UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

SHORT-TERM STOCK MARKET PREDICTION WITH A HYBRID MACHINE LEARNING APPROACH

Ljubljana, March 2026 ROK STANIČ

AUTHORSHIP STATEMENT

The undersigned Rok Stanič, a student at the University of Ljubljana, School of Economics and Business, (hereafter: UL SEB), author of this written final work of studies with the title Short-term stock market prediction with a hybrid machine learning approach, prepared in collaboration with mentor red. prof. dr. Jurij Jaklič.

D E C L A R E

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the UL SEB Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the UL SEB Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the UL SEB in accordance with the relevant UL SEB Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the UL SEB Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published;
11. that I have verified the authenticity of the information derived from the records using artificial intelligence tools.

Ljubljana, March 5, 2026 Author's signature: Rok Stanič

ABSTRACT

Investing in the stock market offers opportunities for long-term capital growth, yet short-term price prediction remains inherently difficult despite decades of research in financial economics and machine learning. This thesis investigates whether combining sentiment scores from financial news headlines with technical indicators can enhance predictive accuracy and support better investment decision-making for short-term traders. Six LSTM models were developed across 5, 10, and 20-day horizons. We compared models that were based only on technical indicators with sentiment-augmented models, trained on more than 100,000 observations from 212 securities. In our setup, sentiment-augmented models proved particularly useful for managing downside exposure and capturing the magnitude of price movements, whereas technical models achieved higher directional accuracy. Sentiment effectiveness varies by horizon: it adds disturbances at short timeframes but provides optimal value at medium-term horizons. By analyzing multiple predictive dimensions across different time horizons, this thesis shows that in our setup sentiment integration introduces measurable trade-offs between directional accuracy and magnitude estimation, thereby offering a more nuanced understanding of model performance.

KEY WORDS: machine learning, stock prediction, sentiment analysis, financial forecasting, technical analysis

SUSTAINABLE DEVELOPMENT GOALS

POVZETEK

Vlaganje na delniški trg ponuja priložnosti za dolgoročno rast premoženja, a napovedovanje kratkoročnih gibanj cen ostaja zelo zahtevno, kljub več desetletnemu raziskovanju na področju finančne ekonomije in strojnega učenja. Magistrsko delo raziskuje, ali lahko kombinacija sentimentih ocen iz naslovov finančnih novic s tehničnimi indikatorji izboljša napovedovalno natančnost in podpre boljše odločanje za kratkoročne vlagatelje. Na podlagi nevronskih mrež LSTM je bilo razvitih šest modelov za napovedna časovna obdobja 5, 10 in 20 dni. Primerjali smo modele, ki temeljijo samo na tehničnih kazalnikih, z modeli, nadgrajenimi z indikatorji sentimenta, naučenimi na več kot 100.000 opazovanjih iz 212 vrednostnih papirjev. Naše ugotovitve razkrivajo, da so modeli razširjeni z indikatorji sentimenta posebej uporabni pri obvladovanju tveganja padcev, in zajemanju obsega gibanja cen, medtem ko imajo tehnični modeli boljše rezultate pri klasičnem napovedovanju samo smeri. Učinek sentimenta je tudi odvisen od dolžine napovednega časovnega obdobja: na krajših intervalih lahko vnese motnje, medtem ko se njegova dodana vrednost najboljše pokaže pri srednjeročnih napovedih. Z analizo več napovednih dimenzij skozi različna časovna obdobja magistrsko delo pokaže, da integracija sentimenta v našem okviru uvaja merljive kompromise med natančnostjo napovedovanja smeri in ocenjevanjem obsega gibanj cen, s čimer ponuja bolj poglobljeno razumevanje zmogljivosti modelov.

KLJUČNE BESEDE: strojno učenje, napovedovanje delnic, analiza sentimenta, finančno napovedovanje, tehnična analiza

CILJI TRAJNOSTNEGA RAZVOJA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

sl. – Slovene

ANN – (sl. umetna nevronska mreža); Artificial Neural Network

API – (sl. vmesnik za programiranje aplikacij); Application Programming Interface

ATR – (sl. povprečni dejanski razpon); Average True Range

BERT – (sl. dvosmerne predstavitve iz transformatorjev); Bidirectional Encoder Representations from Transformers

CMF – (sl. Chaikinov denarni tok); Chaikin Money Flow

CRISP-DM – (sl. medpanožni standardni postopek za podatkovno rudarjenje); Cross-Industry Standard Process for Data Mining

CSV – (sl. vrednosti, ločene z vejicami); Comma-Separated Values

DAX – (sl. nemški borzni indeks); Deutscher Aktienindex

EMA – (sl. eksponentno drseče povprečje); Exponential Moving Average

ETF – (sl. borzno trgovalni sklad); Exchange-Traded Fund

FinBERT – (sl. finančne dvosmerne predstavitve iz transformatorjev); Financial Bidirectional Encoder Representations from Transformers

FTSE – (sl. borzni indeks Financial Times); Financial Times Stock Exchange

GRU – (sl. vratna rekurentna enota); Gated Recurrent Unit

IPO – (sl. prva javna ponudba delnic); Initial Public Offering

LSTM – (sl. dolgokratkoročni spomin); Long Short-Term Memory

MA – (sl. drseče povprečje); Moving Average

MACD – (sl. konvergenca in divergenca drsečih povprečij); Moving Average Convergence Divergence

MAE – (sl. povprečna absolutna napaka); Mean Absolute Error

ML – (sl. strojno učenje); Machine Learning

NLP – (sl. obdelava naravnega jezika); Natural Language Processing

NYSE – (sl. Newyorška borza); New York Stock Exchange

OBV – (sl. obseg na podlagi bilance); On-Balance Volume

OHLC – (sl. odprtje, najvišja cena, najnižja cena, zaprtje); Open, High, Low, Close

OHLCV – (sl. odprtje, najvišja cena, najnižja cena, zaprtje, obseg); Open, High, Low, Close, Volume

R^2 – (sl. koeficient determinacije); Coefficient of Determination

RMSE – (sl. koren povprečne kvadratne napake); Root Mean Square Error

RNN – (sl. rekurentna nevronska mreža); Recurrent Neural Network

ROC – (sl. stopnja spremembe); Rate of Change

RSI – (sl. indeks relativne moči); Relative Strength Index

S&P 500 – (sl. indeks StandardandPoor's 500); StandardandPoor's 500

SMA – (sl. preprosto drseče povprečje); Simple Moving Average

SPDR – (sl. potrdilo o pologu StandardandPoor's); StandardandPoor's Depositary Receipt

SPY – (sl. sklad SPDR S&P 500); SPDR S&P 500 ETF Trust

SVM – (sl. stroj podpornih vektorjev); Support Vector Machine

TF-IDF – (sl. frekvenca izraza in inverzna frekvenca dokumenta); Term Frequency-Inverse Document Frequency

VIX – (sl. indeks volatilnosti); Volatility Index

VWAP – (sl. z obsegom uteženo povprečje cene); Volume-Weighted Average Price

XGBoost – (sl. ekstremno gradientno povečanje); Extreme Gradient Boosting

1 INTRODUCTION

Investing in the stock market is commonly regarded as one of the primary mechanisms for long-term wealth accumulation in modern economies. The idea of it is simple, a marketplace of securities, stocks and bonds facilitated by institutions which can turn an average investor's monthly deposit into a fortune within a given time period. It represents a real opportunity for wealth creation, but the challenge and opportunity for some involves the almost impossible way of predicting the way the stock market will move. Conventional wisdom tells us that building a diverse portfolio meaning investing in multiple companies or exchange traded funds is the most rational and logical step to take but looking at human nature this idea might not appeal to everyone. Long term investors are faced with a far less volatile and uncertain landscape compared to short term investors. The prediction of market prices in the short term is notoriously difficult, markets can move on to an infinite number of variables, they are influenced by historical data, investors sentiment, macro and microeconomic events and unforeseen shocks like global pandemics. Short term prediction is a worldwide effort because it can result in higher gains faster than a long-term strategy and is more dynamic combining multiple economies, behavioral finance, computer science and data science (Barberis and Thaler, 2003; Fischer and Krauss, 2018).

The Efficient Market Hypothesis, a widespread idea in financial economics argues that asset prices reflect all available information meaning all stocks are always valued fairly on exchanges, making it not only impossible to purchase undervalued stocks but also implying that systemic prediction is not possible (Fama, 1970). While the hypothesis might be widespread it is highly contested as there have been many who have outperformed the market regularly, but the opposite is also true, there are decades of empirical research that shows investors frequently underperform market benchmarks due to a number of behavioral biases and decision-making errors (Barber and Odean, 2000). For example, overconfidence often leads to excessive trading resulting in greater losses than profits, looking at day trading the majority of retail investors lose money (Barber et al., 2014). Events against efficient valuation are also present within the history of the stock market, in 1987 the stock market crashed, and the Dow Jones Industrial Average (DJIA) fell by over 20% in a single day, there have been many more bubbles that have burst which would not happen in a market described by the hypothesis. A paradox is created if markets are efficient as the hypothesis claims, prediction is futile, but if inefficiencies exist due to human behavior, structural anomalies or other reasons, then prediction especially through advanced analytical methods not only remains possible but also viable.

Given that human behavior for most investors negatively affects their decision making we take the human out of it. Machine learning (ML) has emerged as a promising alternative to human judgment in stock prediction. On paper it makes sense because such algorithms detect complex, nonlinear patterns in data, exactly what we would need for financial forecasting tasks. The question is proposed if these algorithms are so viable, why are they not more widespread. There are multiple answers, they are not as accessible to the average investor, but they are definitely prevalent among investment funds which like to keep their algorithms and predicting strategies private so there is little to compare to, but another big reason is the lack of universal strategy for positive results, there is no clear path to beating the stock market which again highlights the vast complexity of this system. Early research and iterations presented by Zhang (2003), implemented hybrid models which combine classical statistical methods with neural networks. This approach also served as a foundation for further research, that traditional models can be improved using neural networks, which are capable of recognizing nonlinear influences in data. Later works, such as Fischer and Krauss (2018), highlighted the effectiveness of Long Short-Term Memory (LSTM) networks in capturing sequential dependencies in financial time-series data. These contributions have established machine learning as a legitimate, though still developing, method for addressing the challenges of short-term forecasting.

However, numerical data alone, such as stock prices, trading volumes, and technical indicators, often fails to capture the complete picture of market behavior. The role of qualitative data, particularly news and sentiment, has been increasingly recognized. Loughran and McDonald (2011) demonstrated that negative news coverage has measurable effects on stock returns, while Bollen et al. (2011) showed that collective mood signals from Twitter correlate with market movements. More recently, advances in Natural Language Processing (NLP) such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019) have enabled the transformation of textual data into dense numerical embeddings, which can subsequently be used as inputs for sentiment analysis and predictive models, creating new opportunities to integrate financial and textual information.

Recent research suggests that predictive success in short-term markets depends not only on model architecture, but also on horizon selection, feature composition, and evaluation methodology (Fischer and Krauss, 2018). This thesis therefore adopts a comparative, multi-horizon framework, explicitly contrasting price-based and sentiment-augmented models across multiple predictive dimensions, including directional accuracy, correlation, and trading signal quality.

The purpose of this master's thesis is to enhance predictive accuracy, improve investment decision making and contribute to the development of stock market predicting models that leverage both structured numerical data and numerical sentiment features derived from unstructured text, with particular focus on understanding how sentiment-based features differ from technical indicators in their predictive characteristics and practical trading applications. Machine learning models have proven to be capable of helping predict price movements, with this approach an investor would have help with when to buy a stock and when to sell when talking about short term purchases, given that the combination of all time sensitive variables and numeric variables derived from news headlines provide a deeper insight and to the discovery of unseen influences between certain indicators that effect the markets movement. Investors spend more time researching which microwave to buy and then buy stocks worth much more with a fraction of the research (Lynch, 1989). Using a hybrid model can give instant feedback without relying only on a person's own knowledge, ability and instincts. The main motivation is providing something that transcends traditional analysis with a comprehensive view of hidden market dynamics to regular investors. A hybrid method enhances predictive accuracy and deepens the understanding of market behavior.

The main goal of this thesis is the design, implementation, and optimization of at least one recurrent neural network specifically trained to predict short-term stock price movements based on custom input data. The objective is to evaluate whether combining historical financial data, technical indicators, and sentiment information extracted from news headlines can improve predictive performance. In addition, the thesis aims to identify suitable prediction targets for short-term forecasting and to compare multiple model configurations in order to determine which combinations yield the most reliable results. Ultimately, the goal is to assess whether sentiment- and history-driven trading signals can enhance forecasting accuracy and support improved decision-making strategies for short-term investors.

To achieve these goals, several machine learning models will be designed, implemented, and experimentally evaluated, with particular emphasis on recurrent neural networks such as Long Short-Term Memory (LSTM) architectures due to their ability to model sequential time-series data and preserve temporal dependencies. A data acquisition and management system will be developed to collect and organize historical financial data, technical indicators, and news headlines for selected companies. In addition to standard technical indicators, custom indicator combinations will be constructed. Sentiment information derived from news headlines will be integrated with numerical data, and multiple models will be trained and compared to evaluate their relative performance.

While prior studies often focus on a single prediction horizon or evaluation metric, this thesis adopts a comparative perspective. By evaluating multiple forecasting horizons and contrasting price-based and sentiment-augmented models, the study aims to assess not only predictive accuracy, but also the practical usefulness of predictions in a trading-oriented context. This approach allows differences in model behavior to emerge that may remain hidden under more aggregated evaluation frameworks.

2 STOCK MARKET

In the simplest terms the stock market is a global network or rather a marketplace where investors buy shares in publicly traded companies, which in turn help the companies raise capital and grow. In return the investors are given stocks, a unit of ownership in a company which are issued by the same public company (Bodie et al., 2018). The first stocks sold are usually through an initial public offering (IPO), after which the company’s stocks are traded on an exchange platform. The price is determined like the prices of all things, by the supply and demand for it. Anyone, both large institutions and individuals, can participate in this system. The investors find this appealing because they can increase the value of their assets and even participate in the profits over time. (Malkiel, 2019).

2.1 Overview of Stock Market Dynamics and Challenges in Short-Term Stock Prediction

Stock markets are an ever-changing, highly adaptive, complex system where prices can move dramatically in minutes as countless investors, institutions and algorithms all interact with each other.

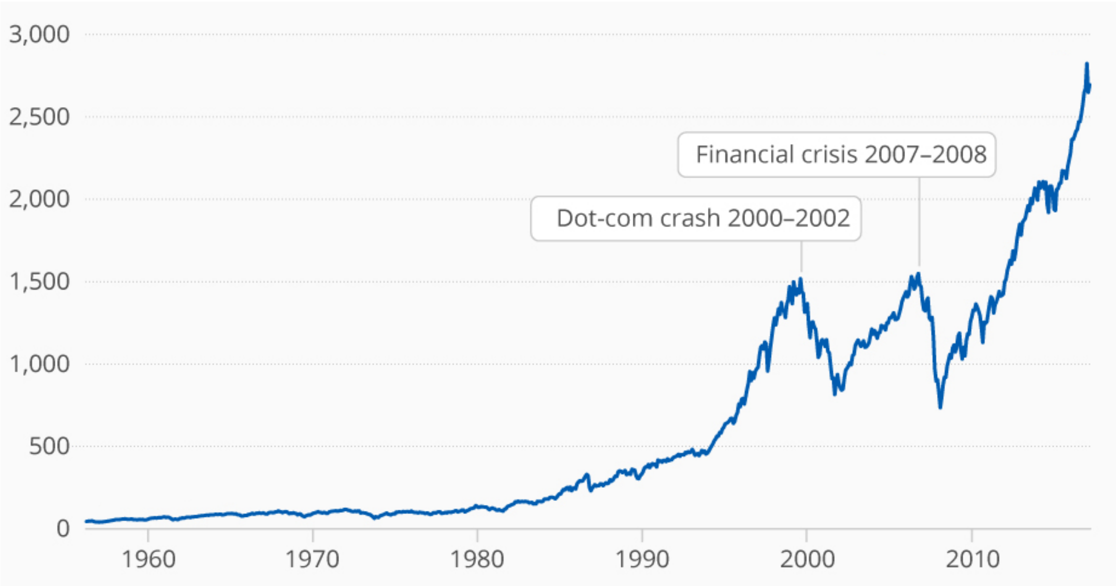
Key drivers of short-term price movements include (Chen et al., 1986; Bodie et al., 2018; Barberis and Thaler, 2003; Brock et al., 1992):

- Earnings announcements
- Macroeconomic data releases (inflation, employment, GDP)
- Central bank interest rate decisions
- Liquidity conditions
- Order book depth and market microstructure
- Momentum effects and technical trading activity
- Geopolitical shocks
- Unexpected corporate events (mergers, scandals, guidance changes)

These dynamics produce a highly nonlinear environment where shocks like geopolitical events, regulatory decisions or unexpected corporate announcements can generate major short term volatility (Fama, 1970). Research suggests that anomalies such as momentum, short-term reversals and calendar effects can be used to make short term predictions contradicting the Efficient Market Hypothesis. While it is possible to recognize anomalies for successful trading it is difficult to know not only which ones will significantly affect the prices, knowing when they will happen and by how much is also a challenge.

Time in the market beats timing the market is a quote attributed to Kenneth Fisher, but it represents a much greater sentiment among many notable investors one of the most famous being Warren Buffet, who preaches that investing for the long term in companies you understand is the best strategy rather than trying to predict short term market movements. And it is a reliable way to increase wealth from the scope of long-term investments only focusing on indexes such as the Dow Jones or S&P 500 would have proved beneficial as they have both grown by over 300 percent in the last twenty years, which is even clearly visually presented in Figure 1. (Yahoo finance, n.d.). Reduced transaction costs, tax benefits, compounding, riding out short term volatility, not dealing with news everyday helping with peace of mind, rather uncomplex and low maintenance are all benefits of long-term investing (Malkiel and Ellis, 2012).

Figure 1: Performance of S&P 500



Source: Loesche (2018)

While long-term investing strategies often appear advantageous for the average investor, short-term fluctuations remain economically significant. Hedge funds, traders, and risk managers closely monitor short-term price movements because they affect liquidity conditions, portfolio risk metrics, and trading opportunities. Sharp price swings can influence margin requirements, rebalancing strategies, and market stability. As a result, understanding short-term volatility remains an important analytical challenge for financial research and trading strategies. (Lo, 2004).

Despite the risks associated with short-term trading, many market participants remain attracted to these strategies. The primary motivation is the possibility of rapid gains generated by daily price fluctuations. Empirical evidence shows that retail investors often underperform market benchmarks due to excessive trading and overconfidence (Barber and Odean, 2000). Nevertheless, the appeal of quick profits and the psychological excitement associated with speculation continue to draw individuals toward active trading (Leslie et al., 2024).

Institutionally short-term horizons are not merely for speculating the health of the market but often necessary. High frequency trading is such a competitive environment where milliseconds can determine profitability. Research by Aldridge (2013) and Brogaard et al. (2014) demonstrates that algorithmic and high-frequency traders contribute significantly to liquidity provision while simultaneously extracting profits from transient price discrepancies. Institutional actors exploit arbitrage opportunities, react to earnings announcements, or hedge exposures. In this sense short term trading is a critical component of modern market microstructure (Kirilenko et al., 2017). Even in the presence of evidence favoring long-term horizons, the economic incentives for institutional short-term activity remain powerful.

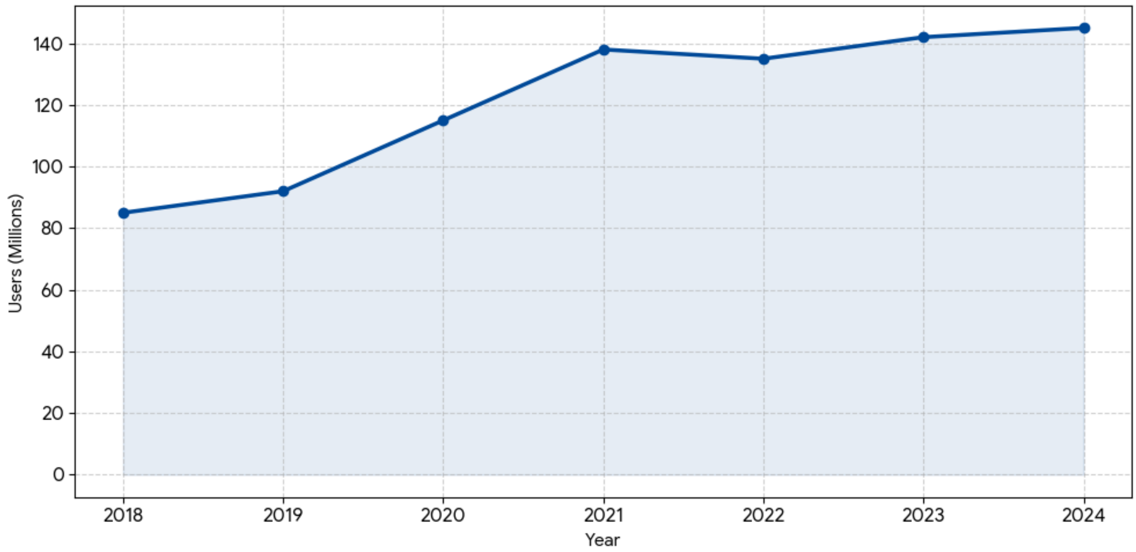
Behavioral finance research helps explain why retail investors continue to engage in short-term trading despite the risks. Cognitive biases such as herding, representativeness, and overconfidence encourage frequent trading, often reducing long-term returns (Barberis and Thaler, 2003; Dorn and Sengmueller, 2009). Studies also emphasize emotional drivers such as fear of missing out (FOMO) and the excitement associated with speculative gains, which can further reinforce active trading behavior (Hoffmann and Shefrin, 2014).

One of the most famous cases of investing with almost no fundamental backing is the GameStop phenomenon. It was widely portrayed as a David versus Goliath narrative: retail investors uniting to take down "elite" hedge funds that held massive short positions. Indeed, hedge funds such as Melvin Capital and Citron Research faced dramatic losses. Melvin lost approximately 53% of its assets under management in January 2021 and required a $2.75 billion capital infusion from Citadel and Point72 to remain solvent (Eavis and Goldstein, 2021). Citron Research, one of the most vocal short-sellers of GameStop, also announced that it would exit short-selling altogether following the event (Franck, 2021). Despite these high-profile losses, the broader hedge fund sector was barely affected. Many institutional players managed their exposures, and some firms, particularly algorithmic and high-frequency traders, profited from the volatility through liquidity provision and arbitrage (SEC, 2021). The Securities and Exchange Commission's post-mortem report highlighted that while certain short sellers incurred substantial losses, other professional investors, especially quantitative funds, benefited significantly from the unprecedented trading activity. While the narrative of retail investors defeating Wall Street captured public attention, the institutional sector at large remained resilient, and in some cases, opportunistically profited. It is a perfect example of the power of social media driving collective action in shaping short-term market dynamics, creating a surge in demand with this sentiment-driven coordination producing price movements detached from fundamentals and sustained volatility that rational pricing models could not explain. (Eaton et al., 2022; Lucchini et al., 2021).

Technological innovations have significantly lowered barriers to entry for retail investors. Commission-free trading platforms, fractional share ownership, and real-time market data have expanded participation in financial markets. Research shows that platforms such as Robinhood increased trading activity, particularly among younger investors, while lower transaction costs encouraged more frequent trading (Barber et al., 2022; Gempesaw et al., 2022).

The expansion of retail participation accelerated dramatically during the COVID-19 pandemic. Lockdowns, stimulus payments, and the rapid adoption of commission-free trading platforms encouraged millions of new investors to enter financial markets. As a result, the number of individuals using mobile trading applications increased substantially, contributing to the surge in speculative trading activity observed during the “meme stock” period.

Figure 2: Retail Trading User Growth



Source: Own work based on FINRA, Robinhood Markets annual reports, and SEC market structure reports.

The opportunity of an accessible stock market has undeniably broadened participation as seen in Figure 2, but it has also exposed inexperienced traders to significant risk. Most of these new investors lack the skills or financial literacy to manage such volatile assets, assess risks and produce results. Often, they make impulsive decisions and concentrated positions. Frequent investments are made into trending stocks with little to nonfundamental backing, what is more akin to noise trading than informed investing (Eaton et al., 2022). The design of modern trading platforms has been gamified which further encourages excessive trading activity (Ryll and Seidens, 2019). Technological innovation has reshaped the market and lowered the barriers for entry, but it also shined a bright light on the vulnerabilities of retail investors, reinforcing why short-term speculation persists despite most knowing its risks. Capturing and recognizing this persistence against academic and practical wisdom is part of successful short-term stock prediction. Integrating financial indicators with sentiment analysis and machine learning approaches can lead to models being able to capture the very dynamics that drive retail trading and institutional responses. Insights into a reshaped market environment are achievable with these methods, where long-term valuation methods are often unsuccessful.

These characteristics imply that short-term prediction models must be robust to noise, regime shifts, and asymmetric market reactions. As a result, a focus on short forecasting horizons and emphasizing directional and probabilistic evaluation might prove more fruitful rather than precise point forecasts.

2.2 The Role of Financial Indicators and Sentiment in Stock Prediction

Financial indicators serve an essential role in both short-term and long-term stock prediction, analyzing indicators is one of the first steps of evaluating whether to make an investment decision. They provide quantitative information that reflects market conditions and investor behavior. Broadly the indicators can be categorized into technical, macroeconomic and sentiment indicators.

2.2.1 Technical Indicators

Technical indicators are at the forefront of short-term investing strategies. Most of them are derived from historical price and volume data. They are used to forecast future price movements.

Common technical indicators include (Brock et al., 1992; Lo et al., 2000):

- Moving averages: These indicators help to smooth or level the price data over a specific period by creating a constantly updating average price. There are many examples, the most popular being the Simple Moving Average (SMA), which is a calculation of the average price over a set number of periods, typically using closing prices. Another is the Exponential Moving Average (EMA) similarly tells us the average price of the stock over a period, but it is weighted giving greater importance to the price in more recent days, as such it is more responsive to new information
- Relative Strength Index (RSI): Indicates momentum, it measures the speed and change of price movements, indicating overbought or oversold conditions
- Moving Average Convergence Divergence (MACD): Specifically designed to identify trends, it shows the relationship between two moving averages of a securities price
- Bollinger Bands: Indicate volatility, they help determine if a stock is under or overvalued, if the price is closer to the upper band, it might be overbought and oversold when the price is near the lower band. There is another band or line which is usually a SMA. Many strategies rely on understanding the three lines.

Every year studies show the effectiveness of these indicators regarding short-term stock price prediction. Technical indicators can also improve predictive performance in algorithmic models (Wolff and Echterling, 2024). Even much earlier research showed that these measures play a central role in shaping trader behavior and market dynamics. Brock et al. (1992) tested SMA and trading-range break rules on U.S. stock data and found that following these strategies produced returns that could not be attributed to only chance, suggesting that technical indicators carry valuable information for short-horizon investors. Another group Lo et al. (2000) recognized patterns in moving averages and momentum and hence significantly increased predictive power, indicating that these elements capture aspects of investor psychology and market microstructure which is not reflected in fundamental data.

Bollinger bands and other measures of intraday volatility also guide trading decisions by signaling overbought or oversold conditions. Trading strategies based on momentum and RSI strategies generate positive return in foreign exchange markets, which share some similar speculative dynamics with equities (Neely et al., 1997). Momentum investing is also considered viable as many past short-term winners often continue outperforming in the near term, a closely followed and exploited effect using indicators like the RSI and MACD. (Jegadeesh and Titman, 1993). Given the positive result produced overtime by different parties, it is likely that technical indicators are not only abstract calculations but with them we can build a framework for systematic trading strategies but are also capable of identifying ever persistent behavioral biases among the numerous market players. With these calculations trend-following, overreaction, and mean reversion can be quantified, helping navigate short-term uncertainty and exploiting temporary inefficiency in financial markets.

2.2.2 Macroeconomic Indicators

As mentioned, short-term changes can significantly influence the market in the long-term but it being the dynamic phenomenon that it is the reverse is also true. Macroeconomic indicators offer insights into how the economy is doing, which can influence stock prices. Key macroeconomic indicators include (Chen et al., 1986):

- Interest rates: The policies of central banks and interest rate changes can affect stock valuations, investor sentiment and their behavior towards investing.
- Inflation rates: Rising inflation usually limits purchasing power and impacts corporate earnings.
- Gross Domestic Product (GDP): One of the most famous indicators of an economies health is the GDP, the variance in its growth rate can influence investor sentiment.
- Unemployment rates: High unemployment can signal economic distress, creating reservations in investing into the market.

Recognizing how macroeconomic variables affect the market has resulted in positive stock returns (Chen et al., 1986) while recent studies confirm Machine learning techniques combined with macroeconomic data has improved accuracy in stock market predictions (Wolff and Echterling, 2024).

2.2.3 Sentiment Indicators

The overall attitude of investors can be described as market sentiment their attitude toward a given security or toward investing in general. Sentiment can be defined as many things, but it is usually derived from various sources, such as (Tetlock, 2007; Bollen et al., 2011):

- Social media: Platforms like Twitter or X and Reddit provide real-time reactions to the market providing valuable investor sentiment.
- News Articles: Most of the time specifically financial news, can influence investor perceptions and market movements
- Analyst Reports: Financial experts report and analyze different stocks which sometimes do sway investor decisions

Sentiment has been recognized as a key factor in influencing stock price movement for a long time, even outside machine learning applications. It has been demonstrated that fluctuations in sentiment explain market anomalies and temporary mispricing, particularly in assets that are difficult to value, such as small-cap and growth stocks (Baker and Wurgler, 2006). Also, the tone of daily financial news is important and correlated when compared to the next day market returns, providing proof that sentiment is not just noise but contains predictive information (Loughran and McDonald, 2011). Social media has exponentially magnified this dynamic, aggregated sentiment from different platforms can anticipate short-term market movements (Rao and Srivastava, 2012). There are behavioral and informational signals to be found in sentiment that have tangible market consequences, what is important to recognize that some of these signals are independent of technical or macroeconomic indicators.

In recent years natural language processing (NLP) has made great advancements which in turn has improved how sentiment is quantified leading to even further discoveries of short-term prediction. Technological advancements have broadened the range of sentiment analysis as it is converted to real-time metrics used as an essential supplement to conventional indicators.

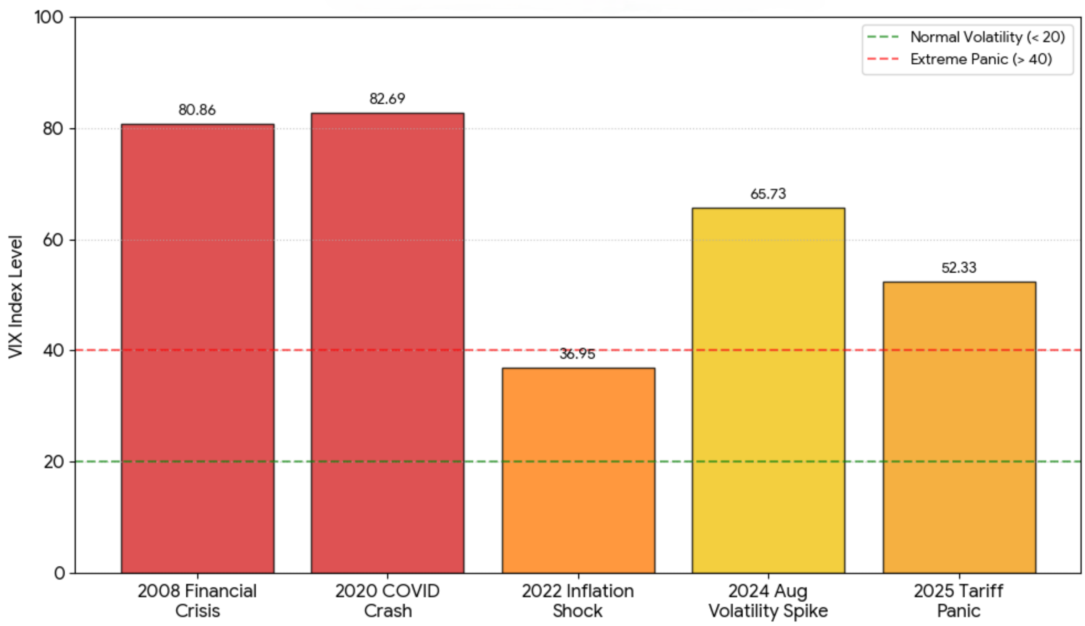
Financial indicators have proven time and time again to be essential for understanding and navigating financial markets, even more so within the context of short-term prediction. While there are many the main ones are technical, which capture price and volume dynamics, with which it is possible to recognize trader behavior, momentum and volatility, while also revealing patterns of overreaction and mean reversion. Macroeconomic indicators focus more on the broader economic conditions linking market outcomes to interest rates, inflation and unemployment. And finally, sentiment indicators further compliment this tapestry of almost unquantifiable information by quantifying the psychological and social dimensions of investing, exposing how the mood of investors, media narratives and digital communities can exert meaningful influence on market movements at least in the short-term. (Brock et al., 1992; Chen et al., 1986; Tetlock, 2007)

2.2.4 Market-Wide Indicators and Systemic Risk Factors

For the most part the core of prediction models is formed by individual stock analysis, while universal models must account for more than just one stock's dynamics but focus on the market as a whole. Two indicators in particular stand out as crucial for understanding systemic market behavior: The VIX index and the S&P 500 ETF.

The VIX or the "fear index" as it is often referred to measures the market's expectation of 30-day volatility implied by S&P 500 index options. The index is unique because it provides a forward-looking estimate of market uncertainty as opposed to looking backwards as some historical measures of volatility do. When the market is in a period of stress, the VIX typically spikes as investors rush to purchase protective options, making it a valuable measurement of market sentiment. VIX levels and research into them has consistently correlated inversely with market returns, which means that high VIX readings often coincide with market bottoms while extremely low readings may signal complacent and potential reversals. This phenomenon is clearly visible in Figure 3 (Whaley, 2000).

Figure 3: VIX spikes during major market stress events



Source: Own work based on data from Yahoo Finance (2025)

One of the best ways to estimate the U.S. equity market is the liquid SPY, the SPDR S&P 500 ETF. On average the trading volume exceeded 75 million shares and assets under management surpassing 400 billion dollars, SPY opens the way to observe the broad market dynamics. Using the SPY one can distinguish between stock-specific movements and market wide trends, which is a critical distinction for risk management.

2.2.5 Return Calculations and Volatility Adjustment

The most common way to measure a stock's performance is comparing "todays" close to "yesterdays" close, we get a picture of the movement of the stock but measuring return in the stock market is much more than just simple price changes. People can use multiple return definitions to capture different aspects of price dynamics. One way is combining traditional close-to-close returns with intraday volatility measures, which provide a more comprehensive view of daily price actions. The intraday measurement is valuable because it recognizes that two stocks with identical closing returns may have vastly different risk profiles based on their intraday behavior (Garman and Klass, 1980).

A key innovation in return calculation is volatility adjustment. We are able to compare movements across different market regimes by scaling returns relative to recent volatility, which are often measured with rolling windows (Engle, 1982). Choosing a window is often a trade-off, the choice is between sensitivity and stability. Shorter horizons are more sensitive to rapid regime shifts but can overreact, while longer horizons provide smoother estimates but lag in fast changing conditions. A mid-range window around 20 days is often a practical compromise. It aligns with monthly trading cycles and often produces more reliable forecasts than either extreme (Sueppel, 2022; Liu et al., 2024).

The coexistence of technical indicators and sentiment-driven effects motivates a hybrid modelling approach. Rather than assuming that one information source dominates, this thesis explicitly compares models trained on price-based features alone with models augmented by sentiment signals, allowing their strengths and weaknesses to emerge empirically.

3 MACHINE LEARNING IN STOCK PREDICTION

The solution to the equation of the stock market some believe might be machines, specifically machine learning which has become one of the most widely applied approach for stock market prediction as it offers tools that can capture complex, non-linear relationships between financial indicators, investor sentiment and price movements (Gu et al., 2020). Where they have the advantage is that traditional econometric models rely heavily on linear assumptions whereas machine learning algorithms can identify subtle patterns in large-scale data that would otherwise stay hidden.

The biggest advantage is the flexibility provided by machine learning which is increasingly more valuable in both academic research and professional finance, where everyone is looking for a critical competitive edge (Chen et al., 2003; Krauss et al., 2017).

3.1 Machine Learning

ML is a subfield of artificial intelligence with a focus on creating algorithms capable of learning patterns and making predictions or decisions based on data, while not being explicitly programmed for every scenario (Mitchell, 1997). ML fundamentally relies on statistical and computational methods to identify complex relationships that are often too intricate for traditional modelling. We can see ML everywhere today in healthcare, image recognition, natural language processing and increasingly in finance.

The more amount of exposure to data ML system gets the more improvement we see with refinement over time, because adaptability is one of the main strengths of these systems. There are three main types of learning regarding ML models the first one is supervised learning where algorithms learn from labelled data to predict outcomes.

The second is unsupervised learning, which goal is to identify hidden structures in unlabeled datasets. Lastly there is reinforcement learning, which optimizes its decisions based on feedback from an environment (Jordan and Mitchell, 2015). Different learning methods are used for different tasks but with them we are equipped to tackle a wide range of predictive and analytical problems.

3.2 Machine Learning Models

ML models are the foundation block through which algorithms learn from data and make predictions. Over the years many models have been developed from simple, interpretable approaches to increasingly complex architectures, each with its own strengths and weaknesses.

Decision tree-based algorithms and models that use them are one of the most widespread. With decision trees data is split according to feature thresholds, which creates a set of rules that can be followed to classify or predict outcomes (Breiman et al., 1984). As understandable as they are, single trees are prone to overfitting. Random forests were developed which aggregate multiple trees to improve stability and accuracy (Breiman, 2001). Random forests can capture non-linear relationships, making them useful for noisy financial datasets. Gradient boosting models which combine multiple weaker models have also improved predictive performance by training trees sequentially, each new tree corrects the errors of the previous ones. One well known example is Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), which has become popular in financial forecasting thanks to its ability to handle large feature sets, reduce bias and variance, and capture subtle patterns in high-dimensional data. Gradient boosting methods have also shown to outperform more complex neural networks in certain financial prediction contexts, most often when the dataset is structured and tabular (Fischer and Krauss, 2018).

A big step forward was the development of neural networks which are inspired by the human brain, consisting of nodes called neurons and arranged in layers. The most basic are called artificial neural networks ANNs, the so called neurons transform inputs into increasingly abstract representations (Rumelhart et al., 1986). Standard ANNs are powerful but not suited for sequential or time-series data such as stock prices, where the order of information is crucial.

Recurrent neural networks (RNNs) were introduced to address the issues with time series, they incorporate feedback loops allowing information from prior time steps to influence current predictions, which makes them more appropriate for modelling sequential dependencies (Elman, 1990). Long-term dependencies are still an issue for RNNs due to many issues like vanishing gradients, but this was mitigated with the development of Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). They introduced a new mechanism called gates which control how information is stored, forgotten or passed forward. Thanks to their unique architecture they are one of the most widely used networks for financial time-series forecasting, as they can capture both short-term fluctuations and longer temporal dependencies in stock price movements.

The evolution from simple decision trees to advanced neural networks like LSTMs shows the continuous effort to better understand the complex and noisy nature of financial markets, with each model building upon its ancestors' limitations, providing improved accuracy, flexibility and interpretability for stock market prediction.

3.3 Structured Numerical Data vs. Unstructured Text Features in Machine Learning

In machine learning textual data cannot just be inserted into the models as a training set, there is a clear distinction of structured numerical data and unstructured textual data, as each type requires different preprocessing strategies. Variables such as prices, returns and technical indicators are continuous and directly compatible with predictive algorithms and can be standardized, normalized and easily scaled (Kuhn and Johnson, 2013).

On the other hand, textual information like news headlines is unstructured, it lacks a predefined numerical format. For machine learning models to recognize it as viable data it must be converted into a numerical format, which usually involves Natural Language Processing (NLP). These are methods that transform textual information into numerical vector representations that can be processed by machine learning models. Early approaches relied on simple frequency-based techniques such as bag-of-words and Term Frequency–Inverse Document Frequency (TF-IDF), which represent text based on word occurrence while ignoring semantic relationships. More recent methods use semantic word embeddings, such as Word2Vec, which capture contextual and semantic similarities between words by learning dense vector representations from large text corpora. (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or contextual models like BERT (Devlin et al., 2019). These embeddings capture meaning and sentiment relationships between words, enabling successful integration of textual sentiment into quantitative models.

3.4 Overview of Machine Learning in Stock Prediction

As computational progress has evolved, being a part of every aspect of our lives, so did the applications of machine learning (ML) in finance, just rather modestly at first. The earliest application of neural networks for financial forecasting was White (1988), who demonstrated that stock returns could be predicted more effectively with ML methods rather than relying on random walk models. Support vector machines, decision trees, and generic algorithms were experimented with through the 1990s achieving moderate success in capturing non-linear dynamics of stock returns (Kim, 2003).

Over time ensemble learning approaches became more common combining random forests and boosting produced better results. A combined SVM, random forest and artificial neural networks with technical indicators reported improved predictive performance compared to classical statistical models (Patel et al., 2015). The next big step was in the 2010s with the rise of deep learning, which enabled models to learn hierarchical future representations from large-scale data. Again, the next step specifically LSTM networks showed that when applied to stock prediction significantly outperformed traditional benchmarks (Fischer and Krauss, 2018).

While ML models applied to stock prediction often report high levels of accuracy it does not necessarily mean high levels of success, even the results are massively varied across different contexts and must be interpreted cautiously. There is the question of what exactly the model predicts and the success of predictions is often dependent on it, most common is directional movement and price level, one must also consider is it predicting for an individual stock, an index or multiple stocks, the model's generalization ability is also important (Gu et al., 2020; Fischer and Krauss, 2018).

The highest accuracy achieved is usually with directional prediction, making it as simple as possible the stock will either go down or up, but of course how exactly to predict the movement is predicted is also debated. Still this approach yields better results than predicting price levels. A study across major indices like NYSE 100, DAX 30 and FTSE 100 compared ML algorithms and found that ANNs achieved over 70% accuracy in predicting directional movements, often outperforming Logistic Regression and SVM in certain markets (Ayyildiz and Iskenderoglu, 2024).

Another way is predicting regression of price levels which tend to have mixed results. Combining Gated Recurrent Unit (GRU) a variant of RNN and XGBoost performed the best in predicting short-term price values outperforming models like LSTM and standard RNNS based on metrics like MAE, RMSE and R^2 (Chang et al., 2024). Many models are also trained on specific stocks and again the performance varies significantly. A study on some of the biggest companies like Apple and Amazon found directional accuracy averaging around 75 percent reaching as high as 89 percent (Khanna et al., 2022). While this may sound amazing stock specific models are prone to overfitting, a common issue in the financial forecasting world. In these scenarios models capture noise instead of meaningful signal, which results in poor results out of the models' sample. High frequency models showed high accuracy with technical indicators but only in-sample performance and failed to maintain accuracy when tested with out-of-sample data, particularly in volatile environments (Deep et al., 2024). These results also align with general machine learning theory, which acknowledges overfitting as a significant obstacle to financial prediction and emphasizes the importance of methods to reduce these problems with feature selection, regularization, and cross-validation. (Ryll and Seidens, 2019; Deep et al., 2024).

3.5 LSTM Architecture for Financial Time Series

LSTM networks made fundamental breakthroughs in sequential data processing, particularly concerning financial time series where both short-term patterns and long-term dependencies coexist. (Hochreiter and Schmidhuber, 1997).

An LSTM is made up of three gates: forget, input and output. They collectively manage information flow through the network. The forget gate determines which information from previous time steps should be discarded, which is crucial for adapting to regime changes in financial markets. The input gate selects information to store in the cell state, which allows the network to identify and retain relevant market signals. Finally, the output gate manages what information passes to the next layer, enabling selective feature extraction based on current context (Greff et al., 2017).

The long-term memory is simulated in the cell state, carrying information across extended sequences without the degradation that limits standard RNNs (Bengio et al., 1994). It enables the models to remember important events that may influence prices days or weeks later. Meanwhile the hidden state serves as the network's working memory, maintaining a compressed representation of recent market dynamics. LSTM architectures most often use multiple layers to capture hierarchical patterns in data. In stock prediction the first layer might identify basic price patterns and technical formations while deeper layers combine them into complex trading signals (Fischer and Krauss, 2018).

Because a lot of information is not necessarily reflected in only price data Natural language Processing (NLP) methods are commonly integrated into financial time-series architectures in order to capture information from news and other textual sources. In recent times financial NLP pipelines rely on contextual language models based on Transformer architectures, which generate representations that adapt to semantic context of each word. Bidirectional Encoder Representations from Transformers (BERT) introduced this paradigm by enabling bidirectional context modelling, allowing linguistic meaning to be derived from both preceding and following words within a sentence (Devlin et al., 2019). These models have proven particularly effective for sentiment analysis tasks, where contextual interpretation is critical for correctly identifying positive, negative, or neutral market signals.

FinBERT is an extension of BERT architecture which is specifically designed for financial applications by pretraining on large-scale financial texts, improving the model's ability to interpret financial terminology, event-driven language, and sentiment polarity (Araci, 2019).

4 METHODOLOGY

In order to achieve a hybrid stock prediction model, we had to integrate both numerical price data and textual sentiment information into one model. Careful planning and consideration of options was explored to devise the appropriate workflow taking into account different research and practices for each step taken. With a combination of statistical analysis, machine learning and natural language processing we extracted meaningful features, engineered targets and evaluated model performances. The main architecture was built with the programming language Python utilizing different libraries and resources necessary for our goal. Key libraries used were:

- Torch (PyTorch): Main library for building deep learning architectures, such as LSTM for advanced time series, providing dynamic computation graphs and efficient gradient flow through complex temporal sequences.
- Finbert: A pre-trained NLP model that is used to analyze sentiment of specifically financial text. Built upon the widely used BERT language model except for the financial domain, fine tuned to its vocabulary.
- Yfinance: A tool used to fetch financial data such as historical OHLCV price data for all relevant tickers.
- Numpy: Used for fast numerical operations for feature calculations, particularly essential for rolling calculations and statistical transformations that form the backbone of technical indicators.

Overall, the approach of this thesis follows the principles of the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. First is the problem formulation and objective definition, followed by data collection and understanding of both financial time-series data and textual news sentiment. Extensive data preparation and feature engineering are applied, after which multiple LSTM-based models are developed and trained across different temporal horizons. Model performance is then evaluated using a comprehensive set of statistical and trading-oriented metrics. While CRISP-DM was not applied rigidly, its iterative and data-centric structure provides a suitable foundation for the hybrid machine learning methodology adopted in this thesis.

4.1 Data preprocessing

In building a hybrid model that incorporates sentiment features there were two main data sources we needed. One that included all relevant stock price information and another non-numerical that included headlines for as many stocks as we could find. We experimented with scraping news headlines from various outlet websites but ultimately abandoned this approach. We rely on two main datasets:

- Price Data: Daily OHLCV (Open, High, Low, Close, Volume) data sourced from the public API yfinance which was made to match the timeframe of the news data source.

With price data a basis for computing technical indicators, volatility-adjusted returns and volume-related metrics was formed and executed. The choice of daily frequency balances information richness with computational efficiency.

- News data: The main source was the public dataset “Daily Financial News for 6000+ Stocks” uploaded to Kaggle by Miguel Aenlle. It includes over 4 million articles or rather headlines for stocks from 2009 to 2020. Each headline is connected to a specific ticker or stock with the date and time also marked. The data has three main columns a title which includes the news headline, a date which is the date and timeframe the news was posted and of course the stock column which is the ticker like AAPL an abbreviation of a company name used for simpler recognition. (Aenlle, 2020)

The stocks or tickers that we would need daily OHLCV data for were determined based on the companies mentioned in the main news dataset, so each ticker also had news data as well as price data. For the news data source, a more robust path was taken. Since ML cannot work with text data all the headlines had to be transformed into a numerical dataset. This was achieved by passing each headline through a sentiment analysis model specifically FinBERT. The final dataset merges price and sentiment data into a comprehensive ticker-date frame, with news headlines mapped to trading days and gaps filled with trading days (missing sentiment features are handled appropriately).

A third dataset was also created to robustly test the given models with out of sample data, scraping news headlines from sites like nasdaq.com and getting the financial information again from yfinance.

4.1.1 Filtering and Representativeness

Given the large dataset and the challenges in building a universal predictive model, not all stocks were used. Because there was no real structure as to which stocks were picked such securities could introduce excessive noise due to wide bid-ask spreads and sporadic trading, while short-lived listings would provide insufficient historical coverage for training temporal models.

Each ticker had to pass two main conditions, first was a coverage criterion. Tickers with fewer than 250 valid trading days are excluded, meaning that they also have at least that many news headlines so at least one per day. This ensures adequate history for computing rolling features and because there are no short sequences it prevents forming biases in models. A liquidity filter was also applied meaning that if the average daily volume did not exceed 100,000 the ticker was filtered out. This ensured alignment with realistic trading assumptions, making this strategy economically feasible.

After both filters were applied, further refinement was needed to ensure consistency and reliability. Rows with missing price or news data were excluded, if duplicates were found they were also removed. Dates had to be checked for consistency, ensuring that only trading days with both price and news data were retained. Extreme price movements, anomalous volumes and other outliers were identified and either capped or removed to prevent distortion of the model's learning process.

Another challenge was the inconsistency on which days news was available, even if an asset had more than 300 rows of news headlines it does not mean the news days were continual meaning a lot of days were skipped. Continual time data is nonnegotiable with the stock prediction models we are trying to achieve so we needed to fill in the gaps between news days with trading days, but that meant we would have less days that included sentiment features. We aimed for the most concentrated unbroken sequences of news days with fewest gaps as possible, ultimately building the final data frame of more than one hundred thousand rows of data.

The final structure of the dataset that was used to train the model resulted in each row representing a valid ticker-date pair, with price and sentiment information (after natural language processing conversion).

4.1.2 Target and Return Definitions

One of the most essential and relevant parts of any prediction model is the target variable and return calculations meaning what exactly are we predicting and what are we using to predict them. It is the essential piece; the return calculations determine what the model learns and then predicts. It directly impacts performance, interpretability and practical utility of architectures.

$$Enhanced\_Return_t = 0.7 \cdot \frac{Close_t - Close_{t-1}}{Close_{t-1}} + 0.3 \cdot \frac{High_t - Low_t}{Close_{t-1}} \quad (1)$$

The first return calculation we made was that of an enhanced return which is something that combines the standard daily price return and intraday volatility. The way it is calculated is seen in formula (1). The daily price return is just today's closing price ($Close_t$) minus yesterday's closing price ($Close_{t-1}$) divided by yesterday's closing price. We use this because it is the most common way to measure return in finance and essential for understanding the general direction and magnitude of price changes. But because it only looks at the closing prices, we have little idea what happened during trading hours. So, to account for that we use a weighted formula which also acknowledges the intraday volatility capturing the highs and the lows of that day. We do this by dividing today's high ($High_t$) minustoday's low (Low_t) by yesterday's close). By combining both we create a more informative formula that reflects not only the direction but the intensity of market movements.

The next return calculation is formula (2), which we call Rolling volatility it is the standard daily deviation or the estimated short-term volatility calculated based on the last twenty trading days, so a month. A higher rolling volatility means bigger price moves, giving a sense of how "noisy" the asset has been lately, while lower rolling volatility means steadier prices.

$$RollingVol20_t = \sigma(Enhanced\_Return_{t-20:t-1}) \quad (2)$$

How it works we calculate it for day *t* we take the daily returns from day t-20 to day t-1 (we do not include "today"), calculate their average, then measure how much each return deviates from that average. After squaring each difference and adding them up we divide by the number of days minus one and take the square root of the result to get rolling volatility. The 20-day window balances responsiveness to changing conditions with statistical stability.

$$Normalized\_Return_t = \frac{Enhanced\_Return_t}{RollingVol20_t} \quad (3)$$

Formula (3) represents normalized return which is just the enhanced return of today divided by the rolling volatility of the past twenty days. This makes it easier to compare across different stocks and time periods, because many have different volatilities and price jumps. This formula became the foundation of all comparison, target selection and prediction of our model. By scaling returns relative to recent volatility, we ensure that our predictive signals remain statistically consistent even as market regimes shift from calm to volatile.

Using the calculations we chose different targets:

- Multiclass target: Trading actions, a determination was made based on weekly returns set to a threshold categorized into SELL, HOLD and BUY.
- Binary target: Directional movement categorized into UP and DOWN based on the normalized return calculation.
- Big move target: Used to identify significant price movements usually moves over 2%, enabling focused evaluation of extreme event prediction capability. This threshold serves as a benchmark for assessing the model's accuracy during periods of high market volatility

We use this set of targets to try to evaluate the efficiency and performance of our model across different trading scenarios. The binary target is the most basic one, simplifying trading recommendations of only up or down, while it may be simple it is the most essential one, knowing whether the market will rise or fall is the centerpiece in any trading strategy. The multiclass captures the nuanced intensity of market movements. Finally, we needed a way to account for significant price moves, somehow to take into consideration the models' ability of perceiving and responding to major events.

4.2 Feature Engineering

Another important step is the collection of indicators and features one chooses for their project because they directly affect the results. As mentioned, there are many ways to categorize features, we used them from a variety of contexts. All features are designed to be stationary, normalized, and comparable across different assets and time periods, ensuring that learned patterns generalize effectively.

Table 1: Technical and Volatility Feature Set

Feature	Description	Rationale and Predictive Value
Log_Return	Daily logarithmic return of the stock	Captures percentage change in price while ensuring stationarity, essential for time-series modelling. Standard measure in financial econometrics (Fama, 1970; Bodie et al., 2018)
Enhanced_Return	Volatility-adjusted return combining close-to-close returns (70%) with intraday range (30%)	Accounts for both directional movement and trading intensity, providing more comprehensive daily performance measures than simple returns. Improves magnitude estimation by incorporating intraday volatility (Garman and Klass, 1980)
MA_Diff	Difference between short-term and long-term moving averages	Indicates trend direction and momentum strength. Proven effective in systematic trading strategies (Brock et al., 1992; Lo et al., 2000)
MACD	Moving Average Convergence Divergence	Trend-following momentum indicator that captures the relationship between moving averages. Widely validated for identifying trend reversals and continuation patterns (Brock et al., 1992; Wolff and Echterling, 2024)
ATR14	14-day Average True Range	Measures market volatility, providing context for interpreting price movements. Essential for volatility regime detection (Lo et al., 2000)
RollingVol20	20-day rolling standard deviation of returns	Estimates short-term volatility and enables volatility regime detection. The 20-day window balances responsiveness with stability, aligning with monthly trading cycles (Engle, 1982; Sueppel, 2022; Liu et al., 2024)

To be continued

Table 1: Technical and Volatility Feature Set (cont.)

Feature	Description	Rationale and Predictive Value
Volume_Ratio	Ratio of current volume to average volume	Highlights unusual trading activity and serves as proxy for liquidity and market attention. Abnormal volume often precedes significant price movements (Chen et al., 1986)

Source: Own work

Each feature seen in table 1 was carefully chosen in accordance with research and available information to maximize the model's ability to learn meaningful patterns in financial data while minimizing noise and bias given our timeframe setup. Machine learning models perform best when data distribution is stable meaning that the statistical properties remain consistent over time, which is why we used normalized, and volatility adjusted features such as rolling volatility and z-scores ensuring stationary inputs.

Raw price features such as Open, Low, High or Close are non-stationary, their scale and distribution can differ greatly over time and across different assets, making it more difficult for models to learn consistent patterns leading to poor generalization. As mentioned, we use technical indicators derived from OHLC data, summarizing relevant information making it more stable and comparable.

Features like RSI, MACD and Volume_Ratio have also been proven to capture momentum, trend direction and liquidity shifts, helping detect overbought or oversold conditions, trend reversals and unusual trading activity, all of which can help in future price prediction.

Table 2: Market-Wide Indicator Features

Feature	Description	Rationale and Predictive Value
VIX_Close	Closing value of the VIX index	Represents implied 30-day volatility and investor fear. Provides forward-looking estimates of market uncertainty rather than historical measures. High VIX readings often coincide with market bottoms while low readings may signal complacency (Whaley, 2000)
VIX_Log_Return	Logarithmic return of VIX	Captures daily changes in market volatility expectations. Sharp VIX increases signal stress, while decreases indicate calming conditions. Essential for detecting regime shifts (Whaley, 2000)

To be continued

Table 2: Market-Wide Indicator Features (cont.)

Feature	Description	Rationale and Predictive Value
Delta_VIX	Daily absolute change in VIX	Highlights shifts in market sentiment and volatility regime changes. Large deltas often precede or accompany significant market moves (Whaley, 2000)
VIX_z	Z-score of VIX relative to historical distribution	Standardizes volatility relative to historical mean and variance, enabling comparison across different market regimes. Identifies unusually high or low fear periods (Engle, 1982)
SPY_Log_Return	Logarithmic return of SPY (S&P 500 ETF)	Captures broad market performance, distinguishing stock-specific movements from market-wide trends. Critical for understanding systematic versus idiosyncratic risk
SPY_RollingVol20	20-day rolling volatility of SPY	Measures recent risk in the broader market. Helps contextualize individual stock volatility and identify market-wide regime changes. The 20-day window balances responsiveness with stability (Macrosynergy, 2022; Liu et al., 2024)
SPY_z	Z-score of SPY returns	Standardizes broad market movements for comparability across different volatility regimes. Identifies extreme market conditions relative to recent history (Engle, 1982)

Source: Own work

VIX and SPY features seen in table 2 provide direct insight into market-wide sentiment and risk, making them a perfect fit for our universal prediction model. One reflects the markets fear and volatility while the other represents the behavior of the most valuable assets two key market benchmarks. With their inclusion the model can adjust its prediction based on broader market conditions which go often hand in hand with individual asset performances.

The next important features were the news headlines from which we would extract sentiment. The transformation of raw news headlines into numerical sentiment features involved a multi-step process using FinBERT, a BERT-based model pre-trained specifically on financial text.

Sentiment and News Features:

- Sentiment_Positive: Probability or score from FinBERT indicating the positive sentiment in a financial headline.
- Sentiment_Negative: Score from FinBERT expressing negative sentiment.

- Sentiment_Neutral: Score from FinBERT expressing neutral sentiment.
- Sentiment_Compound: Aggregated score combining positive, negative and neutral probabilities into a single value, summarizing the overall sentiment of headlines.
- News_Count: The total number of news headlines available for a stock given one day. It serves as a proxy for market attention of something happening, of more investor interest and activity.

Sentiment features exist to capture the psychological and information impact of headlines. News sentiment can drive rapid changes in price, especially during major events. With sentiment the hope is to capture the shifts in investor mood and information flow improving the model's ability to anticipate short-term market moves. Each headline from the Kaggle dataset was cleaned to remove special characters and standardized for input to FinBERT. Each headline was passed through the FinBERT model, which outputs three probability scores: Positive probability (0-1), Negative probability (0-1) and Neutral probability (0-1). For instance, the headline "**The 8 Most Undervalued Stocks Fund Managers Are Buying**" (dated 2017-11-28, ticker: NFLX) produced: Sentiment_Positive: 0.987, Sentiment_Negative: 0.012, Sentiment_Neutral: 0.001. The Sentiment_Compound feature was derived by calculating Compound = Positive – Negative. This provides a single metric ranging from -1 (most negative) to +1 (most positive). When multiple headlines existed for a single stock-date pair, we calculated the mean of each sentiment score across all headlines for that day. This approach would smooth out contradictory signals, weigh each piece of news equally and preserve the overall sentiment trend. Simultaneously, we recorded the total number of headlines per stock-date, creating the News_Count feature as a proxy for market attention and information flow. For trading days without news coverage, sentiment features were set to neutral values (0.5 for positive/negative/neutral, 0.0 for compound), and News_Count was set to 0, explicitly indicating low market attention rather than missing data.

Each feature was curated to balance financial relevance, statistical influence and practical utility allowing for robust experimentation of finding the best preforming models.

4.3 Model Architecture

Our deep learning framework is based around the LSTM architecture, specifically tailored for financial time series prediction, incorporating different strategies to ensure robust learning from real-world data.

4.3.1 Temporal Infrastructure and Data Alignment

Firstly, a custom calendar was constructed for extra validation to take into account weekends, holidays and non-trading days. Each prediction had to be aligned with actual trading days not just calendar days, preventing distortions from non-trading periods. It is used throughout all the model phases ensuring consistency. Our dataset is chronologically split into three distinct subsets seen in table 3.

Table 3: Dataset split

Dataset	**Timeframe**
Training	2009-01-01 to 2016-12-31 (8 years)
Validation	2017-01-01 to 2018-12-31 (2 years)
Testing	2019-01-01 to 2020-06-10 (1.5 years)

Source: Own work

This chronological split ensures that models are evaluated on genuinely future data, preventing any form of temporal leakage that could artificially inflate performance metrics.

4.3.2 Multi-Horizon LSTM and Validation Framework

The key innovation approach we used is the simultaneous generation of multiple sequence lengths. The lengths were 5, 10 and 20 trading days. This allows models to capture different kinds of market dynamics: short-term momentum effects (5 days), medium-term trend developments (10 days), and long-term structural changes (20 days). Each sequence is validated through strict calendar constraints to ensure temporal integrity:

- Maximum gap factor (2.0): A gap factor is set, so that in each sequence the total calendar span (from the first to the last trading day in the sequence) cannot exceed N (number of trading days) times 2 calendar days. For example, in a 10-day sequence the first and last trading day must be within 20 calendar days of each other. This allows for weekends and short holidays but prevents the sequence from stretching across long periods with missing trading days.
- Maximum span requirement (60 days): No sequence can span more than 60 days in total, regardless of the number of trading days.
- Hard gap limits (5 days): The gap between any two consecutive trading days in a sequence must not be greater than 5 calendar days.

Our preprocessing ensures that sequences consist of strictly consecutive trading days, so no sequences are created with missing trading days. While the constraints might seem redundant preserving the chronological order and realistic market conditions were deemed a must for financial time series modelling.

Features are standardized individually for each ticker using only the training data in order to prevent data leakage while accommodating the heterogeneous nature of different securities. Missing sentiment data is treated as informative (indicating low market attention) with explicit coverage indicators rather than simple imputation.

4.3.3 LSTM Architecture Design and Model Variants

The LSTM encoder uses multi-layer architecture with hidden dimensions of 128 units, incorporating batch normalization and dropout for regularization. This design allows the model to learn both short-term momentum and longer-term trend patterns simultaneously.

For comprehensive research we trained models with different setups, features and sequences. There are two main categories technical and sentiment. The difference is that the technical uses none of the news or sentiment features. This separation gives us a perspective on the effects news might have on predictions. The final design of the models is presented in table 4.

Table 4: Model configurations

Features	Model	Timeframe	Design focus
Technical	5d_core	5 days	Designed to capture very short-term momentum and rapid market shifts by focusing on recent price behavior.
Technical	10d_core	10 days	Balances sensitivity to short-term fluctuations with the ability to recognize emerging medium-term patterns and reversals.
Technical	20d_core	20 days	Optimized for identifying longer-term trends by smoothing out short-term noise and emphasizing persistent market behavior.
Sentiment and Technical	5d_sent	5 days	Extends the short-term technical model by incorporating sentiment features to capture immediate psychological reactions to news events.
Sentiment and Technical	10d_sent	10 days	Combines medium-term technical structure with sentiment information to detect sustained market movements influenced by news.
Sentiment and Technical	20d_sent	20 days	Integrates long-term technical trends with aggregated sentiment to reflect both market structure and investor psychology.

Source: Own work

4.3.4 Multi-Task Learning Architecture

Our models use a multihead approach: one for regression and one for classification. The regression head predicts the upcoming timeframe's continuous normalized return, a straightforward prediction of the anticipated price change. The classification head predicts a specific trading decision (SELL, HOLD or BUY), by using thresholds set on the normalized return. The two heads share the representation from the LSTM and an intermediate layer but then branch into separate specialized networks with their own layers and regularization, allowing the model to learn general market patterns while optimizing for the distinct requirements for regression and classification improving overall performance and robustness.

During training the models use focal loss for classification tasks to address class imbalance, this ensures a bigger focus on moves that are harder to classify like the SELL signals. For big move detection (returns > 2%), the loss function is further enhanced with curriculum learning and weighted terms (big_move_alpha = 5.0), so the model gradually focuses more on extreme events as training progresses. This approach allows the model to achieve higher accuracy for common cases but also effective at identifying and predicting rare, high-impact market moves.

4.4 Training Protocol and Evaluation Strategy

The training framework is rigorously validated and ensures robust out of sample performance to prevent overfitting to specific market periods. Hyperparameters and model selection are tuned using only the training and validation sets, with the test set reserved for final unbiased evaluation. The training employs Adam optimization and gradient clipping to prevent gradient explosion in long sequences. Multiple regularization techniques prevent overfitting: dropout reduces overfitting to specific temporal patterns, weight decay provides L2 regularization, early stopping prevents overtraining, and learning rate scheduling with warmup ensures stable convergence. SELL class weighting addresses the natural imbalance in trading signals without creating excessive false positives that would be impractical in real trading.

The evaluation framework focuses on both statistical accuracy and practical trading relevance it is designed with the hope of providing a comprehensive and realistic assessment of model performance. For regression outputs, error magnitude is measured using RMSE and MAE, while correlation metrics (Pearson and Spearman) are used to quantify the relationship between predicted and actual returns. Spearman correlation is especially given special attention, as it better captures non-linear dependencies common in financial markets. Classification performance evaluation is done with precision, recall, F1-score and balance accuracy assesses the model's ability to generate actionable trading signals.

The most basic trading measurement we do is UP and DOWN it is assessed with ROC-AUC, directional accuracy, and Matthews Correlation Coefficient, which provides a comprehensive assessment of the fundamental task. The classification of SELL HOLD and BUY predictions are evaluated through balanced accuracy, per-class precision/recall, and confusion matrix analysis, enabling the assessment of practical trading implementation. For big move detection we use a specialized evaluation of extreme event prediction with precision and recall, directional accuracy during crisis periods and magnitude prediction accuracy for large movements addresses the critical trading requirements of identifying periods of heightened risk and opportunity.

As mentioned, we built another dataset which had nothing to do with the original Kaggle news dataset. The new dataset included additional news headlines from external sources and time periods that do not present in the original dataset. To rigorously validate the generalization capability of our models, we evaluated them using this dataset, which was entirely separate from the main training and testing sets. It ensures that the models are exposed to unseen market conditions and information flows.

The test was designed to mimic real world conditions, where models must adapt to new assets, news sources and an evolving environment. With the test the robustness, transferability and practical utility were assessed. Similar performance metrics were calculated as for the main test set, allowing for direct comparisons.

Our process integrates rigorous data preprocessing, multi-horizon LSTM architectures, and comprehensive evaluation frameworks to predict stock movements. The workflow combines technical indicators with sentiment analysis, using normalized and volatility adjusted features for comparability. To capture different temporal dependencies and market behaviors we experiment with six distinct LSTM models with varying sequence lengths (5d, 10d, 20d) and separate feature sets (technical only vs. added sentiment). Multi-task learning enables the prediction of continuous returns and specific trading actions through shared LSTM encoders with specialized prediction heads. To ensure acceptable performance across market regimes advanced training techniques are used, ensemble strategies including volatility aware weighting and focal loss for class imbalance. Temporal integrity is maintained through strict chronological validation and rigorous sequence validation with custom trading calendar alignment, providing a comprehensive framework for financial time series prediction that balances statistical rigor with practical applicability.

5 RESULTS

We conducted a comprehensive evaluation of six LSTM-based models, some incorporated sentiment analysis for stock prediction and others only used technical features. We carried out our analysis across multiple dimensions of model performance, including correlation, directional accuracy and others. Besides the feature composition of the models the main difference is the sequence length of which we analyze rigorously how it affects the model's performance across different metrics. The primary research objectives were:

- Assess the predictive power of LSTM architectures with varying sequence lengths.
- Quantify the incremental value of sentiment features derived from financial news headlines.

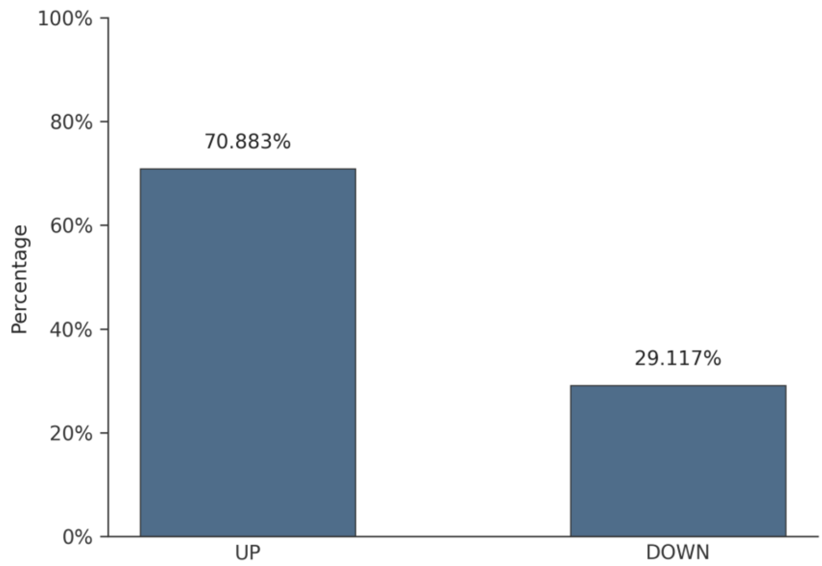
- Evaluate model performance across multiple metrics.
- Develop insights for portfolio risk management by analyzing the models' ability to detect market downturns and big moves.

One fundamental challenge we need to address is posed by the class imbalance in financial prediction, because most stocks rise over time, resulting in a predominance of positive outcomes in most datasets. At its core financial prediction is defined by directional forecasting, determining whether prices will move up or down. A binary decision is the foundation of all trading decisions and represents the minimum viable capability for any practical trading system. The upward bias adds complexity and imbalance to classification tasks, making it challenging to identify actionable signals and assess model performance.

5.1 Dataset Characteristics and Methodological Foundation

The test dataset covers 476 calendar days from January 30, 2019, to June 10, 2020. These predictions span 212 distinct securities that passed liquidity and coverage filters: each ticker maintains at least 250 trading days of history and an average daily trading volume exceeding 100,000 shares. The sequence model variants produce different numbers of predictions due to varying sequence length requirements.

Figure 4: Average distribution of classes UP and DOWN across models



Source: Own work

What is immediately clear in Figure 4 is the distribution of actual market outcomes during this period. Almost 71% of all stock-day observations exhibited positive returns (UP) while only 29% showed negative returns (DOWN).

Table 5: Distribution of actual and predicted classes UP and DOWN

Model	Actual UP %	Actual DOWN %	Predicted UP %	Predicted DOWN %
5d core	71.3	28.7	72.4	27.6
5d sent	71.3	28.7	68.3	31.7
10d core	71.1	28.9	72.1	27.9
10d sent	71.1	28.9	69.7	30.3
20d core	70.8	29.2	73.5	26.5
20d sent	70.8	29.2	71.2	28.8
Overall	71.1	28.9	71.3	28.7

Source: Own work

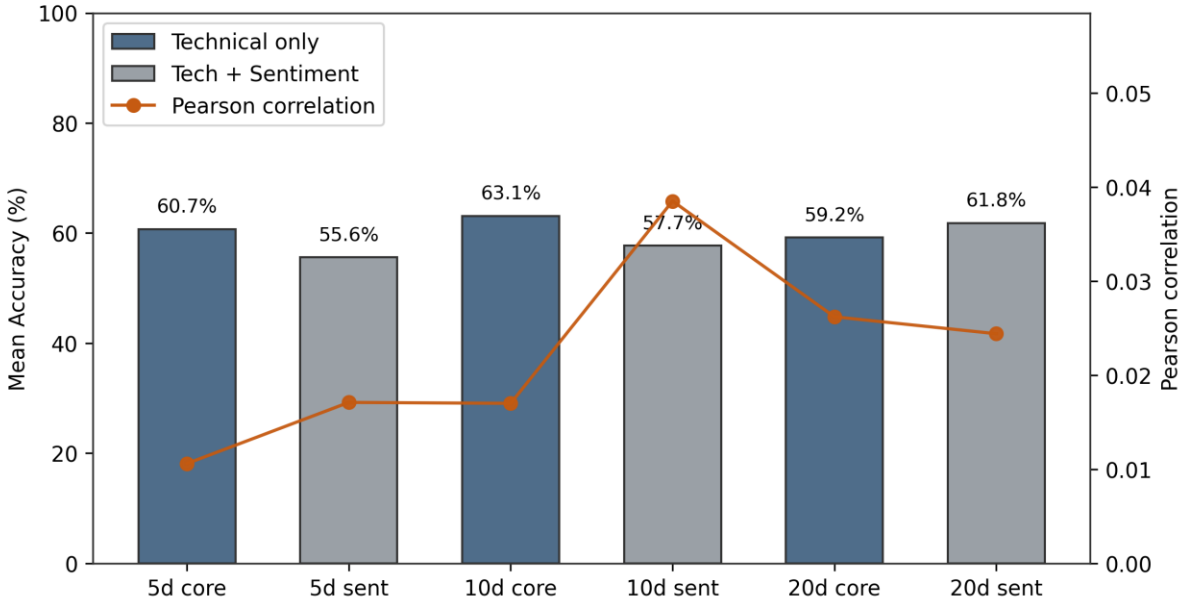
Across all configurations, approximately 71% of observations exhibited positive returns (UP) while a bit less than a third showed negative returns (DOWN), presented in Table 5. Core-only models predict UP almost three out of four times, while sentiment-enhanced are more conservative, by a very small margin.

5.2 Accuracy–Correlation Dynamics: Evaluating Predictive Quality

While predictive performance is often assessed using directional accuracy alone, this metric does not fully capture the quality or stability of model predictions. In financial time-series forecasting, a model may achieve high accuracy by correctly predicting the sign of returns yet still fail to align with the magnitude or temporal structure of actual price movements. Correlation therefore provides a complementary perspective by measuring the consistency between predicted and realized returns over time. Examining accuracy and correlation jointly allows for a more nuanced evaluation of predictive behavior across different model configurations.

In figure 5 we see multiple dimensions of the accuracy-correlation trade-off. The highest correlation is with the 10-day sequence that incorporates sentiment. Comparing it to the same sequence, but with no sentiment features, we see that the inclusion produces a 126% increase in correlation while it reduces accuracy by 8.6 percentage points. The 5-day sequence is similar to the 10-day as it shows sentiment harming accuracy from by about 5.1 pp and correlation performance does improve but only in absolute terms by 61%. The 20-day horizon presents an opposite view, where sentiment features improve accuracy from a marginal amount, an increase of 2.6 pp. Correlation is only slightly reduced by 6.9%.

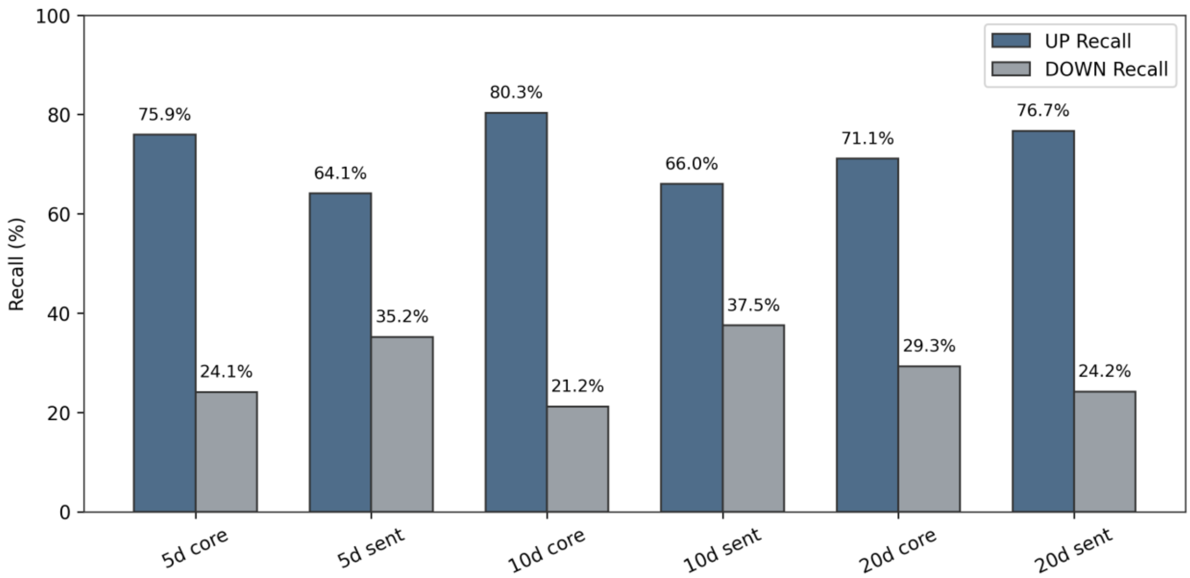
Figure 5: Directional accuracy and correlation of each model



Source: Own work

Examining core models alone, accuracy peaks at 10-day sequence (63.1%), exceeding both 5-day (60.7%) and 20-day (59.2%). Correlation also rises in turn with longer sequences. Looking at sentiment models, accuracy rises in correspondence with sequence length, with 5d being the lowest then 10d and finally the highest 20d with 61.8%.

Figure 6: UP and DOWN recall of each model



Source: Own work

Figure 6 shows the mechanisms and truth behind accuracy differences. The 10-day core gets its highest of all the models accuracy by being exceptional at catching UP moves (80.3% recall) but quite worse at catching DOWN moves (21.2% recall).

The 10-day sentiment sequence is a bit different and the most balanced with 66.0% UP recall and 37.5% DOWN recall. The 20-day sequence is again an exception where sentiment affects the prediction reversibly compared to shorter sequences. The inclusion of sentiment improves overall UP recall by around 5 pp but slightly lowers DOWN recall.

5.3 Trading Signal Classification: BUY, HOLD, and SELL Performance

Directional accuracy is often limited in isolation as it only measures if the model correctly predicts UP or DOWN movements, in practice successful trading requires more signals. The models generate a tertiary class recommendation of SELL, HOLD and BUY, which are built on predicted return magnitudes. Based on the classifications actionable trading decisions could be made, at least that is the idea of liquidating positions or shorting (SELL), maintain current holdings (HOLD), or initiate new positions (BUY).

The three classes are defined by predicted return thresholds:

- SELL (0): predicted_return ≤ -0.5%
- HOLD (1): -0.5% < predicted_return < 0.5%.
- BUY (2): predicted_return ≥ 0.5%.

These thresholds create a realistic decision framework. Small, predicted movements (±0.5%) fall into HOLD, avoiding excessive trading on marginal signals. Larger predicted movements trigger active position changes. The actual returns are labelled using identical thresholds, enabling direct comparison between predicted and actual trading signals.

In figure 7 recall metrics show complementary patterns. SELL recall ranges from 7.4% to 16.8%, meaning models detect at most one in six actual SELL opportunities. BUY recall is higher with a much wider range and models capturing between one-third to two-thirds of actual buying opportunities. HOLD recall also varies widely (30.6-60.8%), reflecting models' different strategies for managing the neutral zone.

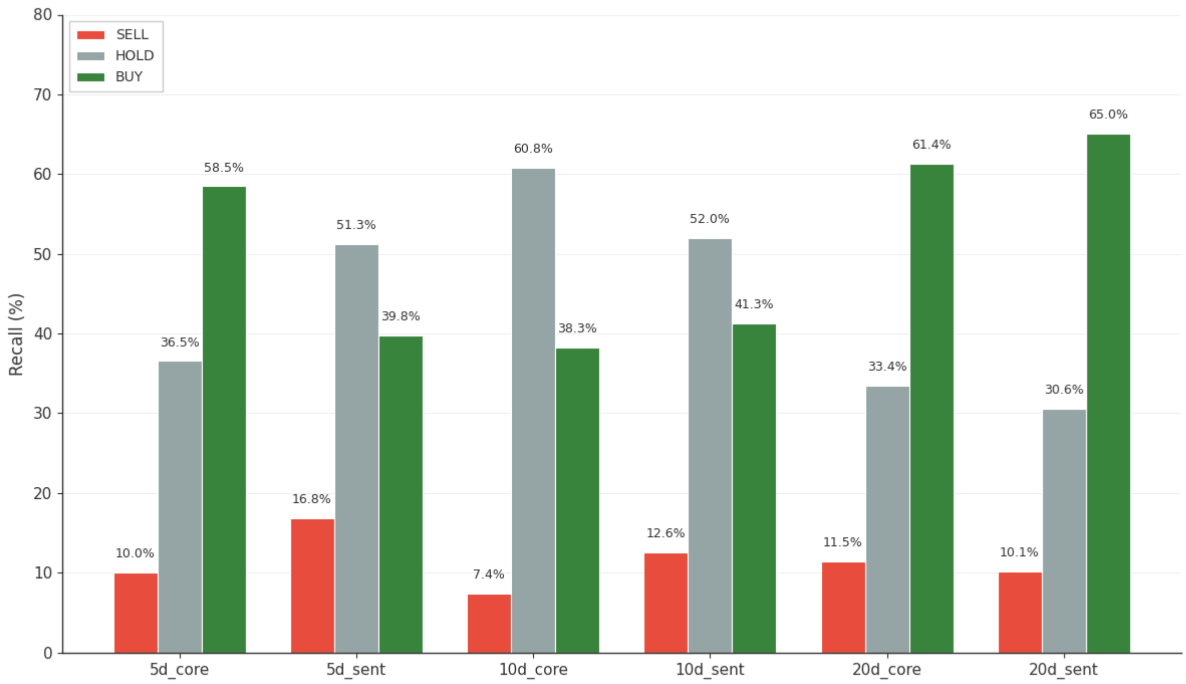
The models that lead in directional accuracy from chapter 5.2 do not have the same results here. The 10d model with only technical features, which had the highest accuracy of 63.1%, has the lowest SELL and BUY recall.

Figure 7: Precision and recall heatmap for tertiary classification of each model

Source: Own work

On the other hand, the HOLD detection is this model's strong point with 60.8% recall, meaning at least it identifies when the market is not volatile. By contrast, the 20-day sentiment model achieves the highest BUY recall (65.0%) but one of the lowest SELL recalls (10.1%), making it good for identifying entry points but poor at signaling exits.

Figure 8: Recall of HOLD, BUY, SELL for each model



Source: Own work

Figure 8 presents the recall performance across all three trading signal classes, revealing distinct model specializations and trade-offs. The 20-day models demonstrate the strongest BUY detection, with 20d_sent achieving 65.0% recall and 20d_core reaching 61.4%, substantially outperforming shorter horizons where 10-day models capture only 38-41% of BUY opportunities. Interestingly, sentiment features improve BUY recall at the 20-day horizon but reduce it at shorter timeframes, where 5d_core (58.5%) outperforms 5d_sent (39.8%).

HOLD signal detection shows the opposite pattern. The 10d_core model achieves the highest HOLD recall at 60.8%, reflecting its conservative stance of predicting HOLD in 57.0% of cases compared to the actual 39.5% rate. The 20-day models take an aggressive approach, predicting HOLD in only about a third of cases, which generates more actionable BUY and SELL signals but reduces HOLD recall to approximately 30-33%. The sentiment models at 5-day and 10-day horizons strike a middle ground with HOLD recall around 51%, balancing actionable signals with conservative positioning.

SELL signal detection remains weak across all models, with recall ranging from only 7.4% to 16.8%. The 5d_sent model performs best at 16.8%, while the 10d_core model captures merely 7.4% of actual SELL opportunities. This persistent weakness reflects both the scarcity of significant downward moves in the test period (approximately 1% of observations) and the inherent difficulty in predicting market declines, which arise from diverse catalysts that create inconsistent patterns.

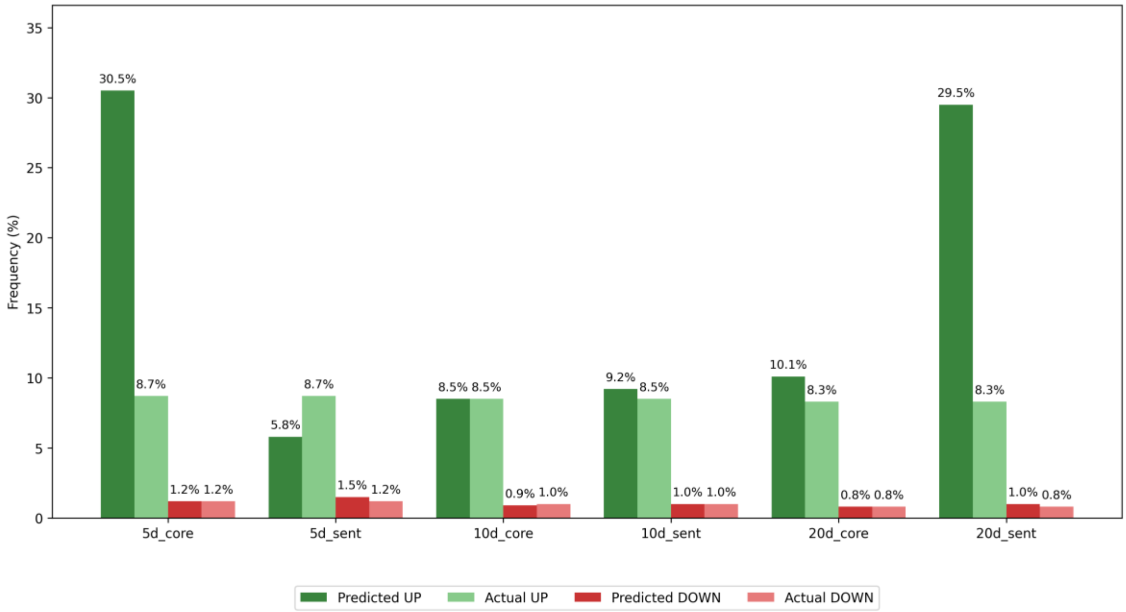
5.4 Big Move Detection and High-Volatility Event Analysis

While there is value in overall directional accuracy and correlation, which measure model performance on typical market days, in terms of practical profitable trading often depends on correctly identifying High-Volatility Events or "big moves". Large price movements in either direction can create the greatest profit opportunities but also pose risks. A model that would be accurate 60% of the time on direction but completely fails to predict major upward moves or crashes can lead to major losses. On the other hand, a model able to predict significant price movements is a much more valuable and practical tool for real-world trading.

5.4.1 Big move frequency

The question remains as what constitutes as a big move or high-volatility event, even though it is somewhat arbitrary we chose more than 2% or less than -2% of predicted normalized return, because it seemed like a move that would generate profit depending on the scenario, but also small enough for a notable sample size in the data. This threshold captures roughly the top 10-15% of price movements, which again represents the class imbalance issue.

Figure 9: Big UP and DOWN move predicted and actual frequency



Source: Own work

Figure 9 shows the frequency of big moves (price movements exceeding ±2%) across all predictions. Dark bars represent predicted frequencies, while lighter bars show actual occurrence rates in the test data.

All models reflect the natural market imbalance toward positive returns. In the actual test data, big UP moves occurred in approximately 8.5% of cases, while big DOWN moves occurred in only 1% of cases. However, several models significantly over-predict extreme movements.

The 5d_core model demonstrates the most aggressive bullish bias, predicting big UP moves in 30.5% of cases compared to the actual 8.7 %. Its sentiment counterpart, 5d_sent, shows better calibration at 5.8%. Interestingly, the 20d_sent model exhibits a similar aggressive bias (29.5% predicted vs. 8.3% actual), while the 20d_core model is more moderate at 10.1%.

Big DOWN predictions were well-calibrated across all models, ranging from 0.8% to 1.5%, closely matching actual occurrence rates of 0.8% to 1.2%. This suggests models are conservative and accurate in extreme downside predictions but struggle with upward movement magnitude calibration.

5.4.2 Profitability of big moves

While frequently and accurately predicting actual big moves would prove to be a significant achievement it is not necessarily a failure if at least the direction of the price is predicted correctly as it would most often not result in a loss of capital, to explore profitably the directional accuracy of big moves was observed.

Table 6: Directional accuracy of big moves

Model	Directional accuracy when Predicted UP	Directional accuracy when Predicted DOWN
5d_core	71.0%	24.2%
5d_sent	68.3%	29.2%
10d_core	70.2%	33.7%
10d_sent	70.8%	30.2%
20d_core	70.6%	29.9%
20d_sent	72.2%	32.6%

Source: Own work

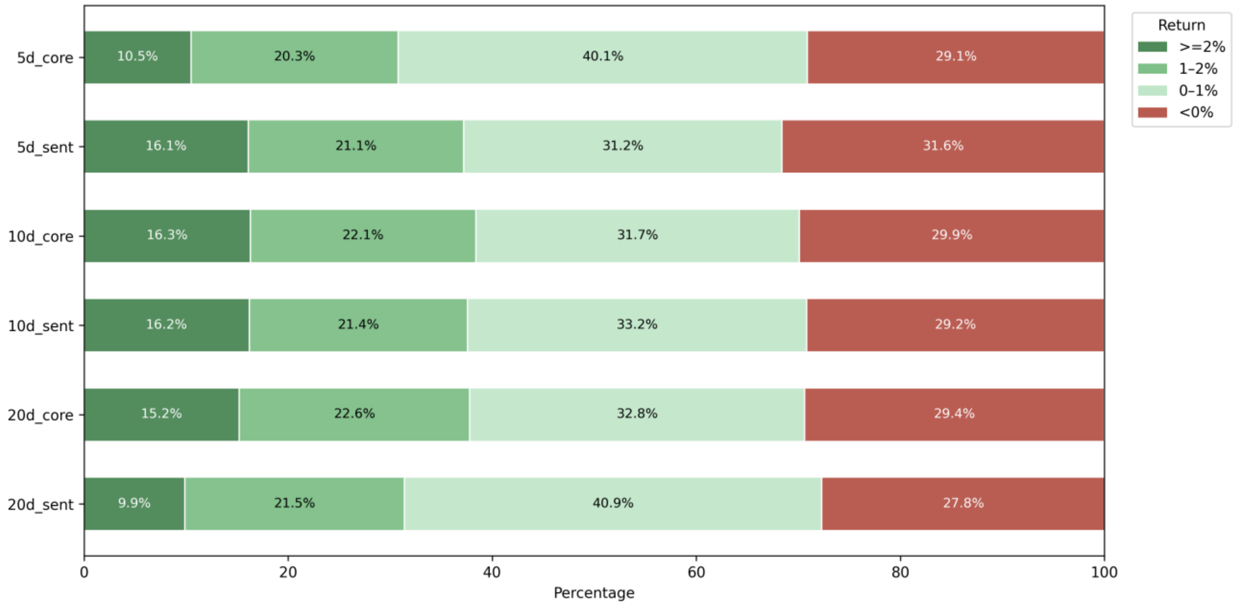
Across all models a consistent and clear asymmetry is seen in Table 6. When predicting UP the models are correct roughly seven times out of ten (68.3%–72.2%). When they predict a big DOWN, their directional accuracy is far lower, concentrated in the mid-20s to low 30s percent (24.2%–33.7%). As seen in previous metrics, UP signals are relatively reliable, DOWN signals much less so.

Looking at the UP column first, the models are tightly clustered. The 20-day sentiment model records the highest UP accuracy (72.2%), followed closely by the 5 day (71.0%). The weakest UP performer is the 5-day sentiment model at 68.3%, but even that is not far below the cluster mean. This degree of conditional accuracy for UP signals is notable.

While big up predictions make decent predictions the big DOWN are not nearly as accurate and more variable. The 10-day core model attains the best DOWN accuracy (33.7%), while the 5-day core model is the weakest (24.2%). Sentiment variants tend to sit in between with 20-day sentiment at 32.6%, and 5-day sentiment at 29.2%. As seen even here, the DOWN signal for these models is a relatively weak indicator.

Figure 10 presents the distribution of actual outcomes when each model predicts a big UP move (≥2%). Each horizontal bar is divided into bins: the darkest green shows the percentage of predictions where the actual return was indeed ≥2%, the next lighter green for 1–2%, then 0–1%, and finally red for cases where the actual return was negative (<0%). A clear pattern emerges while all models achieve some true big UP hits (dark green), which is also precision, most of the predicted big UP moves result in more modest gains (0–2%) or even losses.

Figure 10: Actual distribution of predicted big UP moves



Source: Own work

The 10d_core and 10d_sent models stand out with the highest share of true big UP moves (16.3% and 16.2%), and the largest portion of their predictions falling in the 1–2% and 0–1% bins, indicating a more balanced and realistic calibration. The 5d_core and 20d_sent models, in contrast, have the lowest precision for true big UP moves (10.5% and 9.9%), with a significant share of their predictions resulting in small or negative returns highlighting their aggressive, recall-maximizing nature from figure 9.

Across all models, a substantial portion of predicted big UP moves still end up as losses (red), underscoring the challenge of magnitude prediction even when direction is often correct. 5d_sent has the highest percentage of negative actual returns at 31.6%. The other models all have around the same around 29.5% with 20d_sent being the moderate exception at 27.8%.

5.5 Feature engineering impact: Core vs Sentiment deep dive

One of the central focuses driving this research is quantifying the differences and effect sentiment features derived from financial news headlines, how it compares to predictions solely relying on technical indicators.

In previous sections we saw that sentiment features change the models in complex ways, they improve correlation, harm accuracy, boost DOWN detection while sometimes reducing UP capture, and show being dependent on horizon effects. We try to explore the results in more detail and the reasons behind such effects.

5.5.1 Aggregate Performance Comparison

Overall performance differences between models relying on core technical features and those augmented with sentiment information were also explored. The focus is on identifying consistent patterns in predictive behavior across evaluation metrics.

Table 7: Core vs Sentiment Performance Comparison

Metric	Core Average	Sentiment Average
Directional Accuracy	61.02%	58.37%
Pearson Correlation	0.0179	0.0267
Spearman Correlation	0.0214	0.0306
UP Recall	75.75%	68.91%
DOWN Recall	24.88%	32.30%

Source: Own work

When comparing directional accuracy as seen in table 7, the core or technical features have a slight advantage with a 61.02% average accuracy around 2.5 percentage points higher than the sentiment average accuracy with 58.37%. The only other metric where core features are better at is the UP recall with 75.75% of all UP moves being successfully predicted compared to the sentiment with 68.91%. Where the sentiment outperforms its counterpart is everywhere else, both correlations see an increase of almost 50%, also as opposed to UP recall DOWN is better with sentiment features by around 6% with sentiment features having 32.30% and core with 24.88%.

Further confirmation of sentiment effects was done by observation-level analysis across the predictions, not only the models, because predictions provide a much larger sample size. We determined which features produced a prediction closer to actual return.

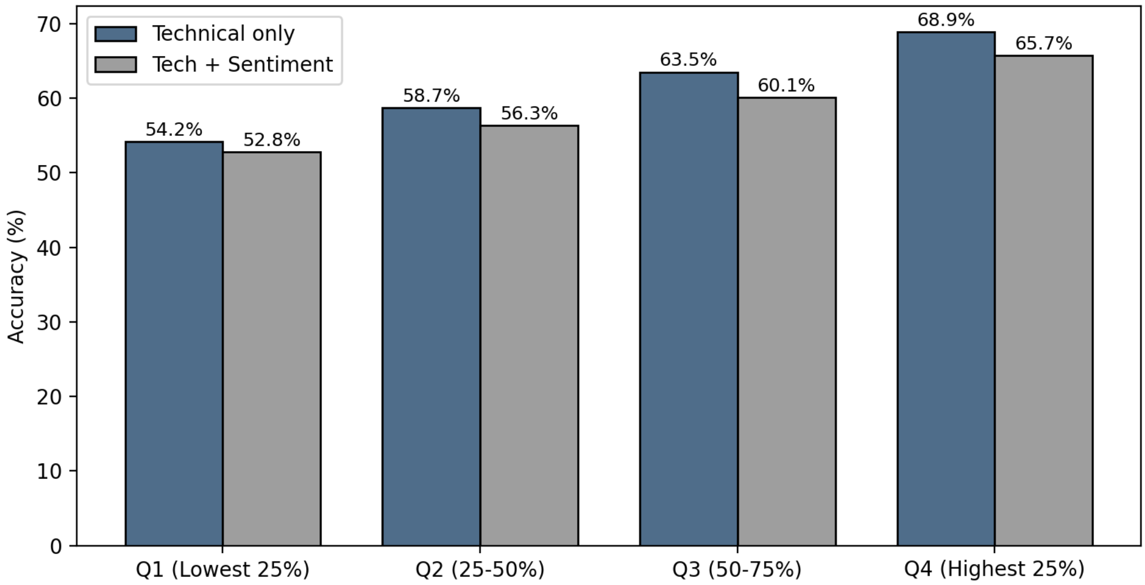
Table 8: Core vs Sentiment Performance Comparison on predictions

Metric	Value
Sentiment Win Rate	54.61%
Core Win Rate	45.39%
Error Reduction	4.03%

Source: Own work

Sentiment-based predictions achieved lower absolute error in 54.61% of cases compared to 45.39% for core price-only predictions and reduced mean absolute error by approximately 4.03%, clearly seen in table 8. Meaning sentiment features produced more accurate predictions than core financial features, with this advantage proven statistically significant through both binomial and paired t-tests.

Figure 11: Confidence comparison of models



Source: Own work

Beyond accuracy an examination of how confident the models are at predicting and how right they were when they were more confident. Both sentiment and technical show the proper calibration meaning that higher confidence predictions are indeed more accurate. However, core models in Figure 11 demonstrate superior confidence differentiation: the gap between Q1 and Q4 is 14.7 percentage points for core versus 12.9 for sentiment. The highest core accuracy is in the highest 25% of confidence with 68.9% compared to 65.7% from the sentiment. The same pattern is in all quartiles with core feature accuracy outperforming the sentiment. In the lowest quartile core has an accuracy of 54.2% and sentiment 52.8% a difference of 1.4%, it is the smallest gap of all the quartiles. This is in line with what we have observed so far where only technical features are better at predicting direction.

5.5.2 Volatility Regime Performance

Markets are quite dynamic and uneven; conditions vary dramatically; the most basic distinctions are calm and chaotic periods. To measure the performance, we partition the test period into three volatility regimes based on 20-day rolling standard deviation:

Table 9 shows that in low volatility environments, core models achieve the highest accuracy with 64.3%, with 4.2 percentage points ahead of sentiment, which has 60.1%.

Table 9: Performance by Market Volatility Regime

Volatility Regime	Core Accuracy	Sentiment Accuracy	Core Correlation	Sentiment Correlation
Low Vol (<20th)	64.3%	60.1%	-0.008	+0.012
Medium Vol (20-80th)	61.5%	58.9%	+0.019	+0.024
High Vol (>80th)	57.8%	55.2%	+0.038	+0.061

Source: Own work

Medium volatility represents the majority of trading days, and both model types perform a bit worse than in low volatility. Core accuracy is at 61.5%, which is more than 2 percentage points higher than sentiment with 58.9%. What is more important is that both models achieve a positive correlation performance, with sentiment having a 26% improvement over core.

In high volatility periods (>80th percentile), both model types experience an additional lowering of accuracy, but sentiment models demonstrate superior correlation performance. While core accuracy with 57.8% still edges out sentiment (55.2%) by 2.6 percentage points, what is more insightful is that sentiment correlation dramatically outperforms core by 61%.

6 DISCUSSION

The results presented in Chapter 5 indicate that the impact of sentiment on predictive performance is nuanced rather than uniformly positive or negative. While models based solely on technical features tend to achieve higher directional accuracy, sentiment-augmented models show stronger alignment with realized returns and improved detection of downside movements. These differences suggest that the usefulness of sentiment depends on the specific predictive objective, rather than on a single aggregate performance measure. Gaining a clearer understanding of when sentiment adds value therefore requires considering how news-driven effects manifest across different forecasting horizons and market conditions.

6.1 Overview of the main findings

This thesis set out to evaluate how incorporating sentiment from financial news headlines could affect the predictions of LSTM-based models for short-term stock return forecasting. Although it was expected that the inclusion of sentiment would lead to improved results, specifically in our setup the findings suggest a more nuanced relationship that cannot be captured by a simple affirmative claim. Sentiment reshapes model behavior in multiple ways, contributing effectively, but does not improve performance across all metrics unilaterally.

Across all configurations a pattern emerges: core-feature models achieve higher directional accuracy, averaging 61% compared to sentiment models' 58.4%. However, sentiment-enhanced models demonstrate a stronger correlation with actual returns, with both Pearson and Spearman correlations increasing by approximately 50%, they also perform better looking at the mean absolute error.

At first glance it seems the core models perform better and while they do, based on accuracy alone, they seem more optimized for binary classification. They benefit from clear technical patterns and the markets structural upward bias. Sentiment models, by contrast, achieve better proportional alignment with return magnitudes, even when their directional timing is less precise. When a sentiment model predicts a large move, its magnitude estimate tends to be more accurate than that of a core model, despite occasionally misidentifying the direction.

The influence of sentiment features varies systematically with the prediction horizon, revealing important dynamics about how news information propagates through markets. At the 5-day horizon, sentiment reduces directional accuracy by 5.1 percentage points while providing only modest correlation improvements. The short temporal window appears insufficient to distinguish meaningful sentiment signals from noise. Many headlines spike sentiment scores without lasting impact, and the model lacks sufficient context to learn which sentiment changes lead to sustained price movements.

At the 10-day horizon, a critical balance point emerges. The 10d_sent model achieves the highest correlations while maintaining acceptable accuracy (57.7%) and demonstrates substantially better DOWN detection with a 77% relative improvement. This horizon appears long enough to distinguish persistent narratives from single-event noise, yet short enough to remain responsive. These findings support Barberis and Thaler's (2003) framework on how behavioral biases manifest in markets. The sentiment advantage at this timescale may reflect gradual information diffusion: sentiment captures building negative perception while prices are still adjusting, providing a window to anticipate movement before it fully materializes in technical indicators.

At the 20-day horizon, sentiment's effects reverse. Directional accuracy improves by 2.6 percentage points while correlation declines, suggesting sentiment acts as a trend-reinforcing measure rather than a predictive signal. The most likely explanation involves reverse causality: over a one-month horizon, news tends to follow prices rather than lead them, aligning with Fama's (1970) efficient market hypothesis where longer timeframes allow prices to fully incorporate available information. This transformation from leading indicator at 10 days to lagging indicator at 20 days implies that traders seeking anticipatory signals should favor medium-term horizons.

What is a common trope across all models is their asymmetry between predicting upward and downward movements. UP recall consistently exceeds 70%, while DOWN recall remains below 38% even in the best configurations. This pattern reflects the empirical distribution of returns (71% UP days in the test period) and the structural characteristics of equity markets.

One of the most notable benefits of sentiment in our research is the superior downside detection. The 10d_sent model's 37.5% DOWN recall represents a meaningful improvement over the 10d_core model's 21.2%, effectively increasing the proportion of detected downturns from approximately 1 in 5 to nearly 2 in 5. While this still means that a large number of DOWN moves are missed, the improvement is still relevant.

When predictions are structured as tertiary trading signals (SELL/HOLD/BUY), performance varies considerably across all classes. SELL signals demonstrate weak detection rates (7.4-16.8% recall) and low precision (13.7-15.2%), which indicates very limited reliability as standalone indicators. BUY signals achieve a moderate performance (38.3-65.0% recall, 48.5-50.3% precision), while HOLD detection varies substantially (30.6-60.8% recall) depending on model configuration.

Despite poor SELL recall, these signals do appear to contain information about market conditions. When models predict SELL, actual returns are consistently worse than average, suggesting that the signals, while infrequent, identify genuinely unfavorable conditions.

Both model types demonstrate reasonable calibration: predictions made with higher confidence are indeed more accurate. Core models show slightly stronger confidence differentiation (14.7 percentage point accuracy gap between lowest and highest confidence quartiles versus 12.9 for sentiment). This difference likely reflects the more consistent patterns in technical indicators compared to sentiment signals, which may be influenced by contradictory or ambiguous news. The calibration finding has important practical implications: a trader could reliably use the model's confidence scores to scale position sizes proportionally, for instance taking larger positions when confidence exceeds 75% and smaller or no positions below 60%, thereby systematically managing risk based on the model's own uncertainty estimates.

At the observation level, sentiment-enhanced models consistently outperformed core price-only models. Across all out-of-sample predictions, sentiment-based predictions achieved lower absolute error in 54.61% of cases and reduced mean absolute error by approximately 4.03%. These improvements were confirmed as statistically significant using both binomial and paired t-tests, indicating that sentiment information provides reliable predictive benefits when evaluated at the individual prediction level.

Importantly, while model-level correlation comparisons did not reach a major statistical significance due to the limited number of model variants, observation-level evaluation reveals a clear and statistically robust performance advantage for sentiment-enhanced predictions. This distinction highlights that correlation metrics alone may understate the practical value of sentiment features, which instead manifest through consistent reductions in prediction error across individual observations.

No single model configuration dominates across all evaluation criteria. The 10d_core model achieves the highest directional accuracy (63.1%) and lowest error metrics. The 10d_sent model leads in correlation and downside sensitivity. The 20d_sent model demonstrates the strongest BUY signal detection (65.0% recall), while the 5d_core model best identifies big moves (39.3% detection rate).

This diversity suggests that model selection should be guided by specific application requirements rather than overall superiority. Different configurations demonstrate distinct strengths that may be suitable for different trading strategies or risk management objectives.

6.2 Directional Accuracy and Magnitude Estimation as Competing Aims

The trade-off between the directional accuracy and magnitude estimation requires a deeper look, because it reflects a fundamental difference in terms of financial forecasting. As we mentioned technical, models achieve an average directional accuracy of around 61%, while sentiment models stay closer to 59%. Where sentiment models have the edge is in correlation, which is almost double compared to technical models.

When evaluating only directional accuracy it is a binary classification, the only concern is with if the model correctly predicts the price movement. What we saw is that this metric is heavily influenced by the class distribution of our dataset, with around 71% observations being UP days, models can achieve relatively high accuracy by favoring upward predictions, even if their understanding of underlying return dynamics is limited. A model that predicts UP for all observations would achieve 71% accuracy while providing no useful information.

The direction of where the price will move is arguably the most important, where a high correlation might not necessarily mean the right direction is predicted often. What it does achieve is that the model predicts a larger movement, actual returns also tend to increase. Magnitude estimation can be even more important than direction in position sizing strategies and such.

The results indicate that sentiment features improve this proportional alignment at the cost of directional accuracy. The mechanism is likely informational, when we get two strikingly opposed headlines the difference in tone conveys information about the likely movement intensity. This nuance is not directly present in technical indicators, which measure only past price action.

Technical models perform better in binary classification precisely because they lack such contextual information. Conditions for them are a bit more unambiguous so they can produce more consistent patterns, which the models learn reliably. It does come at a cost, when other models correctly interpret the tone around a stock, it can more accurately estimate the excepted magnitude of price movement.

What we can conclude is that sentiment adds a layer of context that enables better magnitude estimation but simultaneously introduces ambiguity. As news can cause more doubt and conflicting signals. The model must interpret which signal is more important, which it can get wrong from time to time. However, the upside is the better estimation of magnitude mentioned.

6.3 Temporal Horizon as a Structural Moderator of Sentiment Effects

What is not surprising is that sentiment affects the price predictions differently depending on the time frame we are dealing with.

At 5-day sequences, sentiment reduces performance across nearly all metrics. Short-term financial news is highly variable. Many headlines that seem important in the moment, for example, an analyst comment, product speculation, an isolated management statement usually has minimal lasting impact but may spike sentiment. The issue with so many signals, or too few is that often it is just noise that can obscure patterns learned from price action.

The problem is not a lack of information in the news but a lack of context to distinguish signal from noise. Because models that deal with longer timeframes are actually improved by sentiment, we can speculate that for our set up five days is too little time to differentiate the beginning of genuinely important developments. The model does not have enough observations to learn or learn wrongly which types of sentiment leads to lasting price changes and which disappear within a day or two.

At the 20-day horizon, sentiment affects it differently compared to shorter timeframes where directional accuracy improves but correlation declines. UP recall increases while DOWN recall decreases. These patterns indicate that sentiment at this timeframe acts more as a validating measure rather than a predictive signal.

The most likely explanation is a sort of reverse causality. Over a one-month horizon, 20 days which are working days so exactly one month, news tend to follow prices rather than lead them. Stocks that go up receive more positive coverage, while those that fall attract negative news. As a result, sentiment mostly reflects what has already happened instead of predicting future moves.

At the 10-day horizon which is a sort of middle ground it seems the reverse causality is not a concern. It appears the horizon is long enough to distinguish persistent narratives from single-time events, yet short enough to remain responsive in an effective manner. Consistent negative tone over two weeks more likely signals genuine problems than one negative headline. Similarly, sustained positive coverage indicates confidence that may be reflected in price appreciation.

The 10-day horizon achieves the highest Pearson and Spearman correlation, with the cost of some accuracy, but is much better aligned in terms of magnitude. It also represents a great improvement regarding DOWN detection, almost doubling the recall compared to technical-only models.

Negative news creates a self-reinforcing cycle. A few concerning headlines draw attention to potential problems. Investors usually react and most often sell, which causes prices to fall. The price decline generates more negative coverage, which prompts more selling, prices fall further. This cycle repeats and intensifies. Sentiment picks up each turn before it fully shows in the price. Technical indicators only measure what already happened, so they lag behind. Sentiment catches the rising negative perception while prices are still catching up, giving a short window to see where things are heading before they get there.

6.4 Market Asymmetry, Downside Risk, and the Limits of Prediction

All models struggle with the same problem, which is that they are much better at predicting UP movements than DOWN movements, some models have so much trouble they miss most of DOWN moves. While this can be minimized, it is also a reflection of how equity markets work and what causes prices to fall.

Positive movements are common and follow similar patterns. Companies grow gradually through revenue increases, margin expansion, and market penetration. These processes create recognizable technical signatures: sustained higher highs and higher lows, consistent volume increases, favorable moving average crossovers. The abundance of positive examples in training data lets models learn these patterns effectively. UP recall consistently exceeds 70% across all models, reaching as high as 80.3% in the 10d_core configuration.

Negative movements are quite different. They are less frequent and more varied in how they happen. Declines can come from earnings disappointments, regulatory actions, competitive disruptions, macroeconomic shocks, financial disclosures, or management changes, sometimes declines or even upswing happen for reasons that may seem contradictory. Each catalyst creates a different pattern in both technical indicators and sentiment. The scarcity of DOWN examples combined with their diversity limits how well models can generalize. A model might learn that negative earnings news often leads to declines but struggles to apply this knowledge to regulatory shocks, which look completely different. DOWN recall ranges from 21.2% to 37.5%, much lower than UP detection.

One of the biggest problems is genuinely unexpected events. The largest declines, like the initial COVID-19 crash in March 2020, come from events that could not be anticipated with available data. No amount of historical data would have enabled models to predict a pandemic effectively.

Despite the many challenges, sentiment provides incremental improvement. What we can speculate seems to be that media coverage overreacts to bad news. Journalists and analysts emphasize negative developments, creating situations where sentiment becomes negative faster than prices adjust. This gives sentiment models a brief time window where they can anticipate downward movement before it fully materializes. The 10d_sent model achieves 37.5% DOWN recall compared to 21.2% for the 10d_core model, a 77% relative improvement.

However, this advantage is one-sided. Positive news often lags behind price increases. Companies announce good news only when results are confirmed, journalists report successes only when they are obvious. This means sentiment offers less anticipatory value for upward movements, where technical indicators already capture momentum. There is also a psychological component, such as the fear of having already missed the upward move or the market has already "priced in" the news as opposed to negative news that could even further deteriorate the company.

Why sentiment seems to improve DOWN detection more can be argued for in multiple ways, but the results show that the fact our downward predictions correspond to substantially worse returns indicates the signals are not worthless. Even with our SELL prediction for the 5d_sent model, actual median returns are around -0.85%, compared to +0.07% overall, a full 0.92 percentage point difference. This suggests that while SELL signals have low recall, they contain genuine information. When the model is confident enough to predict SELL, conditions are genuinely worse than average.

Table 10 illustrates that each model configuration exhibits distinct strengths and limitations, reinforcing the view that sentiment integration leads to targeted improvements rather than uniform performance gains. The 10d_core model delivers the most reliable directional predictions, confirming the effectiveness of technical indicators for binary up/down classification. In contrast, the 10d_sent model trades directional accuracy for stronger alignment with realized returns and substantially improved downside detection, addressing a well-documented challenge in equity prediction where negative movements are harder to identify than positive ones (Gu et al., 2020).

The 20d_sent model further highlights the role of sentiment in identifying sustained trends and favorable entry points, while the 5d_core model demonstrates that short temporal windows are particularly effective at capturing extreme price movements, despite lower overall consistency.

Table 10: Summary of the best models

Model	Core advantage	Main trade-off	Practical implication
5d_core	Highly responsive to short-term volatility and extreme price movements.	Directional instability reduces reliability for sustained positioning.	Best suited for event-driven or speculative strategies during high-volatility periods.
10d_core	Most reliable directional predictions across models, particularly in stable markets.	Weak detection of downside risk limits usefulness during market stress.	Appropriate for conservative, consistency-focused trading and long-only strategies.
10d_sent	Improved downside detection and stronger alignment with return magnitudes.	Lower directional accuracy compared to technical-only models.	Valuable for risk management, hedging, and trading under elevated uncertainty.
20d_sent	Captures sustained trends and sentiment regimes over longer horizons.	Slower response to emerging market stress.	Suitable for longer-term positioning and trend-following approaches.

Source: Own work

Taken together, these results emphasize that the value of sentiment lies in enhancing specific predictive capabilities, such as magnitude sensitivity and risk detection during adverse market conditions, rather than improving all evaluation metrics simultaneously.

7 CONCLUSION

One of the main aims of this master's thesis was to assess whether incorporating sentiment information derived from financial news headlines can improve LSTM-based models for short-term stock price prediction using a hybrid machine learning model. Building on the research objectives defined in the introduction, the study focused on developing recurrent neural network architectures, selecting appropriate short-term prediction targets, comparing alternative feature configurations across multiple horizons, and evaluating whether sentiment-driven signals offer practical benefits for short-term investors.

These objectives were addressed through the systematic evaluation of multiple LSTM model configurations that differed only in the information used for training. Models based on core technical features and models augmented with sentiment signals were assessed under identical architectures and evaluation procedures, allowing differences in predictive behavior to be attributed directly to feature composition. The results suggest that sentiment integration contributes measurable improvements in key performance dimensions, especially in magnitude alignment and risk detection, even though its effects vary across evaluation metrics.

Across the configurations examined, technical-only models tend to deliver more consistent directional predictions, particularly in stable market conditions. In contrast, sentiment-augmented models exhibit stronger alignment with realized returns and improved sensitivity to negative market movements. This pattern indicates that technical indicators are effective at capturing directional bias, while sentiment information contributes additional context that becomes particularly relevant when markets deteriorate or react to adverse news. As a result, the relative usefulness of each feature set depends on the specific predictive objective rather than on a single aggregate metric.

The analysis further demonstrates that prediction performance varies systematically with the forecasting horizon. Shorter input windows are more responsive to volatility and extreme price movements, while medium-term sentiment models better capture delayed market reactions to news. Longer-horizon sentiment models, in turn, support the identification of sustained trends and more stable entry points. No single model configuration dominates across all prediction targets, confirming the value of a comparative, multi-horizon evaluation framework.

From a theoretical perspective, these findings are consistent with research in behavioral finance showing that information is incorporated into prices gradually (Hong and Stein, 1999) and that markets often react more strongly to negative news (Tetlock, 2007). Evidence from financial text analysis further suggests that sentiment extracted from news contains information not fully reflected in price-based indicators (Loughran and McDonald, 2011). In this context, the improved downside sensitivity observed when sentiment-based indicators are included appears economically plausible. At the same time, the results indicate that sentiment integration does not uniformly improve predictive performance but instead shifts the balance between directional accuracy and magnitude estimation. By evaluating models across multiple forecasting horizons and performance metrics, this thesis offers a structured comparison of how sentiment-based and technical features affect different aspects of predictive quality. Unlike studies that evaluate sentiment-augmented models primarily through directional accuracy on a limited set of assets (Fischer and Krauss, 2018), this thesis tests six configurations across 212 securities using multiple performance dimensions, revealing trade-offs that single-metric evaluations would not capture.

In practical terms, the results suggest that model selection should be guided by intended use rather than by overall accuracy alone. Technical models are better suited for applications that prioritize stable directional signals, whereas sentiment-augmented models offer advantages for risk-aware strategies, downside protection, and magnitude-sensitive decision-making. The findings also indicate that prediction confidence provides meaningful information, enabling position sizing approaches that scale exposure based on model certainty.

While this study does achieve a certain level of valuable insight, there are many limitations one should be aware of. First, the data and the dataset have inherent limitations. Even though a large amount of data was processed, the number of news articles per stock per day were limited. Only a few companies had multiple news articles written across several consecutive days. Many of them had periods of no news or only a few articles per week. Extensive data preparation was needed to find a balance where sentiment would still be present on enough days to make it significant. If every single stock was included, the days with sentiment would be a small minority, so a balance was necessary, but it came at a cost of congestion and less diversity in stocks. This uneven distribution means models learned from varying amounts of sentiment information across different securities. Stocks with rich news coverage may have benefited more from sentiment features, while those with sparse coverage essentially reverted to technical-only predictions on many days. To maintain continuous sequences required for LSTM models, gaps were filled with neutral sentiment values or sequences were truncated to focus on periods with dense coverage.

Second, what might have even been positive is that features are derived exclusively from the headlines of financial news rather than the full article text. Headlines are designed to capture attention and may use sensational or provocative language that does not reflect the actual content of articles. This limitation may introduce noise into sentiment scores or cause the model to react to tone rather than substance. Full-text analysis might provide more nuanced sentiment signals, though at substantially higher computational cost. But it is also true that many people only read the headlines of articles so it may work both ways.

Third, we capture data based at a daily level so the average prices and sentiment of that day, for the technical factors we minimized this with several optimizations, but for the sentiment we averaged multiple headlines, when they occur on the same trading day. This approach obscures stock dynamics and the potential impact of news timing. For example, if a headline is posted before the market open the effect it may have can be different than one released after close, yet we treat both the same. An initial positive headline followed by a negative correction is indistinguishable from the reverse sequence. We were aware of this limitation from the start as we decided on the daily price, a more ambitious way would be price detection by minutes, but we were limited in our data.

Fourth, the study test data focuses on a specific time period (2019-2020) that includes both normal market conditions and the exceptional volatility of the COVID-19 pandemic. While this provides insight into model behavior during crisis periods, results may not generalize to other market regimes. The models were trained primarily on data from 2009-2018, a period which had a sustained bull market with relatively few severe downturns. Performance during the 2020 test period, particularly during the March crash, may reflect this training bias. Testing on additional out-of-sample periods, including bear markets and different economic cycles, would provide a more complete assessment of generalizability.

LIST OF KEY LITERATURE

1. Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x
2. 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
3. Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, *55*(4), 1705–1765. https://doi.org/10.1111/0022-1082.00265
4. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040
5. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

REFERENCE LIST

1. Aenlle, M. A. (2020). *Massive Stock News Analysis DB for NLP Backtests* [Data set]. Kaggle. https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests
2. Aldridge, I. (2013). *High-frequency trading: A practical guide to algorithmic strategies and trading systems* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
3. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*. https://arxiv.org/abs/1908.10063
4. Ayyildiz, E., & Iskenderoglu, O. (2024). How effective is machine learning in stock market predictions? Heliyon, 10(2), e24123. https://doi.org/10.1016/j.heliyon.2024.e24123
5. Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x
6. Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The Journal of Finance*, *55*(2), 773–806. https://doi.org/10.1111/0022-1082.00226

7. Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention induced trading and returns: Evidence from Robinhood users. *Journal of Finance*, *77*(6), 3141–3190. https://doi.org/10.1111/jofi.13172
8. Barber, B. M., Lee, Y.-T., Liu, Y.-J., & Odean, T. (2009). Just how much do individual investors lose by trading? *The Review of Financial Studies*, *22*(2), 609–632. https://doi.org/10.1093/rfs/hhn046
9. Barber, B. M., Lee, Y.-T., Liu, Y.-J., & Odean, T. (2014). The cross-section of speculator skill: Evidence from day trading. *Journal of Financial Markets*, *18*, 1–24. https://doi.org/10.1016/j.finmar.2013.05.006
10. Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), *Handbook of the economics of finance* (Vol. 1, pp. 1053–1128). Elsevier. https://doi.org/10.1016/S1574-0102(03)01027-6
11. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181
12. Bodie, Z., Kane, A., & Marcus, A. J. (2018). *Investments* (12th ed.). McGraw-Hill Education.
13. 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
14. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
15. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
16. Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance*, *47*(5), 1731–1764. https://doi.org/10.1111/j.1540-6261.1992.tb04681.x
17. Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, *27*(8), 2267–2306. https://doi.org/10.1093/rfs/hhu032
18. Chang, V., Xu, Q. A., Chidozie, A., & Wang, H. (2024). Predicting economic trends and stock market prices with deep learning and advanced machine learning techniques. *Electronics, 13*(17), Article 3396. https://doi.org/10.3390/electronics13173396
19. Chen, A.-S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index. *Computers & Operations Research, 30*(6), 901–923. https://doi.org/10.1016/S0305-0548(02)00037-0
20. Chen, N. F., Roll, R., & Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business*, *59*(3), 383–403. https://doi.org/10.1086/296344
21. Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge*

discovery and data mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

22. Deep, A., Monico, C., Shirvani, A., Rachev, S., & Fabozzi, F. J. (2024). Risk-Adjusted Performance of Random Forest Models in High-Frequency Trading. *arXiv*. https://arxiv.org/abs/2412.15448
23. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423
24. Dorn, D., & Sengmueller, P. (2009). Trading as entertainment. *Management Science*, *55*(4), 591–603. https://doi.org/10.1287/mnsc.1080.0962
25. Eaton, G. W., Green, T. C., Roseman, B., & Wu, Y. (2022). Retail trader sophistication and stock market quality: Evidence from brokerage outages. *Journal of Financial Economics, 146*, 502–528. https://doi.org/10.1016/j.jfineco.2022.08.002
26. Eavis, P., & Goldstein, M. (2021, January 27). Melvin Capital, a hedge fund targeted by GameStop traders, lost 53% in January. *The New York Times*. https://www.nytimes.com/2021/01/27/business/melvin-capital-gamestop.html
27. Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1
28. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773
29. Fama, E. F. (1970). Efficient capital markets: A review of empirical work. *Journal of Finance, 25*(2), 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x
30. Financial Industry Regulatory Authority. (2024). *Retail investor participation and trading trends*. https://www.finra.org
31. 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
32. Franck, T. (2021, January 29). Citron Research to stop publishing short-seller reports after GameStop chaos. *CNBC*. https://www.cnbc.com/2021/01/29/citron-research-to-stop-publishing-short-seller-reports-after-gamestop-chaos.html
33. Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, *53*(1), 67–78. https://doi.org/10.1086/296072
34. Gempesaw, D., Henry, J. J., & Velthuis, R. (2022). Piecing together the extent of retail fractional trading. *Global Finance Journal*, *54*, Article 100757. https://doi.org/10.1016/j.gfj.2022.100757
35. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

36. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009
37. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
38. Hoffmann, A. O. I., & Shefrin, H. (2014). Technical analysis and individual investors. *Journal of Economic Behavior & Organization*, *107*, 487–511. https://doi.org/10.1016/j.jebo.2014.04.002
39. Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance, 54*(6), 2143–2184. https://doi.org/10.1111/0022-1082.00184
40. Jame, R. (2019). *Seeking alpha and losing ground: The unintended consequences of commission-free trading*. University of Kentucky. https://russelljame.com/seeking_alpha_oct_2019.pdf
41. Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x
42. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
43. Khanna, M., Kulshrestha, M., Singh, L. K., Thawkar, S., & Shrivastava, K. (2022). Performance evaluation of machine learning algorithms for stock price and stock index movement prediction using trend deterministic data prediction. *International Journal of Applied Metaheuristic Computing*, *13*(1), 1–30. https://www.researchgate.net/publication/363209368
44. Kim, K.-J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1–2), 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2
45. Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, *72*(3), 967–998. https://doi.org/10.1111/jofi.12498
46. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702. https://doi.org/10.1016/j.ejor.2016.10.031
47. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3
48. Leslie, R. D., Shaw, C. A., & McGrath, D. S. (2024). Correlates of gambling behaviours among day traders: Evidence from a national study. *Journal of Gambling Studies*, *40*, 1751–1771. https://doi.org/10.1007/s10899-024-10314-6
49. Liu, Y. (2024). CNN-GRU-XGBoost stock price prediction model under hyperparameter-based optimisation. *Financial Engineering and Risk Management*, *7*(5), 151–158. https://doi.org/10.23977/ferm.2024.070519
50. Liu, Y., Qian, Z., & Wang, Z. (2024). Volatility forecasting with machine learning and intraday commonality. *International Journal of Forecasting*, *40*(2), 689–703. https://doi.org/10.1016/j.ijforecast.2023.06.003

51. Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, *30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611
52. Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, *55*(4), 1705–1765. https://doi.org/10.1111/0022-1082.00265
53. Loesche, D. (2018). *A bigger picture of the stock market*. Statista. https://www.statista.com/chart/12825/standard-poors-500-performance-since-its-launch-in-march-1957/
54. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance, 66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
55. Lucchini, L., Radaelli, L., Sirbu, A., & Giordano, G. (2021). From Reddit to Wall Street: The role of social media in the GameStop short squeeze. *arXiv preprint arXiv:2107.07361*. https://arxiv.org/abs/2107.07361
56. Lynch, P. (1989). *One up on Wall Street*. New York: Simon & Schuster.
57. Malkiel, B. G., & Ellis, C. D. (2012). *The elements of investing: Easy lessons for every investor* (Updated ed.). Hoboken, NJ: John Wiley & Sons.
58. Malkiel, B. G. (2019). *A random walk down Wall Street* (12th ed.). New York: W. W. Norton & Company.
59. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119). https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
60. Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
61. Neely, C. J., Weller, P. A., & Dittmar, R. F. (1997). Is technical analysis in the foreign exchange market profitable? A genetic programming approach. *Journal of Financial and Quantitative Analysis*, *32*(4), 405–426. https://doi.org/10.2307/2331232
62. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040
63. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162
64. Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using Twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 119–123). IEEE Computer Society. https://dl.acm.org/doi/abs/10.5555/2456719.2456923
65. Robinhood Markets. (2023). A*nnual report*. https://investors.robinhood.com

66. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0
67. Ryll, L., & Seidens, S. (2019). Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey. *arXiv*. https://doi.org/10.48550/arXiv.1906.07786
68. Sueppel, R. (2022, June 11). Six ways to estimate realized volatility. *Macrosynergy*. https://macrosynergy.com/research/six-ways-to-estimate-realized-volatility/
69. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x
70. U.S. Securities and Exchange Commission (SEC). (2021). *Staff report on equity and options market structure conditions in early 2021*. Washington, D.C. https://www.sec.gov/news/studies-reports/reports
71. Whaley, R. E. (2000). The investor fear gauge. *Journal of Portfolio Management*, *26*(3), 12–17. https://doi.org/10.3905/jpm.2000.319728
72. White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns. In *Proceedings of the IEEE International Conference on Neural Networks* (Vol. 2, pp. 451–458). IEEE. https://doi.org/10.1109/ICNN.1988.23959
73. Wolff, D., & Echterling, F. (2024). Stock picking with machine learning. *Journal of Forecasting*, *43*(5), 1181–1196. https://doi.org/10.1002/for.3021
74. Yahoo Finance. (n.d.). *CBOE Volatility Index (VIX)*. https://finance.yahoo.com
75. Yahoo Finance. (n.d.). *Historical stock price data*. https://finance.yahoo.com
76. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0