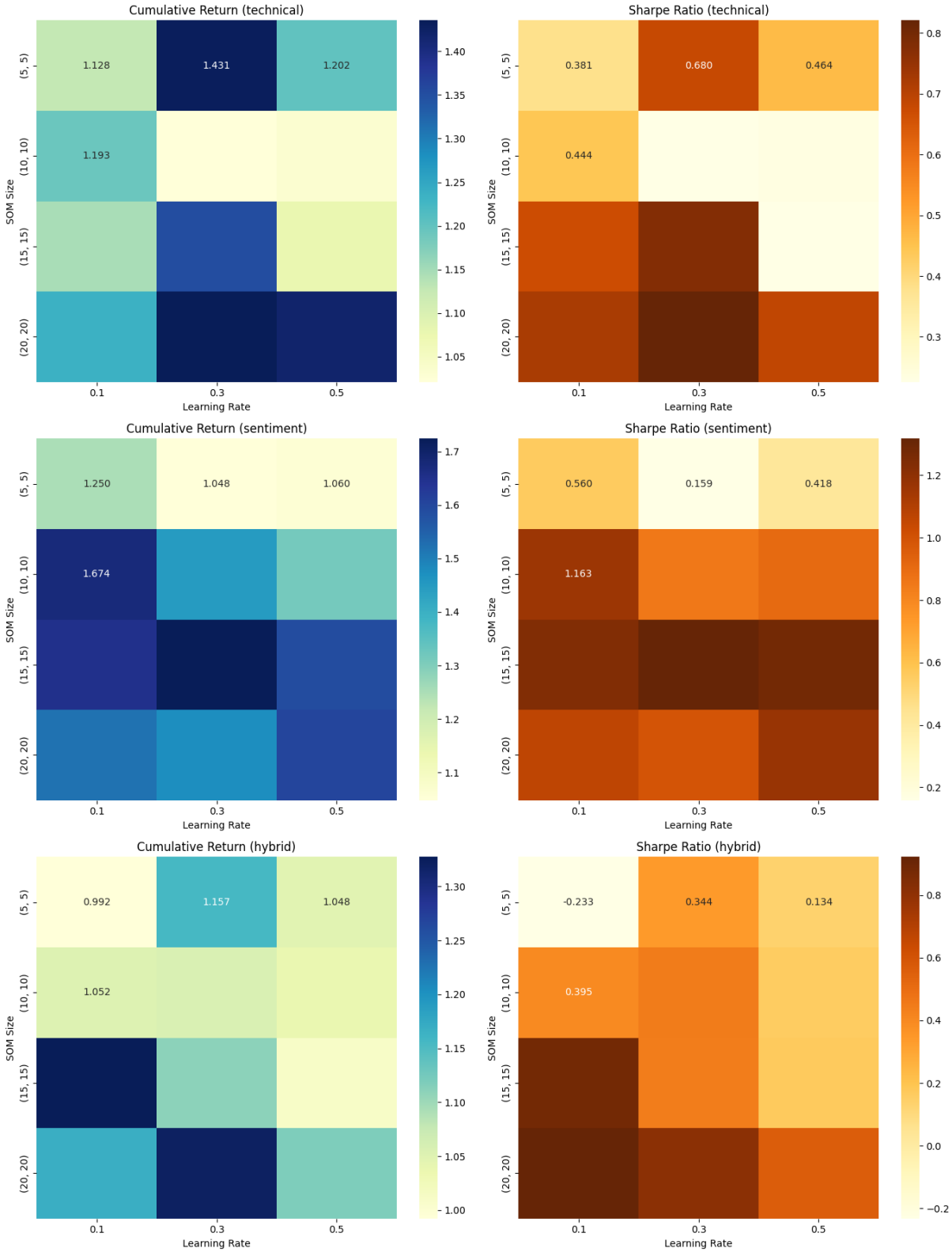

Figure 14. Comparison of Cumulative Return and Sharpe Ratio across SOM configurations (varying SOM Size and Learning Rate) for each SOM model type

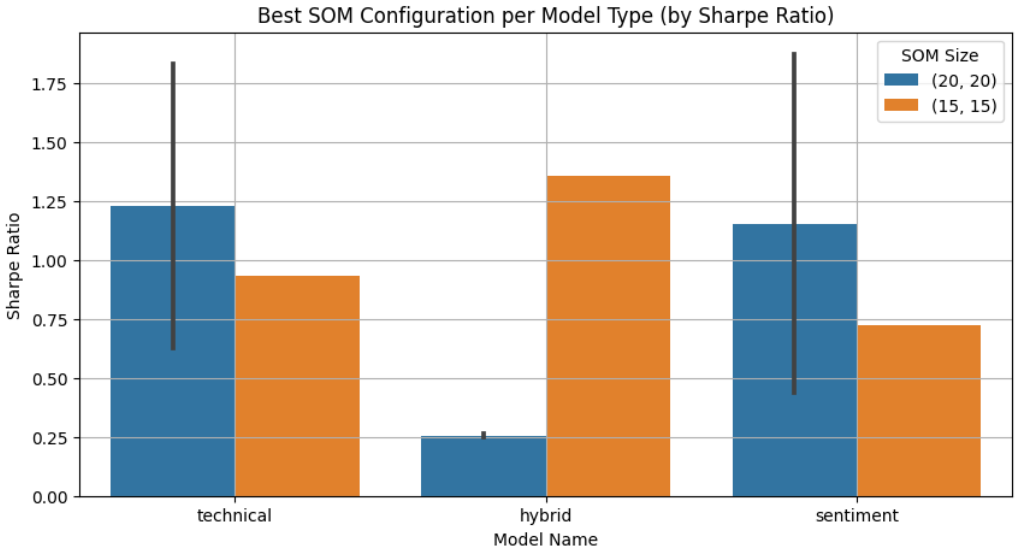

Figure 15 - Sharpe Ratio per SOM hyperparameter configuration for the technical SOM

APPENDIX C - SOM DENSITY HEATMAPS

This appendix provides additional visual support for the interpretation of SOMs, displaying the density distribution of input vectors across the map. Each heatmap shows the number of data instances mapped to each neuron, complementing the U-Matrix interpretation in Chapter 3.3.

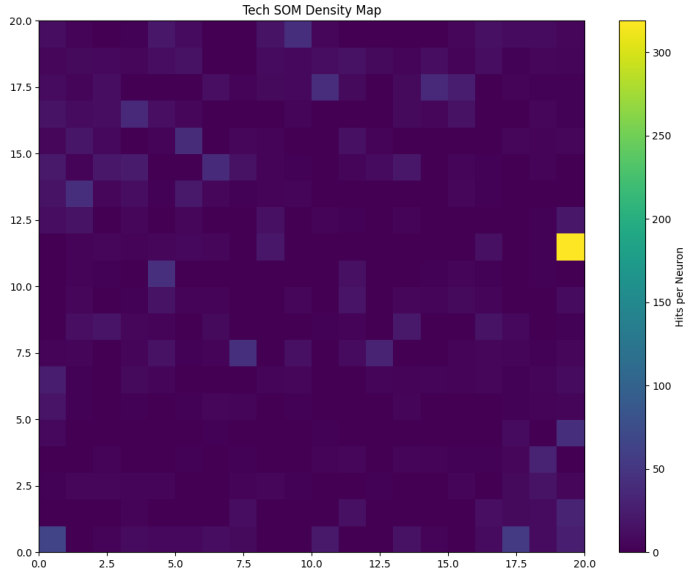

Figure 16. Density heatmap for the Technical SOM

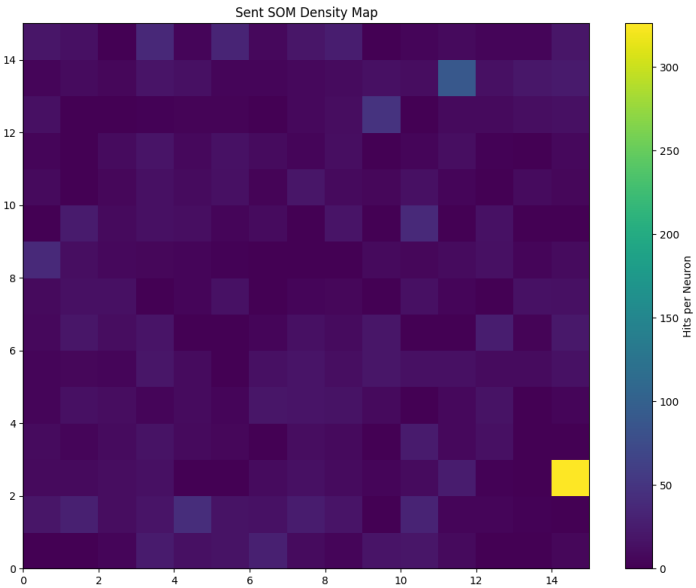

Figure 17. Density heatmap for the Sentiment SOM

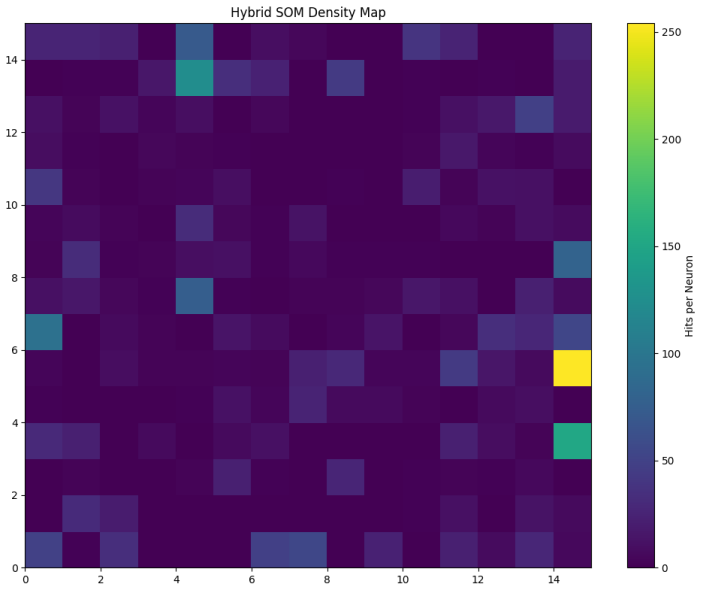

Figure 18. Density heatmap for the Hybrid SOM

APPENDIX D - ADDITIONAL SOM SIGNAL VISUALIZATIONS

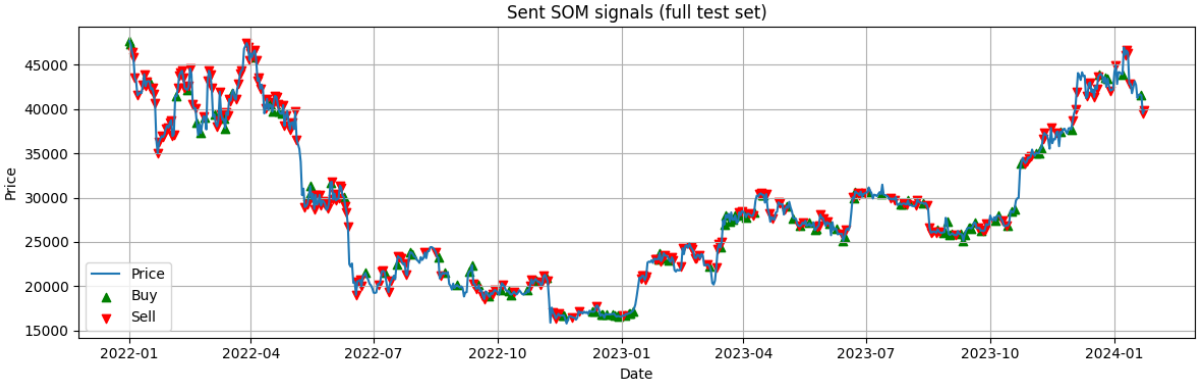

Figure 19 - Signals generated from the clusters outputted from the SOM trained with sentiment data

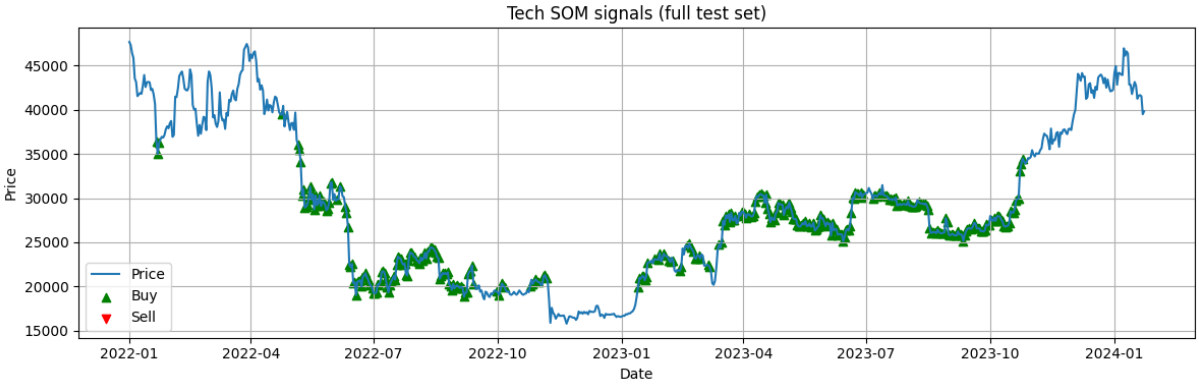

Figure 20 - Signals generated from the clusters outputted from the SOM trained with technical data

APPENDIX E - LABELED U-MATRIX VISUALIZATIONS FOR TECHNICAL AND HYBRID SOMS

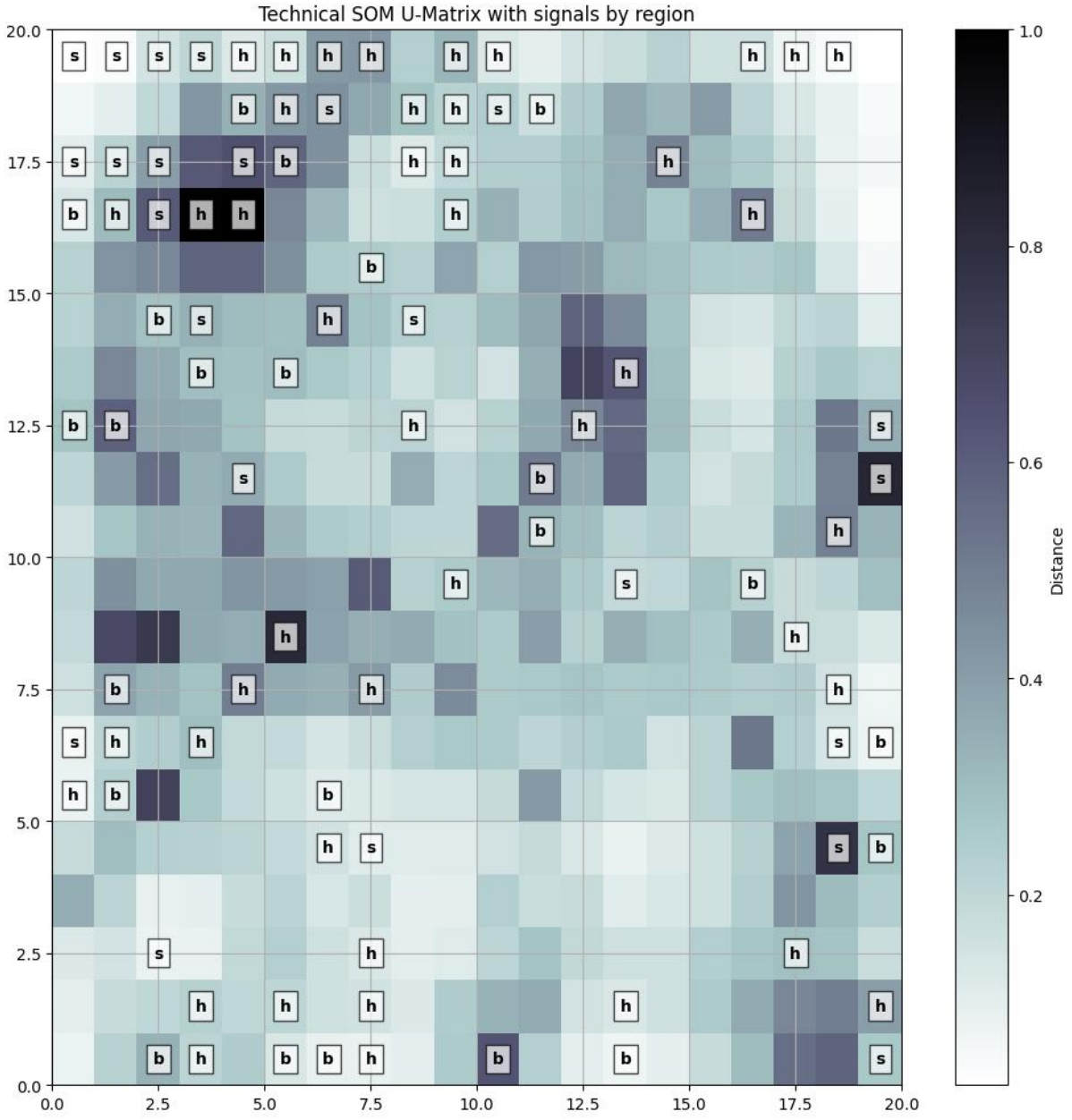

Figure 21 - Labeled U-Matrix of the Technical SOM with Buy/Hold/Sell Signals

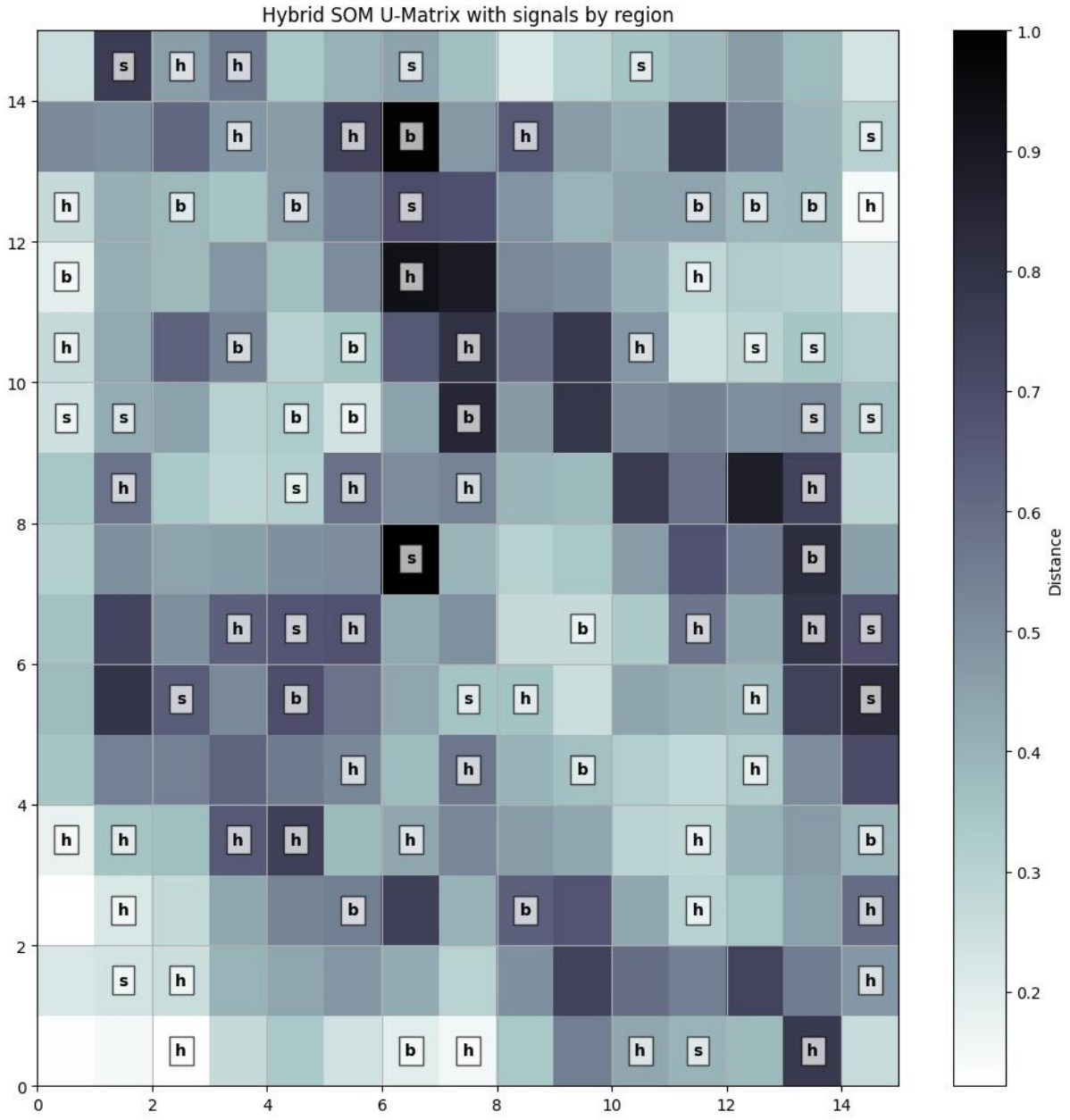

Figure 22 - Labelled U-Matrix of the Hybrid SOM with Buy/Hold/Sell Signals

APPENDIX F - SIGNAL CORRELATION ANALYSIS

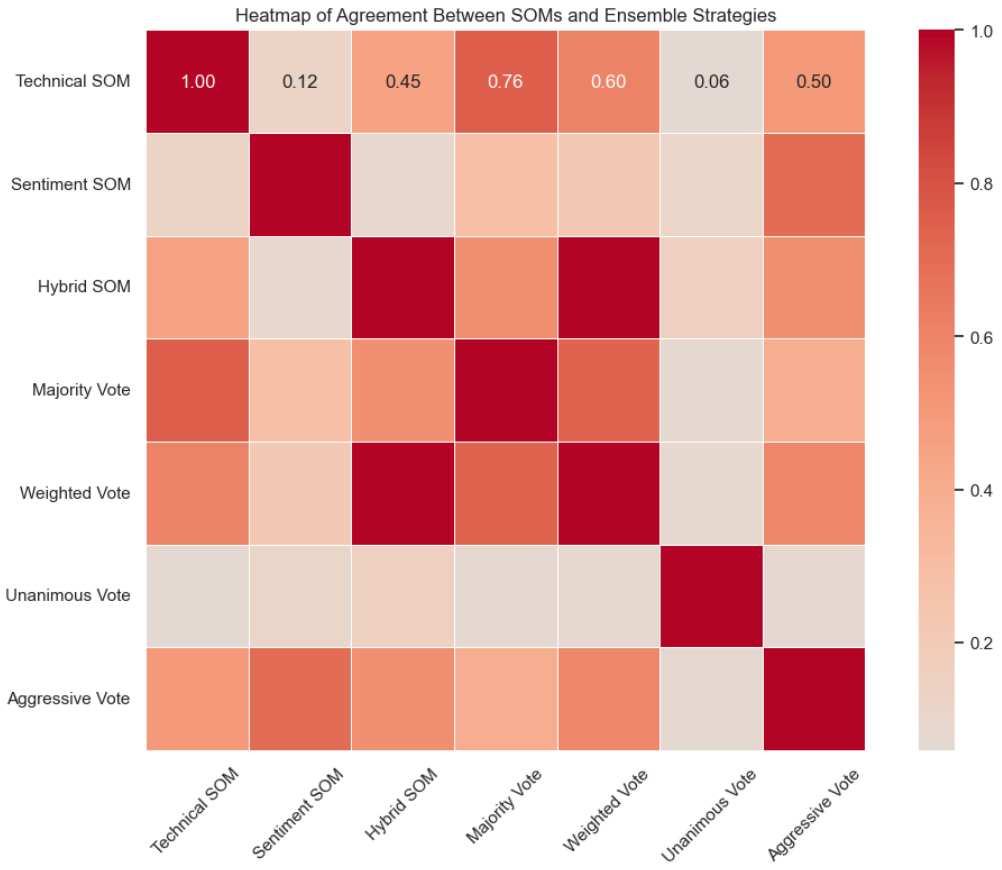

Figure 23 - Correlation Between Trading Signals from SOM Models and Ensemble Strategies

APPENDIX G - CODE REPOSITORY

The full implementation of the methodology, including data preprocessing, training of SOMs, ensemble strategies, evaluation scripts, and an executable Jupyter Notebook main.ipynb demonstrating the full workflow, is publicly available in the following GitHub repository: https://github.com/catarinagoliveira/som_for_trading

ANNEX A - TECHNICAL INDICATOR FORMULAS

1. Simple Moving Average (SMA)

$$SMA_n = \frac{1}{n}\sum_{i=1}^{n} Price_i$$

2. Exponential Moving Average (EMA)

$$EMA_t = \alpha * Price_t + (1 - \alpha) * EMA_{t-1}, \text{ where } \alpha = \frac{2}{n+1}$$

3. Relative Strength Index (RSI)

$$RSI = 100 - \left(\frac{100}{1+RS}\right), \text{ where } RS = \frac{Average\ Gain}{Average\ Loss}$$

4. Moving Average Convergence Divergence (MACD)

$$MACD = EMA_{12} - EMA_{26}$$

$$Signal\ Line = EMA_9(MACD)$$

Note: The MACD uses EMAs of closing prices over 12 and 26 periods to capture short and long-term momentum. The signal line, a 9-period EMA of the MACD, is used to identify potential buy/sell signals when it crosses the MACD line.

5. Commodity Channel Index (CCI)

$$CCI = \frac{(TP - SMA_n(TP))}{0.015 * Mean\ Deviation}, \text{ where } TP = \frac{High+Low+Close}{3}$$

6. Williams %R (WILLR)

$$WILLR = \frac{Highest\ High_n - Close}{Highest\ High_n - Lowest\ Low_n} * (-100)$$

7. Bollinger Bands (Upper & Lower)

$$Middle\ Band = SMA_n$$
$$Upper\ Band = SMA_n + k * Standard\ Deviation$$
$$Lower\ Band = SMA_n - k * Standard\ Deviation$$

where: k is usually 2, which covers approximately 95% of the expected price variation, assuming a normal distribution.

8. Average True Range (ATR)

$$TR_t = \max(High_t - Low_t, |High_t - Close_{t-1}|, |Low_t - Close_{t-1}|)$$

$$ATR_n = EMA_n(TR)$$

9. On-Balance Volume (OBV)

$$OBV_t = \begin{cases} OBV_{t-1} + Volume_t, if\ Close_t > Close_{t-1} \\ OBV_{t-1} - Volume_t, if\ Close_t < Close_{t-1} \\ OBV_{t-1}, otherwise \end{cases}$$

10. Chaikin Money Flow (CMF)

$$CMF = \frac{\sum_{i=1}^{n} \left(\frac{(Close_i - Low_i) - (High_i - Close_i)}{High_i - Low_i} * Volume_i \right)}{\sum_{i=1}^{n} Volume_i}$$

11. Momentum

$$Momentum_n = Close_t - Close_{t-n}$$

12. Rate of Change (ROC)

$$ROC_n = \left(\frac{Close_t - Close_{t-n}}{Close_{t-n}} \right) * 100$$

13. Ultimate Oscillator (UO)

$$BP_t = Close_t - \min(Low_t, Close_{t-1})$$

$$TR_t = \max(High_t - Low_t, |High_t - Close_{t-1}|, |Low_t - Close_{t-1}|)$$

$$UO = 100 * \frac{4*AVG_7 + 2*AVG_{14} + AVG_{28}}{4+2+1}, \text{where } AVG_n = \Sigma BP_n / \Sigma TR_n$$

Note: UO blends short-, medium-, and long-term momentum by calculating buying pressure over three time frames (7, 14, and 28 periods). The weights (4, 2, 1) give greater importance to more recent price action, aiming to reduce false signals common in single-period oscillators.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa